ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
"Национальный исследовательский университет
"Высшая школа экономики"

НЕГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
"РОССИЙСКАЯ ЭКОНОМИЧЕСКАЯ ШКОЛА"(ИНСТИТУТ)

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Вероятностное прогнозирование событий на основе новостей с применением больших языковых моделей

Программа Бакалавр экономики

Совместная программа по экономике НИУ ВШЭ и РЭШ

Автор:
А. А. Петухов

Руководитель ВКР:
Салихов Марат
Радикович, PhD

Москва, 2025 г.

News-Based Probabilistic Forecasting with Large Language Models

Аннотация

В работе предложен улучшенный подход к прогнозированию реальных бинарных событий с помощью больших языковых моделей. Система строится из нескольких этапов: подключение актуальной информации через RAG, объединение предсказаний разных моделей и последующая калибровка результатов. Такой подход позволил значительно повысить точность на датасете из 1000 вопросов с Polymarket (Brier score — 0.178). Я показываю, что крупные модели и агрегации дают прирост качества и устойчивости, а также намечаю векторы для будущих исследований — расширение данных, доработка моделей и адаптация под конкретные области.

Abstract

This thesis presents an optimized multi-stage pipeline for forecasting binary real-world events using large language models (LLMs). The pipeline combines Retrieval-Augmented Generation (RAG), ensemble methods, and post-hoc calibration to significantly improve predictive accuracy on 1000 Polymarket questions, achieving a Brier score of 0.178. Key contributions include the use of RAG for incorporating current news, aggregation of diverse LLM predictions, and logistic calibration for output refinement. Larger LLMs and ensemble strategies improved baseline performance and robustness. The results highlight the potential of LLMs in forecasting and suggest future research directions, including dataset expansion, model refinement, and domain specialization.

Contents

Chapter 1

Introduction

Forecasting is an important task in the modern world. Accurate forecasts allow for better planning, resource allocation and risk management. There are two main types of forecasting. *Statistical forecasting* primarily uses time series models and relies on high-quality data with stationary patterns. *Judgmental forecasting*, on the contrary, is performed primarily by expert human forecasters, who usually rely on diverse news sources and their accumulated knowledge (Tetlock and Gardner (2015)). It is applicable when past data is scarce or is subjected to distributional shift. In this work I focus on judgmental forecasting, simply called "forecasting" in the rest of this paper. Special platforms for forecasting called prediction markets allow users to make bets on the outcomes of events in various spheres, sports, and other spheres.

Human forecasting requires expertise and thorough studying of the topic, making it expensive and slow. Recent surge in Large Language Models (LLMs) development has created new possibilities to use them in complex tasks. These models are trained on large amounts of data, accumulating knowledge across various domains. LLMs can process text very quickly and are available through Application Programming Interface (API), making them both fast and cheap. Standard LLMs, despite their general capabilities, may not perform optimally or show reliable calibration in tasks like event forecasting, where precise probabilistic estimates are important. These models without special tools have no access to recent news and might be producing uninformed forecasts. To address these limitations, different strategies like Retrieval Augmented Generation (RAG), prompt engineering, ensembling and probabilities calibration can be employed. The effectiveness of these components, however, needs to be studied systematically, and I address this gap in current research in my thesis.

The main question of this research is whether a pipeline supplemented with real-time news through RAG, multiple LLMs ensembling, and calibration can achieve better predictive performance than individual models and simpler methods. I also compare LLM-generated base rate estimates (the "outside view") with predictions based both on news and base rates (the "inside view"). This research also studies how the time of prediction relative to an event's resolution date affects the Brier score, accuracy and calibration of predictions. Finally, I compare different methods for aggregating LLM predictions (mean, median, trimmed mean) and estimate calibration using logistic regression. The main goal is to present a tested framework and find better setups for using LLMs to improve automated event forecasting.

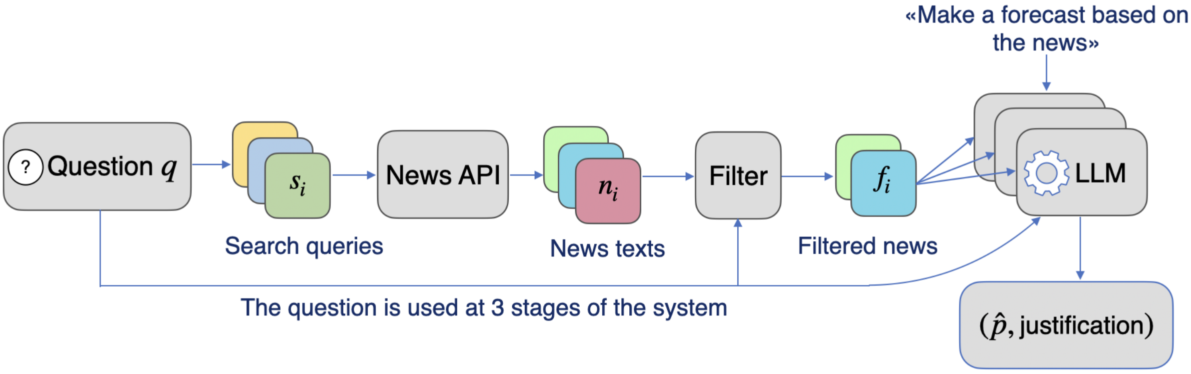

«Make a forecast based on the news»
? Question q
s_i
Search queries
News API
n_i
News texts
Filter
f_i
Filtered news
LLM
$(\hat{p}$, justification)
The question is used at 3 stages of the system

Figure 1.1: **High-level pipeline diagram**: search queries are generated from the question text and its description with the aid of a lightweight LLM. The queries are passed to the news API (GDELT). The news texts are filtered with a text ranking model and the remaining texts are passed on to an ensemble of LLMs.

I build an LLM-based forecasting system for binary ("Yes" or "No") questions, which yields probabilistic estimates of the future events. Given a question, the system starts by collecting relevant data from diverse news sources from the Global Database of Events, Language, and Tone (GDELT). Then, it filters the collected data with a pretrained ranking model to leave only the most relevant news. After that, it queries LLMs for probabilities and justifications, then aggregates their forecasts into a single probability and calibrates it.

To optimize and evaluate the pipeline, I collect a large dataset of binary questions from the worlds largest prediction market called Polymarket. The test set contains only questions published after all the models' knowledge cutoff dates, starting with August 2024 and ending in February 2025. This ensures LLMs do not know the outcomes of any events at the moments of forecasting. The dataset also contains the aggregated crowd predictions for various timestamps across the predictive horizon.

The implemented system makes an important step towards automated event forecasting. The system significantly improves on the previous work and simple baselines as measured by Brier Score and accuracy, standard forecasting metrics. It systematically outperforms the random baseline and the approach of standard prompting strategies.

To summarize my main contributions:

1. I propose a fully automated LLM-based system for forecasting, which shows significant improvements compared to baselines and existing results in the field.

2. I investigate the effects of RAG, ensembling and calibration on forecasting precision.
3. I create a large dataset containing most recent real-world forecasting questions with parsed and ranked by relevance news texts.

Chapter 2

Literature Review

Forecasting. Automated systems have recently started to be explored for the task of event forecasting. Shi et al. (2023) showed that LLMs can outperform the event sequence models due to their reasoning capabilities. They, however, do not use the news texts to better inform the LLM, and use the model for reasoning rather than forecasting. The model they produce, however, is domain-specific, and does not fully leverage the ability of LLMs to process vast amounts of textual data.

ForecastQA by Jin et al. (2021) was the first dataset for event forecasting, which covered a wide range of topics. However, it included some vague and unclear questions since they were written by non-experts. ForecastQA allows forecasting only at one point in time, so models cannot update predictions as new information comes in. These shortcomings motivated the creation of AutoCast by Zou et al. (2022). It was another dataset presented for event forecasting, which comprised questions from professional forecasting competitions held up to 2022. They build their own system and showed that machine learning algorithms could not compete with humans on the task of forecasting. The dataset includes a prepared list of news sources for each question, while the system presented in this thesis collects sources on the fly, making it suitable for real-time usage. Moreover, the newest models' knowledge cutoff dates have moved past 2023, making them unsuitable for evaluation on Autocast due to data leakage in the training process. In this work, the system is evaluated on a set of real-world questions from 2024 to 2025, allowing to apply the most recent LLMs.

Autocast++ by Yan et al. (2024) was the system that improved the original Autocast accuracy results by 8% on binary questions. Yan et al. built a system which used retrieval of the most relevant articles using zero-shot ranking, while my pipeline adopts a pre-trained text ranking model. On binary questions, they evaluate their system by accuracy, since it only predicts the binary outcomes instead of probabilities and justifications. For professional forecasters, Brier score is a more important metric, since it incentivizes calibration of the probabilities as well as accuracy. In this thesis, I use Brier score and accuracy as key metrics for the pipeline evaluation. More details on metrics are given in section 4.

Several studies show that LLMs can be applied to forecasting real-world events. Lopez-Lira and Tang (2023) found that ChatGPT, even without financial training or fine-tuning, could employ news headlines predict stock price movements and forecast next-day stock returns. Their model is domain-specific, and they, therefore, report metrics like Sharpe Ratio. Schoenegger and Park (2023) evaluate GPT-4 on a dataset of 23 questions from one forecasting tournament. They show that the model falls significantly behind human benchmarks on this task and attribute it to the fact that outcomes are not known beforehand, like, for example, in school exams. However, neither Lopez-Lira and Tang, nor Schoenegger and Park explore how LLMs perform when given extensive relevant news data. In this thesis I use a dataset totaling 5774 real-world binary questions, and test the system on 1000 questions, which allows for precise estimation.

Schoenegger et al. (2024b) conduct a research on aggregating LLMs forecasts on a dataset of real forecasting questions of 31 binary samples. They found that aggregating forecasts of multiple language models yields better prediction accuracy with respect to Brier score. They show that median aggregation improves the results compared to the baseline of individual models and sometimes outperforms human predictions. Their ensemble of 12 LLMs yields similar predictions to the aggregate of 925 human forecasters. In my research, I improve by testing my pipeline on a larger sample size and by adding RAG for informing LLMs.

A concept in forecasting psychology, described by Tetlock and Gardner (2015), differentiates between the “inside view” and the “outside view”. The inside view focuses on the specifics of a current case, and the outside view considers the base rate of similar past cases called reference classes. The principle of referring to them proved to be important in training human forecasters (Chang et al. (2016)). Karvetski et al. (2022) also found that using reference classes in reasoning leads to better forecasting. LLMs can potentially serve as generators of outside views, since they are trained on large amounts of text, they implicitly contain statistical information about event frequencies. In forecasting, researchers have started to use this tendency of LLMs. The prompting strategy of Schoenegger et al. (2024a) asks the LLM to produce a base rate and then adjust it. However, they do not report the calibration quality of base rates or how accurate they are.

Calibration in forecasting refers to the match between predicted probabilities and actual outcome frequencies. A well-calibrated model’s predictions should align with observed results. For example, if a model predicts a 70% chance for every event in a group, about 70% events from that group should occur in reality. Good calibration is important for decision-making and is rewarded by the Brier score. LLMs and other AI models can show overconfidence or underconfidence, requiring calibration. Logistic regression calibration is a standard post-processing method. More details on this calibration in 4. In deep learning, multiple calibration techniques are recognized, which improve raw probabilities to reflect the actual chances of events (Wang (2024)).

Retrieval. Real-world events require current information for accurate forecasting, whilst LLMs have static knowledge and lack access to recent news. Having access to recent data is crucial for high quality performance in forecasting competitions Tetlock and Gardner (2015). Retrieval-Augmented Generation (RAG) allows LLMs to fetch and use relevant external documents before generating a prediction. In forecasting, RAG helps provide real-time context. Several recent works design RAG pipelines for forecasting. Yang et al. (2024) proposed TimeRAG, which retrieves relevant time series segments from past data and provides them to an LLM, improving its forecasting.

In this work, retrieval is organized through querying the news dataset with the use of LLM-generated keywords. The retrieved articles' texts are assessed by relevance to the given question and fed to an ensemble of LLMs.

Prediction markets. Prediction markets are recognized for their ability to aggregate information and provide accurate forecasts of future events through crowd beliefs Wolfers and Zitzewitz (2004). These markets allow participants to buy and sell contracts that pay off depending on the outcome of future events. They, therefore, translate collective beliefs into prices that can approximate true probabilities. The aggregated beliefs of the crowd are considered to be a strong benchmark for forecasting, not yet surpassed by machine learning algorithms and automated systems (Schoenegger and Park (2023), Yan et al. (2024)). In this work, I compare the performance of my model to the crowd's performance by Brier Score and accuracy.

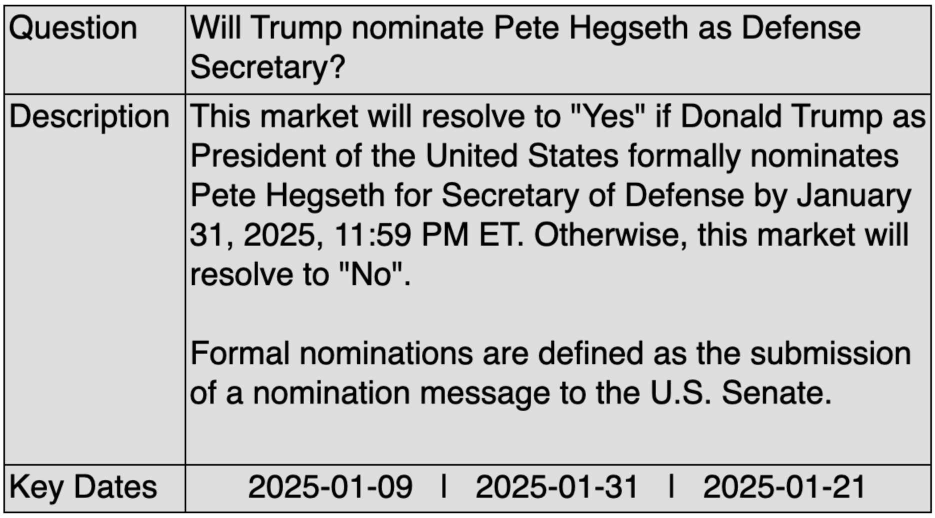
Chapter 3

Data

This section describes the data parsing and curation processes employed in this study. The primary data sources consist of prediction market questions and corresponding news articles, which are processed to form a structured dataset for subsequent analysis.

Polymarket. The core of my dataset is derived from Polymarket, one of the largest prediction markets. Polymarket was selected due to several advantageous characteristics: it provides an official API for data retrieval, stands as one of the largest prediction markets by volume and participation, and includes markets on events of significant interest. Importantly, the outcomes of Polymarket questions are determined by real-world events, independent of market participants' trading activities. The initial collection phase yielded 5,774 binary questions listed on the platform between January 1, 2024, and February 1, 2025. The average resolution time of a single question is 17.05 days. Each question includes its title, full description with resolution criteria, and the originally specified start and end dates of the question. Additionally, resolution date was collected for every question, to avoid data leakage and forecasting after the event had already resolved, which the model could find out about through fresh news. Each question is also supplemented with its outcome, binarized as 1 for "Yes" and 0 for "No". A sample question from the dataset is given in figure below.

Each Polymarket question was also enriched a list of relevant categories (e.g., "Politics", "Finance", "Technology") and geographical locations relevant to it. These were obtained by prompting an LLM with the question's title and description and parsing its structured output. Categorical data allows to create features for aggregation algorithms described in 4. They also allow to see how well models perform across different categories. Geographical data allows to see which locations are mostly covered in the dataset. The dataset contains mostly questions which are related to the United States (for a map, see appendix 7.1). Polymarket does not allow participants from the US to make bets on their prediction market.



Question	Will Trump nominate Pete Hegseth as Defense Secretary?
Description	This market will resolve to "Yes" if Donald Trump as President of the United States formally nominates Pete Hegseth for Secretary of Defense by January 31, 2025, 11:59 PM ET. Otherwise, this market will resolve to "No". Formal nominations are defined as the submission of a nomination message to the U.S. Senate.
Key Dates	2025-01-09 I 2025-01-31 I 2025-01-21

Figure 3.1: **Sample question:** in a dataset, it contains start date, end date and resolution date. Apart from that, description text with resolution criteria is provided.

GDELT. To supplement these prediction market questions with real-world information, we gathered relevant news articles using the Global Database of Events, Language, and Tone (GDELT). GDELT was chosen because it offers free access to a vast, continuously updated repository of global news from diverse sources. Its update rate is 15 minutes, so it is suitable for timely information retrieval. Unlike some event-focused databases, GDELT provides links to original news articles, which is important for accessing the full text needed for LLM analysis. The process for associating news with Polymarket questions, inspired by Halawi et al. (2024), involved first extracting relevant keywords with an LLM. For each Polymarket question, its title and description were provided as input to the GPT-4o-mini language model, which was prompted to identify key terms and named entities in the question. These keywords, after parsing, were used to query the GDELT API. The dates for retrieval varied depending on the question duration: news from the preceding 1 day was retrieved for events lasting under 48 hours, 2 days for events lasting 2-5 days, 3 days for events lasting 6-14 days, and 5 days for events with longer durations. Retrieving news from before the question start date ensures the information reflects context that existed prior to the market's creation. This is important because Polymarket questions are often triggered by real-world events, and early news coverage provides background that likely influenced the market's formulation. For each query, up to 75 articles were retrieved, with the publication date of each article recorded to ensure fairness in experiments involving different retrieval dates.

Retrieved news articles underwent a systematic curation process. Initially, articles were parsed using the Newspaper3k library to extract the main text, not only the titles. Only articles in English were left, since multilingual content lead to inconsistencies in LLM output formatting during early research stages. The extracted texts were then cleaned to simplify the input for LLMs. This cleaning involved several steps: **(i)** whitespace normalization; **(ii)** removal of residual HTML tags and artifacts not fully processed by the parsing library; **(iii)** Unicode normalization, replacing unusual symbols. Finally, duplicate articles were removed. This was achieved by computing a hash of the cleaned text content for each article and removing articles with the same hash. This ensured that news fed to the model was unique.

Evaluation subsample. From the initial set of 5,774 Polymarket questions, a test set of 1,000 questions was constructed for the evaluation experiments. This was necessary due to budgetary constraints, as large-scale evaluations across multiple LLMs require substantial funding. The LLMs were queried via APIs such as OpenRouter and the OpenAI API. The test set comprises questions whose resolution dates fall between August 1, 2024, and February 1, 2025. These 1,000 questions were selected by randomly sampling from all available markets in the collected dataset that resolved after August 1, 2024.

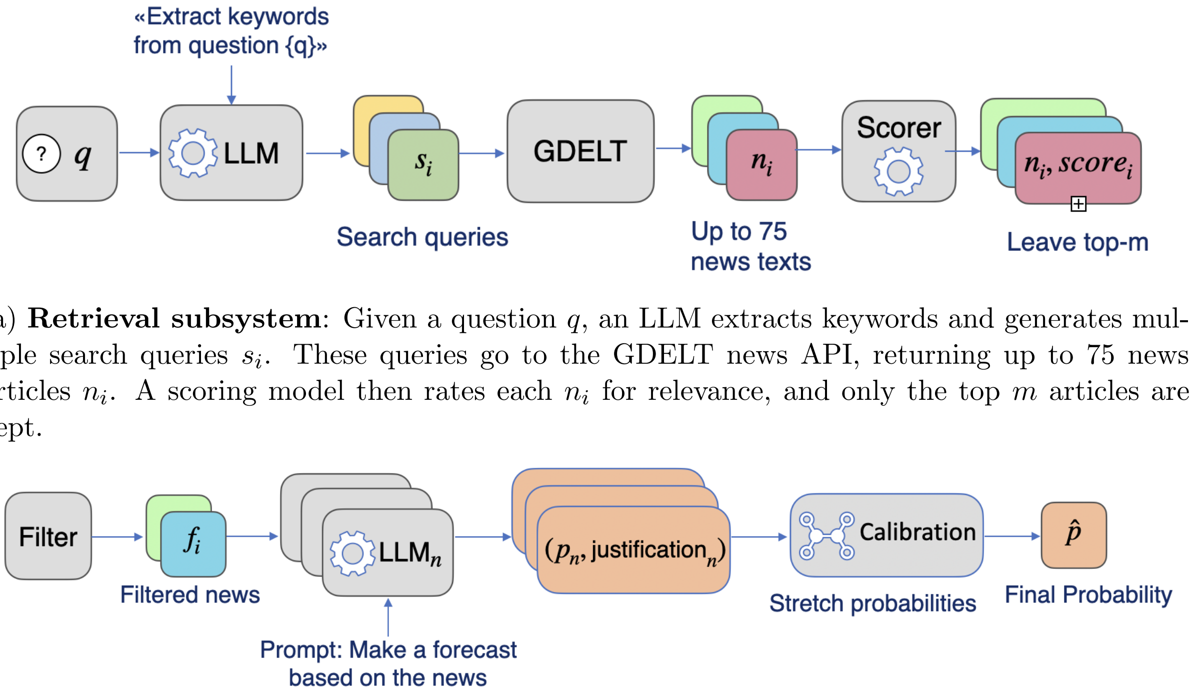
To assess the representativeness of this 1,000-question test set, I compared its characteristics to the broader population of Polymarket questions resolving in the same period (August 1, 2024, to February 1, 2025). First, the proportion of questions that resolved to "Yes" was examined: in the overall dataset of 5,774 questions (which covers a wider range of resolution dates), 42.32% resolved to "Yes", while in the 1,000-question test set, this figure was 42.5%. Z-test for proportions finds no statistically significant difference in the two samples with $p_{\text{value}} = 0.91$. Furthermore, to check for representativeness in terms of topics, the distribution of questions across the LLM-generated categories in the test set was compared against the category distribution of all questions from the dataset resolving within the August 1, 2024, to February 1, 2025, timeframe. Results of the Pearson chi-squared test indicate no statistically significant difference between the category distributions with a $p_{\text{value}} = 0.79$, supporting the test set's representativeness. Distributions over topics on prediction markets can naturally evolve over time, so comparison is done on the same period.

Finally, for the development of a trainable aggregation model, this 1,000-question test set was further partitioned by the date of question resolution. Specifically, 30% of questions for training the aggregator, 10% for validating its hyperparameters (learning rate), and the remaining 60% were reserved as a final hold-out set to evaluate the performance of the complete forecasting system incorporating the trained aggregator.

Chapter 4

Methodology

This section describes the pipeline of the final system, the models used at different stages of this system, post-processing techniques of the forecasts from LLMs, experimental setups and evaluation metrics. The system diagram is given in the figure below.



(a) **Retrieval subsystem**: Given a question q, an LLM extracts keywords and generates multiple search queries s_i. These queries go to the GDELT news API, returning up to 75 news articles n_i. A scoring model then rates each n_i for relevance, and only the top m articles are kept.

(b) **Reasoning subsystem**: The filtered news summaries f_i are fed into several LLMs, each of which outputs a probability p_n and justification for the forecast. Those raw probabilities are passed through a calibration module that "stretches" them into a final, well-calibrated probability $\hat{p}$.

Figure 4.1: Overview of the pipeline. Retrieval and Reasoning subsystems.

The pipeline begins with an information retrieval stage designed to supply relevant context for each forecasting question. For each question, key terms and named entities are first extracted from its title and description using the GPT-4o-mini model. These keywords are then used to query the GDELT API, a comprehensive and continuously updated repository of global news articles. The time window for news retrieval is dynamically adjusted based on the question's duration, as described in the Data section.

Simple keyword-based retrieval can yield noisy or irrelevant results. To tackle this issue, a second-stage re-ranking process is employed to identify the most relevant articles. At this step, I use the Cross-Encoder for MS Macro from HuggingFace, a lightweight yet powerful text-ranking model. As a cross-encoder, it processes the prediction market question and each candidate news article simultaneously. It produces relevance scores for the articles' texts. The model was chosen for its specific training on the MS MARCO passage ranking dataset, which is designed for a similar task of finding relevant text snippets for a given query. This makes it well-suited for filtering the most informative articles to serve as context for the LLMs. The top m ranked articles, with m being a hyperparameter tested at values of 5, 10, and 15, are then passed to the next stage.

An important part of the forecasting process involves prompting a diverse set of eight LLMs: DeepSeek-R1, DeepSeek-V3, Mistral-3, Gemini Flash, GPT-4.1-mini, GPT-4o-mini, Claude Haiku, and Llama-4. For each question, these models are provided with the question's details and the curated news context. Predictions are extracted using several prompting strategies, including a basic prompt for a direct probability estimate, and more complex strategies such as the "CHAMPS KNOW" and "two stages" prompts, which are designed to structure the model's reasoning process. The "two stages" prompt, first asks the model for an "outside view". The model therefore estimates the base rate of similar events. The inside view forecast takes the base rate and adjusts it using the relevant news. The "CHAMPS KNOW" prompt relies heavily on the research by Chang et al. (2016). It stimulates the model to employ advanced forecasting rationales from human superforecasters training.

To analyze the impact of information accumulation over time, predictions for each question are generated at four distinct temporal snapshots, calculated as

$$t_k = t_{\text{start}} + (t_{\text{end}} - t_{\text{start}} - 1) \cdot \frac{k}{4}$$

for $k \in \{1, 2, 3, 4\}$, where t_{start} is the question's start date and t_{end} is the minimum of resolution date and initial end date, i.e. $t_{\text{end}} = \min(\text{end}, \text{ resolution})$. The main evaluation of interest is done at $k = 3$, which is on average 4.26 days prior to the event actual resolution date. The results section suggests that adding new information over time significantly boosts performance.

The raw probabilistic outputs from the LLMs are then subject to aggregation and calibration to enhance their accuracy and reliability. Several non-trainable aggregation methods are tested, including the arithmetic mean, median, and a trimmed mean. The latter removes the single highest and lowest predictions to improve robustness to outliers.

In addition to non-trainable ensembles, a trainable aggregation pipeline was evaluated. Since model forecasts were available on only 1,000 questions (split by time into 30% train, 10% validation, 60% test), logistic regression was chosen to mitigate overfitting. For each depth $k \in \{1, 2, 3\}$, the forecasts of a single model at stages p_1, p_2, p_3 are aggregated as

$$P\big(Y = 1 \mid p\big) \;=\; \sigma\big(\beta_0 + \beta_1 p_1 + \beta_2 p_2 + \beta_3 p_3\big), \quad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

Similarly, forecasts from n different models $\{p_1, \ldots, p_n\}$ are combined as

$$P\big(Y = 1 \mid p_1, \ldots, p_n\big) \;=\; \sigma\Big(\beta_0 + \sum_{i=1}^{n} \beta_i p_i\Big).$$

To enrich the feature set, one-hot encoded categorical variables from categories $x_1, \ldots, x_m$ (e.g., “Politics”, “Finance”) are appended. The resulting logistic-regression model is

$$P\big(Y = 1 \mid p_1, \ldots, p_n, x_1, \ldots, x_m\big) \;=\; \sigma\Big(\beta_0 + \sum_{i=1}^{n} \beta_i p_i + \sum_{j=1}^{m} \gamma_j x_j\Big).$$

Hyperparameters are tuned on the 10% validation set, and final evaluation uses the 60% hold-out.

For the non-trainable aggregations, calibration is applied to individual or aggregated raw forecasts via

$$P(Y = 1 \mid p) \;=\; \frac{1}{1 + \exp\big(-(\beta_0 + \beta_1 p)\big)},$$

where p is the uncalibrated probability.

The performance of all models and configurations is evaluated using two standard metrics. The Brier score measures the mean squared error between predicted probabilities and actual outcomes, providing a comprehensive assessment of both calibration and resolution. It is calculated as

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2$$

where p_i is the predicted probability for event i and o_i is its binary outcome (1 for “Yes”, 0 for “No”). The uninformed random baseline for this metric is 0.25. It is obtained by randomly guessing the outcome of each question. Accuracy is used to measure binary classification performance at a 0.5 probability threshold.

Chapter 5

Results and Discussion

This section summarizes and interprets the primary empirical findings from the evaluation of different LLM-based event forecasting pipelines presented in this thesis. Emphasis is placed on the relative performance of individual models, aggregation strategies, and calibration approaches. The results are presented in a compact form through two main tables reporting Brier score and accuracy. Figures for calibration and temporal dynamics supplement the analysis.

Table 5.1 presents the main quantitative results for the best individual language models, as well as the performance of ensemble aggregation methods, both before and after the application of logistic regression calibration. The baselines for comparison are **(1)** the uninformative random guess with Brier score of 0.25; accuracy of 0.5, and **(2)** the results of prompting without news from Schoenegger and Park (2023). I run DeepSeekR1, prompted with their prompt, on the same set of 1000 questions I evaluate my models on to properly compare the results. These results are listed for the moment of retrieval corresponding to $k = 3$ out of 4. This is on average 4.26 days before resolution with the mean question duration being 17.05 days. Ensembling is done over DeepSeekR1 at $k = 3$ and $k = 2$, DeepSeek V3 at $k = 3$ and $k = 2$, Mistral-3 at $k = 3$ and Gemini-Flash at $k = 3$. Detailed results across all models and prediction dates are given in appendix 7.1.

As shown above, the DeepSeek R1 stands out as the strongest individual model, achieving a Brier score of 0.184 and a raw accuracy of 69.9%. Temporal analysis confirms this progressive improvement, with Brier scores for DeepSeek R1 improving monotonically over time (from $k = 1$ to $k = 4$), as detailed in Appendix 7.1 and visualized in Figure 7.3.

Ensemble aggregation leads to consistent improvements beyond the best single model. The mean and trimmed mean aggregation methods, when applied across predictions from diverse models and previous time horizons, yield the best raw Brier scores of 0.182. Further applying logistic regression calibration to aggregates produces the lowest observed Brier score of 0.178 with accuracy reaching 72.1%. This is a substantial improvement over the best individual model and even more significant improvement over the uninformed baseline. These results confirming the value of combining model diversity with

Table 5.1: Forecasting performance of models and aggregations.

Model / Method	**Raw**		**Calibrated**	
	Brier ↓	Acc. ↑	Brier ↓	Acc. ↑
DeepSeek R1 (best indiv.)	**0.184**	**0.699**	**0.184**	**0.713**
DeepSeek V3	0.194	0.687	0.191	0.689
Mistral-3	0.201	0.653	0.98	0.655
Gemini Flash	0.205	0.651	0.199	0.664
Ensemble, trimmed mean	**0.182**	0.710	**0.178**	**0.721**
Ensemble, mean	**0.182**	0.718	**0.178**	0.718
Ensemble, median	**0.182**	0.704	0.179	0.716
Ensemble, trainable	0.216	0.634	—	—
Uniform baseline	0.250	0.500	0.25	0.5
Schoenegger & Park prompt	0.208	0.634	—	—

statistical post-processing. This result is also significantly better than the uninformed prompting approach by Schoenegger & Park. These results can also be compared to the ones reported in Yan et al. (2024) by the metric of accuracy. They report the accuracy of 67.9% on a different set of questions, but their questions also concern real world events. The improvement after calibration suggests that ensemble predictions are systematically underconfident, a bias that calibration corrects successfully.

Aggregation leads to more accurate predictions because different models have different strengths and tend to make different errors. This is demonstrated in the table in appendix 7.2. Some models are be better at interpreting certain topics or types of events, others may systematically over- or underestimate probabilities for specific questions. When forecasts are averaged, especially through trimming, the extreme predictions of individual models are offset by the opinions of others. This helps to correct for systematic mistakes and reduces the likelihood that a single model's error will dominate the final result. In addition, aggregation decreases the overall noise and improves the stability of predictions. This results in better accuracy and calibration compared to any single model.

Table 5.2 isolates results for the two-stage prompting approach, which separates the estimation of base rates (outside view) from news-informed forecasts (inside view). The analysis focuses on predictions from DeepSeek R1 at $k = 3$.

Table 5.2: Comparison of predictions using outside and inside view approaches (DeepSeek R1, k=3).

Prediction Mode	Brier ↓	Acc. ↑
Outside View (base rates)	0.210	0.630
Enhanced Inside View	0.202	0.677
Standard Inside View (news only)	**0.184**	**0.699**

Contrary to expectations, incorporating the predicted base rates does not improve performance. Base rate forecasts are roughly calibrated (see 7.2), they suffer from low informativeness. This results in both lower accuracy and higher Brier scores relative to news-based predictions. The combination of base rates with news-informed inside view only marginally outperforms pure outside view, and remains inferior to the standard inside view pipeline. The events from dataset are rare and difficult to be forecasted by base rates, which is why base rates might mislead the model even with news.

Calibration curve analysis offers further results. Figure 7.4 demonstrates that the best individual models (DeepSeek R1, DeepSeekV3) are naturally well-calibrated, with only minor overconfidence in the high-probability range. By contrast, ensemble aggregates show systematic overconfidence, particularly in the 60-80% range, which is effectively corrected with logistic regression calibration. Weak baselines, such as GPT-4.1-mini and GPT-4o-mini, are sharply overconfident, a pattern matching their comparatively poor Brier and accuracy scores.

Additional ablation experiments reveal that neither increasing the number of retrieved news articles beyond a moderate threshold nor using more intricate prompt formats produces meaningful improvements (details in Appendix 7.1). Similarly, attempts at feature-based or learned aggregation methods do not reliably outperform simpler ensembling methods like mean, median and trimmed mean, presumably due to the small available sample and high event heterogeneity.

To summarize, these results demonstrate that **(1)** careful aggregation and calibration of state-of-the-art LLMs outperform single-model and naive baselines in event forecasting, and **(2)** incorporating timely news content remains critical, while base rate supplementation appears less effective. These findings show that advanced LLM pipelines can be useful for automated forecasting, but also point to their current limitations and suggest areas for future improvement.

Chapter 6

Conclusion and Future Work

In this thesis, I designed, implemented, and evaluated an optimized multi-stage pipeline for LLM-based forecasting of binary real-world events. The research systematically assessed the impact of base LLM capabilities, Retrieval Augmented Generation, ensemble methods, and post-hoc calibration on 1000 Polymarket questions. The optimal pipeline was incorporating RAG for timely news context, trimmed mean aggregation of predictions from diverse LLM configurations, and subsequent Logistic Regression calibration of ensemble outputs achieved a Brier score of 0.178. This performance significantly surpassed individual LLMs and uncalibrated ensembles.

The key findings are that RAG, aggregation and calibration significantly improve forecasting accuracy. Apart from that, larger LLMs generally exhibit stronger baseline forecasting capabilities. RAG-informed "inside views" from current news proved more effective than LLM-generated "outside view" base rates. Furthermore, ensemble methods, such as mean aggregation, improve forecast robustness.

The established pipeline emphasizes a systematic approach: selecting strong base models, enriching inputs via RAG with current information, robustly aggregating diverse outputs, and calibrating ensemble predictions. Building upon these findings, several directions for future research can further advance LLM-based forecasting.

First, any forecasting algorithms can be used for trading on platforms like Polymarket. One of directions for future research could include profitability test, which can be performed on binary questions (Sethi et al. (2022)).

Data enhancement and diversification can be considered for stronger results on larger samples. Expanding the dataset beyond 1000 questions to include a larger volume and wider variety of event types. Sourcing questions from multiple platforms and encompassing diverse formats (for example, multi-outcome, continuous questions) will test and improve pipeline generalizability.

Methodological Refinements is another direction to look at. Continuously integrating and evaluating newer, more capable LLM architectures as they emerge. One of the directions for future research might be developing domain-specialized agents by fine-tuning

LLMs on forecasting data from special categories. Investigating more sophisticated RAG techniques, such as fine-tuned retrievers for news, graph-based knowledge retrieval, or advanced hybrid search systems. Examination of LLM capabilities for extracting structured event data or key predictive features from news articles to serve as additional model inputs.

Pursuing these directions will build upon this thesis's foundations, further advancing the potential of LLMs to deliver accurate, reliable, and timely forecasts.

References

Welton Chang, Eva Chen, Barbara Mellers, and Philip Tetlock. Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, 11(5):509–526, 2016. doi: 10.1017/S1930297500004599.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models, 2024. URL `https://arxiv.org/abs/2402.18563`.

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, 2021. URL `https://arxiv.org/abs/2005.00792`.

Christopher W. Karvetski, Carolyn Meinel, Daniel T. Maxwell, Yunzi Lu, Barbara A. Mellers, and Philip E. Tetlock. What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? *International Journal of Forecasting*, 38(2): 688–704, 2022. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2021.09.003. URL `https://www.sciencedirect.com/science/article/pii/S0169207021001473`.

Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *SSRN Electronic Journal*, 2023. doi: 10.2139/ssrn.4412788. URL `https://ssrn.com/abstract=4412788`.

Philipp Schoenegger and Peter S. Park. Large language model prediction capabilities: Evidence from a real-world forecasting tournament, 2023. URL `https://arxiv.org/abs/2310.13014`.

Philipp Schoenegger, Peter S. Park, Ezra Karger, Sean Trott, and Philip E. Tetlock. Ai-augmented predictions: Llm assistants improve human forecasting accuracy, 2024a. URL `https://arxiv.org/abs/2402.07862`.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, and Philip E. Tetlock. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy, 2024b. URL `https://arxiv.org/abs/2402.19379`.

Rajiv Sethi, Julie Seager, Emily Cai, Daniel Benjamin, Fred Morstatter, Olivia Bobrownicki, Yuqi Cheng, Anushka Kumar, and Anusha Wanganoo. Evaluating prediction mechanisms: A profitability test. In *CI '24: Proceedings of The ACM Collective Intelligence Conference*, January 2022. Available at SSRN: https://ssrn.com/abstract=3767544 or http://dx.doi.org/10.2139/ssrn.3767544.

Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Y. Zhang, Jun Zhou, Chenhao Tan, and Hongyuan Mei. Language models can improve event prediction by few-shot abductive reasoning, 2023. URL https://arxiv.org/abs/2305.16646.

P. Tetlock and D. Gardner. *Superforecasting: The Art and Science of Prediction*. Random House, 2015. ISBN 9781448166596. URL https://books.google.ru/books?id=450mCQAAQBAJ.

Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art, 2024. URL https://arxiv.org/abs/2308.01222.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, June 2004. doi: 10.1257/0895330041371321. URL https://www.aeaweb.org/articles?id=10.1257/0895330041371321.

Qi Yan, Raihan Seraj, Jiawei He, Lili Meng, and Tristan Sylvain. Autocast++: Enhancing world event prediction with zero-shot ranking-based context retrieval, 2024. URL https://arxiv.org/abs/2310.01880.

Silin Yang, Dong Wang, Haoqi Zheng, and Ruochun Jin. Timerag: Boosting llm time series forecasting via retrieval-augmented generation, 2024. URL https://arxiv.org/abs/2412.16643.

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, 2022. URL https://arxiv.org/abs/2206.15474.

Chapter 7

Appendix

Table 7.1: Aggregated table with individual models' results across 1000 questions test sample

setup (model — prompt type — number of articles — k)	brier	accuracy
deepseek_r1 — basic_prompt — 10 articles — 4	0.180035	0.712425
deepseek_r1 — basic_prompt — 10 articles — 3	0.185012	0.699000
deepseek_v3 — champs_know — 15 articles — 4	0.187841	0.691000
deepseek_r1 — basic_prompt — 10 articles — 2	0.192241	0.696697
deepseek_v3 — champs_know — 15 articles — 3	0.193553	0.687000
mistral_3 — basic_prompt — 10 articles — 4	0.196263	0.667335
deepseek_v3 — champs_know — 15 articles — 2	0.196573	0.678715
gemini_flash — basic_prompt — 10 articles — 4	0.198881	0.689000
gemini_flash — basic_prompt — 10 articles — 3	0.201011	0.658000
mistral_3 — basic_prompt — 10 articles — 3	0.201207	0.654000
deepseek_r1 — two_stages — 10 articles — 3	0.202184	0.676677
deepseek_r1 — basic_prompt — 10 articles — 1	0.202382	0.664665
deepseek_v3 — champs_know — 15 articles — 1	0.204329	0.663655
gpt_4.1_mini — basic_prompt — 5 articles — 4	0.212308	0.660508
gpt_4o_mini — basic_prompt — 15 articles — 4	0.213563	0.676471
gpt_4o_mini — basic_prompt — 10 articles — 4	0.215991	0.666667
gpt_4o_mini — champs_know — 10 articles — 4	0.218002	0.658427
gpt_4.1_mini — champs_know — 5 articles — 4	0.220064	0.652174
gpt_4.1_mini — basic_prompt — 5 articles — 3	0.220565	0.654891
gpt_4o_mini — basic_prompt — 5 articles — 4	0.221788	0.642032
claude_haiku — basic_prompt — 10 articles — 3	0.223624	0.617225

Continued on next page

Table 7.1: Table with all the results across models and aggregations

setup	brier	accuracy
gpt_4.1_mini — basic_prompt — 15 articles — 4	0.224792	0.650575
gpt_4o_mini — champs_know — 15 articles — 3	0.225272	0.649457
gpt_4o_mini — champs_know — 5 articles — 4	0.228564	0.635135
gpt_4.1_mini — basic_prompt — 10 articles — 3	0.228721	0.634146
gpt_4o_mini — champs_know — 5 articles — 3	0.229527	0.642667
gpt_4.1_mini — champs_know — 10 articles — 4	0.230366	0.624742
gpt_4o_mini — champs_know — 15 articles — 4	0.233250	0.634884
gpt_4o_mini — champs_know — 10 articles — 3	0.233315	0.619178
gpt_4.1_mini — basic_prompt — 15 articles — 3	0.234741	0.635870
gpt_4o_mini — champs_know — 5 articles — 2	0.235528	0.608142
gpt_4o_mini — champs_know — 10 articles — 2	0.236543	0.599490
claude_haiku — basic_prompt — 10 articles — 4	0.236874	0.604828
gpt_4o_mini — champs_know — 15 articles — 2	0.238175	0.624679
gpt_4.1_mini — basic_prompt — 15 articles — 2	0.238420	0.624679
gpt_4.1_mini — champs_know — 15 articles — 3	0.239319	0.606742
gpt_4.1_mini — champs_know — 15 articles — 4	0.239598	0.619590
gpt_4.1_mini — basic_prompt — 5 articles — 2	0.240073	0.613232
gpt_4o_mini — champs_know — 5 articles — 1	0.240581	0.587500
gpt_4o_mini — champs_know — 15 articles — 1	0.241218	0.601010
gpt_4.1_mini — basic_prompt — 10 articles — 1	0.242045	0.587500
gpt_4.1_mini — basic_prompt — 15 articles — 1	0.242091	0.594458
gpt_4o_mini — basic_prompt — 5 articles — 3	0.243113	0.577465
gpt_4o_mini — basic_prompt — 15 articles — 3	0.243160	0.595506
gpt_4.1_mini — basic_prompt — 5 articles — 1	0.243808	0.590452
gpt_4o_mini — basic_prompt — 10 articles — 3	0.244770	0.604396
gpt_4.1_mini — basic_prompt — 10 articles — 4	0.245540	0.590389
gpt_4.1_mini — champs_know — 5 articles — 1	0.246678	0.590909
gpt_4.1_mini — basic_prompt — 10 articles — 2	0.247285	0.583756
gpt_4o_mini — basic_prompt — 10 articles — 2	0.247833	0.571795
gpt_4.1_mini — champs_know — 10 articles — 3	0.247906	0.601637
gpt_4o_mini — basic_prompt — 15 articles — 1	0.248373	0.580402
gpt_4o_mini — basic_prompt — 5 articles — 2	0.250543	0.562176
gpt_4.1_mini — champs_know — 15 articles — 1	0.251137	0.607143
gpt_4o_mini — champs_know — 10 articles — 1	0.251644	0.544081

Continued on next page

Table 7.1: Table with all the results across models and aggregations

setup	brier	accuracy
gpt_4.1_mini — champs_know — 5 articles — 3	0.251732	0.591667
gpt_4o_mini — basic_prompt — 10 articles — 1	0.254248	0.571429
gpt_4.1_mini — champs_know — 10 articles — 2	0.254869	0.577320
gpt_4.1_mini — champs_know — 5 articles — 2	0.255707	0.583979
gpt_4.1_mini — champs_know — 10 articles — 1	0.257945	0.567089
gpt_4o_mini — basic_prompt — 15 articles — 2	0.259158	0.559585
gpt_4.1_mini — champs_know — 15 articles — 2	0.259189	0.573298
gpt_4o_mini — basic_prompt — 5 articles — 1	0.263264	0.532828

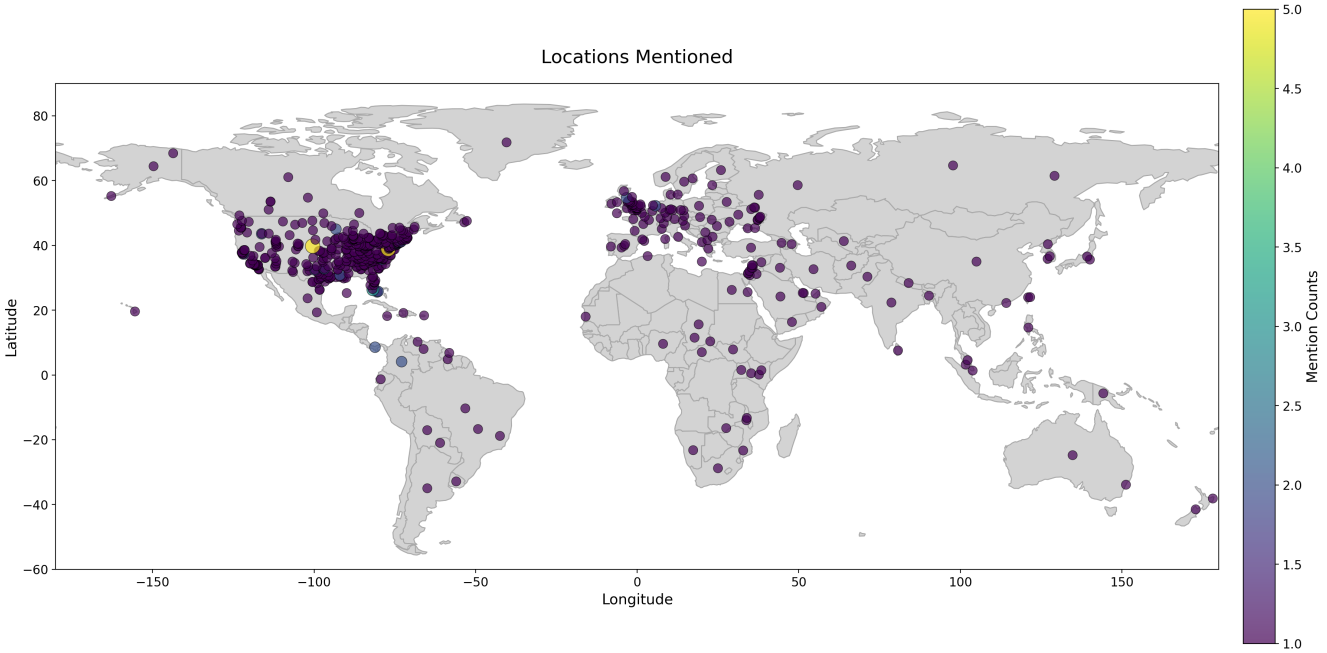


Figure 7.1: **Locations distribution across dataset**: We can see that locations mostly cover the United States.

Table 7.2: Brier scores by setup and category

category	DeepSeek R1	"Inside view"	DeepSeek V3	Gemini Flash	Mistral 3
Business	0.090439	0.093254	0.107807	0.084389	0.100789
Crypto	0.120072	0.138955	0.123317	0.125219	0.118732
Culture	0.092544	0.116189	0.120879	0.121191	0.137196
Economics	0.198542	0.135208	0.120000	0.128333	0.211875

Continued on next page

Table 7.2: Brier scores by setup and category

category	DeepSeek R1	"Inside view"	DeepSeek V3	Gemini Flash	Mistral 3
Education	0.016250	0.025000	0.012500	0.002500	0.006250
Election	0.123907	0.132527	0.119297	0.107473	0.120253
Environment	0.360000	0.202500	0.490000	0.490000	0.090000
Finance	0.129491	0.132963	0.134028	0.135650	0.139954
Geopolitics	0.152400	0.136230	0.119900	0.158506	0.183650
Healthcare	0.040833	0.190833	0.082500	0.005433	0.032500
Law	0.041667	0.135000	0.082500	0.009167	0.045000
Mentions	0.028333	0.052933	0.062500	0.116422	0.075844
Military	0.130143	0.105614	0.120857	0.165646	0.189786
Music	0.022500	0.640000	0.562500	0.722500	0.490000
Other	0.102917	0.126183	0.089653	0.124106	0.126256
Politics	0.124061	0.133767	0.122072	0.110592	0.126178
Research	0.289000	0.137000	0.235500	0.230000	0.262000
Science	0.128056	0.119444	0.127778	0.089722	0.112500
Sports	0.218960	0.240870	0.232485	0.244787	0.238384
Technology	0.158981	0.144911	0.137552	0.108148	0.115648
War	0.167407	0.146167	0.146296	0.214259	0.235278

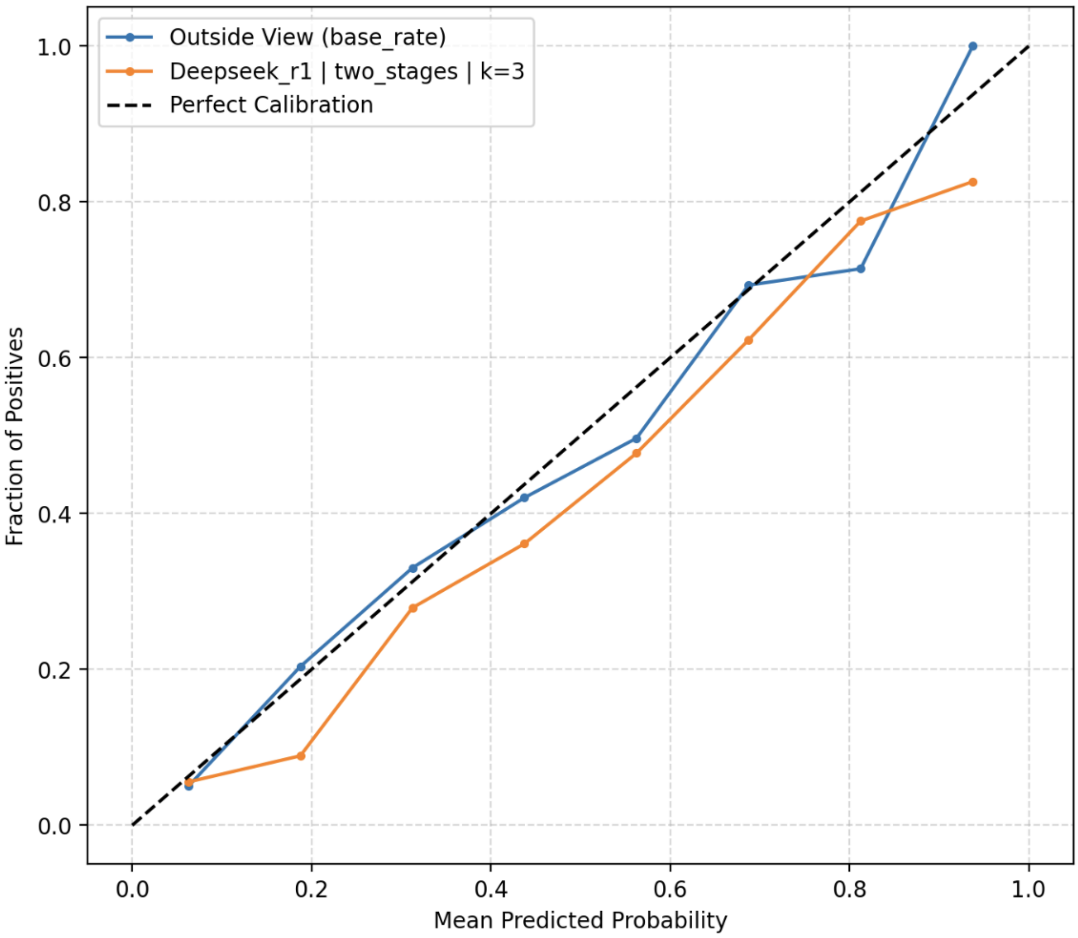


Figure 7.2: **Calibration curve for inside & outside view**: Forecasts are fairly well calibrated.

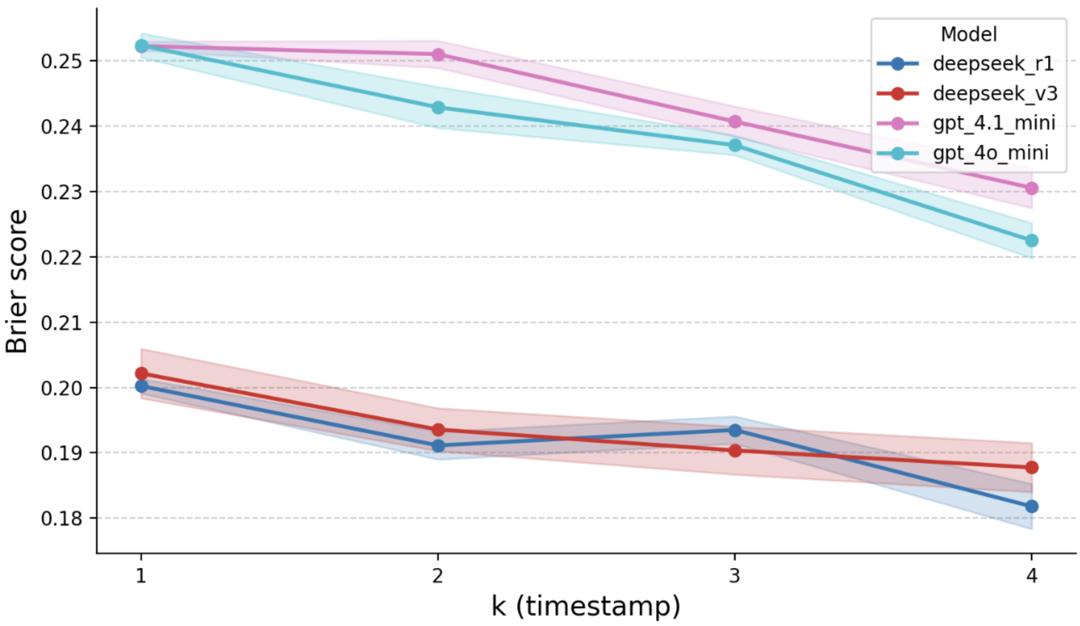


Figure 7.3: **Prediction accuracy over time for best and worst single models: mini GPTs and DeepSeeks**

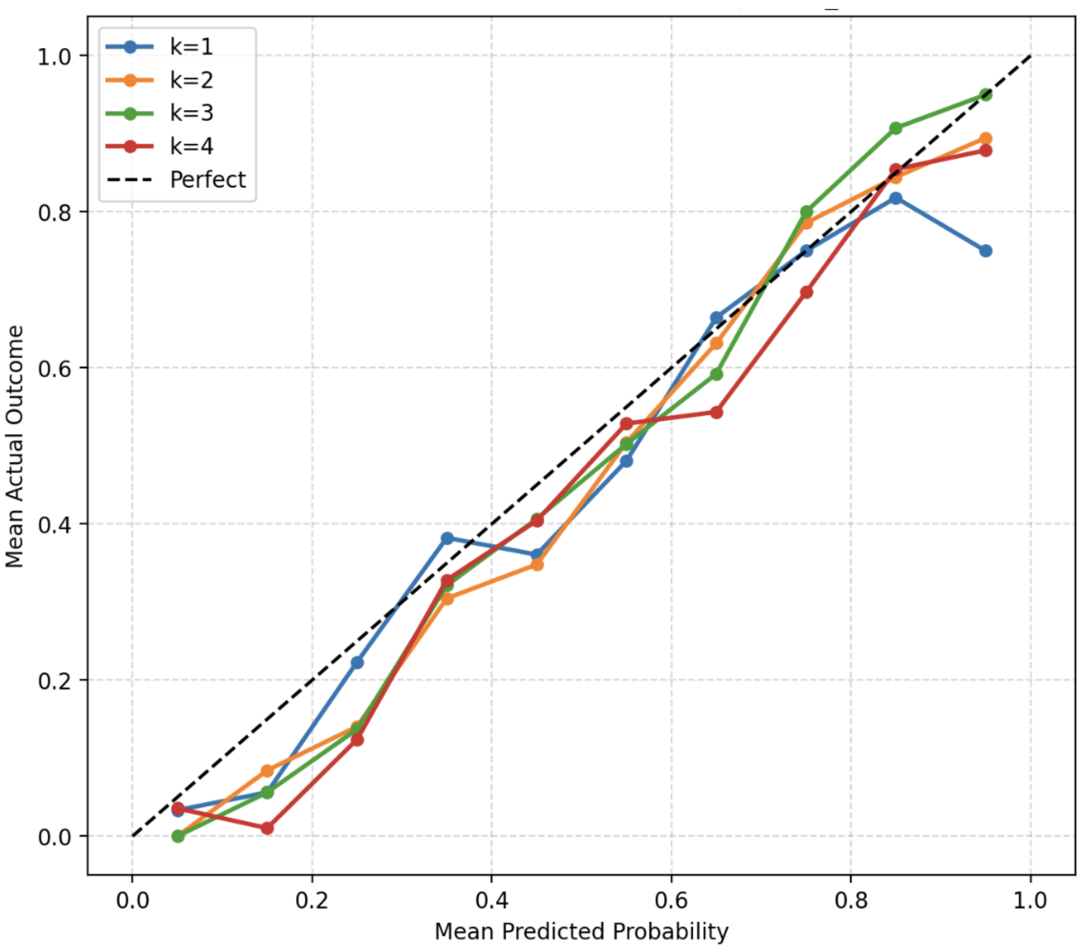


Figure 7.4: **DeepSeek R1 is naturally well calibrated**

Comparison	95% CI	99% CI
Best vs DeepSeek R1	$(-0.016566, 0.002685)$	$(-0.022948, 0.005525)$
Best vs DeepSeek V3	$(-0.026273, -0.006122)$	$(-0.032750, -0.002980)$
Best vs Gemini Flash	$(-0.038296, -0.014241)$	$(-0.047781, -0.010599)$
Best vs Mistral-3	$(-0.039643, -0.016786)$	$(-0.047298, -0.013113)$