Faculty of Economic Sciences

HSE-NES joint
programme in Economics

Moscow, 2025

News-Based Probabilistic Forecasting with Large Language Models

Andrey Petukhov

Advisor: Marat Salikhov

Reviewer: Ivan Stelmakh

Introduction to Forecasting

- High-stakes choices in politics, finance, and security require probability estimates
- Accurate forecasts allow for better planning, resource allocation and risk management
- Can a pipeline supplemented with real-time news, multiple models, and calibration make good predictions?

Approaches to Forecasting

- Statistical Forecasting
  - Time series models;
  - Needs a lot of high-quality data with stationary patterns;
  - Fast once trained, but applied in narrow domain.

«EURO/USD rate tomorrow?»

- Judgmental Forecasting
  - Experts assign probabilities from diverse data;
  - Works when history is short or the world has just changed;
  - Slow, costly, subject to cognitive biases;
  - On platforms like Polymarket*, Metaculus, etc.

«Will Trump win the election in 2024?»

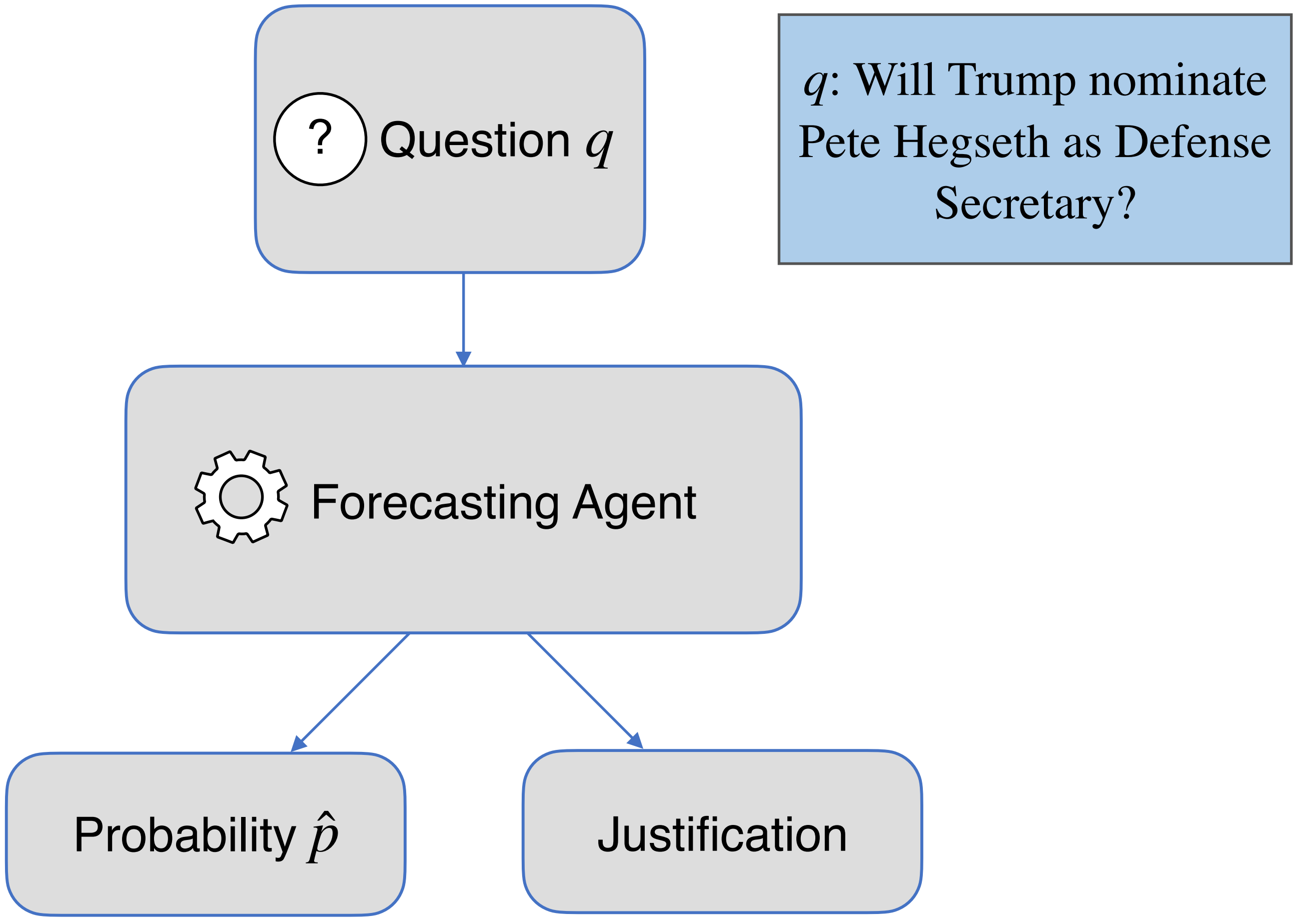
**Polymarket - world's largest prediction market*

Motivation

Create a forecasting agent that is:

- Well-informed
- Fast
- Cheap
- Scalable
- Interpretable

Can Large Language Models (LLMs*) be used for this?



*LLM - language model trained on large amounts of textual data

Related Work

- ForecastQA (Jin et al., 2022), Autocast (Zou et al., 2022) - early forecasting datasets.
- Machine learning systems can be trained to predict the outcomes of events from forecasting competitions (Yan et al., 2024).
- Retrieval improves LLM forecasting accuracy (Yan et al., 2024).
- Large Language Models with no additional data are significantly inferior to the crowd (Schoenegger & Park, 2023, Schoenegger et al., 2023).

My contributions

- I propose a fully automated LLM-based system for forecasting, which shows significant improvements compared to baselines and existing results in the field.
- I investigate the effects of Retrieval Augmented Generation (RAG*), ensembling and calibration on forecasting precision.
- I create a large dataset containing most recent real-world forecasting questions with cleaned and ranked by relevance news texts.

**RAG - a technique of providing LLM with additional relevant data*

Data

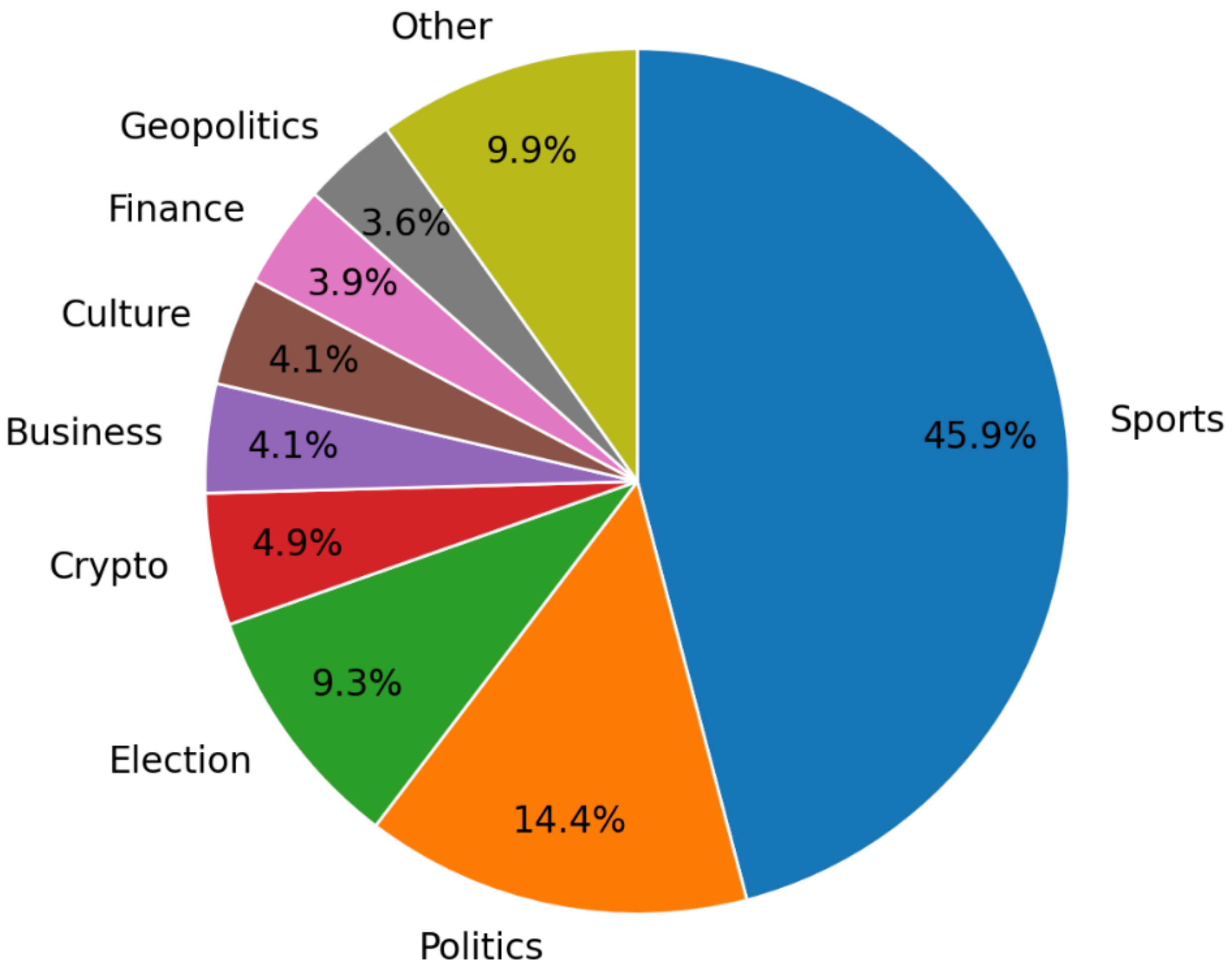
- 5774 binary events from Polymarket
  - Question
  - Description
  - Key dates: start, end, resolution
  - Date range: Jan 2024 - Feb 2025
  - Average duration: 17.05 days
- News corpus: GDELT
  - Updates every ≤ 15 minutes
  - 100k+ news outlets in 65+ languages
  - Has API, returns articles URLs

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

Example of a question from Polymarket

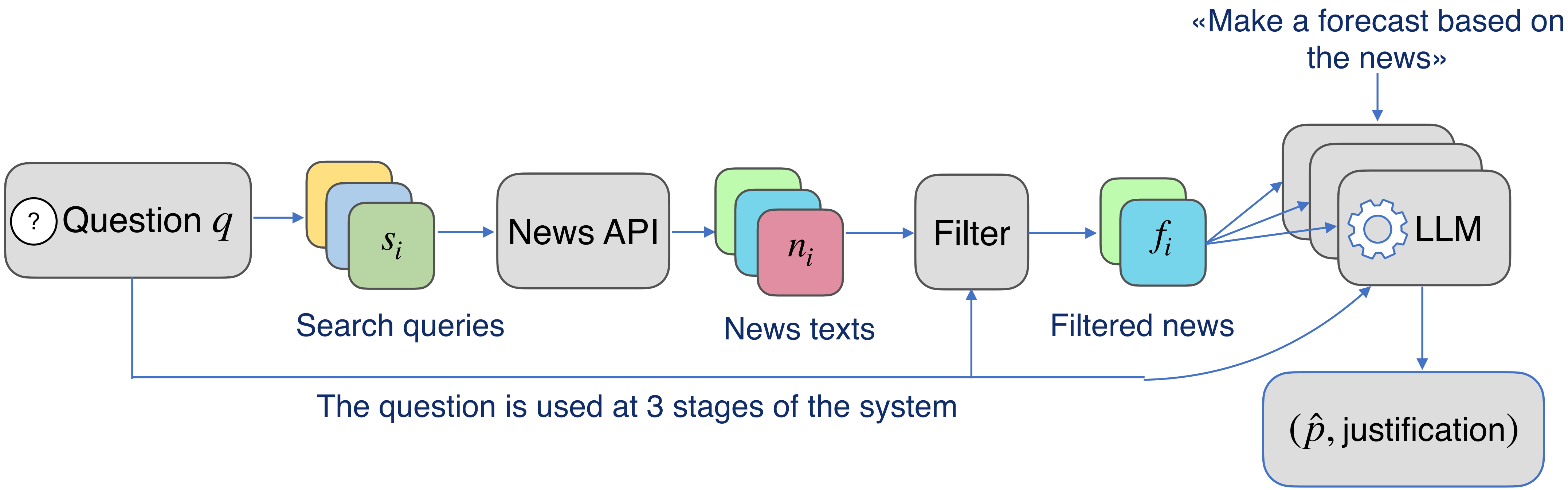
Evaluation subsample

- 1000 events from original dataset
  - Date range: Aug 2024 - Feb 2025
  - Randomly chosen
  - Representative by categories of questions
  - Representative by classes balance (42.32% «Yes» and 42.5% «Yes»)
- Aggregation split
  - Train 30% | Validation 10% | Test 60%

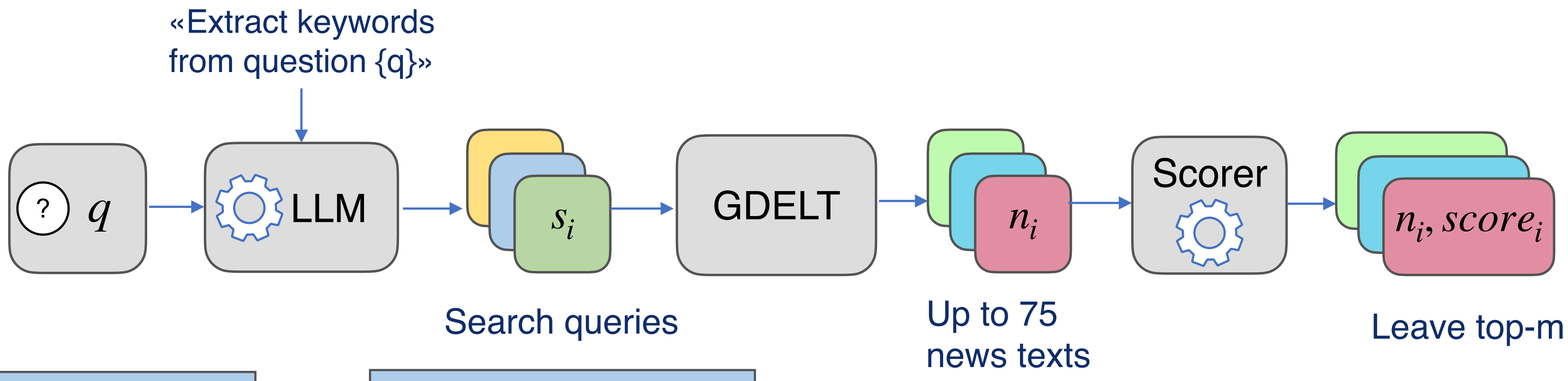


Distribution of question topics

System Architecture. High-level.



System Architecture. Retrieval subsystem



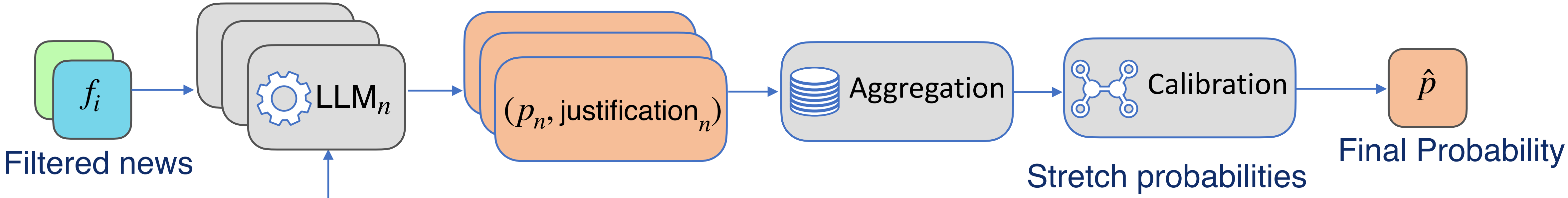
q: Will Trump win the election?

- «Donald Trump»
- «USA election»
- «Republican»
- «Trump nomination»

q: Will Trump win the election?

- («Trump leads in Ohio polls», 5.41)
- («Hurricane hits Florida coast», -4.1)

System Architecture. Reasoning subsystem



Prompt

You are a forecasting assistant. Estimate the probability that the event resolves as "Yes".

EVENT:
Question: {question}
Description and resolution conditions: {description}
Date range: {start_date} to {end_date}

REASONING STEPS: {Instructions}

Methodology

- 8 individual LLMs: DeepSeek-R1, DeepSeek-V3, Mistral-3, Gemini Flash, GPT-4.1-mini, GPT-4o-mini, Claude Haiku, and Llama-4.
- Choose $k \in \{1, 2, 3, 4\}$ for forecasting horizon:

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

- Metrics of interest reported at $k = 3$, ~4.26 days before resolution
- Choose $m \in \{5, 10, 15\}$ for number of news articles
- Three prompting techniques

- Metrics: Brier Score and Accuracy

Brier Score $= \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2$, where

p_i - forecasted probability,

o_i - binary outcome (1 or 0),

The lower - the better

Accuracy $= \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{o}_i = o_i)$, where

$\hat{o}_i = \mathbb{I}(p_i \geq 0.5)$ - forecasted outcome

Aggregation and Calibration

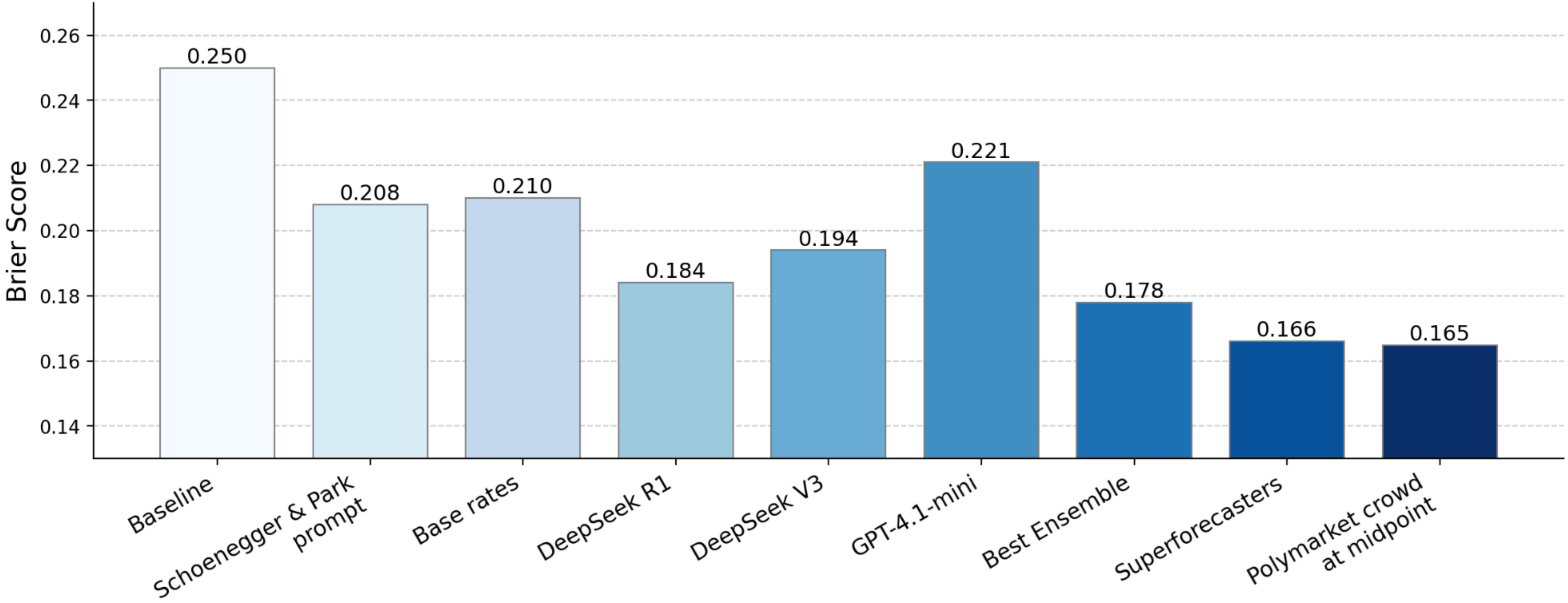
Main Results

- Best approach: trimmed mean and calibrating over DeepSeek R1, DeepSeek V3, Mistral-3 and Gemini-Flash.
- Best ensemble statistically significantly outperforms all individual models, except for DeepSeek R1.
- RAG improves performance
- Trainable aggregation yields no improvements at this scale

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

Forecasting performance of models and aggregations.

Main Results



Conclusion

- RAG, aggregation and calibration significantly improve forecasting accuracy,
- Systematic approach: selecting strong base LLMs, enriching inputs via RAG with current information, aggregating outputs, and calibrating ensemble predictions
- Gap between human crowd and automated system is approximately the same as between the two strongest individual models

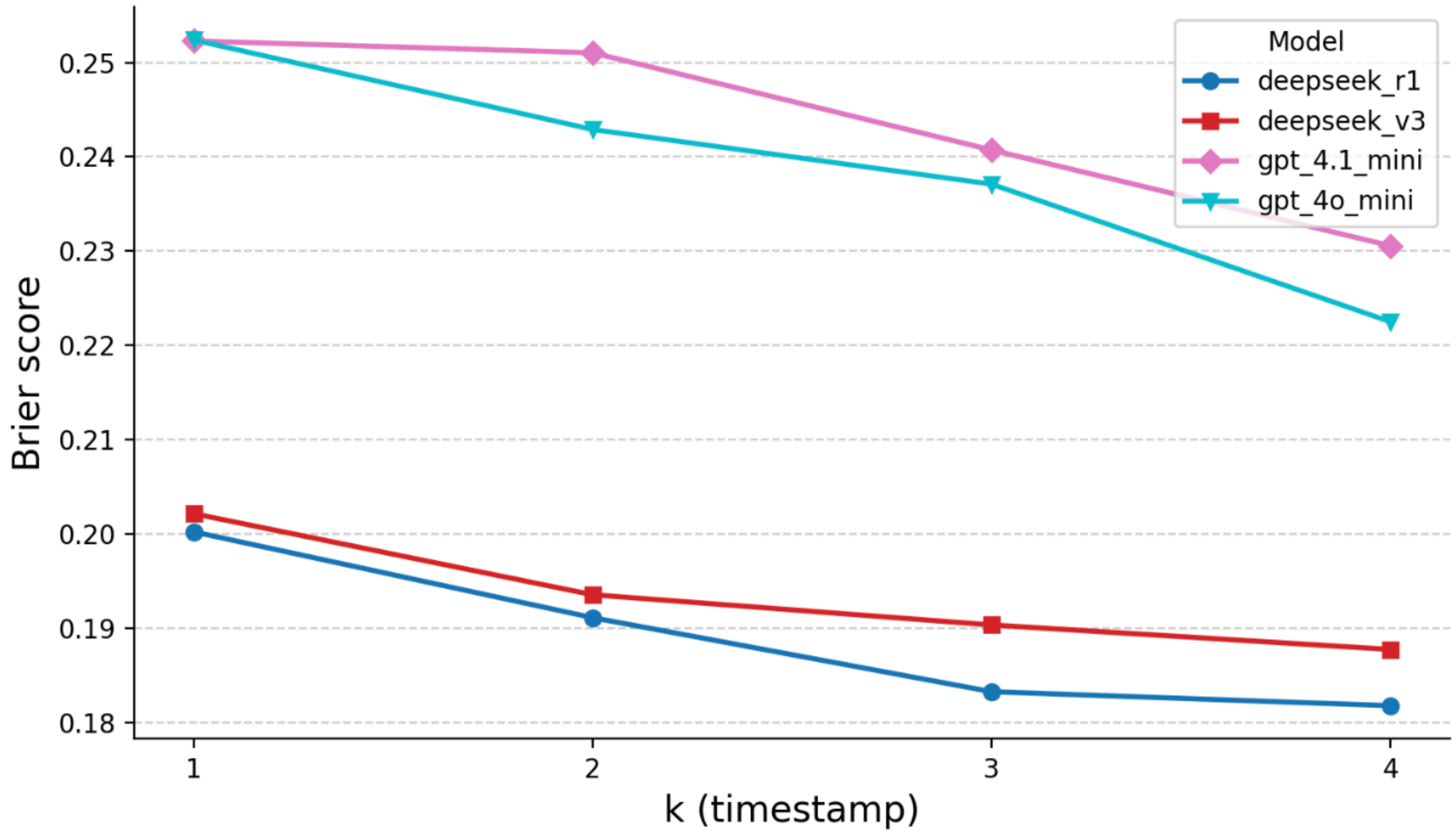
Thank you

Appendix. Statistical significance

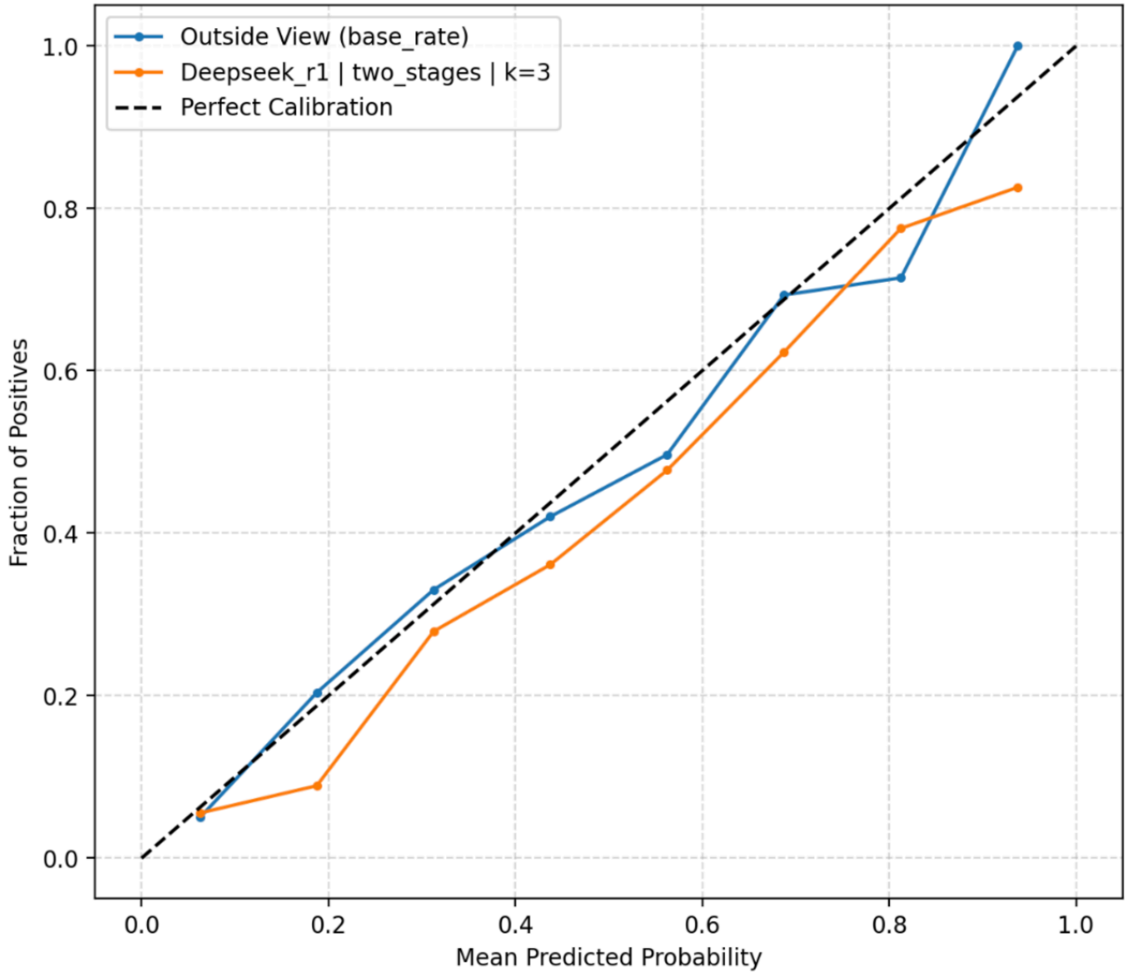
Comparison	95% CI	99% CI
Best vs DeepSeek R1	$(-0.016566, 0.002685)$	$(-0.022948, 0.005525)$
Best vs DeepSeek V3	$(-0.026273, -0.006122)$	$(-0.032750, -0.002980)$
Best vs Gemini Flash	$(-0.038296, -0.014241)$	$(-0.047781, -0.010599)$
Best vs Mistral-3	$(-0.039643, -0.016786)$	$(-0.047298, -0.013113]$

Confidence intervals for difference of Brier Scores

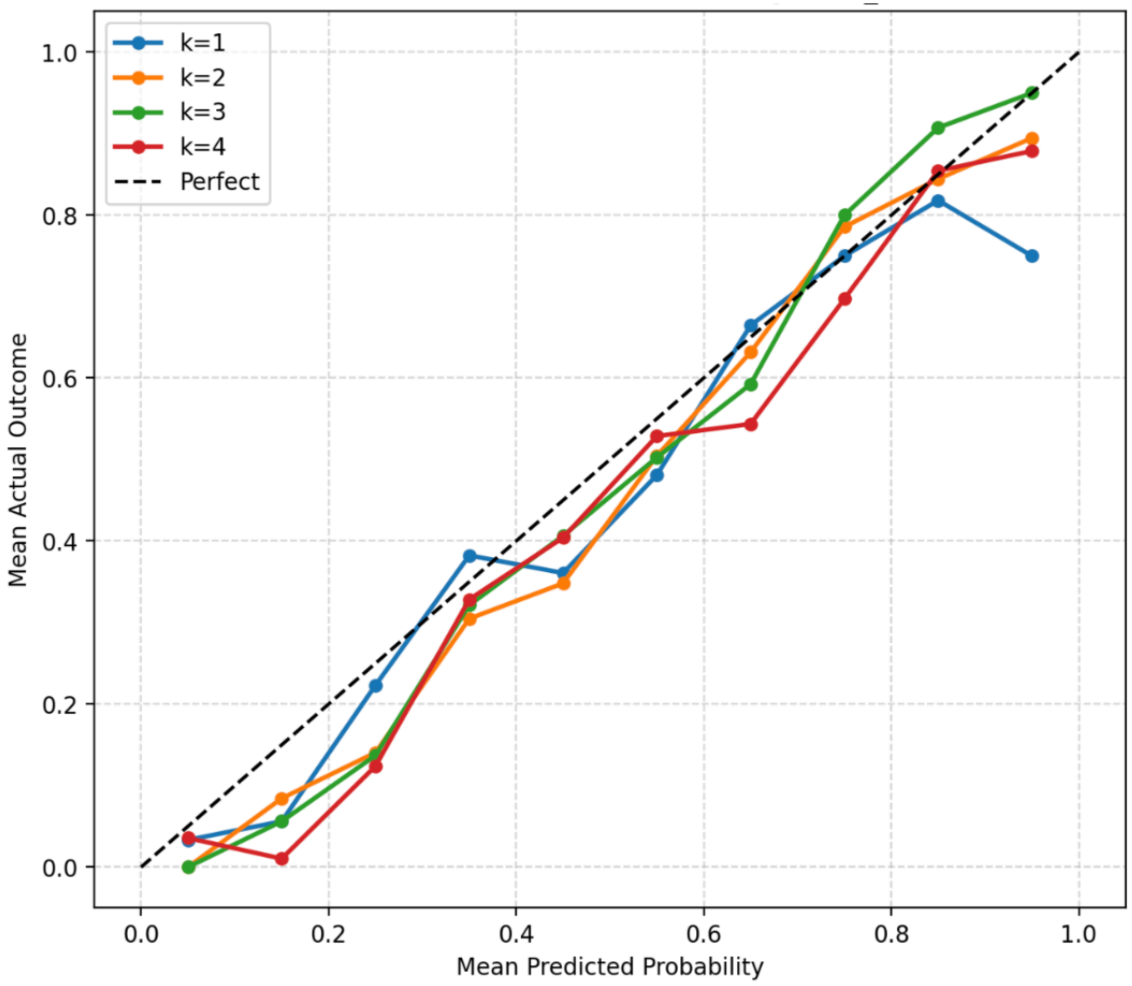
Appendix. Metrics improvement over time



Appendix. Calibration curves



Calibration curves for base rates and enhanced «inside view»



Calibration curves for DeepSeek R1 at different k.