

# Future Is Unevenly Distributed: Forecasting Ability of LLMs Depends on What We’re Asking

Anonymous ACL submission

## Abstract

Large Language Models (LLMs) demonstrate partial forecasting competence across social, political, and economic events. Yet, their predictive ability varies sharply with domain structure and prompt framing. We investigate how forecasting performance varies with different model families on real-world questions about events that happened beyond the model cut-off date. We analyze how context, question type, and external knowledge affect accuracy and calibration, and how adding factual news context modifies belief formation and failure modes. Our results show that forecasting ability is highly variable as it depends on what, and how, we ask.

## 1 Introduction

Large Language Models (LLMs) have increasingly saturated a variety of benchmarks, demonstrating near-superhuman abilities in programming, mathematics, and scientific reasoning (OpenAI, 2025b; DeepSeek-AI et al., 2025; Comanici et al., 2025; Anthropic, 2025b). The forecasting abilities of LLMs remain underexplored. With the growing integration of AI systems into high-stakes decision making, it is essential to assess whether such models can meaningfully predict real-world outcomes and to understand their systematic failure modes. Recent investigations into LLM forecasting such as the Metaculus AI Forecasting Benchmark (Metaculus, 2025a) and ForecastBench (Karger et al., 2025) examine whether models can perform in real-world prediction markets like Polymarket, Metaculus, and Kalshi (Polymarket, 2025; Metaculus, 2025b; Kalshi, 2025), and whether they can generate profit with external reasoning or retrieval tools. These studies, however, do not deeply analyze what types of questions LLMs answer confidently or where they fail. This gap in understanding limits our ability to interpret LLMs’ apparent forecasting success.

We address this gap by analyzing multiple LLM families, including both reasoning-optimized and non-reasoning models, and by evaluating them through a combination of qualitative and quantitative metrics such as accuracy and brier score. We also explore how model behavior changes when factual news context is added prior to prediction.

## Contributions

- **Comprehensive evaluation:** We perform a qualitative and metric-based analysis of LLMs’ forecasting performance across multiple domains, both with and without contextual news inputs.
- **Failure mode taxonomy:** We identify and categorize recurrent failure modes that emerge during forecasting, particularly when contextual information is introduced, highlighting where reasoning and calibration diverge.

## 2 Related Work

The predictive reasoning capabilities of large language models have recently become a topic of growing interest. Early evidence from real-world forecasting tournaments showed that unassisted models such as GPT-4 underperformed relative to aggregate human forecasters (Schoenegger and Park, 2023). Subsequent efforts have sought to improve this gap through large-scale fine-tuning and reinforcement learning on temporal reasoning tasks. Studies such as (Halawi et al., 2024; Lee et al., 2025; Lu, 2025) demonstrate human-comparable accuracy, large-scale event forecasting training pipelines, and direct benchmarking of LLMs against expert forecasters, respectively. Collectively, these works indicate that iterative improvements in reasoning and retrieval alignment yield measurable forecasting gains.

Several recent initiatives have formalized AI forecasting evaluation through structured bench-

marks. The *Metaculus AI Forecast Benchmarking Tournament* (Metaculus, 2025a) and *ForecastBench* (Karger et al., 2025) present dynamic leaderboards using real prediction market questions drawn from platforms such as Polymarket and Metaculus (Polymarket, 2025; Metaculus, 2025b). Prophet Arena (Yang et al., 2025) further examines the theoretical grounding of "LLM-as-prophet" predictive intelligence, emphasizing calibration and model uncertainty. Alongside these developments, studies discuss key pitfalls in evaluating LLM forecasters, including logical leakage, unreliable news retrieval, and data contamination due to excessive reliance on model training cutoffs (Paleka et al., 2025).

Complementary datasets extend this line of inquiry toward temporal and contextual reasoning. *ForecastQA* (Jin et al., 2021), *Autocast* (Zou et al., 2022), *ExpTime* (Yuan et al., 2024), *FOReCAST* (Yuan et al., 2025), and *FutureX* (Zeng et al., 2025) each evaluate long-horizon prediction and confidence estimation under streaming updates. Mutschlechner and Jatowt (2025) analyze contextual cues in prompt design, finding that LLMs' sensitivity to framing influences both calibration and directional correctness (Mutschlechner and Jatowt, 2025).

Parallel to academic benchmarks, open-source infrastructures such as the *Metaculus Forecasting Tools* (Metaculus, 2024) and *ManifoldBot* (Microprediction, 2024) enable autonomous LLM agents to interact directly with market-style systems, bridging probabilistic modeling with real-time trading and decision aggregation. Together, these works frame forecasting as an emerging dimension of LLM evaluation spanning Human-AI comparison, contextual robustness, and dynamic market participation.

**Our evaluation differs from these works as we study the failure modes of these models** with news added as context and also **show the clear difference in model performance** according to question category.

### 3 Methodology

In this section we detail the methodology behind our data processing and evaluation pipeline.

#### 3.1 Data Processing

We began by collecting approximately 10,000 forecasting questions from various prediction markets such as Polymarket, Metaculus, and Manifold Mar-

kets (Polymarket, 2025; Metaculus, 2025b; Manifold Markets, 2025), covering the period from January to July 2025. This period was chosen so that all questions selected were beyond the model's cut-off date. Many of these questions were noisy, that is, their context was hyper-localized or failed to test the forward-looking reasoning ability of large language models in a meaningful way.

Some examples include:

"Daily coinflip"

"Will the % chance of 'YES' on this market close above 50%?"

"Will I get a Donation/Payment of 10,000 or more Mana before 2025?"

These questions do not provide any real signal of forecasting competence or reveal systematic failure modes. To extract a meaningful subset, we designed a three-stage filtering and classification pipeline (Figure 1).

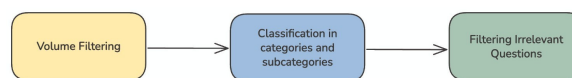


Figure 1: Overview of the data-processing pipeline used to construct the filtered forecasting benchmark.

First, we applied **volume filtering** to remove low-liquidity markets, which typically correspond to hyper-personalized or creator-specific questions. Next, we employed *Gemini 2.5 Flash* (Comanici et al., 2025) as an LLM-as-a-Judge (Zheng et al., 2023) with the following prompt (see Appendix A.3) to classify each question into six primary categories, each with five sub-categories:

- **Politics:** Domestic Policy, Elections & Campaigns, Political Parties & Ideologies, Government Structure, Public Policy & Social Issues
- **Entertainment:** Movies & Television, Music & Audio, Gaming, Celebrity & Pop Culture, Books & Literature
- **Sports:** Professional Sports, International Competitions, Individual Sports, Team Sports, Sports Culture & Recreation
- **Technology:** Computing & Software, Internet & Digital Services, Mobile & Consumer Electronics, Emerging Technologies, Tech Industry & Business

- **Finance:** Personal Finance, Banking & Financial Services, Markets & Trading, Economic Indicators, Corporate Finance
- **Geopolitics:** International Relations, Global Conflicts, Trade & Economics, Regional Affairs, Global Governance

Questions that did not align with any of the above were tagged as *irrelevant*, reducing the corpus to roughly 700 items after aggressive filtering. Despite this reduction, certain residual questions remained non-event-based and failed to meaningfully test predictive reasoning. For instance:

*"Will @Soaffine be active on Manifold again before April?"*

To address these, we performed a second LLM-based filtering pass using a refined judging prompt (see Appendix A.4) to exclude localized or non-forecasting items. The final curated dataset contained 392 questions, evenly distributed across the categories and sub-categories listed above. For each retained question, we also preserved metadata such as `creationTime`, `resolutionTime`, and final resolution probability.

### 3.2 Evaluation Methodology

We start by sampling a uniform subset of 150 questions with seed 42 from the final corpus, ensuring an equal number of questions per category to construct a balanced evaluation dataset. This subset enables consistent cross-category comparison while maintaining representativeness of the larger filtered corpus.

We evaluate a mixture of reasoning-focused and non-reasoning large language models: GPT-5 (OpenAI, 2025b), GPT-4.1 (OpenAI, 2025a), DeepSeek-R1 (DeepSeek-AI et al., 2025), and Claude 3.7 Sonnet (Anthropic, 2025a). We sample from the models at a temperature of 0.0, max token budget as 4500 tokens to ensure that models have enough buffer to express their reasoning, as well as for deterministic sampling with 0.0 temperature.

Each model is prompted using a standard forecasting prompt (see Appendix A.1) along with the question text and its creation date, to provide temporal grounding. Apart from this contextual timestamp, the models have no access to external tools, retrieval systems, or web search capabilities.

For every prompt, each LLM outputs two required fields:

`<ans>YES/NO</ans>`

`<conf>0–1 confidence score</conf>`

We evaluate predictions using three key metrics: **accuracy**, the **brier score**, and the **Expected Calibration Error (ECE)**.

**Accuracy.** Measures whether the model’s predicted resolution (`<ans>`) matches the actual market resolution for each question. A correct match contributes 1, and an incorrect match contributes 0; the mean across all samples yields the final accuracy score.

**Brier Score.** Quantifies probabilistic calibration by penalizing confidence errors and is formally defined as:

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2, \quad (1)$$

where  $f_i$  denotes the model’s predicted probability (confidence) for a "YES" outcome, and  $o_i \in \{0, 1\}$  represents the ground-truth outcome. Lower values indicate better probabilistic accuracy.

**Expected Calibration Error (ECE).** Measures the discrepancy between predicted confidence and empirical accuracy across probability bins. Predictions are partitioned into  $M$  bins based on confidence, and ECE is computed as:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (2)$$

where  $B_m$  is the set of predictions whose confidence scores fall into bin  $m$ ,  $\text{acc}(B_m)$  is the average accuracy within that bin, and  $\text{conf}(B_m)$  is the mean predicted confidence. Lower ECE values indicate better calibration between confidence and correctness.

### 3.3 Evaluation with News Context

For the second evaluation condition, we augment each forecasting question with external context retrieved from contemporary news sources. This ensures that LLMs have the same set of information that a human forecaster would have, when the event was created as a question in the prediction market. We collect ten news snippets per question by querying *Exa* (Exa, 2025) with the question text, using the question’s creation date as the upper bound for publication time. Despite the temporal cutoff, we occasionally observed leakage in the form of articles published after the creation date. Such snippets were removed completely to ensure temporal purity in all model inputs.

Category	Model	Accuracy	Brier	ECE
Entertainment	Claude-3.7	68.00%	0.23	0.18
	DeepSeek-R1	68.00%	0.28	0.20
	GPT-4.1	64.00%	0.33	0.26
	GPT-5	52.00%	0.24	0.20
Finance	Claude-3.7	44.00%	0.31	0.31
	DeepSeek-R1	48.00%	0.35	0.29
	GPT-4.1	40.00%	0.33	0.27
	GPT-5	56.00%	0.26	0.20
Geopolitics	Claude-3.7	84.00%	0.12	0.12
	DeepSeek-R1	84.00%	0.32	0.36
	GPT-4.1	88.00%	0.40	0.45
	GPT-5	84.00%	0.14	0.09
Politics	Claude-3.7	68.00%	0.22	0.25
	DeepSeek-R1	64.00%	0.27	0.29
	GPT-4.1	72.00%	0.38	0.42
	GPT-5	64.00%	0.21	0.17
Sports	Claude-3.7	48.00%	0.28	0.33
	DeepSeek-R1	48.00%	0.26	0.23
	GPT-4.1	60.00%	0.45	0.49
	GPT-5	52.00%	0.28	0.26
Technology	Claude-3.7	68.00%	0.25	0.27
	DeepSeek-R1	64.00%	0.27	0.35
	GPT-4.1	72.00%	0.42	0.47
	GPT-5	68.00%	0.24	0.23

Table 1: Category-wise metrics *without news context*. Each category contains 25 questions. Accuracy (%) is shown alongside Brier and Expected Calibration Error (ECE) averaged per category.

Each model is then re-evaluated on the context-augmented prompt (see Appendix A.2) using the same scoring metrics, *Accuracy*, *Brier score* and *ECE*, to measure how additional factual context influences forecasting calibration and directional correctness.

## 4 Results by Category

Table 1 presents category-wise model performance *without news context*, evaluated using Accuracy, Brier Score, and Expected Calibration Error (ECE). ECE captures the deviation between model-predicted probabilities and observed outcomes, offering a finer measure of calibration quality beyond raw accuracy. Results show that GPT-5 and Claude-3.7 achieve strong calibration on structured domains such as *Geopolitics* and *Politics*, while DeepSeek-R1, GPT-4.1 display higher ECE in noisier domains like *Entertainment* and *Technology*.

## 5 News-Augmented Forecasting

We next evaluate the same set of models when each question is supplemented with up to ten time-bounded news snippets retrieved prior to the question’s creation date. Table 2 shows the corresponding metrics. While certain domains such as *Finance*

Category	Model	Accuracy	Brier	ECE
Entertainment	Claude-3.7	56.00%	0.27	0.19
	DeepSeek-R1	40.00%	0.34	0.42
	GPT-4.1	44.00%	0.36	0.35
	GPT-5	56.00%	0.27	0.25
Finance	Claude-3.7	56.00%	0.31	0.30
	DeepSeek-R1	52.00%	0.31	0.27
	GPT-4.1	68.00%	0.29	0.23
	GPT-5	60.00%	0.23	0.14
Geopolitics	Claude-3.7	80.00%	0.15	0.21
	DeepSeek-R1	80.00%	0.31	0.40
	GPT-4.1	76.00%	0.46	0.54
	GPT-5	84.00%	0.13	0.16
Politics	Claude-3.7	64.00%	0.26	0.26
	DeepSeek-R1	68.00%	0.29	0.33
	GPT-4.1	68.00%	0.33	0.31
	GPT-5	72.00%	0.18	0.12
Sports	Claude-3.7	60.00%	0.22	0.19
	DeepSeek-R1	64.00%	0.24	0.31
	GPT-4.1	56.00%	0.27	0.26
	GPT-5	56.00%	0.23	0.25
Technology	Claude-3.7	52.00%	0.29	0.36
	DeepSeek-R1	48.00%	0.33	0.40
	GPT-4.1	64.00%	0.43	0.53
	GPT-5	68.00%	0.23	0.30

Table 2: Category-wise metrics *with news context*. Each category contains 25 questions. Accuracy (%) is shown alongside Brier and Expected Calibration Error (ECE) averaged per category.

and *Sports* benefit from context (lower brier, improved ECE), others such as *Entertainment* and *Technology* show declines, consistent with recency bias and noise amplification effects introduced by the additional text. We detail the explanation of these results in further sections.

## 6 Analysis and Failure Modes with Context

Analyzing the models’ responses and reasoning traces from our evaluation reveals several recurring failure modes. When incorporating news as additional context, we observe issues consistent with those reported by Paleka et al. (2025) (Paleka et al., 2025), particularly those concerning unreliable news retrieval. Despite enforcing explicit temporal bounds on article publication dates through *Exa* (Exa, 2025), we find that articles published after the question’s cutoff sometimes containing information that effectively resolves the question can still appear in the retrieved set when filtering is insufficient.

The addition of news as context improves the model in certain aspects such as clarifying the time scope of the question and latching onto proper signals, but it is also highly prone to various issues.

311	We detail some of them below.		
312	<b>Recency Bias.</b> Models tend to overweight recent		
313	news over historical trends encoded during pretrain-		
314	ing. This often leads to situations where the model		
315	changes a correct resolution into an incorrect one.		
316	<i>Question: "S&amp;P 500 above 6050 on June</i>		
317	<i>13?"</i>		
318	<b>Raw model (a):</b> NO, 0.34 confidence,		
319	reasons that the index is near resistance		
320	at 6000 and mean reversion plus limited		
321	trading days make a breakout unlikely.		
322	(Correct)		
323	<b>News model (b):</b> YES, 0.54 confidence,		
324	reads snippets from the few days before		
325	June 13 describing the S&P "flirting with		
326	6000," "record highs," and "strategist up-		
327	grades targeting 6100." (Wrong)		
328	The model allowed the most recent headlines		
329	to dominate its prior, turning a correct mean-		
330	reversion call into an overconfident breakout bet.		
331	The complete reasoning trace is provided (see Ap-		
332	pendix B.1).		
333	<b>Rumour Overweighting.</b> Models frequently		
334	anchor to unverified information or speculation		
335	present in retrieved snippets, causing them to		
336	switch from a correct to an incorrect resolution.		
337	<i>Question: "Tariffs on China above 150%</i>		
338	<i>by end of June?"</i>		
339	<b>Raw model (a):</b> NO, high confidence		
340	(0.85), cites precedent and policy friction.		
341	(Correct)		
342	<b>News model (b):</b> YES, high confidence		
343	(0.65), flips after reading late-April and		
344	May headlines suggesting tariffs were		
345	"likely" to rise to 150%. (Wrong)		
346	Headlines indicated possibility rather than policy.		
347	The correct resolution required actual implemen-		
348	tation by the deadline, which did not occur. Ru-		
349	mour anchoring overweighted momentum of cov-		
350	erage and underweighted institutional lag, shifting		
351	the model from a cautious, process-aware NO to		
352	an overconfident, headline-driven YES. Reasoning		
353	trace (see Appendix B.2).		
354	<b>Definition Drift.</b> Models sometimes misinter-		
355	pret acronyms or context when additional news		
356	shifts their semantic grounding, leading to incor-		
357	rect predictions.		
	<i>Question: "Will MATS applications open</i>		358
	<i>in March?"</i>		359
	<b>True resolution:</b> YES		360
	<b>Raw model (a):</b> YES, 0.58 confidence,		361
	interprets MATS as a recurring academic		362
	program that historically opens applica-		363
	tions each March, referencing prior cy-		364
	cles. (Correct)		365
	<b>News model (b):</b> NO, 0.35 confidence,		366
	reinterprets MATS as the Mid-America		367
	Trucking Show, where registrations open		368
	months before March. (Wrong)		369
	The model with news context was exposed to		370
	snippets dominated by the trucking expo the most		371
	search visible meaning of MATS and thus shifted		372
	semantic grounding from an academic program to a		373
	trade event. This altered both the reference domain		374
	and the expected timeline, leading to a confident		375
	but misplaced "NO." It underweighted contextual		376
	cues from the original question (application cycle,		377
	academic phrasing) and overtrusted frequency in		378
	retrieved snippets, effectively letting entity salience		379
	override contextual fit. Complete examples are		380
	provided (see Appendix B.3).		381
	An additional behavior observed in <i>DeepSeek-R1</i>		382
	( <i>DeepSeek-AI et al., 2025</i> ) is that it does not pro-		383
	vide any reasoning traces even when explicitly men-		384
	tioned to verbalise its reasoning process through		385
	the prompt. The model does output the final tags		386
	in the form of <ans></ans> and <conf></conf>		387
	tags, but does not verbalise its reasoning process.		388
	For examples (see Appendix B.4)		389
	<b>7 Conclusion</b>		390
	Forecasting competence in LLMs is highly uneven,		391
	reflecting not only data coverage but the cogni-		392
	tive framing embedded in prompts. While we may		393
	expect adding recent news should improve forecast-		394
	ing accuracy, we find that sometimes it does while		395
	at other times it makes it worse because of defini-		396
	tion drift, rumour anchoring and recency bias etc.		397
	The findings underscore that future-reasoning abil-		398
	ity is conditional, not emergent, and invite design		399
	of benchmarks that disentangle knowledge recall		400
	from probabilistic inference.		401
	<b>8 Limitations</b>		402
	Our study relies on a heavily filtered subset of		403
	prediction-market questions to ensure that the		404
	benchmark targets genuine forward-looking events		405

rather than platform-specific or trivial outcomes. While this filtering improves signal quality, it also limits the scale and diversity of the evaluation set, and our findings may not fully generalize to noisier or less structured forecasting environments.

Second, our evaluation considers only two inference conditions forecasting with and without contemporaneous news context. This binary setup does not capture a broader spectrum of information access patterns, such as iterative retrieval, selective evidence weighting, or multi-step tool use, which may materially alter model behavior. As a result, our results should be interpreted as characterizing baseline model tendencies rather than upper bounds on achievable forecasting performance.

Third, we restrict evaluation to single-shot, deterministic model outputs at zero temperature. While this choice improves comparability and reduces variance, it may underrepresent the benefits of ensembling, sampling-based uncertainty estimation, or calibration-through-aggregation strategies commonly used by human forecasters.

Finally, although we enforce temporal constraints on retrieved news, residual leakage or salience bias in external information sources cannot be fully ruled out. Such effects may disproportionately influence models that heavily weight recent textual evidence, complicating attribution of errors to reasoning versus retrieval artifacts.

## References

Anthropic. 2025a. Claude 3.7 sonnet system card. <https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf>. Accessed: 2025-10-21.

Anthropic. 2025b. Claude 4.5 sonnet system card. <https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf>. Accessed: 2025-10-21.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 8 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,

Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Exa. 2025. Exa api documentation. <https://exa.ai/>. Accessed: 2025-10-21.

Danny Halawi, Fred Zhang, Yueh-Han Chen, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. *Preprint*, arXiv:2402.18563.

Xinyu Jin, Hantian Zhang, Wenhan Xiong, Yichong Qin, Caiming Xiong, Jingjing Liu, Yiming Wu, and 1 others. 2021. Forecastqa: A question answering benchmark for event forecasting. *Preprint*, arXiv:2104.05441.

Kalshi. 2025. Kalshi event forecasting exchange. <https://kalshi.com/>. Accessed: 2025-10-21.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. 2025. Forecastbench: A dynamic benchmark of ai forecasting capabilities. *Preprint*, arXiv:2409.19839.

Sang-Woo Lee, Sohee Yang, Donghyun Kwak, and Noah Y. Siegel. 2025. Advancing event forecasting through massive training of large language models: Challenges, solutions, and broader impacts. *Preprint*, arXiv:2507.19477.

Janna Lu. 2025. Evaluating llms on real-world forecasting against expert forecasters. *Preprint*, arXiv:2507.04562.

Manifold Markets. 2025. Manifold prediction market platform. <https://manifold.markets/>. Accessed: 2025-10-21.

Metaculus. 2024. Metaculus forecasting tools. <https://github.com/Metaculus/forecasting-tools>. Accessed: 2025-10-21.

Metaculus. 2025a. Ai forecasting benchmark. <https://www.metaculus.com/aib/>. Accessed: 2025-10-21.

Metaculus. 2025b. Metaculus prediction market. <https://www.metaculus.com/>. Accessed: 2025-10-21.

Microprediction. 2024. Manifoldbot: Automated prediction market agent. <https://github.com/microprediction/manifoldbot>. Accessed: 2025-10-21.

Gerrit Mutschlechner and Adam Jatowt. 2025. Analyzing the role of context in forecasting with large language models. *Preprint*, arXiv:2501.06496.

509	OpenAI. 2025a. Gpt-4.1 system card. <a href="https://openai.com/index/gpt-4-1/">https://openai.com/index/gpt-4-1/</a> . Accessed: 2025-10-21.	Maintain all information in the original question.	561
510			562
511		<rephrased_question></rephrased_question>	563
512	OpenAI. 2025b. Gpt-5 system card. <a href="https://cdn.openai.com/gpt-5-system-card.pdf">https://cdn.openai.com/gpt-5-system-card.pdf</a> . Accessed: 2025-10-21.	2. Provide a few reasons why the answer might be no. Rate the strength of each reason.	565
513			566
514			567
515	Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. 2025. <a href="#">Pitfalls in evaluating language model forecasters</a> . <i>Preprint</i> , arXiv:2506.00723.	<no_thoughts></no_thoughts>	568
516			569
517		3. Provide a few reasons why the answer might be yes. Rate the strength of each reason.	570
518	Polymarket. 2025. Polymarket prediction platform. <a href="https://polymarket.com/">https://polymarket.com/</a> . Accessed: 2025-10-21.	<yes_thoughts></yes_thoughts>	572
519			573
520		4. Aggregate your considerations. Think like a superforecaster (e.g. Nate Silver).	574
521	Philipp Schoenegger and Peter S. Park. 2023. <a href="#">Large language model prediction capabilities: Evidence from a real-world forecasting tournament</a> . <i>Preprint</i> , arXiv:2310.13014.	<considerations></considerations>	575
522			576
523		5. Output an initial probability (prediction) given steps 1-4.	577
524			578
525	Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. 2025. <a href="#">Llm-as-a-prophet: Understanding predictive intelligence with prophet arena</a> . <i>Preprint</i> , arXiv:2510.17638.	<initial_probability></initial_probability>	579
526			580
527		6. Evaluate whether your calculated probability is excessively confident or not confident enough. Also,	581
528			582
529	Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. 2024. <a href="#">Exptime: Temporal reasoning benchmark for event forecasting</a> . <i>Preprint</i> , arXiv:2407.06142.	consider anything else that might affect the forecast that you did not before consider.	584
530			585
531			586
532	Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. 2025. <a href="#">Forecast: The future outcome reasoning and confidence assessment benchmark</a> . <i>Preprint</i> , arXiv:2502.19676.	<extra_considerations></extra_considerations>	587
533			588
534		7. Output your answer in <ans>YES/NO</ans> and the confidence in <conf>0-1</conf>.	589
535			590
536	Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, and 1 others. 2025. <a href="#">Futurex: An advanced live benchmark for llm agents in future prediction</a> . <i>Preprint</i> , arXiv:2508.11987.	Output the confidence as a number between 0 and 1 (e.g. 0.85), without a % sign. Do not output anything else.	592
537			593
538		Make sure you follow all instructions, reduce reasoning effort if required. Do not repeat the points mentioned in the prompt.	595
539			596
540		Example (correct format):	597
541	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. <a href="#">Judging llm-as-a-judge with mt-bench and chatbot arena</a> . <i>Preprint</i> , arXiv:2306.05685.	<extra_considerations>	598
542			599
543		While I am reasonably confident in this forecast, unexpected events such as corporate restructuring,	600
544			601
545		internal conflicts, or external pressures could alter the situation. The absence of recent news increases	602
546			603
547	Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. 2022. <a href="#">Forecasting future world events with neural networks</a> . <i>Preprint</i> , arXiv:2206.15474.	confidence, but I remain cautious due to the unpredictable nature of corporate leadership dynamics.	604
548			605
549			606
550			607
551			608
552	<b>A Prompt Templates</b>		609
553	<b>A.1 Core Forecasting Prompt (No News)</b>	</extra_considerations>	610
554	SYSTEM_PROMPT = ""	Example (wrong format):	611
555	Question created on {date_str}):		612
556	{question}	6. Evaluate whether your calculated probability is excessively confident or not confident enough. Also,	613
557	Instructions:		614
558	1. Given the above question, rephrase and expand it to help you do better answering.	consider anything else that might affect the forecast that you did not before consider.	615
559			616
560		<extra_considerations>	617
			618

619 While 0.85 confidence is high, it is  
620 not overly confident because unexpected  
621 events can always intervene.

622 However, no recent news or rumors  
623 indicate instability in leadership. The  
624 media industry can be volatile,  
625 but major leadership changes often come  
626 with early signals. Since none are  
627 evident, the estimate seems  
628 reasonable. It would be prudent to  
629 slightly discount the confidence if any  
630 new information arises during 2024,  
631 but as of now, 0.85 is appropriate.

632 </extra\_considerations>  
633 ""

## 634 A.2 News-Augmented Forecasting Prompt

635 The forecasting prompt used in this condition  
636 is identical to the base prompt described in Ap-  
637 pendix A.1. The only modification is that up to  
638 ten time-bounded news snippets retrieved via the  
639 *Exa* API (Exa, 2025) are appended to the end of  
640 the prompt before model inference.

## 641 A.3 Category Classification Judge Prompt

642 SYSTEM\_PROMPT = ""  
643 # Question Classifier System Prompt

644 You are a classifier that categorizes a  
645 given question into one of the following  
646 categories and their respective  
647 sub-categories. Choose the most  
648 appropriate category and sub-category  
649 that best fits the question's primary  
650 focus.

651 ## Categories and Sub-Categories:

652 ### 1. Politics

- 653 - Domestic Policy: Questions  
654 about internal government policies,  
655 legislation, regulations, and  
656 governance within a country.
- 657 - Elections & Campaigns: Questions  
658 about voting processes, political  
659 candidates, election results, and  
660 campaign activities.
- 661 - Political Parties & Ideologies:  
662 Questions about political movements,  
663 party platforms, political philosophies,  
664 and partisan issues.
- 665 - Government Structure: Questions about  
666 constitutional matters, branches of  
667 government, political systems, and  
668 institutional processes.
- 669 - Public Policy & Social Issues:  
670 Questions about policy debates, social  
671 reforms, civil rights, and politically  
672 relevant social topics.

673 ### 2. Entertainment

- 674 - Movies & Television: Questions about  
675 films, TV shows, streaming content,  
676 actors, directors, and cinema industry.

- Music & Audio: Questions about  
677 songs, artists, albums, concerts, music  
678 industry, and audio entertainment. 679
- Gaming: Questions about video  
680 games, gaming platforms, esports, game  
681 development, and gaming culture. 682
- Celebrity & Pop Culture:  
683 Questions about famous personalities,  
684 entertainment news, awards, and popular  
685 culture trends. 686
- Books & Literature: Questions about  
687 authors, novels, publishing, literary  
688 works, and reading culture. 689

690 ### 3. Sports

- Professional Sports: Questions about  
691 major league competitions, professional  
692 athletes, team performance, and sports  
693 business. 694
- International Competitions: Questions  
695 about Olympics, World Cup, continental  
696 championships, and global sporting  
697 events. 698
- Individual Sports: Questions about  
699 tennis, golf, athletics, martial  
700 arts, and other individual competitive  
701 activities. 702
- Team Sports: Questions about football,  
703 basketball, cricket, rugby, and other  
704 team-based sports. 705
- Sports Culture & Recreation: Questions  
706 about sports history, fan culture,  
707 recreational activities, and sports  
708 lifestyle. 709

710 ### 4. Technology

- Computing & Software: Questions  
711 about computers, operating systems,  
712 applications, programming, and software  
713 development. 714
- Internet & Digital Services: Questions  
715 about websites, online platforms,  
716 digital services, and internet-related  
717 topics. 718
- Mobile & Consumer Electronics:  
719 Questions about smartphones, tablets,  
720 gadgets, and consumer technology  
721 products. 722
- Emerging Technologies: Questions about  
723 artificial intelligence, blockchain,  
724 virtual reality, and cutting-edge  
725 innovations. 726
- Tech Industry & Business: Questions  
727 about technology companies, tech  
728 entrepreneurship, and technology market  
729 trends. 730

731 ### 5. Finance

- Personal Finance: Questions about  
732 budgeting, saving, investing, loans, and  
733 individual financial planning. 734
- Banking & Financial Services:  
735 Questions about banks, credit,  
736 insurance, financial institutions,  
737 and financial products. 738

739	- Markets & Trading: Questions about	Relevant (True):	797
740	stock markets, bonds, commodities,	- About a public, verifiable	798
741	trading strategies, and investment	event/outcome/metric (elections,	799
742	vehicles.	matches, earnings, court rulings,	800
743	- Economic Indicators: Questions about	awards, macro prints, product releases,	801
744	inflation, GDP, unemployment, interest	etc.).	802
745	rates, and macroeconomic metrics.	Irrelevant (False):	803
746	- Corporate Finance: Questions about	- Personal/self-referential (e.g., "Will	804
747	business finance, company valuations,	I meet X...", "Will I score top	805
748	mergers & acquisitions, and corporate	in Kaggle..."), unverifiable private	806
749	financial strategies.	matters.	807
750	### 6. Geopolitics	- Platform/meta mechanics (e.g., "Will	808
751	- International Relations: Questions	all 'YES' traders have a unique max	809
752	about diplomatic relations, treaties,	payout...").	810
753	international cooperation, and	- Vague opinion polls with no verifiable	811
754	bilateral/multilateral agreements.	source.	812
755	- Global Conflicts: Questions about wars,	Edge:	813
756	territorial disputes, military tensions,	- Public figures/events/results/releases	814
757	and international security issues.	are Relevant if objectively checkable.	815
758	- Trade & Economics: Questions about	Return JSON only: {"is_relevant":	816
759	international trade, economic sanctions,	true false, "reason": "<one	817
760	trade agreements, and global economic	sentence>"}.	818
761	relations.	""	819
762	- Regional Affairs: Questions about	<b>B Failure Mode Case Studies</b>	820
763	specific geographic regions, regional	<b>B.1 Recency Bias Example: "S&amp;P 500 above</b>	821
764	organizations, and area-specific	<b>\$6050 on June 13?"</b>	822
765	political developments.	{"question": "S&P 500 above \$6050 on	823
766	- Global Governance: Questions about	June 13?", "true_resolution": "NO",	824
767	international organizations, global	"pred_1": "NO", "pred_2": "YES",	825
768	institutions, international law, and	"correct_1": true, "correct_2": false,	826
769	worldwide policy coordination.	"conf_1": 0.66, "conf_2": 0.54,	827
770	7. If it is not related to any of the	"response_1": "<rephrased_question>Will	828
771	above, classify it as "Irrelevant" for	the S&P 500 index be above 6,050	829
772	both category and sub-category.	on June 13, 2025 (interpreting	830
773	## Classification Instructions:	'above' as the closing level on	831
774	1. Read the question carefully and	that date), as asked on June 11,	832
775	identify its primary focus.	2025?</rephrased_question><no_thoughts>-	833
776	2. Select the most appropriate main	Likely starting level below 6,050;	834
777	category (1-6).	reaching/exceeding it within two	835
778	3. Choose the most relevant sub-category	trading days requires an above-average	836
779	within that main category.	move. Strength: High - 6,000 is a	837
780	4. If a question spans multiple	psychological/resistance area; initial	838
781	categories, choose the one that	breakouts often fail or retest before	839
782	represents the primary or most	advancing. Strength: Medium - Near-term	840
783	significant aspect.	macro/event risk could trigger a	841
784	5. STRICTLY STICK TO THE CATEGORIES	pullback (e.g., data releases, Fed	842
785	MENTIONED. DO NOT MAKE UP ANY CATEGORY.	communications). Strength: Medium -	843
786	For the category field your answer	Mean reversion after recent gains can	844
787	should be SPORTS, ENTERTAINMENT,	cap short-horizon upside. Strength:	845
788	POLITICS, GEOPOLITICS, TECHNOLOGY,	Medium-Low - Valuation headwinds	846
789	FINANCE, or IRRELEVANT. Nothing else.	reduce odds of rapid further gains	847
790	""	over a 2-day window. Strength:	848
791	<b>A.4 Refined Filtering Judge Prompt</b>	Low-Medium</no_thoughts><yes_thoughts>-	849
792	The following is the detailed prompt for stage three		
793	of our data processing pipeline.		
794	""\		
795	You are a strict *relevance* classifier		
796	for prediction-market questions.		

850	Ongoing bull-market momentum,	zone inviting profit-taking/pinning:	902
851	especially in mega-cap tech, can	Medium-Strong - Recent run-up increases	903
852	push the index through round-number	odds of mean reversion/'sell	904
853	levels. Strength: Medium - A clean	the news': Medium - Narrow	905
854	breakout above 6,000 could trigger	leadership/valuation stretch raises	906
855	CTA/stop-buying and extension to	fragility to negative surprises:	907
856	6,050+. Strength: Medium - Positive	Medium-Weak</no_thoughts><yes_thoughts>-	908
857	macro surprises (easing inflation,	Positive macro surprise (benign	909
858	supportive policy) can add 1%+ in a day.	CPI/dovish Fed tone) can produce	910
859	Strength: Low-Medium - Low-volatility	>1% rally, clearing 6,050: Strong	911
860	grind-up regimes increase chances	- Momentum/breakout dynamics after	912
861	of incremental new highs.	reclaiming 6,000 with bullish	913
862	Strength: Low-Medium</yes_thoughts>	sentiment and upgrades: Medium	914
863	<considerations>Without the exact June	- Mega-cap tech leadership can	915
864	11 level, scenario analysis: - If	drive index-level gains quickly:	916
865	5,800: odds to be >6,050 by June 13	Medium - Trend-following/systematic	917
866	are low ( 10%). - If 5,900-5,980:	flows post-upgrade headlines	918
867	moderate ( 25-40%), needing a 1-2% move.	support near-term upside:	919
868	- If already greater than or equal	Weak-Medium</yes_thoughts>	920
869	to 6,000: moderate-to-good ( 50-65%),	<considerations>Base rate: with	921
870	as momentum can extend. Plausible	the index near 6,000, a 0.8% move	922
871	mid-2025 baseline places the index	to >6,050 over two sessions is common,	923
872	more likely below 6,000 than above, so	45-55% territory. Macro catalysts	924
873	weighted odds favor "No." Aggregating	create a bimodal outcome; if CPI/Fed	925
874	scenarios yields an estimate near	are market-friendly, a decisive break	926
875	one-third for "Yes."</considerations>	above 6,050 is plausible; if not, index	927
876	<initial_probability>	may stall or retrace to/below 6,000	928
877	0.34</initial_probability>	due to resistance and positioning.	929
878	<extra_considerations>Uncertainty	Sentiment skew is modestly positive	930
879	about whether 'above' refers to close	given recent upgrades, but event	931
880	versus intraday; using close lowers	risk tempers confidence. Slight	932
881	the probability slightly. Lack of the	edge to "Yes."</considerations>	933
882	exact starting level introduces major	<initial_probability>0.55	934
883	variance; to avoid overconfidence, keep	</initial_probability>	935
884	the estimate near one-third rather than	<extra_considerations>Ambiguity between	936
885	lower. No further strong asymmetries	closing level vs. intraday high	937
886	identified.</extra_considerations>	matters; intraday probability would	938
887	<ans>NO</ans><conf>0.66</conf>",	be higher ( 0.60-0.65). Assuming	939
888	"response_2": "<rephrased_question>Will	close, options pinning around 6,000	940
889	the S&P 500 index finish above 6,050	could cap gains. Event clustering	941
890	on Friday, June 13, 2025? Assume	increases variance, so 0.55 may be	942
891	'above 6,050' refers to the official	a touch high; adjust slightly down	943
892	closing level of the S&P 500. Context:	to 0.54 to account for pinning/mean	944
893	the index recently revisited 6,000	reversion risk.</extra_considerations>	945
894	(first time since February) and	<ans>YES</ans><conf>0.54</conf>"}	946
895	several strategists (e.g., Barclays)		
896	raised bullish year-end targets to	<b>B.2 Rumour Overweighting: "Tariffs on</b>	947
897	6,050.</rephrased_question><no_thoughts>-	<b>China above 150% by end of June?"</b>	948
898	Event risk (CPI/Fed around June 12	{"question": "Tariffs on China above	949
899	could be hawkish, push yields up, hit	150% by end of June", "true_resolution":	950
900	equities): Strong - 6,000-6,050 is	"NO", "pred_1": "NO", "pred_2": "YES",	951
901	a psychological/technical resistance	"correct_1": true, "correct_2": false,	952

953	"conf_1": 0.85, "conf_2": 0.35,	politically salient sectors (autos,	1005
954	"response_1": "<rephrased_question>Will	steel, solar) could spur rapid action,	1006
955	any government impose tariffs on imports	especially if framed as national	1007
956	from China at a rate exceeding 150%	security. Strength: Low-Medium -	1008
957	by the end of June (June 30, 2025)?	Legislative action: a fast-tracked	1009
958	Ambiguity notes: "Tariffs on China above	bill could mandate punitive tariffs,	1010
959	150%" could mean (a) broad headline	though timing is tight. Strength:	1011
960	tariffs by a major economy (e.g., the	Low - Precedent for firm-specific	1012
961	U.S.) on a major product category (such	rates: certain firms can face >150%	1013
962	as EVs), (b) average tariffs across	AD/CVD margins; if the question counts	1014
963	goods, or (c) product-specific duties	these as "tariffs on China," this	1015
964	(including anti-dumping/countervailing	could satisfy the condition. Strength:	1016
965	duties) that sometimes exceed	Medium</yes_thoughts><considerations>Key	1017
966	150%. The timeframe is short	uncertainty is definition. If counting	1018
967	(from April 11 to June 30,	any product-specific AD/CVD, base rates	1019
968	2025).</rephrased_question><no_thoughts>-	for some Chinese goods exceeding 150%	1020
969	Policy/process friction: large headline	exist and new cases could cross 150%	1021
970	tariff hikes (e.g., Section 301/232-type	within months. If the intent is a broad,	1022
971	moves) typically require procedures,	headline tariff (e.g., on EVs) exceeding	1023
972	consultations, and political capital;	150%, the base rate is extremely low in	1024
973	implementing >150% in 2.5 months	peacetime, and the short horizon lowers	1025
974	is unlikely. Strength: High -	odds further. Political incentives	1026
975	Political-economic costs: a 150%+ tariff	favor targeted tools or security-based	1027
976	invites retaliation, supply-chain	restrictions over shock 150%+ tariff	1028
977	disruption, and inflationary	hikes. Absent a major geopolitical	1029
978	risks; leaders may prefer targeted	trigger or imminent trade case known	1030
979	bans/controls instead. Strength:	to produce such margins, the most	1031
980	High - Existing measures already	likely outcome by end-June is no new	1032
981	high: where protection is desired	150%+ headline tariff; niche AD/CVD	1033
982	(e.g., EVs), tariffs were already	outcomes are possible but uncertain	1034
983	raised substantially; further quick	within the timeframe.</considerations>	1035
984	escalation to >150% risks overreach.	<initial_probability>0.12	1036
985	Strength: Medium-High - Alternatives	</initial_probability>	1037
986	to tariffs: governments can use export	<extra_considerations>Main risk is	1038
987	controls, investment restrictions,	definitional: if the market resolves	1039
988	or outright bans (national security)	based on any AD/CVD rate exceeding 150%	1040
989	rather than extreme tariff rates.	on any Chinese product, probability	1041
990	Strength: Medium - International/legal	should be higher. If it requires a	1042
991	constraints: WTO norms and allied	broad, high-profile tariff (e.g., EVs),	1043
992	coordination temper extreme moves	0.12 may be generous. The short window	1044
993	absent acute crisis. Strength:	further reduces odds. My estimate leans	1045
994	Medium</no_thoughts><yes_thoughts>-	conservative; uncertainty remains about	1046
995	Product-specific trade remedies:	ongoing trade cases not captured here,	1047
996	anti-dumping/countervailing duty cases	so confidence in NO is moderate-high,	1048
997	can yield provisional rates well above	not absolute.</extra_considerations>	1049
998	150% on specific Chinese products.	<ans>NO</ans><conf>0.85</conf>",	1050
999	Strength: Medium - Geopolitical shock:	"response_2": "<rephrased_question>By	1051
1000	a sharp escalation (e.g., security	June 30, 2025, will the United	1052
1001	incident) could trigger emergency	States' headline tariff rate applied	1053
1002	national-security tariffs or bans	to merchandise imports from China be	1054
1003	with very high rates. Strength:	increased to greater than 150% (i.e., at	1055
1004	Low-Medium - Domestic industry pressure:	least 151)? Current context suggests	1056

1057	recent hikes variously reported as	administration may prefer enforcement	1109
1058	54%, 125% "reciprocal," and up to	and implementation of April measures	1110
1059	145% per some sources, including	before escalating again. A move	1111
1060	changes targeting de minimis shipments	exactly to 150% (not >150%) is more	1112
1061	(e.g., Temu/Shein). Assess whether a	politically "clean," but the question	1113
1062	further hike above 150% will be enacted	requires greater than 150%, which	1114
1063	by end-June.</rephrased_question>	is less likely unless jumping to a	1115
1064	<no_thoughts>- Political/economic	higher round (e.g., 200%), which seems	1116
1065	blowback: additional hikes above 150%	aggressive by June.</considerations>	1117
1066	could worsen inflation and consumer	<initial_probability>0.42	1118
1067	prices; strong business and retail	</initial_probability>	1119
1068	opposition likely to slow moves.	<extra_considerations>Ambiguity over	1120
1069	Strength: High - Legal/administrative	which tariff measure is referenced	1121
1070	frictions: further increases may	could sway the outcome: certain	1122
1071	require additional proclamations,	product-specific or penalty rates	1123
1072	inter-agency review, and risk	may already exceed 150%, but the	1124
1073	injunctions or litigation; timing	market narrative tends to cite a	1125
1074	to end-June is tight. Strength: Medium	single headline rate. Court challenges	1126
1075	- Recent large increases suggest	or preliminary injunctions could	1127
1076	a pause: after big April moves (to	delay any new proclamation. If	1128
1077	125-145%), the administration may assess	economic data in May-June show	1129
1078	impacts before another escalation.	easing inflation, political space for	1130
1079	Strength: Medium - Diplomatic/strategic	escalation could widen; conversely,	1131
1080	pacing: keeping headroom for future	any price spikes or supply-chain	1132
1081	leverage rather than immediate	disruptions could constrain action.	1133
1082	escalation beyond 150%. Strength:	Given these, 0.42 may be slightly	1134
1083	Low-Medium</no_thoughts><yes_thoughts>-	conservative but avoids overconfidence	1135
1084	Policy posture: the administration's	amid definitional and legal	1136
1085	hawkish stance and willingness to	uncertainties.</extra_considerations>	1137
1086	move rapidly in April imply readiness	<ans>YES</ans> <conf>0.65</conf>"}	1138
1087	for further hikes soon. Strength:		
1088	Medium-High - Executive authority: tools	<b>B.3 Definition Drift: "Will MATS</b>	1139
1089	like Section 301/IEEPA/"reciprocal	<b>applications open in March?"</b>	1140
1090	tariffs" allow swift adjustments	{"question": "Will MATS applications	1141
1091	without Congress, enabling a quick	open in March?", "true_resolution":	1142
1092	bump above 150%. Strength: Medium	"YES", "pred_1": "YES", "pred_2":	1143
1093	- Targeting de minimis/fast-fashion	"NO", "correct_1": true, "correct_2":	1144
1094	imports: further tightening to curb	false, "conf_1": 0.58, "conf_2": 0.35,	1145
1095	Temu/Shein could include pushing rates	"response_1": "<rephrased_question>Are	1146
1096	above 150%. Strength: Low-Medium -	the upcoming MATS program applications	1147
1097	Signaling/politics: a symbolic breach	expected to open in March (as in	1148
1098	above 150% could be touted as toughness	prior cycles), or has the application	1149
1099	on China. Strength: Low</yes_thoughts>	timeline shifted?</rephrased_question>	1150
1100	<considerations>Conflicting reports	<no_thoughts> - The program could	1151
1101	(54%, 125%, 145%) imply definitional	shift its calendar due to funding	1152
1102	ambiguity; assuming the "headline"	or operational changes; March might	1153
1103	broad tariff. Moving from 145% to	move to Feb/Apr (strength: medium). -	1154
1104	>150% within 80 days would be another	They might adopt rolling admissions	1155
1105	significant step with inflation optics	instead of a fixed March opening	1156
1106	and legal risks. While the White House	(strength: medium). - Ambiguity about	1157
1107	can act quickly, successive hikes risk	which "MATS" is meant; some programs	1158
1108	backlash and court challenges. The	with that acronym do not open in	1159
		March (strength: strong). - External	1160
		constraints (venues, visas, partner	1161
		timelines) could push the window later	1162
		(strength: weak-medium). </no_thoughts>	1163
		<yes_thoughts> - Historical pattern:	1164
		recent MATS summer cohorts have	1165
		typically opened applications around	1166

1167	March (strength: strong). - March	months ahead; March openings are	1237
1168	aligns well with a summer-start	atypical because the event is in March.	1238
1169	pipeline, making it a stable choice	Evidence: 2025 reporting places MATS	1239
1170	(strength: medium). - Absent	activity around late March/early April;	1240
1171	contrary announcements, organizations	consistent with prior years, reducing	1241
1172	tend to keep predictable cycles	likelihood of March opening. Ambiguity:	1242
1173	(strength: medium). </yes_thoughts>	acronym uncertainty increases variance,	1243
1174	<considerations>Base rates from prior	but the most common "MATS" is the	1244
1175	cycles favor March for the summer	trucking show. Tail scenarios:	1245
1176	round, but acronym/program ambiguity	late/last-minute windows exist but	1246
1177	and potential operational changes	are not the primary "applications	1247
1178	reduce confidence. With no specific	opening" phase.</considerations>	1248
1179	year given and no fresh signals,		
1180	a modestly favorable probability	<initial_probability>0.30	1249
1181	is warranted.</considerations>	</initial_probability>	1250
1182	<initial_probability>0.6	<extra_considerations>Ambiguity could	1251
1183	</initial_probability>	mean the user is asking about a	1252
1184	<extra_considerations>Ambiguity	different MATS entirely, which modestly	1253
1185	likely makes the estimate slightly	raises the chance of a March opening.	1254
1186	overconfident. Nudging down to	If referring specifically to on-site	1255
1187	reflect possible reference to a	registration or late credentialing,	1256
1188	different "MATS" or a one-off	"opening in March" could be technically	1257
1189	schedule change. Final probability:	true, though not the main phase. To	1258
1190	0.58.</extra_considerations>	avoid overconfidence in an ambiguous	1259
1191	<ans>YES</ans> <conf>0.58</conf>",	acronym, slight upward adjustment	1260
1192	"response_2": "<rephrased_question>	from 0.30 to 0.35 for YES seems	1261
1193	Clarify whether the "MATS" in	appropriate.</extra_considerations>	1262
1194	question refers to the Mid-America	<ans>NO</ans> <conf>0.35</conf>"}	1263
1195	Trucking Show and, if so, whether		
1196	its application/registration window		
1197	(for exhibitors or attendees) is		
1198	scheduled to open in March of the		
1199	relevant year. If "MATS" refers to a		
1200	different program with the same acronym,		
1201	is its application window expected		
1202	to open in March?</rephrased_question>		
1203	<no_thoughts> - The Mid-America Trucking		
1204	Show (MATS) event typically occurs in		
1205	late March, so exhibitor applications		
1206	and attendee registration usually open		
1207	months earlier (Strong). - A recent		
1208	2025 reference shows MATS activities		
1209	happening around late March/early April,		
1210	implying applications would have opened		
1211	before March (Moderate-Strong). -		
1212	"Applications" may not be the right		
1213	term for MATS attendee registration;		
1214	exhibitor applications and media		
1215	credentials typically open well in		
1216	advance (Moderate). - If the user meant		
1217	AIMA MAT (not MATS), those application		
1218	cycles for February/March sessions		
1219	tend to open in January/February, not		
1220	start in March (Weak-Moderate due		
1221	to acronym mismatch). </no_thoughts>		
1222	<yes_thoughts> - Some organizations		
1223	open final or late application windows		
1224	in March for events occurring late		
1225	March/early April; MATS could allow		
1226	late registrations or on-site sign-ups		
1227	(Weak-Moderate). - If "MATS" refers		
1228	to a different program with that		
1229	acronym, its cycle could plausibly		
1230	begin in March (Weak). - Exceptional		
1231	scheduling changes (e.g., policy shifts,		
1232	one-off timelines) could push openings		
1233	into March in a given year (Weak).		
1234	</yes_thoughts> <considerations>Base		
1235	rate: large trade shows like MATS		
1236	open exhibitor/attendee processes		
		<b>B.4 Missing Reasoning Traces in</b>	1264
		<b>DeepSeek-R1</b>	1265
		{ "question": "NFL Week 17 (Dec 28	1266
		2024): Will the Washington Commanders	1267
		(home) beat the Atlanta Falcons	1268
		(away)?", "true_resolution": "YES",	1269
		"pred_1": "YES", "pred_2": "YES",	1270
		"correct_1": true, "correct_2":	1271
		true, "conf_1": 0.7, "conf_2":	1272
		0.7, "response_1": "<ans>YES</ans>	1273
		<conf>0.7</conf>", "response_2":	1274
		"<ans>YES</ans> <conf>0.7</conf>"}	1275