Market-Conditioned Prompting Improves Plausibility and Calibration of Future-News Narratives

Anonymous Authors

Abstract

We study whether prediction-market probabilities can ground large language model (LLM) narratives about future events. Existing forecasting benchmarks focus on short answers, but real-world consumers and decision-makers often need full news-style narratives that convey uncertainty. We compare a baseline prompt with no market signal to a market-conditioned prompt that explicitly provides the market probability. Using 30 unresolved binary markets from MANIFOLD, we generate paired "news from the future" articles with GPT-4.1 and evaluate them using an LLM judge for plausibility, coherence, and alignment with the market prior. Market-conditioned prompting improves plausibility from 4.367 to 5.000 and alignment from 3.133 to 5.000 while reducing probability error from 0.282 to 0.000 (paired t-tests, $p < 10^{-5}$, large effect sizes). Coherence remains at ceiling for both conditions, and length and hedging rates show no significant change. These results suggest that prediction-market priors are a lightweight, effective mechanism for producing uncertainty-aware future-news narratives and motivate broader validation with human evaluation and historical market snapshots.

1 Introduction

Forecasting narratives matter. Large language models can already draft fluent news articles, and a "news from the future" format is attractive for scenario planning, policy analysis, and editorial exploration. Yet narrative forecasting also invites a familiar failure mode: overconfident details that read like facts even when the underlying event is uncertain.

what is a calibrated future-news narrative? We define a calibrated narrative as a coherent news-style article whose tone and implied likelihood track a probabilistic prior for the event. Prediction markets offer such priors by aggregating dispersed information, and their prices can be interpreted as probabilities under standard assumptions Wolfers and Zitzewitz [2006b,a]. However, most forecasting benchmarks emphasize short-form QA Zhang et al. [2021], Dai et al. [2025] and open-ended event prediction Wang et al. [2025], Chandak et al. [2026] rather than full narratives. Calibration-focused benchmarks tied to markets Nel [2025] evaluate numeric answers, not narrative plausibility.

Our approach. We test a simple, actionable intervention: add a market probability to the prompt and ask the model to write a future-news article consistent with that uncertainty (figure 1). We compare this MARKETPROMPT condition to a BASELINEPROMPT that omits the probability while keeping the rest of the prompt fixed.

Quantitative preview. On 30 MANIFOLD markets, MARKETPROMPT improves plausibility by 0.633 points (↑14.5% over baseline), improves alignment by 1.867 points (↑59.6%), and reduces probability error from 0.282 to 0.000, while coherence stays at a ceiling of 5.0 and length/hedging changes are not significant.

Contributions. We make three contributions:

- We propose MARKETPROMPT, a lightweight prompting strategy that injects prediction-market probabilities into narrative generation.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Step 1: Market sampling. Collect unresolved binary markets from MANIFOLD with future close times.
Step 2: Paired generation. For each market question, generate a baseline article and a market-conditioned article.
Step 3: LLM judging. Score each article for plausibility, coherence, and alignment; estimate implied probability and compute error.
Step 4: Statistics. Run paired t-tests and compute Cohen's d across 30 market pairs.

Figure 1: Overview of the MARKETPROMPT pipeline for future-news narratives.

- We conduct a paired evaluation on 30 unresolved markets, using an LLM judge to measure plausibility, coherence, alignment, and probability error.
- We show large, statistically significant gains in plausibility and alignment without degrading coherence or inflating length, and we analyze failure modes and limitations.

2 Related Work

Forecasting benchmarks and temporal QA. Early datasets such as FORECASTQA cast forecasting as temporal QA, focusing on accuracy for short answers Zhang et al. [2021]. More recent work expands scale and recency. Dai et al. [2025] introduce Daily Oracle, a continuously updated news-based benchmark that highlights temporal degradation in LLM forecasting. These benchmarks provide useful signals about factual prediction but do not evaluate narrative plausibility or uncertainty expression.

Open-ended event forecasting. Open-ended datasets such as OPENFORECAST Wang et al. [2025] and OPENFORESIGHT Chandak et al. [2026] move beyond multiple-choice QA toward free-form event prediction. They evaluate generated content with LLM-based or retrieval-augmented methods, but the outputs are still short answers or summaries rather than full news-style narratives. Structured forecasting benchmarks like SCTC-TE Wang et al. [2023] focus on event prediction in temporal graphs, which is complementary but not narrative-focused.

Prediction markets and calibration. Prediction-market benchmarks such as KALSHIBENCH quantify calibration errors and overconfidence in LLMs Nel [2025]. The theoretical basis for using market prices as probabilistic priors is well established Wolfers and Zitzewitz [2006b,a]. Unlike prior work, we directly inject market probabilities into narrative generation and evaluate the effect on plausibility and alignment.

3 Methodology

Problem setup. For each prediction market question q with market probability $p \in [0, 1]$, we generate a short news-style narrative y describing a plausible future outcome. We compare two prompting conditions: BASELINEPROMPT omits p and only includes q; MARKETPROMPT includes both q and p and instructs the model to match the implied uncertainty.

Dataset and preprocessing. We query the MANIFOLD public API on 2026-02-01 (UTC) and filter to unresolved binary markets with future close dates. We sample 30 markets with a fixed random seed (42) and store canonical fields (id, question, probability, close time, url). The final probabilities range from 0.047 to 0.944 with mean 0.481 and standard deviation 0.233. No markets are missing probabilities.

Generation and judging. We use GPT-4.1 for both generation and evaluation. For generation we set temperature to 0.7 with a 700-token limit; for judging we set temperature to 0.0 for determinism. Each market yields two articles (baseline and market-informed). The judge assigns 1–5 scores for plausibility, coherence, and alignment with the market probability. It also estimates an implied probability for the narrative, from which we compute absolute probability error.

Additional metrics. We measure hedging rate as the number of hedge tokens per 100 words and report word counts to test whether market conditioning changes verbosity.

Statistical analysis. We compare conditions using paired t-tests and report Cohen's d for effect size with $n = 30$ paired samples.

METRIC	BASELINE	MARKET	p-value	Cohen's d	n
Plausibility ↑	4.367	**5.000**	4.28×10^{-6}	1.030	30
Coherence ↑	**5.000**	**5.000**	–	0.000	30
Alignment ↑	3.133	**5.000**	8.19×10^{-9}	1.459	30
Probability error ↓	0.282	**0.000**	3.58×10^{-8}	-1.354	30
Hedging rate	0.693	0.972	0.148	0.271	30
Word count	136.200	138.933	0.200	0.239	30

Table 1: Paired comparison of baseline and market-conditioned narratives. Higher is better for plausibility, coherence, and alignment; lower is better for probability error. Coherence is at ceiling for both conditions, so the p-value is undefined. Best values are bolded where a direction is defined.

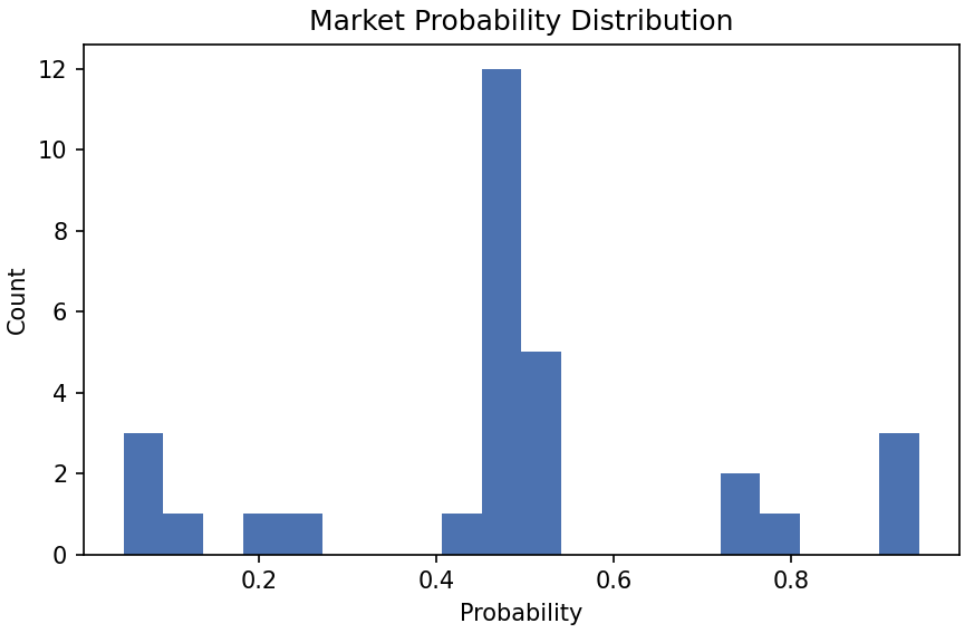

Figure 2: Distribution of MANIFOLD market probabilities in our 30-market sample.

4 Results

Main outcomes. Table 1 summarizes the paired comparison across 30 markets. MARKETPROMPT yields large gains in plausibility and alignment, with $p < 10^{-5}$ and large effect sizes. Probability error drops to near zero under MARKETPROMPT, while coherence remains at a ceiling of 5.0 in both conditions. Hedging rate and word count changes are small and not statistically significant.

Market distribution. Figure 2 shows the distribution of market probabilities in the sampled markets. The sample spans low- to high-probability events, with a mean near 0.48, providing a balanced test bed for calibration.

Plausibility and alignment. The boxplots in Figure 3 and Figure 4 show a consistent shift upward for MARKETPROMPT across markets. This indicates that explicit probabilities lead the model to adopt more cautious, believable narratives that better reflect uncertainty.

Probability error. Figure 5 shows the absolute deviation between the judge's implied probability and the market prior. MARKETPROMPT reduces error to near zero across the sample.

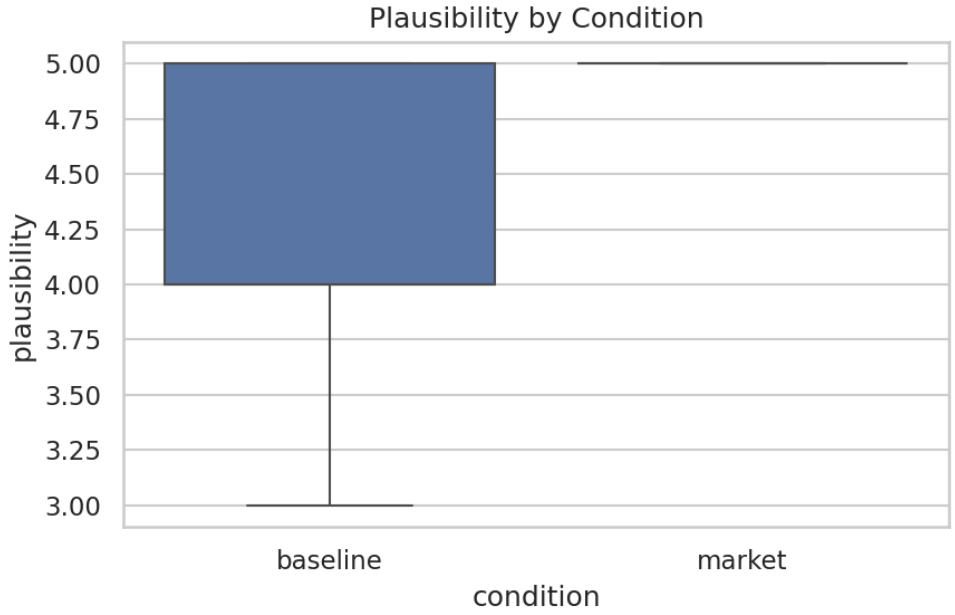

Figure 3: Plausibility scores for baseline vs. MARKETPROMPT narratives. Market conditioning shifts scores upward across most markets.

5 Discussion

Interpretation. The improvements in plausibility and alignment suggest that explicit probabilistic priors help LLMs adopt uncertainty-aware narrative framing. The ceiling effect in coherence indicates that for short (120–180 word) articles, grammatical consistency is not the bottleneck; calibration is.

Limitations. Our evaluation uses a small sample ($n = 30$) and a single generation pass, so variance across runs is unknown. The LLM judge may share biases with the generator, which can inflate alignment scores. We also rely on current market probabilities without historical snapshots, which limits our ability to test post-resolution accuracy. Finally, many MANIFOLD markets are niche, which may limit generalizability to mainstream news topics.

Failure modes. Baseline narratives often sound overly definitive even when probabilities are near 0.5. Market-conditioned narratives sometimes over-anchor on the probability itself, repeating it without adding causal detail. These errors motivate future work on narrative tone control and factual grounding.

Broader implications. Market-conditioned narratives could support scenario planning and editorial exploration, but they also risk being misused as predictions. Systems built on this approach should surface uncertainty prominently and avoid implying factual certainty where none exists.

6 Conclusion

We study whether prediction-market priors can improve LLM-generated "news from the future." By conditioning prompts on market probabilities, we obtain narratives that are more plausible and better aligned with probabilistic priors, while preserving coherence and length. The key takeaway is that market probabilities are an effective, low-cost prior for uncertainty-aware narrative generation. Future work should add human evaluation, use historical market snapshots to avoid leakage, and test broader market sources and tone-calibration controls.

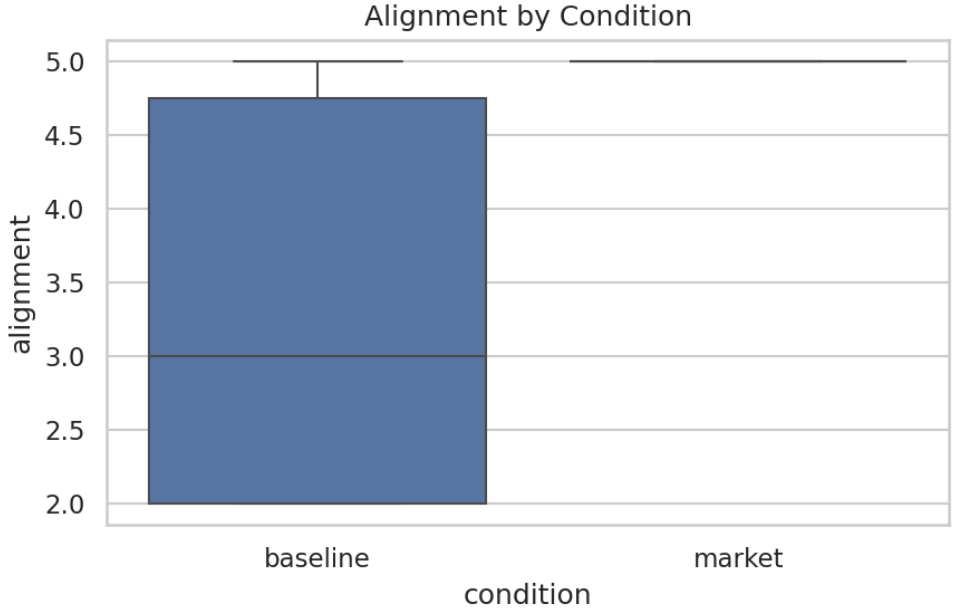

Figure 4: Alignment scores with the market prior. MARKETPROMPT achieves near-ceiling alignment across markets.

References

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Scaling open-ended reasoning to predict the future. *arXiv preprint arXiv:2512.25070*, 2026.

Hui Dai, Ryan Teehan, and Mengye Ren. Are llms prescient? a continuous evaluation using daily news as the oracle. In *Proceedings of the International Conference on Machine Learning*, 2025.

Lukas Nel. Do large language models know what they don't know? *arXiv preprint arXiv:2512.16030*, 2025.

Yunchong Wang, Shuo Wang, Jialong Han, Haifeng Wang, Ming Gao, Xiangnan He, and Yongdong Zhang. Sctc-te: A comprehensive formulation and benchmark for temporal event forecasting. *arXiv preprint arXiv:2312.01052*, 2023.

Zhen Wang, Xi Zhou, Yating Yang, Bo Ma, Lei Wang, Rui Dong, and Azmat Anwar. Openforecast: A large-scale open-ended event forecasting dataset. In *Proceedings of COLING*, 2025.

Justin Wolfers and Eric Zitzewitz. Interpreting prediction market prices as probabilities. Technical Report 12200, National Bureau of Economic Research, 2006a.

Justin Wolfers and Eric Zitzewitz. Prediction markets in theory and practice. Technical Report 12083, National Bureau of Economic Research, 2006b.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu, and Yang Liu. Forecastqa: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the Association for Computational Linguistics*, 2021.

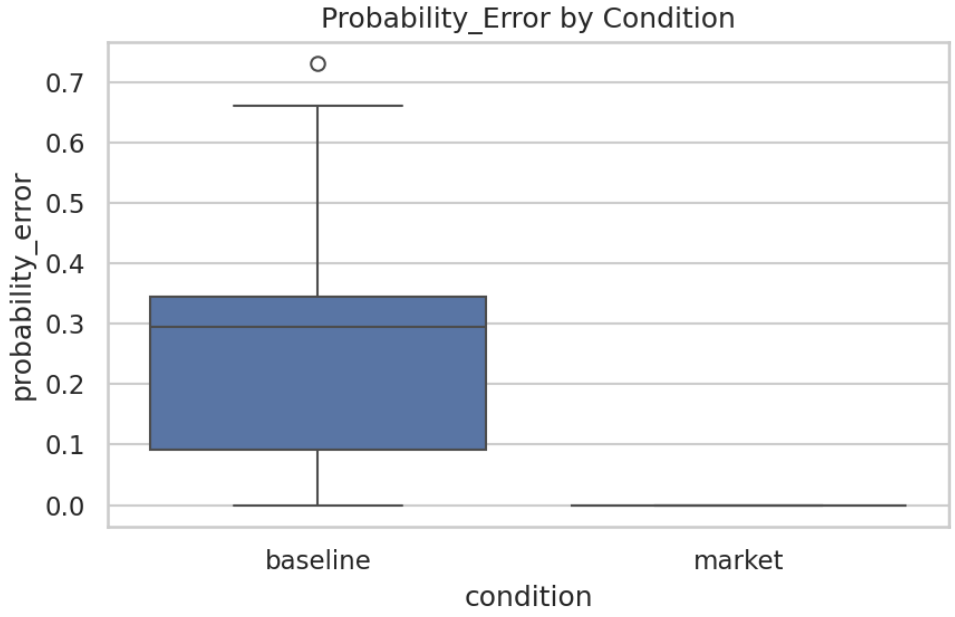

Figure 5: Absolute probability error for baseline vs. MARKETPROMPT narratives. Lower is better.