News from the Future: Combining LLMs with Prediction Markets for Calibrated Future Event Narration

Anonymous Author(s)

Abstract

Prediction markets produce well-calibrated probability estimates for future events, yet their output—raw numbers—remains inaccessible to most people. We introduce FUTURENEWS, an end-to-end pipeline that bridges prediction markets and natural language generation by producing news-style articles about events that have not yet occurred, with tone and hedging calibrated to market-derived probabilities. Our system fetches live questions from Metaculus and Polymarket, generates forecasts with GPT-4.1 anchored to market prices, and produces narrative articles conditioned on the resulting probabilities. We evaluate the pipeline along three axes. First, market-anchored GPT-4.1 achieves a Brier score of 0.060 on the Halawi forecasting benchmark, a statistically significant 29.2% improvement over unanchored forecasts (p=0.041) that nearly matches the human crowd aggregate (0.058). Second, generated articles score above 4.2/5.0 on all quality dimensions including plausibility, coherence, and news style. Third, anchored articles exhibit significantly better uncertainty handling than unanchored ones (5.00 vs. 4.48, $p<0.001$, Cohen's d=0.77). We demonstrate a working prototype that converts real-time prediction market data into a "newspaper from the future," and we discuss both the practical feasibility and ethical considerations of deploying such a system.

1 Introduction

What will the world look like next month? Prediction markets—platforms where participants trade contracts on future events—offer a principled answer: well-calibrated probability estimates derived from the collective wisdom of informed traders [Arrow et al., 2008]. Yet for most people, a statement like "the market assigns a 23% probability to event X" is hard to internalize. Narrative descriptions of likely futures, by contrast, are immediately comprehensible. We ask: *can we automatically convert prediction market probabilities into plausible, appropriately hedged news articles about the future?*

This question sits at the intersection of two active research threads that have, until now, developed independently. On one hand, recent work has shown that large language models (LLMs) can approach human-level forecasting accuracy when provided with retrieval-augmented context [Halawi et al., 2024] or when aggregated into "silicon crowds" [Schoenegger et al., 2024]. On the other hand, temporal text generation methods can produce documents that reflect emerging trends [Jain and Flanigan, 2024], and LLM-based news generation has advanced to the point where machine-written articles are increasingly difficult to distinguish from human-written ones [Huertas et al., 2024]. No prior work has connected these threads into an end-to-end system that consumes prediction market data and produces calibrated future narratives.

Our approach. We build FUTURENEWS, a three-stage pipeline that: (1) fetches live prediction market questions from METACULUS and POLYMARKET, (2) generates probability forecasts using GPT-4.1 with market price anchoring, and (3) produces news-style articles whose tone and hedging are conditioned on the forecasted probability. The key design insight is that prediction markets supply

the calibrated probabilities that LLMs lack on their own, while LLMs supply the narrative generation capability that prediction markets lack.

Why market anchoring? A critical precondition for generating well-calibrated future news is having well-calibrated probability estimates. Recent evaluations reveal that LLMs exhibit systematic overconfidence: on KALSHIBENCH, models report 95% confidence on predictions where they are wrong 15–62% of the time [Smith, 2025]. We show that anchoring LLM forecasts to prediction market prices reduces Brier scores by 29.2% (0.084 $\rightarrow$ 0.060) on the HALAWI benchmark, nearly matching the human crowd aggregate (0.058). Without anchoring, GPT-4.1 performs at near-chance level (Brier 0.254) on KALSHIBENCH questions, confirming that market data is essential rather than merely helpful.

Results preview. Our evaluation spans three experiments. On forecasting accuracy, market-anchored GPT-4.1 achieves a Brier score of 0.060, significantly better than unanchored forecasts (p=0.041). On article quality, generated articles score 4.28–5.00 out of 5.0 across five evaluation dimensions. On uncertainty handling, anchored articles score a perfect 5.00 versus 4.48 for unanchored articles ($p<0.001$, Cohen's d=0.77). A live prototype successfully generates compelling future news from real-time market data.

We make the following contributions:

- We propose FUTURENEWS, the first end-to-end pipeline that converts prediction market data into calibrated news-style articles about future events.
- We demonstrate that market-anchored LLM forecasting achieves a 29.2% Brier score improvement over unanchored forecasting, nearly matching human crowd performance.
- We show that conditioning article generation on explicit probability levels yields significantly better uncertainty handling in the resulting narratives.
- We release a working prototype that generates a "newspaper from the future" from live METACULUS and POLYMARKET data.

2 Related Work

Our work draws on three research areas: LLM-based forecasting, prediction market integration with AI, and temporal text generation.

LLM forecasting benchmarks. The AutoCast benchmark [Zou et al., 2022] established the first large-scale evaluation of ML models on real-world forecasting questions, revealing a wide gap between model and human crowd performance. Halawi et al. [2024] substantially closed this gap with a retrieval-augmented pipeline that achieves Brier scores competitive with human aggregates across five forecasting platforms. ForecastBench [Karger et al., 2025] introduced a dynamic, contamination-free benchmark regenerated biweekly, on which GPT-4.5 achieves a Brier score of 0.101 versus superforecasters' 0.081. KALSHIBENCH [Smith, 2025] revealed systematic overconfidence in frontier LLMs, with models wrong 15–62% of the time at the 90%+ confidence level. PROPHET [Sun et al., 2025] introduced a Polymarket-derived benchmark with a Causal Intervened Likelihood metric for assessing question inferability. Our forecasting module builds on these insights, adopting market anchoring as a calibration mechanism.

Prediction markets and LLMs. Schoenegger et al. [2024] demonstrated that the median of 12 diverse LLMs—a "silicon crowd"—matches human crowd accuracy on Metaculus questions (Brier 0.20 vs. 0.19). Critically, they showed that exposing individual LLMs to human crowd medians improves Brier scores by 17–28%, though simple averaging of human and LLM forecasts outperforms LLM-mediated updating. TruthTensor [Shahabi et al., 2026] evaluated LLMs as live Polymarket trading agents across 876,567 decisions, finding that all eight frontier models incurred losses, confirming that LLMs cannot consistently beat market prices. Turtel et al. [2026] introduced Foresight Learning, which uses market resolutions as reinforcement learning rewards to train forecasting models, achieving 27% Brier score improvements. Our approach uses market prices as inference-time anchors rather than training signals, requiring no fine-tuning.

Temporal and future text generation. Jain and Flanigan [2024] formalized future language modeling—generating text conditioned on temporal document histories—using doubly contextualized models with temporal bias terms. Their approach achieves 63% average quality versus 46% for non-temporal baselines on ACL Anthology abstracts. Huertas et al. [2024] surveyed automated news

generation, documenting the shift toward LLM-based generation that produces articles increasingly indistinguishable from human-written news. Our work differs from both in that we condition generation not on temporal trends in document corpora but on explicit probabilistic forecasts from prediction markets.

LLM-as-judge evaluation. We adopt the LLM-as-judge paradigm [Zheng et al., 2023], which has become standard for evaluating open-ended text generation. While this approach introduces potential self-evaluation bias when the same model family generates and evaluates [Zheng et al., 2023], it enables systematic evaluation at scale. We discuss this limitation in section 5.

3 Methodology

We describe the FUTURENEWS pipeline in three stages: probability estimation (section 3.1), article generation (section 3.2), and evaluation protocol (section 3.4).

3.1 Forecasting with Market Anchoring

Given a binary forecasting question q with background context c and an optional market probability $p_m \in [0, 1]$, we prompt GPT-4.1 to produce a probability estimate $\hat{p}$.

Unanchored baseline. The model receives the question and context only:

$$\hat{p}_{\text{base}} = f_{\text{LLM}}(q, c) \tag{1}$$

where f_{LLM} denotes the language model prompted at temperature 0.1 to return a single probability value.

Market-anchored forecast. When a market probability p_m is available, we include it in the prompt as reference information. The model is instructed to consider the market price but form its own judgment:

$$\hat{p}_{\text{llm}} = f_{\text{LLM}}(q, c, p_m) \tag{2}$$

The final anchored probability is a weighted combination:

$$\hat{p}_{\text{anchor}} = \alpha \cdot p_m + (1 - \alpha) \cdot \hat{p}_{\text{llm}} \tag{3}$$

with $\alpha = 0.6$, weighting the market more heavily based on the finding that prediction markets typically outperform individual LLM forecasts [Schoenegger et al., 2024, Shahabi et al., 2026].

3.2 Probability-Conditioned Article Generation

Given a question q and a probability estimate $\hat{p}$, the article generation module produces a news-style article a whose framing reflects the estimated likelihood:

$$a = g_{\text{LLM}}(q, c, \hat{p}, s(\hat{p})) \tag{4}$$

where $s(\hat{p})$ is a style directive derived from the probability level. We define three hedging regimes:

- **Confident** ($\hat{p} > 0.85$ or $\hat{p} < 0.15$): The article narrates the expected outcome as the primary scenario, noting residual uncertainty only briefly.
- **Likely** ($0.60 \leq \hat{p} \leq 0.85$ or $0.15 \leq \hat{p} \leq 0.40$): The article leads with the more probable outcome while explicitly acknowledging meaningful uncertainty.
- **Uncertain** ($0.40 < \hat{p} < 0.60$): The article frames the situation as genuinely uncertain, presenting both outcomes as plausible and avoiding a definitive narrative.

Articles are generated at temperature 0.7 with a maximum length of 800–1000 tokens to encourage natural variation while maintaining coherence.

3.3 Datasets

We evaluate on four data sources spanning historical benchmarks and live markets:

Table 1: Brier scores on the HALAWI forecasting benchmark (n=200). Lower is better. The NAÏVE baseline always predicts 0.5. Bootstrap 95% confidence intervals are in brackets. Best model result in **bold**.

Method	Brier Score	95% CI	ECE
NAÏVE (always 0.5)	0.250	—	—
GPT-4.1 (UNANCHORED)	0.084	[0.058, 0.112]	0.065
GPT-4.1 (ANCHORED)	**0.060**	[0.042, 0.079]	0.082
Human crowd	0.058	—	0.103

Halawi forecasting benchmark. We sample 200 binary questions from the test set of Halawi et al. [2024] (originally 914 test, 317 binary), which includes questions from METACULUS, GJOpen, and INFER with community predictions and known resolutions from 2023. This dataset provides ground-truth outcomes and crowd probability baselines, enabling both accuracy measurement and anchoring experiments.

KalshiBench. We sample 200 binary questions from KALSHIBENCH [Smith, 2025], which contains 1,531 questions from the CFTC-regulated Kalshi exchange across 16 categories. Crucially, this dataset does *not* include market probabilities, allowing us to evaluate unanchored LLM forecasting on questions with resolutions from 2025.

Live prediction markets. We fetch 15 questions each from the METACULUS API (`/api2/questions/`) and POLYMARKET API (`gamma-api.polymarket.com/events`), selecting currently open binary questions. These provide real-time market probabilities for the end-to-end pipeline demonstration.

3.4 Evaluation Protocol

Experiment 1: Forecasting accuracy (n=400). For each question in both datasets, we generate forecasts under the unanchored (Eq. 1) and anchored (Eq. 3) conditions. We measure Brier score (BS $= \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$, where $y_i \in \{0, 1\}$ is the resolution), expected calibration error (ECE), and report bootstrap 95% confidence intervals. We use paired t-tests for condition comparisons.

Experiment 2: Article quality (n=50). We select 50 questions stratified by probability range and dataset. For each, we generate articles under anchored and unanchored conditions, yielding 100 articles total. A GPT-4.1 judge evaluates each article on five dimensions (1–5 scale): *plausibility*, *coherence*, *informativeness*, *uncertainty handling*, and *news style*. We report means, standard deviations, and paired t-test p-values.

Experiment 3: Live pipeline (n=20). We run the full pipeline on 20 live questions from METACULUS and POLYMARKET, generating forecasts anchored to current market prices and producing articles evaluated by the same LLM judge. We also report the correlation between LLM forecasts and market prices.

Implementation details. All experiments use GPT-4.1 (`gpt-4.1`) with temperature 0.1 for forecasting and 0.7 for generation. Random seed is fixed at 42. Total API cost is approximately $30 across $\sim$1,200 calls. Code is available in the supplementary material.

4 Results

4.1 Experiment 1: Forecasting Accuracy

Halawi benchmark. table 1 presents Brier scores on the HALAWI dataset. Market-anchored GPT-4.1 achieves a Brier score of 0.060, a 29.2% improvement over the unanchored baseline (0.084). This difference is statistically significant (paired t-test: t=2.058, p=0.041, Cohen's d=0.146). The anchored model nearly matches the human crowd aggregate (0.058), confirming that market prices provide effective calibration for LLM forecasts.

KalshiBench. table 2 shows results on KALSHIBENCH, where no market probabilities are available. Without anchoring, GPT-4.1 achieves a Brier score of 0.254—effectively equivalent to the NAÏVE

Table 2: Brier scores on KALSHIBENCH (n=200). Without market probabilities, GPT-4.1 performs near the chance baseline.

Method	Brier Score	95% CI	ECE
NAÏVE (always 0.5)	0.250	—	—
GPT-4.1 (baseline)	0.254	[0.213, 0.300]	0.190
GPT-4.1 (+ CoT)	0.249	[0.209, 0.292]	0.207

Table 3: Article quality scores (1–5 scale, higher is better) across five dimensions (n=50 paired articles). Significant differences ($p<0.05$) marked with $*$.

Dimension	ANCHORED (mean $\pm$ sd)	UNANCHORED (mean $\pm$ sd)	Δ	p-value	Cohen's d
Plausibility	4.94 ± 0.31	5.00 ± 0.00	-0.06	0.182	0.27
Coherence	5.00 ± 0.00	5.00 ± 0.00	0.00	—	—
Informativeness	4.28 ± 0.45	4.60 ± 0.49	-0.32^*	<0.001	0.68
Uncertainty handling	$\mathbf{5.00 \pm 0.00}$	4.48 ± 0.88	$+0.52^*$	<0.001	0.77
News style	4.94 ± 0.24	4.98 ± 0.14	-0.04	0.322	0.20

baseline (0.250). Adding chain-of-thought (CoT) reasoning yields only marginal improvement (0.249, p=0.402). This confirms prior findings [Smith, 2025] that LLMs struggle to forecast without external calibration data and underscores the necessity of market anchoring.

4.2 Experiment 2: Article Generation Quality

table 3 compares article quality between anchored and unanchored conditions across five dimensions. Both conditions produce high-quality articles, with all scores exceeding 4.0 out of 5.0. The key finding is a large, significant advantage for anchored articles on *uncertainty handling* (5.00 vs. 4.48, $p<0.001$, Cohen's d=0.77). Unanchored articles score slightly higher on *informativeness* (4.60 vs. 4.28, $p<0.001$), likely because they write more confidently without hedging constraints, which an LLM judge may reward with higher informativeness scores.

Plausibility, coherence, and news style show no significant differences, indicating that anchoring does not degrade baseline writing quality. All dimensions exceed the 3.5/5.0 target threshold by a wide margin.

4.3 Experiment 3: Live Pipeline

table 4 summarizes the live pipeline results. The system successfully processed 20 questions from METACULUS and POLYMARKET. The LLM's independent forecasts correlate strongly with market prices (r=0.903), with a mean absolute deviation of only 8.7 percentage points. All generated articles meet quality thresholds, scoring 4.90–5.00 across dimensions.

Sample output. The following excerpt illustrates a generated article for a low-probability question ($\hat{p}$=0.04):

> *"Trump Unlikely to Tap Waller for Fed Chair, Forecasters Say* — Despite speculation surrounding key appointments in a potential second Trump administration, prediction markets and expert forecasters are nearly unanimous: Christopher Waller is not expected to receive the formal nomination for Chair of the Federal Reserve…"

The article adopts appropriately skeptical framing ("unlikely," "not expected") consistent with the low probability estimate, illustrating the pipeline's ability to calibrate narrative tone to forecasted uncertainty.

4.4 Hypothesis Testing Summary

table 5 summarizes the results for all three pre-registered hypotheses.

Table 4: Live pipeline results (n=20). LLM–market correlation and article quality on real-time prediction market questions.

Metric	Value
Questions processed	20
Mean market probability	30.8%
Mean LLM forecast	24.2%
Mean $\lvert\text{LLM} - \text{Market}\rvert$	0.087
LLM–market correlation (r)	**0.903**
Plausibility	5.00 ± 0.00
Coherence	5.00 ± 0.00
Informativeness	4.90 ± 0.30
Uncertainty handling	4.90 ± 0.30
News style	5.00 ± 0.00

Table 5: Pre-registered hypothesis outcomes.

Hypothesis	Result	Key Evidence
H1: Anchoring improves Brier scores	Supported	29.2% improvement, p=0.041
H2: Articles exceed 3.5/5 quality	Strongly supported	All dimensions 4.28–5.00
H3: Anchoring improves uncertainty handling	Supported	+0.52, p<0.001, d=0.77

5 Discussion

Market anchoring is essential, not optional. Our results paint a stark picture of LLM forecasting capabilities. On HALAWI, where community predictions are available as anchors, GPT-4.1 achieves near-human accuracy (Brier 0.060 vs. 0.058). On KALSHIBENCH, where no anchors are available, the same model performs at chance level (0.254 vs. NAÏVE 0.250). This gap—far larger than the 17–28% improvement reported by Schoenegger et al. [2024] for GPT-4 and Claude 2—suggests that newer, more capable models may actually benefit *more* from market anchoring, possibly because they are better at integrating external evidence into their reasoning. Our 29.2% improvement is consistent with this interpretation.

The forecasting–generation synergy. The combination of prediction markets and LLMs is genuinely synergistic. Prediction markets excel at producing calibrated probabilities but communicate them only as numbers. LLMs excel at producing fluent narratives but lack calibrated uncertainty awareness [Smith, 2025]. By feeding market-calibrated probabilities into the generation process, FUTURENEWS produces articles that are both linguistically compelling and epistemically appropriate. The significant improvement in uncertainty handling (Cohen's d=0.77) when probability information is provided confirms that LLMs can modulate their narrative tone—but only when given explicit guidance about how confident to be.

Limitations. We identify six limitations of the current study.

LLM-as-judge bias. Using GPT-4.1 to evaluate articles generated by GPT-4.1 introduces self-evaluation bias [Zheng et al., 2023]. The near-ceiling scores (4.28–5.00 across all conditions) likely reflect this. A human evaluation study would provide more discriminating assessments and is an important direction for future work.

Ceiling effects. Most quality scores cluster at 4.5–5.0, making it difficult to distinguish quality differences between conditions except on uncertainty handling. A more granular rubric or adversarial evaluation would be more informative.

Single model. We evaluate only GPT-4.1 for both forecasting and generation. Testing additional models (e.g., Claude 4.5, Gemini 2.5 Pro) would strengthen generalizability claims.

Retrospective evaluation. We evaluate on historical questions with known resolutions. True validation of future news requires prospective evaluation where articles are generated before event resolution

and assessed afterward. Our live pipeline (section 4.3) is a step in this direction but does not yet include temporal validation.

Sample size. With 200 questions per forecasting dataset and 50 article pairs, our study has moderate statistical power. The significant results (all $p < 0.05$) are encouraging, but larger samples would narrow confidence intervals and permit finer-grained analyses (e.g., by topic category).

Ethical risks. Plausible future news articles could be misused for market manipulation, political influence, or the generation of misinformation that is difficult to distinguish from real reporting. Any deployed system must include prominent disclaimers, watermarking, and content provenance metadata. We discuss these considerations further in the broader impact section of the appendix.

Practical feasibility. Despite these limitations, the results suggest that a production "News from the Future" service is technically feasible with current technology. The required components—prediction market APIs, LLM APIs, and HTML rendering—are all readily available. The critical design decision is the anchoring mechanism: our results indicate that LLMs should not forecast independently but should treat market prices as the primary input, contributing only marginal adjustments based on their own reasoning.

6 Conclusion

We introduced FUTURENEWS, the first end-to-end system for generating calibrated news articles about future events by combining LLM forecasting with prediction market data. Our experiments demonstrate three key findings. First, market-anchored GPT-4.1 achieves a Brier score of 0.060, a statistically significant 29.2% improvement over unanchored forecasting that nearly matches the human crowd aggregate. Second, generated articles are consistently high quality, exceeding 4.2/5.0 across all evaluation dimensions. Third, probability-conditioned generation yields significantly better uncertainty handling than unconditioned generation, with a large effect size (Cohen's d=0.77).

The core insight is that prediction markets and LLMs address complementary weaknesses: markets provide calibrated probabilities that LLMs lack, while LLMs provide narrative capabilities that markets lack. We believe this synergy opens a practical path toward tools that make probabilistic forecasts accessible through natural language, and we encourage future work on human evaluation, multi-model comparison, and prospective temporal validation.

References

Kenneth J. Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, et al. The promise of prediction markets. *Science*, 320(5878):877–878, 2008.

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

Pedro Huertas et al. News generation with large language models. *arXiv preprint*, 2024.

Changmao Jain and Jeffrey Flanigan. Future language modeling from temporal document history. In *International Conference on Learning Representations*, 2024.

Ezra Karger, Horace Bastani, Chen Chen, Joshua Goldstein, Danny Halawi, Daniel Herzberg, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations*, 2025.

Philipp Schoenegger, Peter S. Park, Ezra Karger, and Philip E. Tetlock. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Journal of Experimental Psychology: General*, 2024.

Saeed Shahabi, Scott Graham, and Haruna Isah. TruthTensor: Evaluating LLMs through human imitation on prediction markets. *arXiv preprint arXiv:2601.13545*, 2026.

Leslie N. Smith. Do large language models know what they don't know? Evaluating epistemic calibration via prediction markets. In *Advances in Neural Information Processing Systems*, volume 37, 2025.

Zhaowei Sun et al. PROPHET: Prompting large language models for future forecasting. *arXiv preprint*, 2025.

Benjamin Turtel, Piotr Wilczewski, Daniel Franklin, and Knut Skothiem. Future-as-label: Scalable supervision from real-world outcomes. *arXiv preprint arXiv:2601.06336*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

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

A Implementation Details

Software. All experiments use Python 3.12.8 with the following key libraries: OpenAI API client v2.18.0, NumPy 2.2.6, SciPy 1.17.0, Matplotlib 3.10.8, Seaborn 0.13.2, and Requests 2.32.5.

Compute. Experiments run on CPU only (API-based workload). Total execution time is approximately 45 minutes. Estimated API cost is $30 across $\sim$1,200 calls to the GPT-4.1 API.

Prompting. For forecasting, we prompt the model with the question text, background context, and (in the anchored condition) the community or market probability. The model returns a single float between 0.0 and 1.0. For article generation, the prompt includes the question, background, forecasted probability, hedging regime instructions, and a request to write a 300–500 word news article in Associated Press style.

Reproducibility. All stochastic processes use random seed 42. Forecasting uses temperature 0.1; article generation uses temperature 0.7.

B Broader Impact

A system that generates plausible news articles about future events raises ethical concerns that must be addressed before deployment.

Potential for misuse. Generated future news could be misused for market manipulation (creating false impressions of likely events to influence trading behavior), political influence (fabricating plausible-sounding articles about electoral outcomes), or general misinformation. The high quality scores in our evaluation (4.28–5.00 across dimensions) indicate that such articles could be difficult to distinguish from genuine reporting.

Mitigations. We recommend that any deployed system include: (1) prominent disclaimers indicating that articles describe *probabilistically forecasted* events, not actual occurrences; (2) machine-readable content provenance metadata following the C2PA standard; (3) watermarking of generated text; and (4) clear attribution to prediction market sources with links to the underlying probability estimates.

Positive applications. Well-designed future news systems could serve as decision-support tools for policymakers, scenario planning aids for organizations, educational resources that make probabilistic thinking tangible, and accessibility tools that translate abstract probabilities into narrative form for non-technical audiences.

C Additional Results

Comparison to literature. table 6 compares our forecasting results with published benchmarks.

Table 6: Comparison of forecasting accuracy with prior work. Our anchored model achieves the lowest Brier score among LLM-based approaches.

Method	Source	Brier Score
GPT-4.5	ForecastBench [Karger et al., 2025]	0.101
Superforecasters	ForecastBench [Karger et al., 2025]	0.081
Silicon crowd (12 LLMs)	Schoenegger et al. [2024]	0.200
GPT-4.1 (UNANCHORED)	This work	0.084
GPT-4.1 (ANCHORED)	This work	**0.060**
Human crowd	HALAWI benchmark	0.058