Performance of Large Language Models in Prediction Markets

Author: Ketill Hugi Halldórsson

Date: 13/02/2026

TU Delft

Performance of Large Language Models in Prediction Markets

In fulfilment of the requirements for the master's program

Management of Technology

at the Delft University of Technology

By

Ketill Hugi Halldórsson

Student number: 6079881

To be defended publicly on Friday, February 27, 2025, at 14:00.

Graduation committee

Chairperson	: Dr. A.Y. Ding
First Supervisor	: Dr. LL.M S. Renes
Second Supervisor	: Dr. A.Y. Ding
Advisor	: Dr. Ir. R. van Bergem

Executive Summary

Decision-making under uncertainty often relies on access to accurate probabilistic forecasts. In many contexts, such forecasts are scarce or difficult to obtain. Decentralised predictionmarkets are widely regarded as effective tools to aggregate dispersed information that decision-makers can use as forecasts to make better decisions about an uncertain future. Recently, there have been major advances in large language models that have led to claims that large language models could complement or replace market-based forecasting by synthesising information without the need for incentive-driven markets. However, there is limited empirical evidence comparing forecasts generated by large language models to market-based aggregation under real-world conditions. This thesis puts these claims to the test by examining the extent to which large language models can replicate or complement human forecasting as reflected in decentralised prediction markets. Using Polymarket as a benchmark for collective human forecasting, probability forecasts generated by large language models are compared to live market probabilities. Forecasting performance is evaluated across different market conditions, and the decision-making relevance of forecasts generated by large language models is evaluated through trading simulations. The results show that market probabilities are consistently more accurate than the forecasts generated by large language models in terms of predictive accuracy. The findings hold across all evaluated models, model combinations, prompting strategies, market stages, and liquidity levels of markets. A regression-based aggregation model that mixes market probabilities and large language model forecasts achieves predictive performance comparable to that of the market in some cases, but it fails to generalise when put to the test under realistic conditions. The findings suggest that large language models at their current stage cannot substitute prediction markets as information aggregation mechanisms. The results challenge claims that large language models can replicate the performance of prediction markets in the generation of accurate probabilistic forecasts. The results highlight the need for caution when deploying large language models in the context of high-stakes decision-making.

Acknowledgments

I would first like to thank Sander Renes for his supervision and for inspiring the idea that led to this thesis. I left each of our meetings with renewed motivation, a clearer vision of the goal, and a better understanding of how to reach it. His guidance helped me stay focused whenever I lost direction.

I am also grateful to Aaron Ding for his supervision and guidance through this process. His input at every major milestone strengthened the quality of this work, and his encouragement to seek help when needed was invaluable.

Furthermore, I would like to thank my great advisor, Rutger van Bergem, for his great advice, particularly during the early ideation phase. Seeing him around the faculty building was always a pleasure, as were the coffees we shared.

Furthermore, I would like to thank Rob Caulk from AskNews for providing me with AskNewsAPI, a great software that came in good use in the data collection phase.

Lastly, I am deeply grateful to my mother, father, and sister for their constant support and for listening patiently as I spoke endlessly about my research and its challenges. I also thank my brother Krummi for encouraging me to take breaks and step away from the research to go on walks with him. Those breaks were truly special, and I will always cherish the moments we shared.

Contents

List of Tables

List of Figures

Nomenclature

Acronyms

AI	Artificial Intelligence
IQR	Interquartile Range
LLM	Large language model
OOS	Out of sample

1 Introduction

1.1 Forecasting under Uncertainty

Decision-making under uncertainty is an ongoing challenge across economics, policy, and technology management contexts. In such a setting, actors are often required to form probabilistic beliefs about future events, ranging from macroeconomic outcomes to political developments and technological achievements. While probabilistic judgment is central to decision-making under uncertainty, individuals face major cognitive and practical constraints when forming these beliefs, particularly as the volume of available information increases (Roetzel, 2019).

Given these limits to individual forecasting in complex environments, approaches that aggregate multiple beliefs have been considered as a way to form collective probabilistic assessments. Assessing the quality of a forecast requires more than evaluating statistical accuracy alone. In the context of decision-making, forecasting performance must also be considered in terms of its economic or practical relevance. This highlights the need to empirically test whether proposed forecasting approaches deliver meaningful practical value, rather than assuming effectiveness based solely on theoretical principles or statistical relevance. Taken together, these considerations underscore both the difficulty of forecasting under uncertainty and the importance of evaluating forecasting approaches beyond accuracy alone. The next section will introduce the forecasting tools examined in this thesis and outline how their performance is evaluated.

1.2 Prediction Markets as Information Aggregation Mechanisms

One way of aggregating information is through the application of prediction markets. Prediction markets are markets based on contracts that pay out conditional on the outcome of future events, typically binary outcomes where a contract pays one unit if the event occurs and zero otherwise. The price of such a contract can be seen as the market's collective belief about the probability of the event occurring. Prediction markets work on the principle that dispersed information held by individual traders is gathered and reflected in the price of a contract through trading activity. Traders with positive information about

the outcome of an event are incentivised to buy contracts, while those with negative information sell, pushing the price upwards or downwards depending on the situation. Under the right conditions, the market price of a contract is often interpreted as reflecting a synthesis of all available information, provided that assumptions such as sufficient liquidity and rational and self-interested trading hold (Wolfers & Zitzewitz, 2006).

In practice, the assumptions required for prediction markets to effectively aggregate information are not always satisfied. In particular, the extent to which prices reflect dispersed information depends on market liquidity and active participation, which shape incentives for informed trading. In low-liquidity markets, weaker incentives and lower trading activity can result in noisier prices and slower incorporation of information, making aggregation outcomes more sensitive to market-specific conditions.

Taken together, these considerations highlight that prediction markets provide a structured mechanism for aggregating dispersed information, with performance that depends on market conditions such as liquidity and rational trading. As a result, market-implied probabilities serve as a conditional but informative benchmark against which the performance of other forecasting sources can be compared.

1.3 Large Language Models as Probabilistic Forecasters

Large language models (LLMs) are statistical models, trained on a vast amount of text data to identify structures and patterns in languages, and can generate coherent natural language outputs. Unlike prediction markets, LLMs are not incentive-driven and do not update beliefs through payoffs or trading but rely on trends observed in their training data, model architecture, prompting strategies, and information made available to them at any point in time.

Despite these differences, LLMs may still be able to produce probabilistic forecasts when clearly prompted to do so, particularly when provided with external, context-specific information relevant to the forecasting task. In such settings, LLMs can be prompted to synthesise available text information into probability estimates, even though forecasting is not one of their core functionalities. LLMs also differ from prediction markets in terms of the information they can draw on. While market prices may reflect private information held by traders, LLMs do not have access to such private signals and instead rely on

publicly available text and any external information provided during the forecasting process.

An emerging idea is that large language models might be able to replicate the information aggregation achieved by prediction markets at lower cost (Buterin, 2024). This claim, however, is speculative and has not been systematically tested. This thesis examines this idea by comparing LLM-generated probability forecasts to market-implied probabilities, treating prediction markets as a conditional benchmark. The analysis focuses on whether, and under what conditions, LLM forecasts are able to replicate market-level aggregation outcomes.

1.4 From Forecasting Accuracy to Decision-Making Relevance

Forecasting accuracy is a logical starting point when evaluating probabilistic forecasts, although it is not sufficient for assessing their practical relevance. Accuracy can be evaluated both in- and out-of-sample, depending on the evaluation design. They do provide valuable information about how closely forecasts align with realised outcomes. However, accuracy alone is not sufficient to determine whether a forecast approach is useful in a decision-making context.

The economic value of a forecast depends on how the probability assessments are transformed into actions and how the resulting outcomes compare to other relevant alternatives. This typically involves trading strategies and decision rules that evaluate and determine when and how forecasts are used, as well as the opportunity cost associated with alternative strategies or benchmarks. A forecasting approach is only economically valuable if it yields higher net profits than competing options under similar conditions. As a result, forecasting accuracy and economic value capture different dimensions of performance and need to be evaluated separately. Accuracy measures how well forecasts predict outcomes; the economic value of a forecast is reflected in whether those forecasts can lead to improved, more profitable decisions when evaluated against feasible alternatives. This motivates the analysis to be split between evaluating accuracy and economic value separately. Using trading simulations to assess whether probability forecasts generate profits relative to simple benchmark strategies.

1.5 Research Gap and Contribution

Existing research shows that prediction markets can serve as powerful tools for the aggregation of dispersed information. Building on theories such as the Efficient Market Hypothesis and the wisdom of the crowds, previous studies show that prediction market prices can provide accurate probability assessments across a wide range of domains. More recent work on information finance suggests that such aggregation mechanisms can be deliberately designed to extract information as a public good. In this context, emerging technologies such as artificial intelligence are often proposed as a way to reduce the cost of operating such systems and to extend their applicability. While the literature provides a solid theoretical foundation, an apparent knowledge gap persists. Even though large language models are discussed as potential tools for supporting information aggregation, there is limited evidence on whether LLM-generated probability forecasts can approximate the aggregation outcomes shown in decentralised prediction markets. Particularly, it remains unclear how the relative performance of LLM forecasts compares to market-implied probabilities across different market conditions, such as liquidity level, or if LLM-generated forecasts can be translated into economically significant outcomes suited for decision-making. As a result, key conditions underlying info finance are speculative and remain to be empirically tested.

This study adds to this literature by providing an empirical evaluation of probability forecasts generated by large language models relative to market-implied probabilities in decentralised prediction markets. It examines forecasting across multiple models, prompting strategies, and liquidity levels. Furthermore, it assesses economic relevance using trading simulations based on LLM-generated probability forecasts and market-implied probabilities. Doing so, the study offers evidence on the extent to which LLMs can replicate, or fail to replicate, the information aggregation abilities of prediction markets in an economically meaningful forecasting context, thereby identifying both the capabilities and the limitations of LLMs when evaluated against prediction markets as a benchmark for collective forecasting.

1.6 Research Questions

This thesis examines the extent to which large language models can be used as forecasting agents in the context of decentralised prediction markets. Specifically, it examines whether LLM-generated forecasts compare with market-implied probabilities and whether the alignments translate into economic relevance. In this context, human forecasting is represented by prediction markets, which aggregate the beliefs of many individual participants into market-implied probabilities. The main research question is:

"To what extent are LLMs able to replicate or complement human forecasting as reflected in decentralised prediction markets?"

This research question is addressed through the following sub-questions:

1. How much variation is there between the LLMs prediction accuracy in highly liquid markets compared to low liquidity markets?
2. To what extent can prompt engineering enhance the accuracy of the LLMs prediction capabilities?
3. To what extent can LLM forecasts or combinations of them achieve higher predictive accuracy than market-implied probabilities?
4. How do LLM-driven trading strategies perform relative to simple trading algorithms in generating persistent, out-of-sample profits on decentralised prediction markets?

1.7 Research Approach

This study is quantitative and takes an empirical research approach to evaluate the forecasting performance of large language models in the context of decentralised prediction markets. Probability forecasts generated by LLMs are gathered and then compared to market-implied probabilities over a set of prediction markets over time, treating the market as a benchmark for collective human forecasting. The analysis is focused on relative performance, using standard probabilistic scoring rules to enable systematic comparison between LLM forecasts and market-implied probabilities across models, prompting strategies, and market conditions. This evaluation is conducted on multiple levels, examining various models, prompting strategies, and market conditions.

The economic relevance of LLM-generated forecasts is examined separately from statistical accuracy. Economic value is evaluated using trading simulations based on both LLM-generated forecasts and market-implied probabilities. These simulations are then evaluated out of sample to assess whether any observed trading profits would still hold up under more realistic conditions, and that way, evaluate their decision-making relevance in practice.

1.8 Relevance

The relevance of this thesis is both societal and academic. The thesis provides empirical evidence on the current capabilities and limitations of large language models as probabilistic forecasting tools when evaluated against already established tools for collective human forecasting. LLM-generated probability forecasts are evaluated relative to market-implied probabilities from decentralised prediction markets that serve as a benchmark for collective human forecasting. This comparison reveals to what extent LLMs can replicate or complement existing aggregation outcomes.

From an academic point of view, this adds to the literature on decision-making under uncertainty by moving beyond speculative claims about the potential of LLMs and grounding the evaluation in observed performance under realistic conditions. More specifically, the study contributes to the discussion on information aggregation and forecasting by evaluating how an emerging technology performs when applied within an already existing forecasting and decision-making setting, rather than looking at the LLMs performance in isolation. This perspective aligns well with Management of Technology research, as it investigates how new technologies can be adopted, evaluated and integrated into established organisational and market mechanisms.

The findings are also societally relevant as large language models and their applications are highly discussed in today's narratives. Whether those applications are for research, industry or public discourse to support judgment and decision-making. Large language models are often used to summarise information and to express uncertainty, while their suitability for producing probabilistic assessments is largely unknown and untested. By comparing LLM-generated forecasts to collective human forecasts, this thesis aims to provide evidence on the current capabilities of these models in forecasting tasks. This is

particularly relevant in the decision-making context where the application of prediction markets is infeasible due to practical constraints like limited resources, regulatory constraints, or low participation and incentives. In such a setting, LLM-powered forecasting tools may be considered as an alternative or a complementary source of information to market-based aggregation. The results underscore the importance of empirical testing and validation before deploying AI-generated probability forecasts in high-stakes environments. By evaluating and clarifying the current stage of LLMs as forecasting tools, the study supports more cautious and informed use of emerging AI technologies in the decision-making context.

1.9 Thesis Outline

This thesis is structured as follows. Chapter 2 introduces the theoretical background and reviews relevant literature on probabilistic forecasting, prediction markets, and information aggregation. Chapter 3 describes the research design and methodology, including the data sources, forecasting setup, and evaluation procedures used in the analysis.

Chapter 4 presents the empirical results. It first reports the forecasting accuracy of LLM-generated probability forecasts relative to market-implied probabilities across different models and prompting strategies. The chapter then examines how forecasting performance varies across market liquidity conditions. This is followed by an analysis of whether combining multiple LLM forecasts, or combining LLM forecasts with market-implied probabilities, improves predictive accuracy. Finally, the chapter evaluates the economic relevance of LLM-generated forecasts using trading simulations based on model-generated probabilities and contemporary market-implied odds.

Chapter 5 discusses the results in relation to the research questions and existing literature, highlighting key limitations and implications and outlining possible directions for future research. Chapter 6 concludes by summarising the main findings.

2 Theoretical Background

A recent blog post published in November 2024 titled “From Prediction Markets to Info Finance” brings attention to an emerging approach in which financial market mechanisms are purposefully engineered to extract and collect information (Buterin, 2024). Buterin calls this approach “info finance”. He suggests that markets can, in principle, be designed as information engines whose prices and outcomes reveal decision-relevant information. This idea is built on a wide cross-disciplinary foundation from economics, computer science, finance, game theory, political science and policy making. Previous studies have researched prediction markets as tools to capture dispersed knowledge, providing evidence that market-generated forecasts can be very accurate (Wolfers & Zitzewitz, 2004). Scholars have proposed several domains in which prediction market mechanisms may be applied, including policy evaluation and scientific verification (Hanson, 2013; Dreber et al., 2015).

Prediction markets

Prediction markets are a form of market based on contracts that pay out based on the outcome of future events (usually binary events, pay 1 if the outcome is true, 0 if the outcome is false). The price of the contract can be seen as the market’s collective probability forecast of the outcome of the event. Prediction markets work on the idea that information is reflected in the price of the contracts that are offered on the market. As a result, traders with positive information buy contracts, driving the price up, and those with negative information then sell contracts, driving the price down. In a state of equilibrium, the market should reflect a synthesis of available information under conditions of sufficient liquidity and rational, self-interested trading. Under the assumptions of risk-neutral or logarithmic utility traders, a prediction market's price can equal the mean belief of traders in a certain market (Wolfers & Zitzewitz, 2006).

The theoretical foundation for this phenomenon is based on two economic theories: firstly, the Wisdom of the Crowds, and secondly, the Efficient Market Hypothesis.

The efficient market hypothesis states that market prices reflect all available information, making consistent profit from stock trading or in other words, beating the market

impossible. The efficient market hypothesis has received a lot of criticism as it assumes that markets are efficient, which is often not the case, as many factors can hinder market efficiency, namely, low liquidity, transaction costs, market psychology and human emotions (Chen, 2023). Markets are often broken down into three different types. Weak markets, where all past prices of a stock are reflected in the current price, making technical analysis unable to predict the market. Semi-strong markets, where all public information is also reflected in the stock pricing making fundamental and technical analysis incapable to gain an edge on the market and lastly strong markets where all public and private information is reflected in the market, making it impossible for anyone even traders with insider information, to make profits, as the information they have is already reflected in the stock price (Fama, 1970). A recent study that analysed 664 markets found strong evidence supporting semi-strong informational efficiency, that prices reflected 90-100% of public information. Whereas they find little evidence for strong informational efficiency, as they estimate 0-30% of private information to be reflected in the market prices (Page & Siemroth, 2021).

A second foundational theory behind prediction markets is the wisdom of the crowds. Wisdom of the crowds is a phenomenon where aggregating multiple independent judgments, without deliberation or consensus, can produce more accurate outcomes than individual judgments (Hamada et al., 2020). This fundamental aspect of prediction markets poses a problem, as they become inefficient when there is little volume or liquidity. With little volume, there are small gains to be made. This leaves little incentive for any player to go out and spend time and resources to make an educated prediction based on information they have gathered.

These theoretical foundations have also been supported by empirical evidence across a wide range of forecasting contexts. Previous research shows that prediction markets often achieve lower forecasting errors than traditional alternatives such as expert panels or opinion polls, particularly in settings with active participation and sufficient liquidity (Wolfers & Zitzewitz, 2004). Other applications include political forecasting, sales predictions, and demand estimation. In these domains, market-based aggregation has been shown to effectively synthesise dispersed information into accurate probability

assessments. This evidence supports the use of prediction market prices as a demanding benchmark for evaluating alternative forecasting approaches.

That being said, the quality of information in prediction markets depends on who trades, what information they possess and what incentives drive them (Wolfers & Zitzewitz, 2006). Favouring longshot bias, when traders overestimate low-probability events, is common among unskilled traders. Other factors that can influence the accuracy of predictions are poor market conditions, including wealth asymmetry and risk aversion (Manski, 2006). Manski points out that prediction market prices do not always represent the mean belief of the actors trading on the market, which can be partly explained by the previously mentioned limitations. However, experiments show that well-designed prediction markets can aggregate dispersed information effectively and converge toward a rational expectations benchmark, especially if arbitrage is possible to correct incorrect pricing (Plott & Chen, 2002).

Manipulation attempts on prediction markets have been rare and usually self-defeating. Well-informed, rational traders who detect mispricing can trade against it, restoring the price balance (Rhode & Strumpf, 2006). This highlights how the incentive structure helps to keep biases and strategic schemes away, as traders only profit from being more right than others.

Info finance

Info finance is a recent phenomenon that describes a mechanism, like prediction and decision markets, that leverages financial incentives to extract truthful, actionable information (Buterin, 2024). In traditional finance, information is a byproduct of the market, whereas in info finance, information is the main output of the market and is seen as a public good (Buterin, 2024). Buterin frames this as a correct-by-construction ideal, where markets are designed to extract truthful information through incentive alignment. Accurate public predictions benefit the masses and are shareable; they are underproduced due to the free rider problem (Grossman & Stiglitz, 1980). Prediction markets attempt to address this issue by rewarding individuals with accurate predictions through their incentive structure.

Developments in info finance systems have been enabled by recent advances in both blockchain technology and artificial intelligence. Blockchain infrastructures enable decentralised, prediction markets with resolution oracles and programmable contracts. Platforms such as the Polymarket operate on blockchains, using smart contracts and oracles to resolve binary events and ensure payouts without centralised mediation or control (Polymarket Documentation, 2024). These technical features support conditional and combinatorial market designs while maintaining data integrity and resistance to exploitation attempts (Hanson, 2003; Rhode & Strumpf, 2006).

At the same time, recent developments in large language models have increased interest in their potential role as forecasting agents within information markets. LLMs can quickly synthesise large volumes of unstructured text, such as news articles, and output probabilistic judgments. This has motivated proposals to integrate LLMs into prediction and information markets, either as decision-support tools for human traders or as independent, autonomous forecasting agents (Schoenegger et al., 2024). However, whether such models can reliably complement or replace market-based aggregation mechanisms remains a question. This thesis aims to address this question by comparing LLM-generated probabilistic forecasts to market-implied probabilities under real-time market conditions.

3 Methodology

To answer the research question stated in Chapter 1. This research follows a quantitative research design complemented by a detailed literature review. This chapter outlines the research approach taken to compare forecasts generated by a selection of LLMs with market-implied probabilities obtained from the Polymarket. To ensure transparency and reproducibility, the chapter explains the procedures for data collection, selection, and application of LLMs utilised in forecasting, and the design of the data collection workflow. Furthermore, it describes the evaluation metrics and statistical tests employed in the study, including the Brier score, the Wilcoxon signed-rank test, and a pilot-based assessment of statistical power. Finally, considerations of reliability, validity and methodological limitations are discussed. The following sections provide a detailed explanation in each step, beginning with the research design.

3.1 Literature Synthesis Approach

The literature review in this thesis was conducted to provide a conceptual context and a strong theoretical foundation for the analysis. As a starting point, the review was anchored around the info-finance perspective, which views prediction markets as institutional information aggregation mechanisms (Buterin, 2024). This starting point was chosen because it directly aligns with the research question of the thesis. The blog post itself was treated as an inspiration to scour the web to find relevant literature of high quality to base the theoretical background on. To support and accelerate the initial screening and exploration of literature, a large language model (GPT-o1) was used, toggling its “deep research” functionality. This procedure was used to assist in identifying sources and research themes. The prompt used can be found in Appendix A. All sources identified through this process were manually verified. Original papers were consulted, and only credible sources were included in the literature review. Similar uses of LLMs for literature reviews have been documented in recent methodological research (Scherbakov et al., 2025) .

3.2 Prediction Market Data Source

The source of market data used in this study is Polymarket, a decentralised prediction market platform that offers binary outcome contracts. Each contract pays out one unit if the corresponding event occurs and zero otherwise. The market price of a YES contract is interpreted as the market-implied probability of the event occurring at a given point in time. The study is focused on a predefined set of live prediction markets, which were selected based on two criteria. First, they had a minimum liquidity of $1,000, which was done to avoid empty markets. Second, their resolution date had to fall within the period of 30.08.2025 – 15.09.2025, which marks the end of the data collection phase. The final set of evaluated markets represents the subset of markets from the Polymarket that fulfil these criteria. This set provides the empirical basis for all subsequent analyses in the Results chapter.

3.3 Forecasting Time Frame

The data collection followed a forward-looking design, where prediction markets were observed prior to their resolution, rather than evaluating historical markets. Forecasts were generated and recorded while markets were still active. This design was chosen for two reasons. This design ensures that both market- and LLM-generated forecasts were formed under realistic information conditions, restricting information to what is available at the time of prediction. Both market-implied probabilities and LLM-generated forecasts were collected daily throughout the data collection period. For each observation day, LLM-based forecasts were generated for all open markets and paired with the current market price, ensuring that LLM-generated forecasts and market-implied probabilities were based on the same information environment. The data collection period began on 29.07.2025 and continued until the final market resolution on 13.09.2025. Due to technical disruptions in the data collection workflow, observations are missing for a few days in the collection period. These days were omitted from the data analysis, and affected markets were paired using the nearest available observation date when stage-based analyses were conducted. The final dataset consists of paired observations at the market and date level. Data collection was conducted in real time; as a result, the exact number of observations per

market varies based on resolution dates, as the markets are dynamic and can close before their set resolution date if the resolution conditions are fulfilled early.

3.4 Large Language Model Selection

Before the main data collection phase started, a proof of concept was conducted using two models (GPT-3.5-Turbo and GPT-4o-mini) to verify that the forecasting pipeline functioned and to ensure that all functionalities worked as intended. This phase also provided a preliminary set of markets to estimate variance and statistical power. This proof of concept was only used before official data collection started, and no data collected in this period was used in the analysis reported in this thesis.

Large language models were leveraged as probabilistic forecasting agents to generate probability assessments for the selected prediction markets. All models were treated as black-box forecasters. They all received the same input of information and were strictly prompted to only output a single probability between 0-1 for every market every day in over the data collection period. The models did not receive any feedback, did not learn from outcomes and did not interact with markets or with each other over the data collection period. The main constraints in model selection were practical considerations, cost, availability, and allow for repeated daily forecasting. Models were chosen from a variety of providers to avoid over-reliance on a single model provider. Furthermore, we deliberately chose models ranging from reasoning models, designed for deep analytical output, to lightweight models optimised for speed and efficiency. All models were accessed through API applications using fixed prompts and identical information input. Output was restricted to only probabilistic predictions; any non-numeric or invalid outputs were excluded.

Model Overview

The models selected for the experiment are summarised in Table 3.1. Descriptions below are limited to high-level model characteristics:

- GPT-3.5-Turbo: A general-purpose language model designed for cost-efficient text generation and broad task coverage (OpenAI, 2025).

- GPT-4o-mini: A lightweight model optimised for low latency and reduced cost while supporting more structured outputs (OpenAI, 2024).
- Gemma 3 12b: An open-weight model with a moderate parameter count, enabling efficient inference under constrained computational budget (Google AI, 2025).
- Gemma 3 27B: A larger open-weight model with increased capacity relative to the 12B variant (Google AI, 2025).
- DeepSeek Chat V3: A chat-oriented model designed for conversational tasks and instruction following, exploits MoE (deepseek-ai, 2025).
- DeepSeek R1: A reasoning-focused model, designed to support more explicit intermediate reasoning processes, capable of applying Chain-of-Thought on its own, exploits MoE (deepseek-ai, 2025).

Table 3.1: Large language model structure

Model	*Provider*	*Model class*	*Parameters*	*Reasoning model*	*Context length*	*Cost*
GPT-3.5-Turbo	OpenAI	General-purpose	Undisclosed	No	16k	Low
GPT-4o-mini	OpenAI	Lightweight general	Undisclosed	Partial	128k	Low
Gemma 3 12B	Google	Open-weight	12B	No	128k	Low
Gemma 3 27B	Google	Open-weight	27B	No	128k	Free
DeepSeek Chat v3	DeepSeek	Chat-oriented	37B active (617B total)	Partial	128k	Free
DeepSeek R1	DeepSeek	Reasoning	37B active (617B total)	Yes	128k	Free

3.5 Prompting strategies

Two prompting strategies were employed for the LLMs to generate probabilistic forecasts. A Zero-shot baseline and a Chain-of-Thought (CoT) treatment. This was done to assess whether prompting strategy influenced forecasting accuracy. Under the zero-shot condition, models are prompted to directly provide a probabilistic estimate for each event, while under CoT conditions, shot by instructing the LLM to provide intermediate reasoning steps before arriving at a answer, encouraging the model to articulate its thought process (Wei et al., 2022; Prompting Guide, 2025).

Prompt structure and content were held constant throughout the data collection period, except for the inclusion of structured reasoning instructions for the CoT prompting. This design helps ensure that any observed differences in forecasting accuracy can be attributed to the prompting strategy and do not reflect differences in information or model interactions.

3.6 Information Sources and Forecasting Pipeline

Prediction data collection

The LLMs were provided with up-to-date information relevant to each prediction market, public news articles retrieved using AskNewsAPI, a pre-built web scraping and aggregation service. AskNewsAPI collects news articles from a large number of sources from the web across over 15 countries and languages. In this study asknews_searcher function was applied to retrieve articles relevant to each market based on keyword queries created from market descriptions.

For each market and observation date, this function retrieved two sets of articles. First, six recent articles (published within 48 hours) prioritising the most recent of available articles. Second, 10 historical articles dating back to 2023; here, it prioritised articles published within the last 60 days (Metaculus, 2025). The resulting set of articles served as the external information provided to the LLMs for each forecasting task.

External information synthesis

The articles retrieved were preprocessed before being fed to the LLMs. During proof of concept testing, passing a full set of raw articles directly to the models often exceeded

practical token limits, resulting in probabilistic forecasts that resembled a coin flip (50/50 probabilities). This behaviour is consistent with limitations in processing large, complex information inputs, where excessive information can reduce effective reasoning and lead to models resorting to this undesirable behaviour (Levy et al., 2024). In this step, the retrieved articles were summarised into a compact textual representation that preserved factual content while reducing the input length. This summarisation process was applied across all markets, models, prompting strategies, and observation dates. Figure 3.1 shows the Data collection and forecasting pipeline. All code, data, and replication materials used in this study are publicly available at: https://github.com/KetillHafdal/llm-vs-prediction-markets

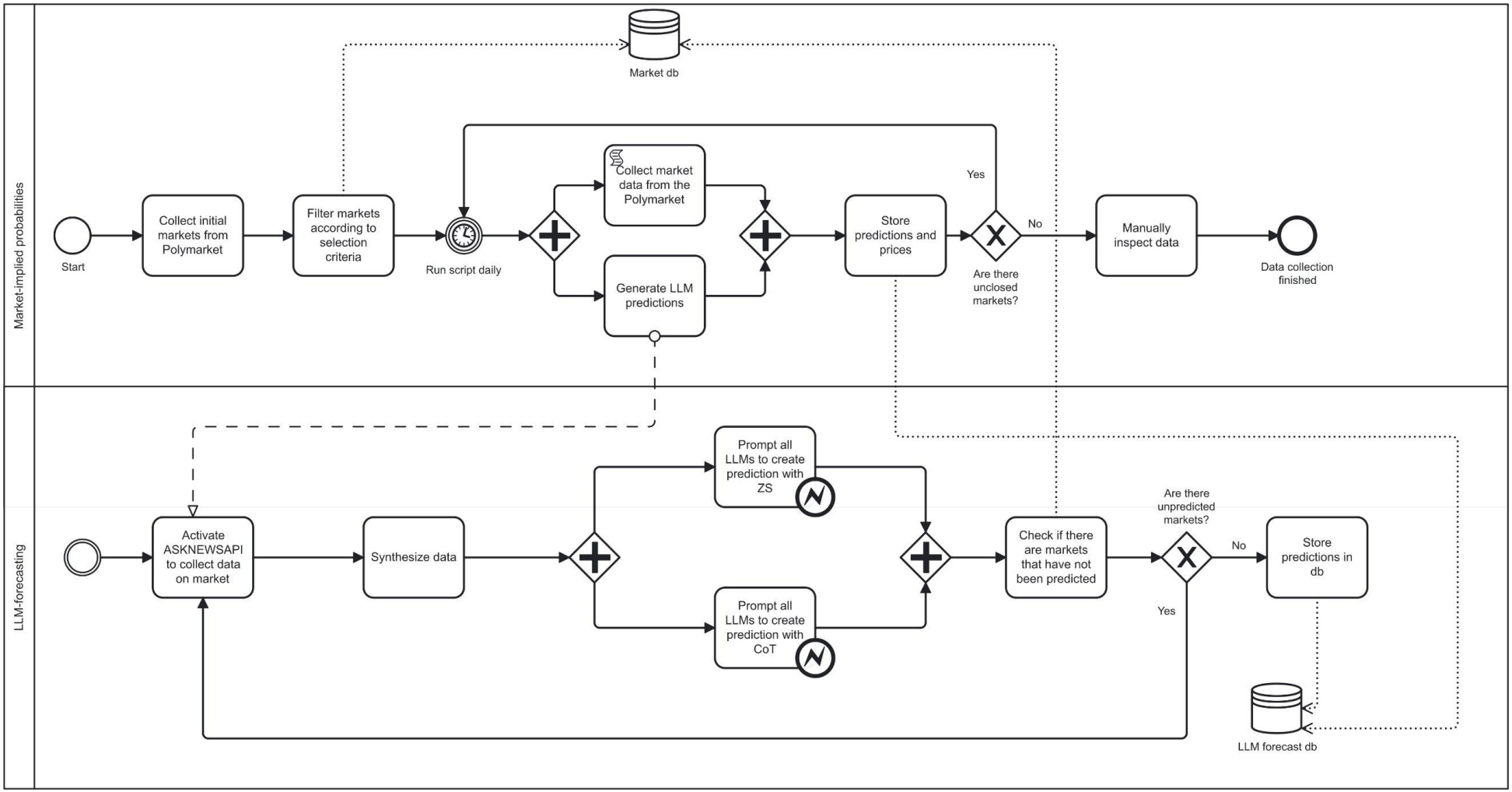

Figure 3.1: Data collection and forecasting pipeline. The figure illustrates the flow of information from Polymarket and AskNewsAPI through preprocessing and synthesis, probabilistic forecasting by large language models, and storage of paired market–model observations

3.7 Market segmentation

To evaluate if forecasting performance varies based on market conditions, prediction markets were split up based on liquidity. Trading volume was used as a proxy for liquidity, as liquidity was not available for all markets, and both measurements reflect the level of trading activity in a market. Markets were segmented into three categories: Low-, medium- and high-liquidity markets. This level of segmentation was selected to offer the chance to investigate what effects market liquidity has on the prediction capabilities of the LLMs while also being careful not to segment the markets into too small portions to prevent overfitting. The markets were segmented using a using terciles of the final observed trading volume. All markets that fulfilled the inclusion criteria described in Section 3.2 were included in one of these liquidity groups.

3.8 Evaluation Metrics and Statistical Testing

Forecasting accuracy

Forecasting accuracy was evaluated using Brier scores, a scoring rule for probabilistic predictions of binary outcomes. The Brier score measures the squared difference between a probabilistic forecast and the realised outcome. With lower values indicating a higher accuracy. The Brier score is defined as:

$$Brier\ score\ = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$$

Where p_i represent the predicted probability and $y_i \in \{0,1\}$ represents the realised outcome and N denotes the total number of evaluated forecast observations. The Brier score was selected as the main evaluation metric because it is widely used in forecasting evaluation and allows for direct comparison between probabilistic forecasts and market-implied probabilities.

Statistical assessment

To compare the forecasting accuracy of LLM-generated probabilities and market-implied probabilities, paired statistical tests were employed. Comparisons were made at the market-date level, pairing each LLM-generated forecast with a market-implied probability

for the same event and the same time. Wilcoxon signed-rank tests were applied to test whether the median difference in Brier scores between two paired forecasts differs from zero. This test works for paired, ordinal data and does not rely on normally distributed data. When multiple paired comparisons were performed, p-values were adjusted using Holm's correction to control for the difference in sample size between individual models.

Initial assessment of statistical power

Statistical power was evaluated using the proof of concept data, collected before the main data collection phase had begun. Monte Carlo simulation was applied to simulate the statistical power for the Wilcoxon signed-rank test. Power was estimated as the proportion of simulated samples yielding a significant result at the 5% level. This was done to assess whether sample sizes within each liquidity group were sufficient.

3.9 Aggregation methods

In addition to individual evaluation of LLM-generated forecasts, aggregation methods were applied to examine whether combining multiple probabilistic forecasts improves predictive accuracy. All aggregation methods were applied after the data collection phase was finished, and evaluated using the same Brier score evaluation metric previously defined.

LLM-only aggregation

To examine whether LLM-generated forecasts could improve predictive accuracy through aggregation, an equal-weight aggregation of all available LLM forecasts was computed for a fixed subset of early, mid, and late market stages. Each aggregation assigns an identical weight to each model, with the following formula:

$$\hat{p}_{\text{LLM-avg},i,s} = \frac{1}{K}\sum_{k=1}^{K} \hat{p}_{k,i,s}$$

Where $p_{k,i,s}$denotes the probability forecast produced by model k for market i at stage s, and K is the number of models included in the aggregation.

Hybrid aggregations

To evaluate whether LLM forecasts provide complementary information beyond market-implied probabilities, hybrid aggregations were created that combine LLM forecasts and market-implied probabilities. Two hybrid aggregations were applied, the first one treating the LLM-generated forecasts and the market implied probabilities as equally predictive forecasting sources, with the following formula:

$$\hat{p}_{50/50\ hybrid} = \frac{1}{2}\,\hat{p}_{LLM-avg} + \frac{1}{2}\,p_{market}$$

The second hybrid aggregation treats the market implied probabilities as an additional forecaster alongside the individual LLMs, and gives them the same weight as shown in the formula:

$$\hat{p}^{(\text{all})}_{\text{hybrid}} = \frac{1}{K+1}\left(\sum_{k=1}^{K}\hat{p}_k + p_{\text{market}}\right)$$

Regression-based aggregation

A regression-based aggregation that combines multiple LLM forecasts and market-implied probabilities using learned weights rather than fixed ones. This aggregation is implemented as a logistic regression that maps contemporary LLM forecasts and market-implied probabilities to a predicted probability of the realised binary market outcomes. For a given market and stage, the regression is defined as:

$$Pr(Y = 1|X) = logit^{-1}\left(\beta_0 + \beta_{mkt}p_{mkt} + \sum_{k=1}^{K}\beta_k p_k\right)$$

where Y is the realised market outcome, p_{mkt} is the market-implied probability, and p_k is the forecast probability produced by LLM k.

3.10 Economic and Decision-Relevance Evaluation

Trading simulations were conducted to evaluate the decision-relevance of LLM-generated probabilistic forecasts. Forecasts were used as trading decisions and evaluated using realised market outcomes. Two strategies were evaluated: a baseline strategy based solely on market-implied probabilities and a strategy based on the regression-based aggregation.

Baseline trading strategy

A simple market-favourite strategy was employed. For each market and stage, a trading decision was determined by using market-implied probabilities according to the decision rule:

$$d_{i,s}^{(\mathrm{mf})} = \begin{cases} \text{BUY YES} & \text{if } p_{\mathrm{mkt},i,s} > 0.5, \\ \text{BUY NO} & \text{if } p_{\mathrm{mkt},i,s} < 0.5, \\ \text{NO TRADE} & \text{if } p_{\mathrm{mkt},i,s} = 0.5. \end{cases}$$

Where, $d_{i,s}^{(\mathrm{mf})}$ represents the decision of the Market Favourite strategy and $p_{\mathrm{mkt},i,s}$ represents the market-implied probability for market i at stage s and $s \in \{\text{first, mid, last}\}$. All positions were held until markets closed at resolution.

No-information strategy

A no-information benchmark trading strategy was implemented to provide a lower bound on trading performance. For each market i and stage s. A trading decision was determined by flipping a coin. If the outcome was heads, a YES position was taken; if the outcome was tails, a NO position was taken. Trades were executed at the contemporaneous market-implied price and held until market resolution. Formally, the decision rule is given by:

$$d_{i,s}^{(\mathrm{coin})} = \begin{cases} \text{BUY YES} & \text{if heads} \\ \text{BUY NO} & \text{if tails} \end{cases}$$

This strategy does not condition on any market prices, model outputs, or external information, and therefore represents a strictly no-information trading baseline.

Iron Man trading strategy

The Iron Man trading strategy compares the probability produced by the regression-based aggregation to the market-implied probabilities. “For each market i and stage s, the following decision rule applies:

$$d_{i,s}^{(\text{iron})} = \begin{cases} \text{BUY YES} & \text{if } p_{\text{iron},i,s} > p_{\text{mkt},i,s} \\ \text{BUY NO} & \text{if } p_{\text{iron},i,s} < p_{\text{mkt},i,s} \\ \text{NO TRADE} & \text{if } p_{\text{iron},i,s} = p_{\text{mkt},i,s} \end{cases}$$

Where, $d_{i,s}^{(\text{iron})}$ represents the decision of the Iron Man strategy, $p_{\text{iron},i,s}$ represents the aggregation probability and $p_{\text{mkt},i,s}$ represents the market-implied probability, all for market i at stage s. All positions were held until markets closed at resolution.

Decision evaluation

Trading performance was evaluated under both in-sample and out-of-sample conditions. In the in-sample setting, the aggregation parameters were estimated and evaluated on the same data. The out-of-sample conditions provide a more realistic deployment like conditions to evaluate the trading decisions made by these models under uncertainty.

3.11 Reliability, Validity, and Limitations

To ensure the reliability and validity of the methods. All forecasts and prediction market data were collected using a fixed, documented, and reproducible data collection pipeline, along with consistency in forecasting prompts, models and evaluation procedures throughout the data collection period. Furthermore, all data analysis was recorded in a systematic manner to ensure reproducibility.

The paired design of directly matching LLM-generated forecasts with market-implied probabilities for the same markets at the same time supports internal validity. External validity is limited due to the research focus on a single prediction market platform and a limited observation window.

Missing observations due to technical disruptions and model failure to produce a valid probability output, causing forecasts affected by these shortcomings to be omitted from the study, is a clear methodological limitation of this study. Although this limitation does not affect the internal consistency of the analysis.

4 Results

This chapter presents the results of the experiment designed to evaluate the forecasting performance of large language models in comparison to forecasts implied by Polymarket prices, representing aggregated human beliefs. The results were gathered following the methods described in Chapter 3 and are structured around the sub-questions introduced in Chapter 1. The chapter proceeds by first examining variation across market conditions, then assessing prompting effects, followed by comparisons between LLM forecasts and aggregated models, and concluding with the economic implications derived from trading simulations. The results are based on a dataset that consists of a time series of market data from the Polymarket paired with daily forecasts generated with the LLMs; Brier scores are used as the main evaluation metric. The full data analysis, all code and replication materials are publicly available at: https://github.com/KetillHafdal/llm-vs-prediction-markets.

4.1 Final evaluation dataset

The data collection produced a large set of forecast observations across all days, markets and prompting strategies. The process consists of generating LLM forecasts for the 53 markets using six different models, two different prompting strategies and collecting market-implied probabilities from the Polymarket daily over a 47-day period, excluding 10 days during which data collection malfunctioned and no observations were recorded. Overall, the data collection phase yielded 15,264 individual LLM-generated forecasts. Not all requested forecasts were successfully generated. Some LLMs failed to return probabilistic predictions for a specific market x day x prompt combination. Due to this incomplete forecast availability, only paired observations where an LLM forecast could be directly paired with a corresponding market-implied probability were retained for evaluation. After applying this restriction, the final analysis is based on 14,099 paired observations.

The empirical analysis splits forecasts in four dimensions:

1 Forecasting model (six LLMs).

2 Market phase (first, mid, last).

3 Market liquidity (low, mid, high).

4 Prompting strategy (Zero-shot, Chain-of-Thought).

This yields a conceptual 6 × 3 × 3 × 2 comparison structure. However, not all tests operate on the full set with all the dimensions. Depending on the research question, dimensions are either stratified explicitly or collapsed through aggregation, resulting in a different number of observations across tests. Therefore, we explicitly state which dimensions are held fixed and which are aggregated over in each subsection below.

4.1.1 Distribution of forecasting accuracy

The distribution of forecasting accuracy provides an overview of the distribution of forecasting errors. A lower Brier score means that the error is low and is therefore a better prediction. Figure 4.1 shows the distribution of all the Brier scores for both the Polymarket prices (red) and the LLM forecasts (blue).

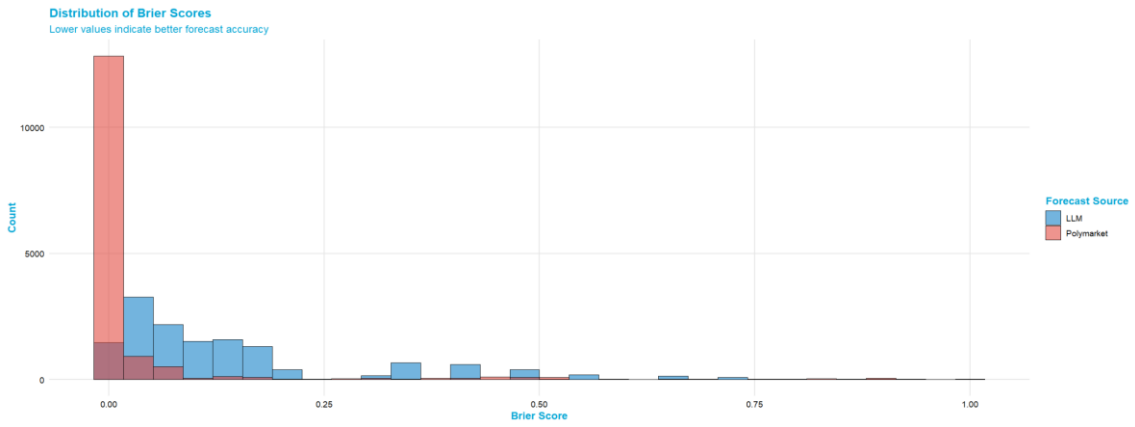

Figure 4.1: Distribution of aggregated Brier scores

From Figure 4.1, we can see that the distribution of the Polymarket Brier scores is severely left-skewed. The distribution of the Brier scores for the LLM forecasts is also left-skewed, although it is not nearly as severe as the one for Polymarket. This indicates that the average error for Polymarket predictions is systematically lower than that of the LLM forecasts. Indicating that at an aggregate level, LLM forecasts do not outperform market-implied probabilities. The distributions above are representative of all observations made in the data collection phase.

4.1.2 Forecast availability and model coverage

The initial dataset was made up of 15.264 observations. After carefully removing the observations that were not paired (LLM-forecast paired with a market price for the same observed day), 14.099 paired observations remained. Table 4.1 depicts the mean Brier scores for each model and prompt variation. It also shows how many individual forecasts each model was able to produce.

Table 4.1: Model overview

Model	Prompt type	N	Mean Brier score	Failure rate
openai_gpt_3_5_turbo_0613	Zero-shot	1272	0.196	0.0%
openai_gpt_3_5_turbo_0613	Chain-of-Thought	1272	0.182	0.0%
openai_gpt_4o_mini	Zero-shot	1272	0.152	0.0%
openai_gpt_4o_mini	Chain-of-Thought	1272	0.152	0.0%
google_gemma_3_12b_it	Zero-shot	1267	0.110	0.4%
google_gemma_3_12b_it	Chain-of-Thought	1263	0.111	0.7%
google_gemma_3_27b_it_free	Chain-of-Thought	1251	0.136	1.7%
google_gemma_3_27b_it_free	Zero-shot	1246	0.134	2.0%
deepseek_deepseek_r1_free	Chain-of-Thought	1196	0.100	6.0%
deepseek_deepseek_r1_free	Zero-shot	1184	0.111	6.9%
deepseek_deepseek_chat_v3_0324_free	Zero-shot	809	0.121	36.4%
deepseek_deepseek_chat_v3_0324_free	Chain-of-Thought	795	0.113	37.5%

Rows are ordered by descending N. Lower Brier scores indicate better probabilistic accuracy.

Forecast availability varies greatly across models; failure rate ranges from 0.0% all the way to 37.5%, meaning that some models were able to produce a probabilistic forecast every time, while others failed at the task more than a third of the time.

In contrast, forecast availability does not practically differ by prompting strategy. Zero-shot prompting results in 7.36% missing forecast, while Chain-of-Thought prompting results in 7.64%. The difference is negligible and is unlikely to affect the results in a meaningful manner.

It is interesting to see how the models with higher failure rates tend to exhibit lower mean Brier scores, which tells us that the models that struggle the most with making the actual

predictions are also the strongest models when it comes to forecasting accuracy of predictions. From these metrics, we can derive that the most suitable models for this kind of analysis are the Google Gemma models and DeepSeek r1, as they provide a low mean Brier score, while also being able to consistently produce a forecast.

4.1.3 Model-level variation in forecasting accuracy

As depicted in Table 4.1, the six Large language models selected for the study put up different levels of predictive performance. To see how the models behave, they have been benchmarked against the Polymarket, once again using our main evaluation metric, the Brier score. To visualise the spread, skewness and outliers of the data set, a box plot was created, Figure 7.1.

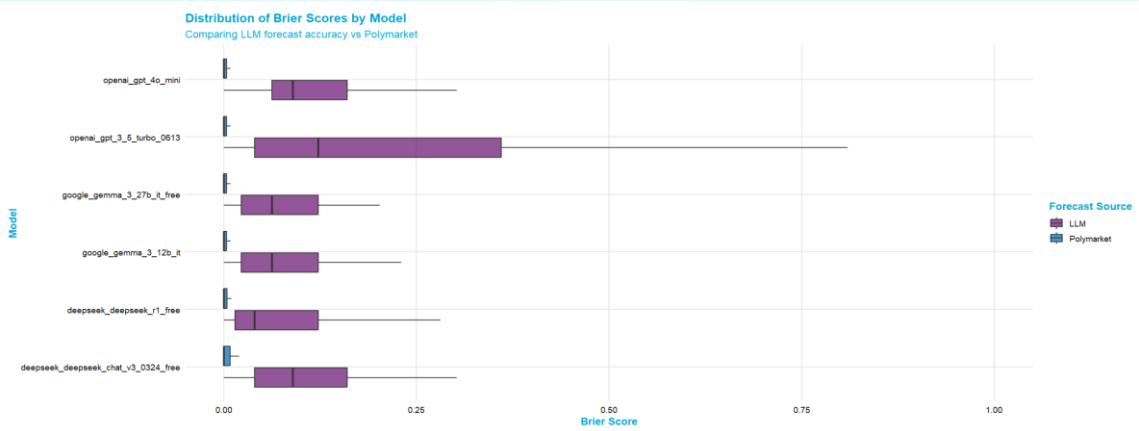

Figure 7.1: Distribution of Brier scores by model

From Figure 7.1, it is evident that across all models, Polymarket Brier scores are tightly concentrated very close to zero, with no meaningful tails. This implies that Polymarket has a high predictive accuracy, with low variance. All the LLMs underperform the Polymarket significantly, with higher medians, wider interquartile range (IQR), and longer tails.

The best-performing LLMs are DeepSeek-R1 and the two Google Gemma models; they have a relatively concentrated IQR and a relatively low median, while it's still considerably higher than that of the Polymarket.

The worst-performing LLM is by far GPT-3.5 Turbo; it has a larger IQR than any other model, accompanied by the highest median and the longest tail. These metrics indicate that the model, on average, exhibits larger forecast errors than the other models. The long

tail reaching upwards of 0.75 means that predictions can be very far off and display a high inconsistency with their probabilistic predictions.

The two remaining models, DeepSeek-V3 and GPT-4o-mini, exhibit intermediate performance, with median Brier scores and interquartile ranges that lie between the strongest and weakest models.

These results suggest that standalone LLM forecasts do not outperform market-implied probabilities in this setting. Building on this baseline comparison and the observed performance differences across models, we now examine whether forecasting accuracy systematically varies across markets, starting with liquidity.

4.2 Liquidity and forecasting accuracy

This section examines if relative forecasting accuracy of LLMs and market-implied probabilities varies with different levels of market liquidity. Liquidity is approximated using final observed trading volume, and forecasting performance is evaluated using market-level differences in Brier scores between LLM forecasts and Polymarket prices. The analysis is focused on relative accuracy differences rather than absolute performance levels.

4.2.1 Liquidity segmentation

To examine if forecasting performance depends on liquidity, we split the original 53 markets into liquidity groups based on their final observed trading volume. Markets are classified into low-, medium-, and high-liquidity groups using terciles of final observed trading volume. Table 4.2 reports the resulting liquidity groups and corresponding volume ranges.

Table 4.2: Markets per-final liquidity group

Liquidity group	Volume range	n	share
Low liquidity	0 – 21,020	18	34.0
Medium liquidity	21,020 – 223,435	17	32.1
High liquidity	223,435 – 1,443,302	18	34.0

After splitting up the markets, we can investigate if there is a significant difference in predictive accuracy based on the market liquidity.

4.2.2 Accuracy differences across liquidity groups

To assess if forecasting accuracy varies systematically with market liquidity, we compare market-level Brier score difference across low-, mid-, and high-volume markets. Table 4.3 reports these results. The mean and median differences are defined as the difference between the LLM forecast and the Polymarket implied probabilities, such that positive values indicate lower accuracy for the LLM relative to the market.

Accuracy comparisons are made at the market level, using Brier score differences between LLM forecasts and the market-implied probabilities. To account for potential changes in forecasting accuracy over the life cycle of the markets, we report these market-level differences separately for the three market stages (first, mid, last) with the liquidity group members being held fixed. For each market, only the first, middle and final forecasts are kept for the stage-wise comparison, with all in-between forecasts being omitted. For each liquidity group and stage comparison, summary statistics and Wilcoxon signed-rank test results are reported and displayed.

Table 4.3: Wilcoxon signed rank results (market-level) by liquidity group & stage.

Liquidity group	Market stage	N markets	Median Δ Brier	Mean Δ Brier	SD Δ Brier	Wilcoxon V	p-value	Rosenthal r	Rank-biserial r	Direction
Low volume	first	18	0.0334	0.0726	0.1494	153	0.0035	0.6878	0.7895	Market better
Low volume	mid	18	0.0291	0.0416	0.1475	154	0.0031	0.6981	0.8012	Market better
Low volume	last	18	0.0516	0.0799	0.1238	156	0.0023	0.7185	0.8246	Market better
Mid volume	first	17	0.0890	0.1284	0.1538	143	0.0018	0.7579	0.8693	Market better
Mid volume	mid	17	0.0760	0.1078	0.1940	138	0.0039	0.7004	0.8039	Market better
Mid volume	last	17	0.0762	0.1097	0.2309	121	0.0066	0.6585	0.7794	Market better
High volume	first	18	0.0635	0.0914	0.0923	170	0.0003	0.8623	0.9883	Market better
High volume	mid	18	0.0896	0.1128	0.1035	152	0.0004	0.8368	0.9869	Market better
High volume	last	18	0.0898	0.0946	0.0669	171	0.0002	0.8725	1.0000	Market better

Across all stages, the median accuracy gap is smallest in the low-volume markets, where the median difference ranges from 0.0029 to 0.0052. In contrast, mid- and high-volume markets provide a consistently higher median difference, typically exceeding 0.07. This pattern indicates that the relative accuracy gap between LLM forecasts and market-implied probabilities increases with market liquidity, while LLM forecasts remain closer to market performance in low-liquidity markets. To formally assess whether the observed accuracy differences across liquidity groups are statistically significant, a Wilcoxon signed-rank test was conducted on the market-level Brier score difference. Testing each liquidity group and market stage combination separately and evaluating whether the median difference between LLM forecasts and market-implied probabilities differs from zero. Table 4.3 shows the test results and a summary of forecasting accuracy. Across all liquidity groups and market stages, the Wilcoxon signed-rank tests reject zero median difference, indicating that LLM forecasts have significantly higher Brier scores than market-implied probabilities. This shows that LLM forecasts consistently carry a higher predictive error than the market-implied probabilities, regardless of market liquidity or stage.

4.2.3 Liquidity and market stages

Building on the liquidity-based comparisons reported in 4.2.2, we investigate whether the observed market-level accuracy differences differ across market stages (first, mid, last). The same market-level Brier score differences reported in Table 4.3 are evaluated by market stage to examine the effect of liquidity on accuracy patterns over the market life

cycle. They show that the order of liquidity groups remains consistent across all market stages. Median and mean differences stay rather stable across the life cycle of the markets.

Descriptively, the mid stage shows a slightly smaller median difference for low- and mid-liquidity markets compared to the first and last stages, while no comparable pattern is seen for high-liquidity markets. These variations are not great and do not change the overall liquidity ordering.

Having shown that market liquidity affects relative accuracy regardless of market stage, we next examine whether model-side interventions, more specifically prompt engineering, can improve LLM forecasting performance.

4.3 Prompt engineering effects

This section evaluates whether prompt engineering affects the forecasting accuracy of LLMs by comparing Chain-of-Thought and Zero-shot prompting under paired forecasting conditions.

4.3.1 Data structure and paired comparison setup

To isolate the effects of prompting strategy, we compare CoT and Zero-shot prompting by directly comparing paired forecasts produced for the same market, day and model. This ensures that the differences in accuracy reflect the prompting strategy rather than differences caused by market compositions. Prompting effects are measured using the difference in Brier score between CoT and Zero-shot forecasts. We first assess the aggregate effect of prompting on forecast accuracy. The analysis uses paired CoT and Zero-shot forecasts generated for the same market, day and model. Observations where either prompt failed to generate a probabilistic forecast are omitted from the sample, resulting in 6,810 paired prompt-level observations.

4.3.2 Overall prompt-effect assessment

To begin with, we evaluate if Chain-of-Thought prompting can increase the forecasting accuracy of LLMs over Zero-shot prompting at the aggregated level. At the aggregated level, the median CoT and Zero-shot difference is zero, indicating no systematic improvement in forecasting accuracy from applying CoT prompting. While a paired Wilcoxon signed-rank test shows significance ($p < 0.001$), this result is driven by a large

number of paired observations (N=6,810) rather than by a meaningful shift in forecasting performance. The distribution of differences is heavily centred around zero, as is indicated by a median difference of zero and very small effect sizes, as is displayed in Table 4.4.

Table 4.4: Pooled CoT vs Zero-shot (paired) median diff (CoT-ZS)

N pairs	Median Δ Brier	Wilcoxon V	p-value	Rosenthal r	Rank-biserial r
6810	0	1928518	0.00031	0.04366	-0.07735

The significant test result does not indicate that either prompting strategy is consistently more accurate than the other. It reflects minor distributional differences that are statistically detectable but practically negligible.

Overall, prompt engineering produces at best a very modest aggregation effect that does not move the median forecasting performance in practice. Because aggregation may hide differences across models, we next investigate prompting effects on the model level.

4.3.3 Model-level prompting effects

Aggregated prompting effects may conceal meaningful differences in how individual models respond to prompt engineering and, therefore, their forecasting accuracy. Therefore, we next look at paired CoT vs Zero-shot differences, separated on the model basis. Figure 4.2 plots the median LLM - Polymarket Brier score difference for each model for both Zero-shot and Chain-of-Thought prompting.

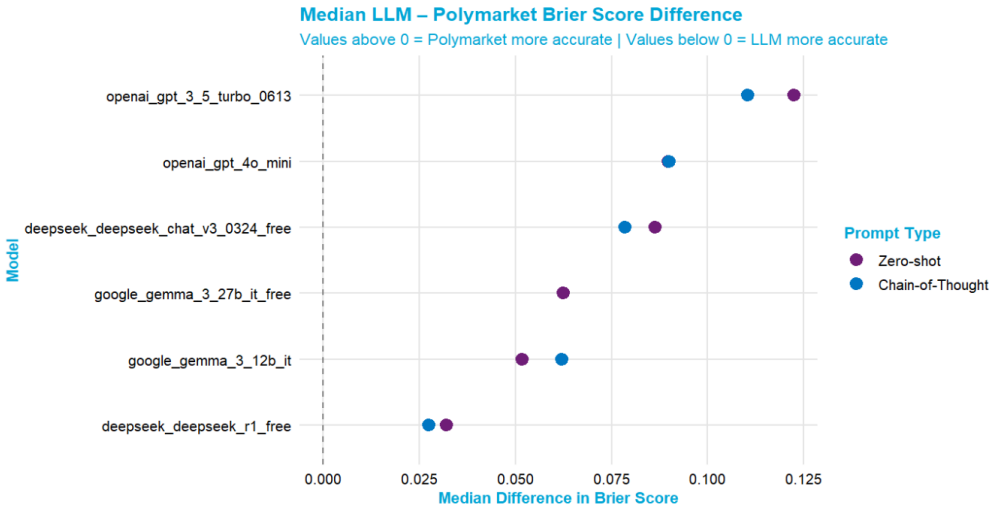

Figure 4.2: Median model-level Brier score difference

Across all models, the mean difference is clustered close to zero, which indicates that the prompting effects are very small in magnitude. Directionally, the models are not uniform, where all the models except for the Gemma ones show marginally better results for the CoT prompting. In contrast, the Gemma models favour Zero-shot prompting ever so slightly. This suggests that prompting effects are model-dependent rather than uniform.

Table 4.5: Per-model CoT vs ZS (paired), (diff=CoT-ZS)

Model	N pairs	Median Δ Brier	Mean Δ Brier	Wilcoxon V	p-value	Rank-biserial r	Rosenthal r	Holm-adjusted p
deepseek_deepseek_r1_free	1158	0	-0.01174	119213.0	0.00000	-0.24647	0.17693	0.00000
openai_gpt_3_5_turbo_0613	1272	0	-0.01358	69050.5	0.00000	-0.26377	0.16882	0.00000
google_gemma_3_12b_it	1263	0	0.00090	117548.0	0.00000	0.24924	0.16942	0.00000
deepseek_deepseek_chat_v3_0324_free	613	0	-0.00245	8669.5	0.04416	-0.16261	0.24318	0.13248
google_gemma_3_27b_it_free	1232	0	0.00156	18338.5	0.56604	0.04063	0.17154	1.00000
openai_gpt_4o_mini	1272	0	-0.00066	41155.0	0.84689	0.01110	0.16882	1.00000

Across all models, the median paired difference in Brier score is zero, indicating no median shift caused by the selection of prompting strategy at the model level. Some models (GPT 3.5, DeepSeek r1 and Gemma 3 12b) show statistically significant Wilcoxon test results. However, effect sizes remain small, and the direction is not consistent across the different models, as is shown in Table 4.5. Overall, model-level analysis indicates that prompt engineering produces uneven and limited edge that does not translate into consistent improvement in forecasting accuracy in practice.

Taken together, the prompt engineering results indicate that Chain-of-Thought prompting at best yields small and uneven accuracy improvements relative to Zero-shot prompting. Aggregated comparisons reveal no improvements in median forecasting performance, and model-based analysis shows that any average gains are extremely small in magnitude and depend on model architecture. These findings suggest that prompt-level treatment alone cannot improve forecasting accuracy in practise.

Overall, the results from the analysis above show that the selection of prompting strategy, model choice, market liquidity, and market stage do not show any case where LLM forecasts consistently outperform market-implied probabilities. This motivates a broader evaluation of whether LLM forecasts, either individual or in any combination, can outperform market-implied probabilities. In the next section, we will compare standalone and aggregated LLM forecasts directly against market prices.

4.4 LLMs and ensembles vs market-implied probabilities

This section compares the forecasting accuracy of LLM predictions to market-implied probabilities, treating the market as the benchmark. Forecasting accuracy is evaluated using Brier scores, with the focus on whether any LLM-based approach can outperform market -implied probabilities in terms of forecasting accuracy. Here we will take a simple to complex approach, starting off by trying out the simplest comparisons and then investigating if more complex solutions can match or outperform the market probabilities prediction accuracy.

4.4.1 LLM forecast vs market-implied probabilities

This subsection compares the forecasting accuracy of individual LLMs to market-implied probabilities. Accuracy is evaluated using market-level Brier scores, with negative differences indicating lower error for the LLM forecast. Across models and market stages, we do not see any case where an LLM achieves a lower median error than the market. Table 4.6 displays paired Wilcoxon signed-rank test results for each model, showing that LLM forecasts carry significantly higher errors than market-implied probabilities. Full stage-wise results are reported in Appendix B.

Table 4.6: Model vs market-implied probabilities Brier median, negative values indicate better model performance

Model	N pairs	Median Delta Brier	Wilcoxon V	p-value	Rosenthal r	Better
deepseek_deepseek_chat_v3_0324_free	1604	0.08430	1181661	0	0.20501	Market
deepseek_deepseek_r1_free	2380	0.02995	2588993	0	0.16830	Market
google_gemma_3_12b_it	2530	0.06090	2877266	0	0.16324	Market
google_gemma_3_27b_it_free	2497	0.06248	2831661	0	0.16431	Market
openai_gpt_3_5_turbo_0613	2544	0.12172	3037991	0	0.16279	Market
openai_gpt_4o_mini	2544	0.08991	2961964	0	0.16279	Market

4.4.2 Equal-weight LLM crowd

Next, we test whether combining multiple LLM forecasts can improve predictive accuracy. The first aggregation experiment is the simplest one, a simple LLM crowd forecast, by taking the average of all available LLM forecasts for each market and stage (first, mid, last).

$$\hat{p}_{\text{LLM-avg},i,s} = \frac{1}{K}\sum_{k=1}^{K} \hat{p}_{k,i,s}$$

The aggregation performance is evaluated at the market level using paired Brier score differences relative to market-implied probabilities, and is statistically tested using a Wilcoxon signed-rank test of median differences, see Table 4.7.

Table 4.7: Wilcoxon signed-rank results comparing equal-weight LLM average to market

Stage	N markets	Median Δ Brier (LLM − Market)	Wilcoxon V	p-value	Rosenthal r	Rank-biserial r
first	53	0.0695	1346	0	0.7661	0.8812
mid	53	0.0729	1324	0	0.7393	0.8505
last	53	0.0734	1279	0	0.6846	0.7876

Across all stages, the median Brier score difference between the aggregation and the market is positive, indicating higher error for the aggregation than the market. With large effect sizes that are consistent and monodirectional across stages and are statistically significant ($p<0.001$), a performance-weighted LLM aggregation was also tested and showed similar results (see Appendix C). Comparing the median-level Brier score difference to those of individual models shown in Table 4.6. The equal weights ensemble with a stage-averaged median difference Brier score of 0.072 performs better than the worst-performing individual model (GPT-3.5-Turbo) wich has a median diffrence brier score of 0.122. But when comparing in with the best performing individual model (DeepSeek-r1), which has a median difference Brier score of 0.030, it performs worse. This indicates that the equal-weight ensemble evens out variance in forecasting errors, but is not able to achieve errors as low as the stand-alone best-performing model. These results show that simple aggregation alone is insufficient to close the performance gap with market-implied probabilities.

4.5 Hybrid LLM-market forecast aggregations

In addition to pure LLM aggregation, several hybrid aggregations were tested that blend LLM forecasts with market-implied probabilities. All hybrid forecasts are created at the market level for each stage (first, mid, last) and evaluated using the same paired

comparison framework as in the previous sections. For every market and stage, a hybrid forecast probability is created and compared to the market-implied probability benchmark using the difference in Brier scores. Statistical significance is calculated using the Wilcoxon signed-rank test of median differences.

4.5.1 Hybrid forecast construction

Let $\hat{p}_{LLM-avg}$ denote the equal-weight LLM average defined in section 4.4.2, and p_{market} denote the market-implied probability.

50/50 Hybrid: LLM average and market

The first hybrid aggregation assigns equal weight to the LLM ensemble and the market-implied probabilities:

$$\hat{p}_{50/50\,hybrid} = \frac{1}{2}\,\hat{p}_{LLM-avg} + \frac{1}{2}\,p_{market}$$

This aggregation treats the LLM ensemble and the market-implied probabilities as equally informative sources.

Equal-weight hybrid: LLMs and market

A second hybrid aggregation treats the market probability as an additional forecaster alongside the individual LLMs, and gives the all the same weight:

$$\hat{p}^{(\text{all})}_{\text{hybrid}} = \frac{1}{K+1}\left(\sum_{k=1}^{K}\hat{p}_k + p_{\text{market}}\right)$$

Where K represents the number of LLM forecasts available for a given market and stage.

Hybrid aggregation results

Table 4.8 displays the Wilcoxon signed-rank test results comparing each aggregation method to the market benchmark across market stages. The equal-weight LLM aggregation from section 4.4.2 is included for comparison.

Table 4.8: Wilcoxon signed-rank results (Brier) across aggregation methods and stages (Brier difference = Method-Market)

Method	Stage	N markets	Median Δ Brier	Wilcoxon V	p-value	Rosenthal r	Rank-biserial r
Equal-weight LLM avg							
Equal-weight LLM avg	first	53	0.0695	1346	0	0.7661	0.8812
Equal-weight LLM avg	mid	53	0.0729	1324	0	0.7393	0.8505
Equal-weight LLM avg	last	53	0.0734	1279	0	0.6846	0.7876
50/50 hybrid							
50/50 hybrid	first	53	0.0224	1323	0	0.7381	0.8491
50/50 hybrid	mid	53	0.0224	1311	0	0.7235	0.8323
50/50 hybrid	last	53	0.0199	1275	0	0.6798	0.7820
Equal-weight LLM + market							
Equal-weight LLM+market	first	53	0.0626	1341	0	0.7600	0.8742
Equal-weight LLM+market	mid	53	0.0629	1320	0	0.7345	0.8449
Equal-weight LLM+market	last	53	0.0625	1279	0	0.6846	0.7876

Across all stages, both hybrid ensembles reduce the median Brier score difference relative to the LLM-only ensemble. The 50/50 hybrid yields the smallest median difference. That indicates that allocating a substantial weight to the market-implied probabilities narrows the performance gap more effectively. Nevertheless, median Brier score differences are strictly positive across all hybrid aggregations and stages, indicating that hybrid forecasts do not outperform market-implied probabilities. Effect sizes are large and statistically significant ($p<0.001$) across all stages and models.

These results indicate that incorporating market-implied probabilities into the aggregations improves forecasting accuracy compared to LLM-only aggregation. However, it can't compete with the market-implied probability forecasting accuracy. The extent of improvement depends on the weight assigned to the market. While LLM forecasts can be incorporated with market-implied probabilities, they do not improve their forecasting accuracy through any of the methods tested so far. This motivates the evaluation of data-driven aggregation methods that allow weights to be learned rather than fixed from the beginning.

4.5.2 Iron Man model

Next, we evaluate a regression-based model that combines multiple LLM forecasts and market-implied probabilities using learned weights rather than fixed ones. This model, referred to as the Iron Man model, is implemented as a stage-specific logistic regression that maps contemporary LLM forecasts and market-implied probabilities to a predicted probability of the realised binary market outcomes. For a given market and stage, the Iron Man model is defined as:

$$Pr(Y = 1|X) = logit^{-1}\left(\beta_0 + \beta_{mkt}p_{mkt} + \sum_{k=1}^{K}\beta_k p_k\right)$$

where Y is the realised market outcome, p_{mkt} is the market-implied probability, and p_k is the forecast probability produced by LLM k. Separate models are estimated for each market stage (first, mid, last).

Relative to the fixed-weight hybrid aggregations evaluated before, the Iron Man model is the most flexible aggregation considered in this study. It allows predictor weights to be learned from the data instead of having them set in the beginning. The predictive accuracy of the Iron Man model outperforms fixed-weight hybrids; it yields a performance comparable to, and in some cases marginally better than, the market benchmarks, see Table 4.9.

Table 4.9: Iron Man vs market-implied probabilities, median Brier difference (Iron Man – Market)

Stage	N markets	Median Δ Brier	Mean Δ Brier	p-value
first	53	-0.0016	-0.0374	0
mid	53	-0.0005	-0.0394	0
last	53	0.0000	-0.0458	0

The Iron Man model includes multiple highly correlated predictors, as is displayed in Table 4.10 for the mid-stage of the markets. The correlation tables for the first and last stages also show highly correlated predictors; they can be found in Appendix D.

Therefore, due to multicollinearity, individual coefficient estimates are unstable and should not be interpreted in isolation.

Table 4.10:Correlation table between market-implied probabilities and LLM forecasts (mid stage)

	p_market	openai_35_zs	openai_35_cot	openai_4o_zs	openai_4o_cot	google_12b_zs	google_12b_cot	google_27b_zs	google_27b_cot	deepseek_chat_zs	deepseek_chat_cot	deepseek_r1_zs	deepseek_r1_cot
p_market	1.000	0.583	0.524	0.679	0.686	0.617	0.718	0.688	0.703	0.769	0.746	0.742	0.740
openai_35_zs	0.583	1.000	0.948	0.757	0.749	0.716	0.642	0.776	0.786	0.773	0.729	0.794	0.814
openai_35_cot	0.524	0.948	1.000	0.756	0.746	0.729	0.629	0.707	0.718	0.715	0.702	0.752	0.763
openai_4o_zs	0.679	0.757	0.756	1.000	0.976	0.853	0.723	0.829	0.840	0.820	0.836	0.870	0.855
openai_4o_cot	0.686	0.749	0.746	0.976	1.000	0.838	0.716	0.845	0.855	0.813	0.812	0.858	0.841
google_12b_zs	0.617	0.716	0.729	0.853	0.838	1.000	0.846	0.845	0.854	0.806	0.825	0.797	0.809
google_12b_cot	0.718	0.642	0.629	0.723	0.716	0.846	1.000	0.761	0.770	0.736	0.747	0.664	0.680
google_27b_zs	0.688	0.776	0.707	0.829	0.845	0.845	0.761	1.000	0.993	0.883	0.853	0.855	0.862
google_27b_cot	0.703	0.786	0.718	0.840	0.855	0.854	0.770	0.993	1.000	0.894	0.869	0.869	0.873
deepseek_chat_zs	0.769	0.773	0.715	0.820	0.813	0.806	0.736	0.883	0.894	1.000	0.915	0.931	0.936
deepseek_chat_cot	0.746	0.729	0.702	0.836	0.812	0.825	0.747	0.853	0.869	0.915	1.000	0.886	0.888
deepseek_r1_zs	0.742	0.794	0.752	0.870	0.858	0.797	0.664	0.855	0.869	0.931	0.886	1.000	0.976
deepseek_r1_cot	0.740	0.814	0.763	0.855	0.841	0.809	0.680	0.862	0.873	0.936	0.888	0.976	1.000

Table 4.11 displays the estimated coefficients for the Iron Man model in the middle market stage as a representative example. Coefficient tables for the first and last stages are similar and are reported in Appendix E. Coefficient estimates exhibit extremely large standard errors and test statistics close to zero, resulting in p-values near one, which is indicative of severe multicollinearity among predictors.

Table 4.11: Iron Man logistic regression coefficients (mid-stage)

Term	Estimate	Std. Error	z value	p-value
(Intercept)	-13.3882	201770.1	-1e-04	0.9999
p_market	158.7058	394251.3	4e-04	0.9997
openai_35_zs	-71.7074	473423.4	-2e-04	0.9999
openai_35_cot	-98.7337	466888.1	-2e-04	0.9998
openai_4o_zs	-335.9154	1657672.6	-2e-04	0.9998
openai_4o_cot	215.7210	2062006.5	1e-04	0.9999
google_12b_zs	-386.5057	930306.4	-4e-04	0.9997
google_12b_cot	528.2524	1341865.5	4e-04	0.9997
google_27b_zs	461.9816	971453.7	5e-04	0.9996
google_27b_cot	-527.3733	1991091.5	-3e-04	0.9998
deepseek_chat_zs	-406.5344	925220.6	-4e-04	0.9996
deepseek_chat_cot	269.9735	384523.3	7e-04	0.9994
deepseek_r1_zs	90.9195	2076541.6	0e+00	1.0000
deepseek_r1_cot	261.3998	2364604.8	1e-04	0.9999

As a result, the individual coefficients are unstable and should not be interpreted on their own. This regression is therefore used as a forecast-combination tool, rather than to draw conclusions from individual coefficients.

Forecasting accuracy alone does not determine economic relevance; we next evaluate whether these predictions can be translated into profitable trading decisions. Therefore, we embed the Iron Man model into a simple trading strategy. We first evaluate performance under in-sample conditions, where parameters of the Iron Man are estimated using the full dataset.

We begin with the simplest possible implementation: a 1-share buy-and-hold strategy. With a simple trading strategy, for each market × stage observation, the Iron Man either buys one YES contract or one NO contract, depending on whether Iron Man assigns a higher probability than the Polymarket price. The model can buy one YES or NO contract for each market at each stage and then hold it until the market is resolved. Under these in-sample, frictionless trading conditions, the Iron Man returns a total profit of $15.581, with average profit per trade being $0.098 as depicted in Table 4.12. However, due to the in-sample setup, these results should be interpreted as indications rather than as an actionable, economically sensible trading strategy. The following section, therefore, investigates whether this apparent profitability persists under out-of-sample conditions.

4.6 Out-of-sample trading performance

This section evaluates whether LLM-based forecasting models can generate meaningful profits under out-of-sample conditions. For each market, the Iron Man model is re-estimated on all remaining markets, with the target market excluded from training. The resulting model is then used to generate forecasts for the held-out market at each stage (first, mid, and last), these forecasts are then translated into trading rules. This way, we ensure that no information from the evaluated market is used during model estimation to prevent information leakage and to approximate more realistic conditions. Trading outcomes are therefore evaluated on out-of-sample forecasts. Further implementation details are provided in the methodology section.

4.6.1 Sample setup

Trading performance is evaluated under an out-of-sample framework where forecasts for each market are generated without using that market during model estimation. This setup prevents information leakage and approximates deployment-like conditions. Implementation details are provided in the methodology section.

4.6.2 Benchmark market favourite strategy

To set a benchmark for model-based trading performance, we evaluate a very simple trading strategy based solely on the market implied probabilities. The Favourite strategy buys a YES token if the market price is greater than 0.5 $ and buys a no token if the market

price is lower than 0.5 $ and holds them until the market is resolved. Table 4.12 displays the Favourite strategy results.

4.6.3 No-information coin flip strategy

To set a no-information benchmark for model-based trading performance, we evaluate a very simple trading strategy based on no information; instead, we flip a coin if it lands on heads, we buy a YES token, and conversely, if it lands on tails, we buy a NO token. Trades are executed at the contemporaneous market price and held until resolution. Table 4.12 displays the coin flip strategy results.

4.6.4 Iron Man trading strategy

The Iron Man-based trading strategy is evaluated using out-of-sample forecasts, with the same decision rule as in the in-sample analysis described in Section 4.5.2. Table 4.12 shows that under the deployment-like conditions of out-of-sample, the Iron Man-based trading strategy generates a total profit of $1.100, with a median profit of $0.009, and a mean profit of $0.007. While profitability remains positive, it is substantially lower than in the in-sample evaluation.

4.7 Comparing trading strategies

Comparing the Iron Man out-of-sample trading strategy to the in-sample trading strategy, we can see that both can produce profit, although the profit is dramatically decreased in the out-of-sample trading strategy compared to the in-sample strategy. The average profit per trade drops from $0.098 per trade to $0.007 per trade.

Table 4.12: Overall trading performance: Iron Man (in-sample) vs Iron Man (out-of-sample) vs Market Favourite vs Coin Flip

Strategy	N observations	N trades	Total PnL	Mean PnL (per trade)	Median PnL (per trade)	Share profitable trades
No-information (coin flip)	159	159	-0.6785	-0.0043	0.000	44.7%
Market favourite	159	159	4.3805	0.0276	0.016	95.6%
Iron Man OOS	159	159	1.1005	0.0069	0.009	67.3%
Iron Man	159	159	15.5815	0.0980	0.019	82.4%

Now, if we compare the out-of-sample Iron Man results to the market benchmark (see Table 4.12), all measured profit metrics favour the simpler non-LLM driven Market favourite strategy, where total profits for the market favourite is almost four times greater

than that of the out-of-sample one, and median profits go from $0.016 for the market favourite to $0.009 for the out-of-sample iron man model.

The no-information coin-flip benchmark provides an additional point of reference: while the Iron Man strategy outperforms a purely random trading rule, it does not close the gap to the market-based benchmark. Therefore, we can derive that there is no economic advantage to using LLM forecasting to enhance or predict market outcomes. Overall, the results show that LLM-driven trading strategies do not outperform even simple market-based rules under out-of-sample conditions.

4.8 Result summary

This section summarises the empirical findings corresponding to the four sub-questions. The results show us that LLMs can generate consistent probabilistic forecasts and are able to support profitable trading strategies. Even though the trading strategies are profitable, they do not provide any advantages over market-implied probabilities when evaluated out-of-sample. Improvements to forecasting accuracy from prompt engineering and model aggregation are limited in magnitude. Market liquidity is systematically associated with differences in forecasting accuracy. The relative accuracy gap between the LLM predictions and the market is smallest in low-liquidity markets and larger in mid- and high-liquidity markets. However, LLM forecasts are never able to consistently outperform or even match the market-implied probabilities, regardless of the liquidity level or other tested conditions. Furthermore, within the set of trading strategies evaluated in this study, simple rules based solely on market-implied probabilities yield a higher profit than the LLM-based strategies tested. This shows that within our experimental setup, a simple strategy based on public market information performs better than the best LLM-based strategy that we tested and evaluated. The following chapter discusses the implications of these findings in a broader perspective.

5 Discussion

This thesis is not concerned with information for information's sake, but rather with information for decision making. The setting of decentralised prediction markets is therefore not random; using their market-implied probabilities as a benchmark was a deliberate decision, as they represent a real-world mechanism that aggregates dispersed beliefs under incentives and updates continuously through trading activity. This benchmark is demanding; it does not reflect a theoretical optimum, rather it represents a functioning institutional mechanism for probabilistic forecasting under uncertainty (Wolfers & Zitzewitz, 2004).

The purpose of the discussion is therefore not to reassess whether prediction markets are "correct" or "efficient" in the theoretical sense. Prediction-market literature emphasises that prices do not necessarily correspond to literal mean beliefs to be informative or relevant for decision making. Manski (2006) cautions against interpreting market prices as direct estimates of average beliefs, although he also recognises their value as information aggregation tools for decision making. What matters in this context is therefore not theoretical perfection but rather comparative performance. Market prices provide one of the best available, low-cost aggregations of information at a given point in time.

Viewed in this way, the analysis aligns with the broader info-finance view that forecasting performance is a property of systems, rather than individual agents or models. Buterin (2024) shows markets as information-processing institutions that transform diverse signals into actionable probabilities through incentives and feedback, not through superior knowledge or cognition at the level of any single participant or forecasting agent. (Buterin, 2024; Hanson, 2003). This perspective motivates comparisons not between individual forecasters or models, but rather between alternative institutional ways for producing actionable decision-relevant probabilities.

In contrast, large language models operate as standalone systems. They generate forecasts without incentives, capital allocation, or continuous improvements. The question is therefore not whether LLMs can generate probabilistic predictions, but rather whether

they can replace, complement or economically improve upon an existing information aggregation mechanism that already occupies this role in practice (Buterin, 2024).

The discussion follows a progression that reflects this institutional framing. First, we ask whether LLM forecasts are a suitable replacement for market-implied probabilities for forecasting purposes. Anchoring the question with Brier score comparisons. Second, we consider whether LLMs can add value when combined with public market information. Third, we evaluate if any complements are meaningful once cost, effort, and simple alternatives based on public market information are considered.

Replacement is the strongest claim; if unsupported, it rules out any even stronger claims. Complementary is weaker, and only interesting if applicable beyond the theoretical realm and can work in practice. Finally, improvements must be assessed relative to other possible improvements, such as simple strategies that rely solely on publicly available market information. This progression ensures a disciplined basis for interpreting the results, making sure that claims about LLM-based forecasting are evaluated relative to existing institutional alternatives rather than in isolation.

5.1 Are large language models a suitable replacement?

The first and strongest claim is whether large language models can substitute for decision markets and their market-implied probabilities for forecasting in a decision-making context. We directly addressed this question in the Results through paired comparisons of LLM forecasts and market prices using Brier scores. In no case did the LLM forecasts consistently achieve a lower predictive error than the market-implied probabilities. This finding, although interesting, is not unexpected when interpreted through the lens of prediction-market theory. Prediction markets are explicitly designed to aggregate dispersed information using incentives, and their price reflects the beliefs of actors participating in the market, who have invested interest and the ability to update their positions as new information arises, private or public (Wolfers & Zitzewitz, 2004). As a result, achieving predictive accuracy comparable to, or better than, the market-implied probabilities requires more than generating plausible probabilistic predictions. It requires access to decision-relevant information that is either unavailable or not yet incorporated by the market and the ability to effectively allocate a probability assessment to it.

The results show that LLM forecasts do not meet this standard. Even though LLMs can produce consistent probability estimates, these estimates do not translate into higher or even comparable accuracy relative to market prices. This suggests that LLMs do not uncover additional decision-relevant information beyond what is already reflected in market prices at the time of forecasting or that they simply cannot translate this information to probabilistic prediction with the same accuracy that the market is capable of.

This conclusion does not rely on strong assumptions about market efficiency. As pointed out in Manski (2006), market prices should not be interpreted as literal averages of individual beliefs, or as guarantees of optimality. Nevertheless, even under these interpretative cautions, market prices remain highly informative aggregation mechanisms. The comparison, therefore, sets a demanding but fair benchmark: If an alternative forecasting method cannot outperform market prices, claims about it replacing institutional forecasting mechanisms must be rejected.

The results show a clear constraint on what LLM-based forecasting can achieve in this setting. Information aggregation mechanism design emphasises that performance gains come from aligning incentives, participation, and feedback rather than from superior individual cognition alone (Buterin, 2024; Hanson, 2003)**.** The results are consistent with this view; without incentives and continuous updating, LLMs are not capable of substituting markets as standalone forecasting solutions.

An extension of the replacement question is whether LLM forecasts can benefit from aggregation, in line with the wisdom of the crowds theory. If individual LLM forecasts contain partially independent errors, then aggregating across models could, in theory, reduce noise and improve predictive accuracy (Hamada et al., 2020). This logic mirrors the motivation behind prediction markets, as they take dispersed beliefs and aggregate them into a single probabilistic measurement. The results show that simple aggregation of LLM forecasts does reduce variance relative to individual models. The reduced variance is not nearly enough to close the performance gap when compared with the market benchmark, as it remains dominant in terms of predictive accuracy. These findings are interesting and suggest that forecast errors across LLMs are not independent enough for aggregation alone to close the performance gap. This could be explained by the fact that

they are all fed the same information from the information web scraper for individual markets and therefore reach similar conclusions.

Taken together, both individual and aggregated LLM forecasts fail to achieve predictive accuracy comparable to market-implied probabilities. This rules out replacement both at the single model level and when applying aggregation-based wisdom of the crowd approaches. This narrows the scope and begs the question whether LLMs can add value when combined with the market information rather than replacing it outright.

5.2 Are large language models complements?

Seeing that replacement has been ruled out, the question remains: can LLMs add value when combined with market-implied probabilities, rather than replacing them? The results show that hybrid aggregations that include the market-implied probabilities can narrow the performance gap and outperform LLM-only-based forecasts. However, improvements relative to the market benchmark are limited; across the evaluated configurations, only a single hybrid aggregation achieves marginally lower errors than the market baseline. That is the Iron Man model operating under in-sample conditions.

The in-sample performance of the Iron Man model indicates that when optimally combining LLM forecasts with market-implied probabilities under optimised conditions, it is possible to construct a forecast that achieves a lower predictive error than the market benchmark. This suggests that LLM-based forecasting signals can contribute useful information when integrated within a combined forecasting system. However, this additional contribution appears to be fragile. These improvements over the market benchmark are confined to the in-sample setting and do not hold up in an out-of-sample one, where the market-implied probabilities remain the most accurate forecast. This limits the interpretation of the Iron Man result: it shows that LLM forecasts can complement market-implied probabilities within a specific sample, but it does not provide evidence that such gains can hold up under realistic forecasting conditions. Interpreted in this way, the Iron Man results support a narrow form of complementarity. LLM forecasts can contribute marginally to predictive accuracy improvements within optimised conditions. These improvements disappear when we move beyond the sample used for construction. As a result, the findings show that different forecasting signals can be combined, but do

not show that LLM-based augmentation of prediction markets would be beneficial in practice.

5.3 Trading as a decision-making test

One way to evaluate whether LLM-based forecasts contribute meaningfully in a practical decision-making environment is to move beyond forecast accuracy and examine trading performance. Trading simulations provide a natural testing environment for this purpose as they translate probabilistic forecasts into concrete, actionable decisions with measurable economic consequences. Instead of asking whether LLMs can generate slightly better probability assessments in an ideal world. This approach examines whether LLM forecasts would be a sensible, worthwhile mechanism when deployed in a realistic decision-making context, when compared with other options. Trading simulations serve two roles. First, they allow us to operationalise forecasting accuracy by making forecasts that determine actual decisions, such as whether to take a position relative to market prices. To anchor the interpretation of trading performance, a no-information coin-flip strategy was included as a lower-bound benchmark. This strategy translates random decisions into trades executed at prevailing market prices, ensuring that profitability above this level reflects the exploitation of information rather than chance alone. Second, they allow performance to be evaluated and compared with other possible improvements to the prediction markets by utilising simple trading strategies based solely on readily available public market information. This shifts the focus from improved statistical accuracy to decision-relevant performance.

The initial trading results were very promising. Trading simulations under in-sample conditions show that the Iron Man model can generate substantial profits by leveraging a slight accuracy edge relative to the market and translating it into a simple decision rule. That rule being if Iron Man forecast is greater than that of the market buy YES token, and conversely if it's lower buy a NO token. This shows that an optimised combination of LLM forecasts and market-implied probabilities can be engineered to exploit patterns in the observed sample. By doing so, demonstrating feasibility in a narrow sense: under favourable conditions, with optimisation on the evaluation sample, it is possible to create a decision-making process that beats the market. However, under more deployment-like,

out-of-sample conditions, the picture changes as profits from the same strategy fall off sharply. Although the Iron Man strategy remains profitable, the profits are less then 10% of the in-sample one. This profit deterioration indicates that in-sample gains do not generalise reliably.

Looking at other possible improvements through the lens of cost-benefit comparisons, a cheap outside option based solely on public market information exists. A simple decision rule that buys a YES position when the market price exceeds 0.5 and a NO position when it falls below 0.5 yields a higher trading profit than the out-of-sample Iron Man strategy. This option requires no LLM implementation, no hybrid aggregations, and no resources that are not publicly available. The comparison between the two reveals that any gains from LLM-based trading are dominated by a cheaper strategy that relies exclusively on publicly available market prices.

This result can be interpreted through a broader lens of information efficiency. Price-based markets have been shown to incorporate almost all available public information in their prices, while falling short of strong-form efficiency and leaving small, structural deviations that can be exploited (Page & Siemroth, 2021; Wolfers & Zitzewitz, 2004). A similar logic possibly applies to prediction markets, which are also price-based aggregation markets. The simple price-based strategy exploits the structure already embedded in market prices, rather than information external to the market. These limited inefficiencies are therefore sufficient for a simple decision rule to generate excess returns, without implying a failure of the market as an information aggregation mechanism.

Taken together, this implies that even though LLM-based trading can beat the market, it is hard to economically justify. If a simple decision strategy operating solely on public market prices performs better in deployment-like conditions than a complex hybrid model. The relevant constraint is no longer model capability but rather opportunity cost: The time and effort, along with the operating cost required for LLM integration, do not buy any advantage over what is possible with available public market information, and therefore should not be pursued.

5.4 Market dynamics

Before we dive into broader implications, we will look at how forecasting performance varies across market stages and liquidity levels, not to identify exceptions from the main findings, but rather to better understand the mechanisms behind them and clarify how differences between LLM forecasts and market-implied probabilities evolve with market dynamics.

The results indicate that the performance gap between LLM forecasts and market-implied probabilities is smallest in the middle stage of a market's life cycle. In the intermediate stages of markets, LLM forecasts performance is relatively close to that of the market-implied probabilities, and as markets develop further as resolution approaches, the relative performance gap widens. This pattern is consistent with market-based information aggregation, where markets become more efficient as markets mature and trading activity increases close to the resolution date, while LLM forecasts stay comparatively stable.

We identify a slightly different pattern when examining liquidity levels. In lower-volume markets, where trading activity is lower, information aggregation of the market is more limited, and therefore, performance gaps are smaller. In deeper markets, with greater liquidity, the prices are updated more frequently, indicating a greater aggregation of dispersed information. In these markets, the comparative predictive advantage of market-implied probabilities becomes higher.

These findings are consistent with existing literature that states prediction markets become more informative as participation increases (Hanson, 2003). The market dynamics analysis further reinforces the explanation developed earlier regarding information aggregation. Markets improve as time passes and new information is incorporated through trading activity, especially as events approach their resolution date and market activity increases, while LLM forecasts stay comparatively static. In this sense, prediction markets learn over time, whereas LLMs remember.

5.5 Implications on info-finance

Recent discussion around info-finance argues that advancements in AI, more specifically the capabilities of large language models, may enable new ways of information

aggregation, potentially replacing or complementing existing prediction market mechanisms. In this perspective, accuracy is treated as a system-level outcome, influenced by how information is processed, combined, and acted upon, rather than by the superior intelligence of any single actor (Buterin, 2024).

The findings of this study place clear restrictions on such claims. While LLMs are capable of producing probabilistic forecasts and can be incorporated into combined forecasting systems, they cannot replicate the core features that make prediction markets efficient information aggregation tools. Across individual, aggregated crowds, and hybrid configurations, LLM-based forecasts fail to achieve predictive accuracy comparable to market-implied probabilities under realistic conditions. Even though in-sample optimisations yield strong performances, these gains do not generalise out-of-sample, where they almost diminish and are dominated by simpler, cheaper strategies that solely rely on public market prices. From an info-finance point of view, the results suggest that the advantages of prediction markets do not originate from superior reasoning or foresight at the level of individual forecasters, but rather from institutional properties: incentives, capital investments, continuous updates and responses to new information. These features allow markets to adapt quickly as events develop and information becomes available. This particularly applies when market activity increases as the market life cycle comes to maturity and resolution approaches. LLM-based systems do not appear to be able to compete with this dynamic market aggregation as they are not able to update as effectively in response to new information. At least not at the same rate as the market, and are not privileged to any private information.

These interpretations are consistent with earlier work, emphasising that effective belief aggregation depends on mechanism design rather than agent intelligence alone (Hanson, 2003). From this point of view, LLMs cannot substitute for the markets, although they can work as inputs that can meaningfully contribute to them, only when embedded within systems that already perform aggregation through incentives and feedback. The in-sample results of the Iron Man model illustrate this narrowly defined form of complementarity: when aggregation weights are estimated on observed data, the model can extract incremental predictive signal from LLM forecasts beyond what is contained in market-implied probabilities. However, the sharp decline in performance under out-of-sample

conditions indicates that this signal does not generalise, and therefore does not translate into a practical alternative to market-based aggregation.

Looking at the results through a wider scope, the findings caution against interpreting LLM forecasts as standalone information aggregators in areas where prediction markets already exist. Claims that AI can replicate market-based forecasting seem to overlook the role of participation, incentives, and dynamic continuous updating in producing decision-relevant probabilities. LLMs can assist and complement in structuring, summarising and interacting with existing information under controlled, carefully defined conditions, but they do not replicate the institutional learning that serves as the base for the performance of prediction markets. This conclusion is specific to environments where prediction markets already exist and provide a benchmark to compare against. In an environment where no such market exists, LLM-based forecasting systems may be relevant for decision-making under uncertainty. However, such applications fall outside the scope of this study and are therefore not considered.

To summarise, the findings from this study suggest that current advances in model capabilities in isolation are not sufficient to reproduce the informational aggregation advantages of market-based systems. For LLMs in the sense of info-finance to offer practical improvements over the already existing prediction markets, it would likely need to address how incentives, updating and feedback are implemented at the system level.

5.6 Limitations

The findings of the thesis should be interpreted in light of several limitations regarding experimental design choices and scope in mind. First, the selection of large language models was limited in numbers to only six models and did not include the most recent models available at the time of data collection. This constraint was primarily driven by practical considerations such as time and cost. As a result, the conclusion drawn from this research applies to the models evaluated and may not generalise to newer or more powerful models.

Second, the prompting strategies considered in this study were restricted to Chain-of-Thought and Zero-shot prompting. More complex prompting approaches were not explored. This approach allowed for a controlled comparison of baseline prompting

effects, although the results are limited to these two prompting strategies and cannot speak to the full range of possible prompting strategies.

Third, the analysis is based on a fixed set of markets on the Polymarket over a limited time window. This setting allowed for a detailed and internally consistent evaluation, although it restricts the external validity of the findings across different prediction platforms, market samples, and time windows.

Finally, the trading simulations used in this study use simplified trading strategies designed for relative performance evaluations rather than realistic deployment. While appropriate for comparing economic outcomes across forecasting approaches under controlled conditions. This design excludes practical constraints that could influence trading performance in a real market setting.

6 Conclusion

The objective of this thesis was to empirically test and evaluate the predictive forecasting performance of large language models relative to market-implied probabilities gathered from decentralised prediction markets, which serve as an approximation of aggregated human forecasting. Applying a quantitative experimental design, LLM-generated probability forecasts were systematically compared to market-implied probabilities across several models, prompting strategies, market stages, and market depths. Forecasting performance was evaluated on two different criteria: statistical accuracy, measured using Brier scores, and economic value, assessed through trading simulations. This chapter summarises the empirical findings presented in the Results chapter to answer the main research question and its sub-questions based strictly on the research findings. Ultimately, this research objective was found to be achieved

6.1 Summary of empirical findings

At the baseline level, forecasts generated by large language models consistently carry a higher forecasting error than market-implied probabilities across all evaluated models and market stages. In comparison to the market benchmark, the LLM-generated probabilities display higher variance and lower stability, resulting in wider error distribution. Individual LLM forecasts can occasionally outperform market-implied probabilities at the observation level, although these instances do not translate into systematic accuracy improvements and cannot be relied upon as a forecasting advantage.

Forecasting Accuracy and Market Liquidity

When comparing forecasting performance across market liquidity stages, the relative accuracy gap between LLM-generated forecasts and the market-implied probabilities varies systematically with liquidity. The markets with the lowest liquidity exhibit the smallest performance gap and median Brier differences compared to the market benchmark. The mid- and high-liquidity markets have considerably larger accuracy gaps. This pattern persists across every market stage, and at no stage is the performance of liquidity groups changed. Some fluctuations in accuracy are observed over the market life

cycle, most notably a slight increase in accuracy at the midpoint of a market's life cycle in the low and mid liquidity groups. These observed variations do not change the overall pattern that LLM-forecasts are closest in predictive accuracy in low-liquidity markets, although at no point do they reach or outperform the accuracy of the market benchmark.

Effects of Prompt Engineering

To verify if prompt engineering improves forecasting accuracy, Chain-of-Thought prompting was compared to Zero-shot prompting under paired forecasting conditions. At the aggregated level, the results show that no systematic difference in median forecasting accuracy was identified between the two prompting strategies. Paired statistical tests revealed a statistically significant result, but no median shift; the significance is driven by the large number of observations rather than by meaningful effect sizes. Evaluating the effect of prompting on the individual model level revealed that prompt-related accuracy differences are small and inconsistent across models, with neither prompting strategy producing a reliable or practically relevant accuracy gain. This indicates that the selection of prompting strategies does not lead to meaningful improvements in baseline predictive performance.

Forecast Aggregation and Hybrid Models

To evaluate whether combining multiple model forecasts could improve predictive performance, multiple aggregation approaches were employed. These approaches combine forecasts across models, in various ways, including simple equal-weight averages across LLM forecasts as well as hybrid combinations that integrate the market-implied probabilities with the LLM forecasts. Across all evaluated aggregation methods, the purely LLM-based forecasts are less accurate than the market benchmark. The hybrid aggregation methods improve upon the forecasting accuracy, although they are also less accurate than the market benchmark. The Iron Man, a flexible logistic aggregation model, was able to produce accuracy comparable to the market benchmark. However, its estimated coefficients are unstable due to multicollinearity between inputs, limiting its reliability as a general aggregation approach.

Economic Significance of LLM-Based Forecasts

To assess whether statistical accuracy translates to economic significance, trading simulations were conducted using a LLM-powered, hybrid regression model(Iron Man) and compared against simple benchmark strategies. In-sample results show that the Iron Man simulation strategy can generate positive returns. Although these returns diminish when evaluated in a more deployment-like out-of-sample setting. Under out-of-sample evaluation, the Iron Man trading strategy underperforms the simple benchmark strategy that is solely based on publicly available market-implied probabilities. More specifically, the benchmark strategy trades on the market-favoured outcome consistently yields a higher profit than that of the Iron Man trading strategy. As a result, the LLM-powered trading strategies evaluated in this thesis cannot outperform or even match the simplest of benchmark strategies and are therefore not be deemed economically significant.

6.2 Answer to the Main Research Question

This thesis evaluated the extent to which large language models can replicate or complement human forecasting as is reflected in decentralised prediction markets. Across all evaluated models, prompting strategies, market conditions and aggregation methods, LLM-generated forecasts are not able to achieve a consistently lower forecasting error than market-implied probabilities. Aggregation methods, more specifically hybrid market/LLM aggregations, were able to considerably reduce the performance gap; none provide a systematic improvement when compared to the market benchmark. Furthermore, LLM-based trading strategies fail to outperform simple benchmark strategies based solely on public market information when evaluated in deployment, such as out-of-sample conditions. These findings indicate that under the conditions evaluated in this study, large language models neither replicate nor complement human forecasting in a way that delivers superior predictive or economic performance relevant for decision-making.

6.3 Future research

The findings of this thesis may spark interest in areas of future research, primarily related to scope, system design and evaluation context.

One area for future research could be to evaluate and examine LLM-based forecasting in decision environments where no existing prediction markets are in place, and their implementation is, for some reason, deemed impractical. This current study evaluates LLM predictive performance relative to a strong benchmark that aggregates information through incentives and continuous updating. In environments where no such mechanisms exist, LLM-based systems may serve a decision-relevant purpose, for example, by serving as provisional forecasting tools or assisting decision-making under uncertainty. Evaluating LLM applicability under these conditions would require alternative benchmarks and other decision criteria. This falls outside the scope of the current analysis.

Another area for future research could be to investigate system-level designs that better replicate the institutional features of prediction markets. The results presented in this study indicate that the marginal gains from hybrid models are sensitive and have limited generalizability out-of-sample. Suggesting that improvements are unlikely to emerge from aggregation alone. Instead, future research could explore the possibilities of incorporating market-like incentives and feedback into the model structure, allowing LLM-based forecasting agents to better respond to new information over time, thereby replicating the core functionalities of prediction markets.

Another area relevant for future research is that of technological advancements. This study evaluated the current stage of LLM-based forecasting at the time of this analysis. Given the rapid pace of AI and LLM development in recent times, it may be informative to re-examine the research question of this thesis in the future, once models have advanced further in the future. Re-evaluation could assess whether more advanced models can better replicate or complement market-based information aggregation. This does not affect the conclusions drawn in this thesis, as they reflect the current state of the technology, and any future technology developments fall outside the scope of this current study.

Future research could also look at LLM-forecasting assisted decision performance in a broader sense. In this study, decision relevance is assessed through trading simulations. These LLM-based simulation strategies fail to match the performance of simple, low-cost approaches that rely solely on publicly available market information. Extending this type of decision-level evaluation to other relevant decision-making settings, such as policy analysis or strategic decision-making within organisations, could help clarify when and

under what conditions LLM-based probability assessments are effective for decision-making under uncertainty.

Final Remarks

This thesis offers an empirical evaluation of the extent to which large language models can replicate or complement human forecasting as seen in decentralised prediction markets. Applying a controlled experimental design, probabilistic forecasts generated by large language models were evaluated and compared against market-implied probabilities across multiple models, prompting strategies, market conditions and aggregation approaches. The results show that under the experiment conditions, LLMs do not reach the predictive forecasting accuracy of prediction markets, and improvements obtained through aggregation do not translate into systematic decision-relevant economic advantages. Under the conditions evaluated, LLM-based forecasting approaches do not provide any meaningful advantages over market-based human aggregation. The contribution of this study is to show to what extent and under what conditions large language models fall short of complementing or replacing human judgment.

7 References

Buterin, V. (2024). *From prediction markets to info finance.* Vitalik.eth.limo. Retrieved from https://vitalik.eth.limo/general/2024/11/09/infofinance.html

Chen, J. (2023). *Efficient market hypothesis (EMH): Forms and criticisms.* Investopedia. Investopedia. Retrieved from https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp

deepseek-ai. (2025). *DeepSeek-R1 (GitHub repository).* GitHub. Retrieved from https://github.com/deepseek-ai/DeepSeek-R1

deepseek-ai. (2025). *DeepSeek-V3 (GitHub repository).* GitHub. Retrieved from https://github.com/deepseek-ai/DeepSeek-V3

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, B., Wilson, B., Chen, Y., . . . Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, 112*(50), 15343-15347. https://doi.org/10.1073/pnas.1516179112

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25*(2), 383-417.

Google AI. (2025). *Gemma 3 model overview.* Google AI. Retrieved from https://ai.google.dev/gemma/docs/core

Grossman, S., & Stiglitz, J. (1980). On the impossibility of informationally efficient markets. *American Economic Review, 70*(3), 393-408.

Hamada, D., Nakayama, M., & Saiki, J. (2020). Wisdom of crowds and collective decision-making in a survival situation with complex information integration. *Wisdom of crowds and collective decision-making in a survival situation with complex information integration, 2020*. https://doi.org/10.1186/s41235-020-00248-z

Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers, 5*(1), 107-119.

Hanson, R. (2013). *Bet on beliefs.* Overcoming Bias. Overcoming Bias. Retrieved from https://www.overcomingbias.com/2013/01/bet-on-beliefs.html

Levy, M., Jacoby, A., & Goldberg, Y. (2024). *Same task, more tokens: The impact of input length on the reasoning performance of large language models.* arXiv. Cornell University. Retrieved from https://arxiv.org/abs/2402.14848

Manski, C. F. (2006). Interpreting the predictions of prediction markets. *Economics Letters, 91*, 425-429. https://doi.org/10.1016/j.econlet.2006.01.004

Metaculus. (2025). *forecasting_tools/helpers/asknews_searcher.py (GitHub code file).* GitHub. Retrieved from https://github.com/Metaculus/forecasting-tools/blob/main/forecasting_tools/helpers/asknews_searcher.py

OpenAI. (2024). *GPT-4o Mini: Advancing cost-efficient intelligence.* OpenAI. Retrieved from https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

OpenAI. (2025). *GPT-3.5 Turbo.* OpenAI. Retrieved from https://platform.openai.com/docs/models/gpt-3.5-turbo

Page, L., & Siemroth, C. (2021). How much information is incorporated into financial asset prices? Experimental evidence. *The Review of Financial Studies, 34*(9). https://doi.org/10.1093/rfs/hhaa143

Plott, C. R., & Chen, K.-Y. (2002). *Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem.* California Institute of Technology. Pasadena, CA: California Institute of Technology. Retrieved from https://www.researchgate.net/publication/4822805

Polymarket Documentation. (2024). *How Are Prediction Markets Resolved?* Polymarket. Retrieved from https://docs.polymarket.com/polymarket-learn/markets/how-are-markets-resolved

Prompting Guide. (2025). *Prompting techniques.* Prompting Guide. Retrieved from https://www.promptingguide.ai/techniques

Rhode, P. W., & Strumpf, K. S. (2006). *Manipulating political stock markets: A field experiment and a century of observational data.* University of North Carolina at

Chapel Hill. Chapel Hill, NC: University of North Carolina at Chapel Hill. Retrieved from https://www.nber.org/papers/w12247

Roetzel, P. G. (2019). Information overload in the information age: A review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research, 12*, 479-522. https://doi.org/10.1007/s40685-018-0069-z

Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A., & Lenert, L. A. (2025). The emergence of large language models as tools in literature reviews: A large language model-assisted systematic review. *Journal of the American Medical Informatics Association, 32*, 1071-1086. https://doi.org/10.1093/jamia/ocaf063

Schoenegger , P., Park, P., Karger, E., Trott, S., & Tetlock, P. (2024). *AI-augmented predictions: LLM assistants improve human forecasting accuracy.* arXiv preprint, arXiv. Retrieved from https://arxiv.org/abs/2402.07862

Wei, J. Z., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., . . . Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* (pp. 24824-24837). Red Hook, NY: Curran Associates, Inc. Retrieved from https://arxiv.org/abs/2201.11903

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives, 18*(2), 107-126.

Wolfers, J., & Zitzewitz, E. (2006). *Prediction markets in theory and practice.* Bonn: Institute of Labor Economics (IZA). Retrieved from https://hdl.handle.net/10419/33384

Appendix A: Reflection on AI use

Seeing how the topic of this research revolves around the current capabilities of large language models, it would have felt off-brand not to see if LLMs could assist me with the research and creation of the thesis. This section will explain how and to what extent AI was used in the process of creating this thesis. AI was used responsibly to assist with the following:

1. **Assistive use**

   Literature review: ChatGPT(o1 pro) was used to create a preliminary literature review that served as a base for the theoretical background of the study. Looking back, the AI-generated preliminary literature review was a good starting point for the theoretical background; it provided a long list of possible sources and exposed many areas that might be of interest for the study. All sources were manually checked for legitimacy and read before and if they were used in the research. The list of sources provided through AI synthesis was not final; many sources were dropped after manual inspection, and others were added to it. AI did not replace the process of exploring and investigating literature, but rather was carefully used as a tool to assist with the process.

   Coding: AI proved to be a valuable tool in generating the data collection and analysis scripts. Through careful prompting and feedback loops when debugging, AI was able to help with creating the code used for these scripts. All code generated by AI was carefully inspected manually, and mistakes in the code were fixed. The AI tool used for this purpose was ChatGPT.

   Spell check: AI proved useful in going over text and identifying errors made by the author regarding the text's integrity. Highlighting errors made by the user and suggesting slight edits to sentences to improve flow, readability and elegance of the text. The AI tool Grammarly was used for this part of the thesis creation process.

2. **Data generation**

   In addition to assistive use listed above, large language models were directly used in the data collection process as data-generating agents for the empirical analysis. LLMs

were prompted to generate probabilistic forecasts for real-world prediction market events. These forecasts, in addition to the market-implied probabilities and the market resolutions, make up the primary dataset used to evaluate LLM forecasting performance relative to market-implied probabilities.

This use of AI is fundamentally different from the assistive applications; the models go from an assistive role supporting the research to objects of the study themselves. All LLM-generated forecasts were produced under a fixed prompting strategy, without feedback, adaptive learning or interaction between models, and were collected systematically over time as a part of the experimental design.

3. Prompting standardisation

To ensure reproducibility and comparability across different models, prompts were standardised. A single research summary prompt, along with a zero-shot forecasting prompt and a Chain-of-Thought prompt, were applied uniformly across models. A separate prompt was used for the initial literature review. The prompts are:

Zero-shot forecasting prompt:

You're a forecasting assistant.
Question: *[question.title]*
Description: *[question.description]*
Research Summary:
[summary]
Based on this information, estimate the probability this event will occur.
Respond only with a number between 0 and 1, and nothing but a number, no text just this number between 0 and 1!.

Chain-of-Thought forecasting prompt:

Question: *[question.title]*
Description: *[question.description]*
Research Summary:
[summary]
Think step by step. Consider all relevant information before giving a final answer.
At the end, estimate the probability this event will occur. Respond only with a number between 0 and 1, It is very important that you don't include any text in you final answer only the number.

Summary prompt:

Extract and synthesize key insights from the following research text.
Focus on trends, arguments, and signals relevant to forecasting.
Do not assign any probabilities or predictions.

Literature review prompt:

I am going to give you a blog post by vitalik buterin on information finance. There is a lot of scientific literature that is behind the ideas presented here. Please write me a scientific analytical and conceptual literature review of those ideas and give me the references from theliterature. Here is the blog post;

https://vitalik.eth.limo/general/2024/11/09/infofinance.html

Overall, I believe AI provided useful assistance in the above-listed parts of the research process. I furthermore recognise that AI is a powerful tool that requires careful provision to be used responsibly and ethically. I hereby state that AI was used responsibly in the ways previously listed in the creation of this thesis.

Appendix B: Wilcoxon signed-rank test

Table 7.1: Wilcoxon Signed-Rank Test Brier

N pairs <int>	Wilcoxon V <dbl>	p-value <dbl>	Estimate <dbl>	Conf.low <dbl>	Conf.high <dbl>
14099	90978873	0	0.07801967	0.07619056	0.08047426

The Wilcoxon signed rank test results show a median difference of 0.078, with a tight confidence interval of 0.076-0.080.

The test has a p-value < 0.001, showing that the difference is statistically significant. Seeing as how the median difference is positive, we can derive the conclusion that Polymarket achieves a lower Brier score than the LLMs in the majority of the paired market comparisons. This suggests that the gap seen in the descriptive results is not due to chance. Rather, it is systematic and holds across the full dataset.

Table 7.2: Wilcoxon Signed-Rank Test Absolute Forecast Errors

N pairs <int>	Wilcoxon V <dbl>	p-value <dbl>	Estimate <dbl>	Conf.low <dbl>	Conf.high <dbl>
14099	93799471	0	0.2334686	0.2294477	0.2365275

The absolute-error comparison is consistent with the previous Brier-score test.

Across the 14,099 paired forecasts, absolute errors are significantly larger for the LLM forecasts than the Polymarket equivalents. The estimated median difference is about 0.23, with a tight confidence interval of 0.229-0.237.

The test has a p-value < 0.001, showing that the difference is statistically significant.

These results, combined with the previous results, show that the Polymarket delivers consistently lower forecasting errors, regardless of whether the accuracy is measured through squared error or absolute deviation.

Table 7.3: Wilcoxon Signed-Rank Onesided

Hypothesis	N pairs	Wilcoxon V	p-value	Method	Alternative
Polymarket better (greater)	14099	90978873	0	Wilcoxon signed rank test with continuity correction	greater
LLM better (less)	14099	90978873	1	Wilcoxon signed rank test with continuity correction	less

Both the one-sided tests point towards the same conclusion as the two-sided results, that the Polymarket predictions are more accurate than those of the LLMs.

On one hand, the Polymarket-better test returns a p-value of 0, suggesting that Polymarket achieves lower Brier scores than the LLMs.

The LLM-better test, on the other hand, fails as it has a p-value of 1, and thus is not supported.

Together, the tests further confirm that the performance of the Polymarket forecasts is greater across the matched forecasts.

Appendix C: Performance-weighted LLM average vs Polymarket

Table 7.4: Wilcoxon signed-rank results comparing performance-weighted LLM average to Polymarket by stage.

stage	n_markets	median_diff_brier	mean_diff_brier	V_brier	p_brier	r_rosenthal_brier	r_rank_biserial_brier	median_diff_abs	mean_diff_abs	V_abs	p_abs	r_rosenthal_abs	r_rank_biserial_abs	direction_brier	direction_abs
first	53	0.0696	0.0949	1345	0	0.7649	0.8798	0.2047	0.2186	1401	0	0.8330	0.9581	Market better	Market better
mid	53	0.0693	0.0819	1321	0	0.7357	0.8463	0.2156	0.2154	1364	0	0.7880	0.9064	Market better	Market better
last	53	0.0725	0.0765	1279	0	0.6846	0.7876	0.2474	0.2495	1331	0	0.7479	0.8602	Market better	Market better

Looking at the results, it is clear that weighting LLMs relative to their performance does not improve their accuracy relative to Polymarket.

On all stages, the performance-weighted LLM crowd average produces higher Brier scores and absolute errors than those of the market. Comparing the weighted averages to the equal-weighted averages reveals that there is only a marginal change in their results. This suggests that the difference in model quality is not large enough to meaningfully change the aggregated forecast.

Overall, the performance weighting does little to nothing in closing the performance gap between the LLM crowd and the market.

Appendix D: Iron Man correlation tables

	p_market	openai_35_zs	openai_35_cot	openai_4o_zs	openai_4o_cot	google_12b_zs	google_12b_cot	google_27b_zs	google_27b_cot	deepseek_chat_zs	deepseek_chat_cot	deepseek_r1_zs	deepseek_r1_cot
p_market	1.000	0.571	0.556	0.621	0.611	0.712	0.637	0.658	0.661	0.730	0.712	0.666	0.707
openai_35_zs	0.571	1.000	0.920	0.836	0.841	0.727	0.702	0.760	0.762	0.759	0.798	0.787	0.776
openai_35_cot	0.556	0.920	1.000	0.854	0.862	0.742	0.690	0.748	0.760	0.701	0.767	0.754	0.769
openai_4o_zs	0.621	0.836	0.854	1.000	0.985	0.803	0.739	0.854	0.852	0.705	0.794	0.777	0.838
openai_4o_cot	0.611	0.841	0.862	0.985	1.000	0.812	0.757	0.857	0.857	0.714	0.796	0.779	0.829
google_12b_zs	0.712	0.727	0.742	0.803	0.812	1.000	0.938	0.851	0.858	0.797	0.879	0.780	0.803
google_12b_cot	0.637	0.702	0.690	0.739	0.757	0.938	1.000	0.799	0.804	0.816	0.884	0.761	0.748
google_27b_zs	0.658	0.760	0.748	0.854	0.857	0.851	0.799	1.000	0.988	0.807	0.822	0.827	0.842
google_27b_cot	0.661	0.762	0.760	0.852	0.857	0.858	0.804	0.988	1.000	0.816	0.833	0.829	0.848
deepseek_chat_zs	0.730	0.759	0.701	0.705	0.714	0.797	0.816	0.807	0.816	1.000	0.916	0.794	0.795
deepseek_chat_cot	0.712	0.798	0.767	0.794	0.796	0.879	0.884	0.822	0.833	0.916	1.000	0.840	0.882
deepseek_r1_zs	0.666	0.787	0.754	0.777	0.779	0.780	0.761	0.827	0.829	0.794	0.840	1.000	0.908
deepseek_r1_cot	0.707	0.776	0.769	0.838	0.829	0.803	0.748	0.842	0.848	0.795	0.882	0.908	1.000

	p_market	openai_35_zs	openai_35_cot	openai_4o_zs	openai_4o_cot	google_12b_zs	google_12b_cot	google_27b_zs	google_27b_cot	deepseek_chat_zs	deepseek_chat_cot	deepseek_r1_zs	deepseek_r1_cot
p_market	1.000	0.591	0.588	0.330	0.413	0.498	0.514	0.523	0.675	0.699	0.856	0.686	0.774
openai_35_zs	0.591	1.000	0.936	0.617	0.650	0.656	0.591	0.649	0.827	0.708	0.718	0.781	0.752
openai_35_cot	0.588	0.936	1.000	0.634	0.649	0.670	0.636	0.654	0.816	0.701	0.709	0.806	0.764
openai_4o_zs	0.330	0.617	0.634	1.000	0.951	0.704	0.682	0.725	0.621	0.485	0.438	0.506	0.486
openai_4o_cot	0.413	0.650	0.649	0.951	1.000	0.731	0.740	0.793	0.661	0.526	0.514	0.568	0.573
google_12b_zs	0.498	0.656	0.670	0.704	0.731	1.000	0.942	0.739	0.663	0.527	0.581	0.710	0.640
google_12b_cot	0.514	0.591	0.636	0.682	0.740	0.942	1.000	0.731	0.638	0.563	0.629	0.739	0.659
google_27b_zs	0.523	0.649	0.654	0.725	0.793	0.739	0.731	1.000	0.850	0.461	0.573	0.646	0.702
google_27b_cot	0.675	0.827	0.816	0.621	0.661	0.663	0.638	0.850	1.000	0.702	0.750	0.823	0.853
deepseek_chat_zs	0.699	0.708	0.701	0.485	0.526	0.527	0.563	0.461	0.702	1.000	0.885	0.748	0.759
deepseek_chat_cot	0.856	0.718	0.709	0.438	0.514	0.581	0.629	0.573	0.750	0.885	1.000	0.798	0.831
deepseek_r1_zs	0.686	0.781	0.806	0.506	0.568	0.710	0.739	0.646	0.823	0.748	0.798	1.000	0.904
deepseek_r1_cot	0.774	0.752	0.764	0.486	0.573	0.640	0.659	0.702	0.853	0.759	0.831	0.904	1.000

Appendix E: Regression coefficient tables

Iron Man logistic regression coefficients (first stage)

Term	Estimate	Std. Error	z value	p-value
(Intercept)	-41.7521	304434.6	-1e-04	0.9999
p_market	138.9134	460519.8	3e-04	0.9998
openai_35_zs	-35.2191	1560572.8	0e+00	1.0000
openai_35_cot	-47.5379	2013788.4	0e+00	1.0000
openai_4o_zs	423.4811	1704477.3	2e-04	0.9998
openai_4o_cot	-383.3635	1726351.7	-2e-04	0.9998
google_12b_zs	-13.6357	1975643.7	0e+00	1.0000
google_12b_cot	94.4300	1234102.3	1e-04	0.9999
google_27b_zs	-228.5825	3654547.9	-1e-04	1.0000
google_27b_cot	291.2590	3830919.2	1e-04	0.9999
deepseek_chat_zs	-123.1271	907744.5	-1e-04	0.9999
deepseek_chat_cot	-29.3141	2303883.6	0e+00	1.0000
deepseek_r1_zs	-14.2692	1275875.9	0e+00	1.0000
deepseek_r1_cot	27.4713	1701877.9	0e+00	1.0000

Iron Man logistic regression coefficients (last stage)

Term	Estimate	Std. Error	z value	p-value
(Intercept)	-696.7920	202766.2	-0.0034	0.9973
p_market	414.8776	227843.2	0.0018	0.9985
openai_35_zs	1677.4393	639093.5	0.0026	0.9979
openai_35_cot	158.4666	223980.1	0.0007	0.9994
openai_4o_zs	1628.2970	804954.6	0.0020	0.9984
openai_4o_cot	-3248.0404	1151373.3	-0.0028	0.9977
google_12b_zs	-1487.7186	590321.6	-0.0025	0.9980
google_12b_cot	1987.7524	722225.8	0.0028	0.9978
google_27b_zs	1155.7275	619640.6	0.0019	0.9985
google_27b_cot	-1196.4057	806931.0	-0.0015	0.9988
deepseek_chat_zs	1024.2113	758029.7	0.0014	0.9989
deepseek_chat_cot	-1357.6356	962974.9	-0.0014	0.9989
deepseek_r1_zs	-311.9199	195128.7	-0.0016	0.9987
deepseek_r1_cot	695.0706	254288.5	0.0027	0.9978