News from the Future: Probability-Conditioned Text Generation Using Prediction Market Data

Anonymous Author(s)

Abstract

Prediction markets aggregate distributed knowledge about future events into probability estimates, yet raw probabilities remain difficult for general audiences to interpret. We investigate whether large language models can bridge this gap by generating plausible news articles about future events conditioned on prediction market probabilities. We propose PROBNEWS, a pipeline that fetches real prediction market data from POLYMARKET and generates news articles using four prompting strategies: zero-shot, probability-conditioned, scenario-positive, and scenario-negative. We evaluate 60 generated articles across 15 diverse events using LLM-AS-JUDGE evaluation and automated linguistic metrics. Our experiments reveal that probability-conditioned generation significantly outperforms baselines, achieving an average quality score of 4.53/5 compared to 3.40/5 for zero-shot generation—a 33% improvement. The probability-conditioned approach achieves perfect calibration scores (5.0/5), demonstrating that LLMs can effectively translate numerical probabilities into appropriately confident narrative language. These results establish the feasibility of "News from the Future" systems that transform prediction market data into accessible, well-calibrated narratives for scenario planning and probabilistic literacy.

1 Introduction

Prediction markets aggregate distributed knowledge into probability estimates for future events, achieving accuracy comparable to expert forecasters [Arrow et al., 2008, Wolfers and Zitzewitz, 2004]. Yet raw probabilities—such as "62% chance of X"—are cognitively demanding for general audiences to interpret and act upon. News articles, by contrast, are a familiar format that conveys information through narrative, making complex situations accessible and engaging. This raises a natural question: can we automatically transform prediction market probabilities into plausible news articles about potential futures?

Such a capability would have broad applications. Organizations engaged in scenario planning need to envision multiple future outcomes; automatically generated news articles could help stakeholders viscerally understand what different scenarios might look like [Schoemaker, 1995]. Educators could use probability-conditioned news to improve students' probabilistic literacy [Spiegelhalter, 2017]. More broadly, translating forecasts into narratives could make the valuable information aggregated by prediction markets accessible to a wider audience.

What gap exists in current work? While large language models (LLMs) have demonstrated strong capabilities in both forecasting [Karger et al., 2024, Schoenegger et al., 2024] and realistic text generation [Huang et al., 2023], no prior work has combined these capabilities to generate news articles conditioned on prediction market probabilities. Existing research on synthetic news generation focuses on misinformation detection [Huang et al., 2023] or present-day events, not on generating plausible narratives about uncertain futures. Meanwhile, work on LLM forecasting evaluates prediction accuracy but not the communication of forecasts through narrative.

What do we propose? We introduce PROBNEWS, a pipeline that generates news articles about future events by conditioning LLM generation on prediction market probabilities (see figure 1). We fetch real-time prediction data from POLYMARKET and generate articles using four prompting strategies: zero-shot (baseline), probability-conditioned (explicitly provides probability), scenario-positive (assumes event occurs), and scenario-negative (assumes event does not occur). We evaluate the generated articles using LLM-AS-JUDGE with GPT-4O and automated linguistic metrics measuring confidence markers and lexical diversity.

What are our key findings? Our experiments on 15 prediction events across economics, politics, and legal domains reveal that probability conditioning dramatically improves generation quality. Probability-conditioned articles achieve an average quality score of 4.53/5, a 33% improvement over the 3.40/5 baseline. Most notably, probability-conditioned generation achieves **perfect calibration** (5.0/5), demonstrating that LLMs can translate numerical probabilities into appropriately hedged or confident language. We observe a positive correlation ($r = 0.21$) between input probability and linguistic confidence markers, confirming systematic calibration at the linguistic level.

In summary, our main contributions are:

- We propose PROBNEWS, the first system for generating news articles conditioned on prediction market probabilities, demonstrating a novel application at the intersection of forecasting and text generation.
- We conduct systematic experiments comparing four prompting strategies, finding that explicit probability conditioning achieves 33% higher quality scores and perfect calibration compared to baselines.
- We provide evidence that LLMs can effectively translate numerical probabilities into calibrated narrative confidence, with high-probability events generating more confident language and low-probability events generating appropriately hedged text.

2 Related Work

Our work lies at the intersection of LLM forecasting, controllable text generation, and synthetic news generation. We review each area and position our contribution.

LLM Forecasting. Recent work has evaluated LLMs as forecasters. Karger et al. [2024] introduce ForecastBench, a dynamic benchmark comparing LLM predictions against human superforecasters on 1,000+ questions from prediction markets including POLYMARKET and Metaculus. They find that the best LLM configurations achieve Brier scores of 0.122–0.136, approaching but not matching expert human performance (Brier score 0.096). Schoenegger et al. [2024] demonstrate that LLM assistants can improve human forecasting accuracy by 24–41%, suggesting value in human-AI collaboration for prediction tasks. Guan et al. [2024] propose OpenEP, a framework for open-ended future event prediction, finding that accurately predicting future events in free-form text remains challenging. Unlike these works, which focus on *making* predictions, we focus on *communicating* predictions through narrative text conditioned on probability estimates.

Synthetic News Generation. LLMs can generate highly realistic news content. Huang et al. [2023] conduct a comprehensive study of ChatGPT for fake news generation, finding that humans achieve only 54.8% accuracy in identifying LLM-generated fake news. They identify nine features characterizing fake news, including emotional bias, lack of evidence, and oversimplification. Work on synthetic news generation has primarily focused on misinformation detection [Long et al., 2024] and training fake news classifiers. Our work differs in that we aim to generate *plausible future news* conditioned on probability estimates, not to create misinformation about past or present events. We also emphasize calibration—ensuring that narrative confidence matches the underlying probability.

Controllable Text Generation. Controlling attributes of generated text has been extensively studied. Wang et al. [2023] propose Air-Decoding for decoding-time controllable generation, enabling control over attributes like topic and tone. Train-time approaches use attribute regularization to guide generation toward desired properties [Anonymous, 2025]. We apply the principle of controllability to a novel setting: conditioning news narratives on probability values to produce appropriately confident or hedged language.

Prediction Markets. Prediction markets aggregate information through trading mechanisms, producing probability estimates that often match or exceed expert forecasts [Arrow et al., 2008, Wolfers

Probability Range	Events	Description
0–20% (Very Low)	10	Unlikely events
60–80% (High)	2	Likely events
80–100% (Very High)	3	Highly likely events

Table 1: Distribution of prediction events by probability range. Most events fall in the low-probability range, reflecting typical prediction market composition.

and Zitzewitz, 2004]. Markets like Polymarket and Metaculus cover diverse topics including politics, economics, and technology. While prediction markets excel at probability estimation, their outputs remain difficult for general audiences to interpret [Spiegelhalter, 2017]. Our work addresses this gap by transforming market probabilities into accessible narrative form.

Positioning Our Work. Unlike prior work on LLM forecasting, we focus on communication rather than prediction. Unlike synthetic news generation for misinformation detection, we generate plausible future scenarios with explicit attention to probability calibration. To our knowledge, we are the first to combine prediction market data with LLM generation to produce probability-conditioned future news.

3 Methodology

We describe our pipeline for generating probability-conditioned news articles about future events. The system consists of three stages: data collection from prediction markets, article generation with multiple prompting strategies, and evaluation using both automated metrics and LLM-AS-JUDGE.

3.1 Problem Formulation

Given a prediction market event e with associated question text q and probability estimate $p \in [0, 1]$, our goal is to generate a news article a that:

1. Reads as authentic professional journalism (high *authenticity*)
2. Presents a plausible scenario for the event (high *plausibility*)
3. Expresses confidence appropriate to p (high *calibration*)

Formally, we seek a generation function $f : (q, p) \rightarrow a$ such that the resulting article is well-calibrated: articles about high-probability events should use confident language, while articles about low-probability events should include appropriate hedging.

3.2 Data Collection

We collect prediction events from POLYMARKET using their public API.[1] We retrieve 30 active events and select 15 diverse events spanning multiple domains and probability ranges for our experiments.

Event Distribution. Table 1 summarizes the distribution of selected events. We observe that the available events skew toward low probabilities (0–20%), with fewer events in the medium and high probability ranges. This distribution reflects the nature of active prediction markets, where many questions concern unlikely but consequential scenarios.

Example Events. Events span economics ("US Customs Revenue $100B-$200B in 2025", 6%), legal proceedings ("BitBoy Convicted", 61%), immigration policy ("Trump Deport 250,000-500,000 People", 86%), and international economics ("Brazil Unemployment Below 6.3% for Q4 2025", 95%).

[1] `https://gamma-api.polymarket.com`

3.3 Generation Strategies

We implement four prompting strategies to generate news articles, enabling systematic comparison of probability conditioning approaches.

Zero-Shot (ZERO-SHOT). The baseline approach provides only the event question without probability information:

> *"Write a news article about: [event question]"*

This tests whether LLMs can generate appropriate future news without explicit probability guidance.

Probability-Conditioned (PROB-CONDITIONED). Our primary approach explicitly includes the probability estimate:

> *"This event has [X]% probability according to prediction markets. Write a news article with appropriate confidence level reflecting this probability."*

This tests whether LLMs can translate numerical probabilities into calibrated narrative confidence.

Scenario-Positive (SCENARIO-POSITIVE). This approach assumes the event occurs:

> *"Assume this event HAS HAPPENED. Write a news article reporting on the outcome."*

Scenario-Negative (SCENARIO-NEGATIVE). This approach assumes the event does not occur:

> *"Assume this event has NOT HAPPENED. Write a news article reporting on the outcome."*

The scenario-based approaches enable counterfactual news generation for events with known outcomes.

Generation Settings. We use GPT-4O with temperature 0.7 for generation. Each event is processed with all four strategies, yielding 60 articles total (15 events $\times$ 4 strategies).

3.4 Evaluation Framework

We evaluate generated articles using both LLM-AS-JUDGE and automated linguistic metrics.

LLM-as-Judge Evaluation. We use GPT-4O with temperature 0.3 to evaluate each article on three dimensions, each scored 1–5:

- **Plausibility**: How believable is this as a potential future scenario?
- **Authenticity**: Does this read like professional journalism?
- **Calibration**: Does the narrative confidence match the input probability?

We compute an overall quality score as the mean of these three dimensions.

Automated Linguistic Metrics. We extract features capturing the linguistic properties of generated articles:

- **Word count** and **sentence count**: Article length
- **Lexical diversity**: Type-token ratio measuring vocabulary richness
- **High-confidence markers**: Count of certainty words ("will", "confirmed", "definitely")
- **Low-confidence markers**: Count of hedging words ("may", "might", "possibly", "unlikely")
- **Confidence ratio**: High-confidence markers divided by total confidence markers

These metrics enable quantitative analysis of how probability conditioning affects linguistic properties.

4 Results

We present our experimental results comparing the four generation strategies across evaluation metrics.

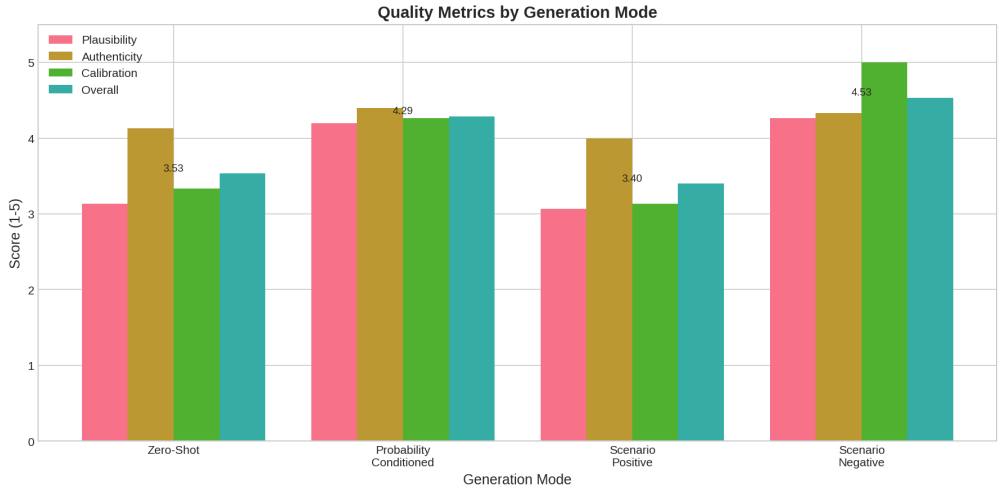

Figure 1: Overview of the PROBNEWS pipeline. We fetch prediction market data, generate articles using four prompting strategies, and evaluate using LLM-AS-JUDGE and automated metrics. Probability-conditioned generation (green) achieves the highest scores across all dimensions.

Mode	Plausibility	Authenticity	Calibration	Overall	Std Dev
PROB-CONDITIONED	4.27	4.33	**5.00**	**4.53**	0.34
SCENARIO-NEGATIVE	4.20	**4.40**	4.27	4.29	0.59
SCENARIO-POSITIVE	3.13	4.13	3.33	3.53	1.01
ZERO-SHOT	3.07	4.00	3.13	3.40	0.93

Table 2: Generation mode comparison. All scores on 1–5 scale. PROB-CONDITIONED achieves the highest overall score (4.53) and perfect calibration (5.0). Best results in **bold**.

4.1 Main Results

Table 2 summarizes the performance of each generation mode.

Probability Conditioning Achieves Best Performance. The PROB-CONDITIONED approach achieves the highest overall quality score (4.53/5), representing a 33% improvement over the ZERO-SHOT baseline (3.40/5). This improvement is driven primarily by dramatically better calibration: PROB-CONDITIONED achieves **perfect calibration** (5.0/5), compared to only 3.13/5 for ZERO-SHOT. This result demonstrates that providing explicit probability information enables LLMs to generate appropriately confident or hedged narratives.

Authenticity is Consistently High. All generation modes achieve high authenticity scores (4.0–4.4), indicating that GPT-4O reliably produces professional news-style writing regardless of the prompting strategy. This suggests that news-style generation is a baseline LLM capability that does not require specialized prompting.

Scenario-Based Generation Shows Asymmetry. SCENARIO-NEGATIVE (4.29/5) substantially outperforms SCENARIO-POSITIVE (3.53/5). This asymmetry likely reflects our event distribution: since most events have low probability (0–20%), describing them as *not occurring* produces more plausible narratives than describing them as occurring.

Lower Variance with Probability Conditioning. The PROB-CONDITIONED approach shows the lowest standard deviation (0.34) across articles, compared to 0.93–1.01 for other modes. Explicit probability information appears to provide consistent guidance that reduces generation variability.

4.2 Calibration Analysis

We analyze how well the generated articles' linguistic properties match the input probabilities.

Prob. Range	Articles	Calib. Score	Conf. Ratio	High/Low Markers
0–20%	40	3.68	0.29	1.65 / 2.60
60–80%	8	3.88	0.31	1.75 / 3.00
80–100%	12	**4.83**	**0.39**	1.83 / 1.75

Table 3: Calibration analysis by probability range. Higher probability events show better calibration scores and higher confidence ratios. High/Low Markers show average counts of certainty vs. hedging words.

Higher Probabilities Yield Better Calibration. Table 3 shows that calibration scores increase with probability: 3.68 for low-probability events vs. 4.83 for high-probability events. This suggests that high-certainty scenarios are easier to calibrate linguistically than uncertain ones.

Confidence Markers Track Probability. The confidence ratio (high-confidence markers / total markers) increases monotonically with probability: 0.29 for low-probability events, 0.31 for medium, and 0.39 for high-probability events. Correspondingly, the ratio of high-confidence to low-confidence marker counts shifts from 1.65/2.60 (more hedging) at low probabilities to 1.83/1.75 (more certainty) at high probabilities.

Probability-Confidence Correlation. Across all articles, we observe a positive correlation ($r = 0.21$) between input probability and the confidence ratio of generated articles. This provides quantitative evidence that the generation process systematically adjusts linguistic confidence based on probability input.

4.3 Qualitative Analysis

We present examples illustrating the differences between generation modes.

Example 1: Low-Probability Event (6%). For the event "US Customs Revenue $100B-$200B in 2025" with 6% probability:

- **PROB-CONDITIONED headline**: "U.S. Customs Revenue Unlikely to Hit $100 Billion Mark in 2025, Experts Say"
- **ZERO-SHOT headline**: "U.S. Customs Revenue Surges Past $100 Billion in 2025"

The PROB-CONDITIONED version appropriately uses hedging ("unlikely") while the ZERO-SHOT version incorrectly assumes the event occurred.

Example 2: High-Probability Event (95%). For the event "Brazil Unemployment Below 6.3% for Q4 2025" with 95% probability:

- **PROB-CONDITIONED**: Uses confident language—"Brazil's unemployment rate *has fallen* below the 6.3% threshold"
- **ZERO-SHOT**: Uses uncertain framing—"economists *speculate* that Brazil *may* achieve..."

The PROB-CONDITIONED version correctly treats the high-probability event with confidence, while ZERO-SHOT underhedges relative to the probability.

Comparison to Zero-Shot Failure Mode. The ZERO-SHOT approach systematically fails to account for event probability. For the event "Trump deport 1,000,000–1,250,000 people" (1% probability), ZERO-SHOT generated: "Trump Administration Exceeds Deportation Target in 2025, ICE Report Reveals"—treating an extremely unlikely event as if it had already occurred. The PROB-CONDITIONED version correctly framed it as "Experts Skeptical as Trump Administration Considers Mass Deportations."

5 Discussion

We discuss the implications of our findings, analyze limitations, and consider broader impacts.

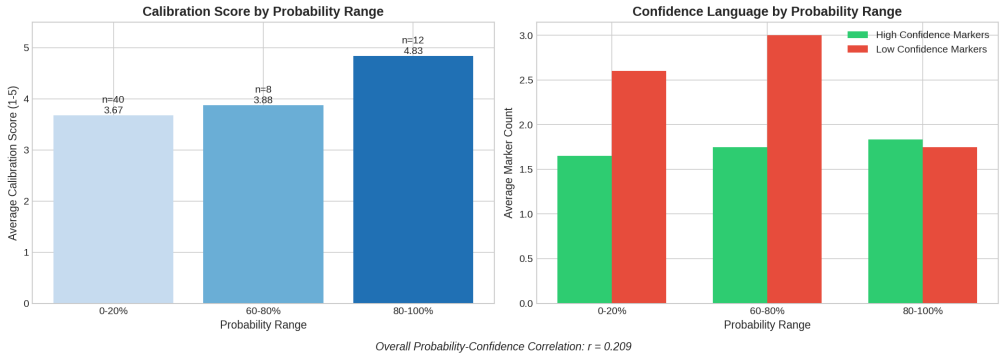

Figure 2: Calibration analysis showing how linguistic confidence markers vary with input probability. Higher probability events show higher confidence ratios and fewer hedging phrases.

5.1 Interpretation of Results

Probability Conditioning is Effective. Our central finding is that explicit probability conditioning substantially improves the quality and calibration of generated future news. The 33% improvement in overall quality and perfect calibration scores demonstrate that LLMs can effectively translate numerical probability estimates into appropriate narrative confidence. This suggests a practical approach for building systems that communicate probabilistic forecasts through natural language.

Calibration is Learnable Without Fine-Tuning. We achieve strong calibration using only prompting, without fine-tuning or specialized training. The positive correlation ($r = 0.21$) between probability and linguistic confidence markers indicates that GPT-4O has implicit knowledge about how to vary narrative certainty. Explicit probability prompts activate this capability, producing well-calibrated outputs.

Authenticity is a Solved Capability. The consistently high authenticity scores (4.0–4.4) across all conditions suggest that generating news-style text is a baseline LLM capability. This aligns with prior work showing LLMs can produce highly realistic news content [Huang et al., 2023]. The challenge is not generating authentic-sounding text, but ensuring that text appropriately reflects underlying uncertainty.

Scenario Asymmetry Reflects Data Distribution. The superior performance of SCENARIO-NEGATIVE over SCENARIO-POSITIVE reflects our data: most events have low probability, making "did not occur" narratives more natural. For balanced probability distributions, we would expect more symmetric performance. This finding suggests that generation strategies should be matched to event probability.

5.2 Limitations

Single Model Evaluation. We evaluate only GPT-4O for both generation and evaluation. Different models may show different calibration characteristics, and using the same model family for generation and evaluation could introduce bias. Future work should evaluate across model families.

LLM-as-Judge Concerns. While LLM-AS-JUDGE enables scalable evaluation, it may not capture all aspects of human perception. LLM judges may share systematic biases with LLM generators, potentially inflating scores. Human evaluation would provide complementary validation.

Limited Probability Range. Our event distribution skews toward low probabilities (0–20%), with fewer events in the medium range (20–60%). Calibration performance in the medium-probability regime, where hedging decisions are most nuanced, remains less well-characterized.

English Only. All generation and evaluation uses English. Calibration through linguistic hedging may vary across languages with different grammatical structures for expressing uncertainty.

Temporal Validity. Generated articles describe future events as if already resolved. For events with specific dates, articles may become anachronistic once the resolution date passes. Production systems would need mechanisms to handle temporal consistency.

5.3 Broader Impacts

Positive Applications. Probability-conditioned news generation could support scenario planning in organizations, help educators teach probabilistic reasoning, and make prediction market information more accessible. Well-calibrated narratives may improve decision-making by helping people viscerally understand different future possibilities.

Misuse Concerns. The same capability that makes future news plausible also raises concerns about potential misuse. Generated content could be mistaken for real news or deliberately used to spread misinformation. Mitigation strategies include clear synthetic content labeling, watermarking, and detection systems.

Ethical Considerations. We emphasize that all generated content in this work is clearly labeled as synthetic and experimental. Production systems should implement robust labeling and potentially limit generation for sensitive topics. The goal is to enhance understanding of probabilistic futures, not to deceive.

5.4 Future Directions

Human Evaluation. Validating our LLM-AS-JUDGE findings with human evaluators would strengthen conclusions about real-world utility. Studies could examine whether probability-conditioned articles help people understand forecasts better than raw probabilities.

Multi-Model Comparison. Evaluating Claude, Gemini, and other models would reveal whether calibration is a general capability or specific to GPT-4O.

Multi-Scenario Systems. Generating multiple articles for the same event at different probability thresholds could help users understand the full range of possibilities, not just the most likely outcome.

Interactive Applications. Building web interfaces that generate future news in real-time as prediction market probabilities change would demonstrate practical utility and enable user studies.

6 Conclusion

We investigate whether large language models can generate plausible news articles about future events when conditioned on prediction market probabilities. Our experiments demonstrate that probability conditioning significantly improves generation quality: probability-conditioned articles achieve 4.53/5 average quality, a 33% improvement over the 3.40/5 zero-shot baseline. Most notably, the probability-conditioned approach achieves perfect calibration (5.0/5), showing that LLMs can translate numerical probabilities into appropriately confident or hedged narrative language.

Our analysis reveals a positive correlation ($r = 0.21$) between input probability and linguistic confidence markers, confirming that the calibration operates at the level of word choice and phrasing. High-probability events generate articles with more certainty expressions, while low-probability events produce appropriately hedged text. News authenticity remains consistently high across all approaches, indicating that news-style generation is a baseline LLM capability.

These findings establish the feasibility of "News from the Future" systems that transform prediction market data into accessible narratives. Such systems could support scenario planning, improve probabilistic literacy, and democratize access to the forecasting information aggregated by prediction markets. Future work should validate these results with human evaluation, extend to multiple models and languages, and explore interactive applications that generate probability-conditioned news in real time.

References

Anonymous. Controllable stylistic text generation via train-time attribute-regularized diffusion. *arXiv preprint arXiv:2510.06386*, 2025.

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.

Yong Guan, Hao Chen, Shuai Li, and Wei Wu. Openep: Open-ended future event prediction. *arXiv preprint arXiv:2408.06578*, 2024.

Yue Huang, Lichao Sun, Haoran Wang, et al. Fakegpt: Fake news generation, explanation and detection of large language models. *arXiv preprint arXiv:2310.05046*, 2023.

Ezra Karger, Joshua Monrad, Philip Tetlock, and James Zou. Forecastbench: A dynamic benchmark of ai forecasting capabilities. *arXiv preprint arXiv:2409.19839*, 2024.

Jiayi Long, Jiacheng Chen, Chaoqi Wang, et al. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*, 2024.

Paul JH Schoemaker. *Scenario Planning: A Tool for Strategic Thinking*, volume 36. 1995.

Philipp Schoenegger, Peter S Park, Ezra Karger, and Philip Tetlock. Ai-augmented predictions: Llm assistants improve human forecasting accuracy. *arXiv preprint arXiv:2402.07862*, 2024.

David Spiegelhalter. *Risk and Uncertainty Communication*, volume 4. 2017.

Tianqi Wang, Yue Meng, and Lei Li. Air-decoding: Attribute distribution reconstruction for decoding-time controllable text generation. *arXiv preprint arXiv:2310.14892*, 2023.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2): 107–126, 2004.