Do Large Language Models Know What They Don't Know? Evaluating Epistemic Calibration via Prediction Markets

Lukas Nel
Lotus AI
lukas@lotus.ai

Abstract

A well-calibrated model should express confidence that matches its actual accuracy—when it claims 80% confidence, it should be correct 80% of the time. While large language models (LLMs) have achieved remarkable performance across diverse tasks, their epistemic calibration remains poorly understood. We introduce **KalshiBench**, a benchmark of 300 prediction market questions from Kalshi, a CFTC-regulated exchange, with verifiable real-world outcomes occurring after model training cutoffs. Unlike traditional benchmarks measuring accuracy on static knowledge, KalshiBench evaluates whether models can appropriately quantify uncertainty about genuinely unknown future events. We evaluate five frontier models—Claude Opus 4.5, GPT-5.2, DeepSeek-V3.2, Qwen3-235B, and Kimi-K2—and find **systematic overconfidence across all models**. Even the best-calibrated model (Claude Opus 4.5, ECE=0.120) shows substantial calibration errors, while reasoning-enhanced models like GPT-5.2-XHigh exhibit *worse* calibration (ECE=0.395) despite comparable accuracy. Critically, only one model achieves a positive Brier Skill Score, indicating most models perform worse than simply predicting base rates. Our findings suggest that scaling and enhanced reasoning do not automatically confer calibration benefits, highlighting epistemic calibration as a distinct capability requiring targeted development.

1 Introduction

The deployment of large language models in high-stakes domains—medical diagnosis, legal reasoning, financial forecasting—demands not only accuracy but also *calibrated uncertainty*. A model claiming "90% confidence" in a diagnosis should be correct approximately 90% of the time on similar cases. Poor calibration manifests as dangerous overconfidence (trusting wrong answers) or unnecessary underconfidence (ignoring correct ones), fundamentally limiting the utility of model predictions for decision-making under uncertainty [Guo et al., 2017].

Despite extensive work on LLM capabilities, epistemic calibration—the alignment between expressed confidence and actual accuracy—remains understudied. Existing evaluations face two critical limitations:

(1) Static knowledge contamination. Traditional benchmarks assess models on questions whose answers existed during training. A model may appear "calibrated" simply by having memorized facts with appropriate confidence, rather than genuinely reasoning about uncertainty.

(2) Lack of verifiable ground truth. Many calibration studies rely on human judgments or synthetic datasets, introducing noise and potential biases in ground truth labels.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Model	Accuracy	ECE ↓
Claude Opus 4.5	69.3%	**0.120**
Kimi-K2	67.1%	0.298
Qwen3-235B	65.7%	0.297
GPT-5.2-XHigh	65.3%	0.395
DeepSeek-V3.2	64.3%	0.284

Key Finding: All models exhibit systematic overconfidence. The gap between confidence and accuracy widens dramatically at high confidence levels, with models averaging 27% error rate even when expressing >90% confidence.

Figure 1: Summary of main results. While accuracy varies modestly (64-69%), calibration error varies dramatically (3× range). Reasoning enhancements (GPT-5.2-XHigh) worsen rather than improve calibration.

We address both limitations through **KalshiBench**, a benchmark leveraging prediction markets—specifically Kalshi, a CFTC-regulated exchange where contracts resolve to verifiable real-world outcomes. By temporally filtering questions to those resolving *after* model training cutoffs, we ensure models cannot have memorized outcomes, providing a clean signal for epistemic calibration.

Our contributions are:

1. **KalshiBench**: A temporally-filtered benchmark of 300 prediction market questions spanning 13 categories with verified ground truth outcomes, designed for rigorous calibration evaluation.
2. **Comprehensive evaluation**: We assess five frontier models across classification (accuracy, F1) and calibration (Brier score, ECE, reliability diagrams) metrics, revealing systematic patterns.
3. **Novel findings**: We demonstrate that (a) all current frontier models are overconfident, (b) reasoning enhancements degrade calibration, (c) only one model beats the base-rate baseline, and (d) calibration and accuracy are largely decoupled.

2 Related Work

Calibration in Neural Networks. Calibration has been extensively studied in classification settings [Guo et al., 2017, Minderer et al., 2021]. Modern deep networks are known to be overconfident [Guo et al., 2017], with various post-hoc calibration methods proposed including temperature scaling [Guo et al., 2017], Platt scaling [Platt, 1999], and isotonic regression [Zadrozny & Elkan, 2002]. However, these methods assume access to held-out calibration data and primarily address discriminative rather than generative models.

LLM Uncertainty Quantification. Prior work on LLM calibration has examined confidence elicitation through verbalized probabilities [Tian et al., 2023, Xiong et al., 2024], multiple sampling [Wang et al., 2023], and logit-based approaches [Kadavath et al., 2022]. Kadavath et al. [2022] found that larger models show improved calibration on factual questions, while Tian et al. [2023] demonstrated that verbalized confidence often diverges from token probabilities. Recent work has explored calibration in specific domains including medical question-answering [Singhal et al., 2023] and mathematical reasoning [Lightman et al., 2023].

Forecasting and Prediction Markets. Prediction markets aggregate collective intelligence to forecast uncertain events [Arrow et al., 2008, Wolfers & Zitzewitz, 2004]. Superforecasters demonstrate that calibration is a learnable skill [Tetlock & Gardner, 2015]. Recent work has begun exploring LLMs as forecasters [Zou et al., 2022, Halawi et al., 2024]. Most relevant to our work, ForecastBench [Karger et al., 2024] introduced a dynamic benchmark evaluating ML forecasting on 1,000 automatically-updated questions, finding that expert human forecasters significantly outperform the best LLMs. However, ForecastBench focuses primarily on accuracy rather than calibration metrics.

Distinction from Prior Work. Unlike existing benchmarks that assess calibration on static knowledge questions, KalshiBench uses temporally-filtered prediction market questions with verified post-training outcomes, eliminating knowledge contamination and providing clean calibration signals. Compared to ForecastBench, we focus specifically on *calibration* rather than raw forecasting accuracy,

providing detailed analysis of reliability diagrams, overconfidence rates, and the relationship between confidence and correctness.

3 KalshiBench Dataset

3.1 Data Source and Collection

KalshiBench sources questions from Kalshi[1], a CFTC-regulated prediction market exchange operating in the United States. Unlike informal forecasting platforms, Kalshi contracts have legally-binding resolution criteria, ensuring unambiguous ground truth. The full KalshiBench dataset contains **1,531 cleaned, deduplicated prediction market questions** spanning from September 2021 to November 2025 across 16 categories, with a 42%/58% yes/no class split.

For our evaluation, we apply temporal filtering based on model knowledge cutoffs and randomly sample **300 questions** (random seed 42) from the filtered set. This sample size balances computational cost against statistical power, and exceeds the 200-question evaluation used in ForecastBench [Karger et al., 2024].

3.2 Temporal Filtering

To ensure models cannot have memorized outcomes, we apply strict temporal filtering based on model knowledge cutoffs:

$$\mathcal{D}_{\text{filtered}} = \{(q, y) \in \mathcal{D} : t_{\text{close}}(q) > \max_{m \in \mathcal{M}} t_{\text{cutoff}}(m)\} \tag{1}$$

where $t_{\text{close}}(q)$ is the resolution time of question q, $t_{\text{cutoff}}(m)$ is the knowledge cutoff of model m, and $\mathcal{M}$ is the set of evaluated models. For our evaluation, the effective cutoff is October 1, 2025 (the latest among all models).

3.3 Dataset Statistics

Table 1: KalshiBench dataset statistics. The full dataset contains 1,531 questions; we evaluate on a temporally-filtered sample of 300 questions (seed=42) resolving after October 1, 2025.

Full Dataset	**Value**	**Evaluation Sample**	**Value**
Total Questions	1,531	Sampled Questions	300
Categories	16	Categories (in sample)	13
Date Range	2021-09 to 2025-11	Date Range	2025-10 to 2025-11
Yes Rate	42.3%	Yes Rate	40.0%
Temporal Span	1,537 days	Temporal Span	46 days

Table 2: Category distribution in KalshiBench. Sports and Politics dominate, but all major forecasting domains are represented.

Category	**N**	**%**	**Yes%**	**Category**	**N**	**%**	**Yes%**
Sports	83	27.7	34.9	Crypto	11	3.7	27.3
Politics	55	18.3	52.7	Climate/Weather	9	3.0	33.3
Entertainment	47	15.7	36.2	Financials	8	2.7	12.5
Companies	30	10.0	60.0	World	6	2.0	50.0
Elections	24	8.0	20.8	Economics	4	1.3	50.0
Mentions	19	6.3	36.8	Social	3	1.0	66.7

[1] `https://kalshi.com`

3.4 Deduplication and Quality Control

Raw prediction market data contains redundant questions (e.g., daily instances of recurring markets). We limit to 2 questions per series ticker to preserve diversity while reducing redundancy. All questions include detailed resolution criteria in the description field, ensuring unambiguous ground truth.

4 Methodology

4.1 Models Evaluated

We evaluate five frontier models representing diverse architectures and training approaches:

Table 3: Models evaluated in KalshiBench. All models have knowledge cutoffs at or before October 2025.

Model	Provider	Knowledge Cutoff	Notes
Claude Opus 4.5	Anthropic	April 2025	Flagship model
GPT-5.2-XHigh	OpenAI	October 2025	Extended reasoning
DeepSeek-V3.2	DeepSeek	October 2025	Open-weight
Qwen3-235B-Thinking	Alibaba	June 2025	Reasoning-enhanced
Kimi-K2	Moonshot	June 2025	Reasoning-enhanced

4.2 Evaluation Protocol

Each model receives a structured prompt containing the prediction market question and resolution criteria. The system prompt explicitly instructs models to be calibrated:

```
System: You are an expert forecaster evaluating prediction
market questions. Given a question and its description,
predict whether the outcome will be "yes" or "no".

You must respond in this exact format:
<think>
[Your reasoning about the prediction, considering base
rates, relevant factors, and uncertainty]
</think>
<answer>[yes or no]</answer>
<confidence>[a number from 0 to 100 representing your
confidence that the answer is "yes"]</confidence>

Be calibrated: if you're 70% confident, you should be
correct about 70% of the time on similar questions.
```

The user message then provides the specific question and description. Notably, the prompt explicitly instructs models to "be calibrated," making the observed miscalibration a failure to follow instructions rather than mere absence of guidance.

We use temperature 0.7 for standard models and temperature 1.0 with extended reasoning for GPT-5.2-XHigh, following provider recommendations.

4.3 Metrics

Classification Metrics. We report accuracy, precision, recall, and macro-F1 for binary classification performance.

Brier Score. The Brier score [Brier, 1950] measures the mean squared error of probability predictions:

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{2}$$

where p_i is the predicted probability and $y_i \in \{0, 1\}$ is the outcome. Lower is better (0 = perfect, 1 = worst possible).

Intuition: The Brier score can be interpreted as follows:

- **0.00**: Perfect predictions—100% confidence on all correct answers
- **0.25**: Random guessing (50% confidence on everything)—the expected score of a completely uninformed predictor on balanced binary outcomes
- **0.20**: Good calibration—roughly equivalent to human forecasters on prediction markets
- **0.33**: Poor calibration—equivalent to always predicting 42% (the base rate) with uniform 75% confidence
- **1.00**: Maximally wrong—100% confidence on all incorrect answers

For context, human superforecasters typically achieve Brier scores of 0.15–0.20 [Tetlock & Gardner, 2015], while the aggregate "wisdom of crowds" on prediction markets often achieves 0.12–0.18.

Brier Skill Score. The Brier Skill Score (BSS) measures improvement over a baseline that always predicts the base rate:

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{climatology}}} \tag{3}$$

where $\text{BS}_{\text{climatology}} = \bar{y}(1 - \bar{y})$ for base rate $\bar{y}$. Positive values indicate improvement over the base rate.

Expected Calibration Error (ECE). ECE [Naeini et al., 2015] measures the average gap between confidence and accuracy:

$$\text{ECE} = \sum_{b=1}^{B}\frac{|B_b|}{N}\,|\text{acc}(B_b) - \text{conf}(B_b)| \tag{4}$$

where predictions are binned by confidence into B bins.

Maximum Calibration Error (MCE). MCE captures the worst-case calibration in any single bin:

$$\text{MCE} = \max_{b\in\{1,\ldots,B\}}|\text{acc}(B_b) - \text{conf}(B_b)| \tag{5}$$

Overconfidence Rate. We define overconfidence rate at threshold τ as the fraction of incorrect predictions among those with confidence $> \tau$:

$$\text{OCR@}\tau = \frac{|\{i : p_i > \tau \wedge \hat{y}_i \neq y_i\}|}{|\{i : p_i > \tau\}|} \tag{6}$$

5 Results

5.1 Main Results

Table 4 presents comprehensive results across all models and metrics.

Key Finding 1: Systematic Overconfidence. All models exhibit substantial calibration errors, with ECE ranging from 0.120 to 0.395. Even the best-calibrated model (Claude Opus 4.5) shows a 12-percentage-point average gap between confidence and accuracy.

Table 4: Main results on KalshiBench (300 questions). Best values in **bold**. Claude Opus 4.5 achieves best performance on both accuracy and calibration metrics. Notably, the reasoning-enhanced GPT-5.2-XHigh shows the worst calibration despite comparable accuracy.

	Classification			Calibration			
Model	Acc	F1	$F1_{yes}$	Brier ↓	BSS ↑	ECE ↓	MCE ↓
Claude Opus 4.5	**69.3**	**0.676**	**0.600**	**0.227**	**0.057**	**0.120**	**0.246**
Kimi-K2	67.1	0.633	0.515	0.347	-0.446	0.298	0.570
Qwen3-235B	65.7	0.607	0.466	0.346	-0.437	0.297	0.479
GPT-5.2-XHigh	65.3	0.599	0.453	0.433	-0.799	0.395	0.622
DeepSeek-V3.2	64.3	0.614	0.507	0.339	-0.407	0.284	0.630

Table 5: Confidence analysis across models. All models show higher confidence when wrong than would be appropriate for well-calibrated predictions. Overconfidence rates at high confidence levels (80%+, 90%+) are alarmingly high.

Model	**Avg Conf**	$\mathbf{Conf_{wrong}}$	**OCR@70**	**OCR@80**	**OCR@90**
Claude Opus 4.5	73.8%	71.0%	27.1%	23.1%	**20.8%**
DeepSeek-V3.2	73.7%	69.2%	**24.7%**	**23.6%**	**14.7%**
Kimi-K2	79.4%	76.3%	25.9%	29.9%	31.1%
GPT-5.2-XHigh	80.1%	76.9%	30.3%	28.3%	27.7%
Qwen3-235B	81.7%	80.4%	32.3%	32.6%	32.4%

Key Finding 2: Most Models Fail to Beat the Base Rate. Only Claude Opus 4.5 achieves a positive Brier Skill Score (0.057), indicating it marginally outperforms simply predicting the 40% base rate. All other models have negative BSS, meaning their probability estimates are *worse than uninformed guessing*.

Key Finding 3: Reasoning Enhancements Hurt Calibration. Counterintuitively, GPT-5.2-XHigh (with extended reasoning) shows the worst calibration (ECE=0.395, BSS=-0.799) despite using 26× more output tokens (∼2M vs ∼138K for Claude). Enhanced reasoning appears to increase confidence without proportional accuracy gains.

5.2 Confidence Analysis

Table 5 reveals troubling patterns in model confidence:

- **High baseline confidence**: Models average 74-82% confidence, far exceeding the 65-69% accuracy range.
- **Confidence when wrong**: Models maintain 69-80% confidence even on incorrect predictions, indicating poor uncertainty awareness.
- **Extreme overconfidence**: At the 90%+ confidence level, models are wrong 15-32% of the time. A well-calibrated model should be wrong <10%.

5.3 Reliability Diagrams

A reliability diagram plots predicted confidence against actual accuracy across binned predictions. A perfectly calibrated model follows the diagonal: when it expresses 70% confidence, it should be correct 70% of the time. Table 6 presents complete reliability data for all five models across all 10 confidence bins.

Several patterns emerge from the reliability analysis:

Claude Opus 4.5 shows the best calibration overall, with relatively small gaps in most bins. However, even Claude becomes overconfident at high confidence levels: at 90%+ confidence (20 predictions), accuracy is only 70%, yielding a +24.6% gap.

Table 6: Complete reliability diagram data for all models (10 bins, 0.1 width). **Conf** = average confidence in bin, **Acc** = accuracy, **N** = count, **Gap** = Conf − Acc (positive = overconfident). Empty cells indicate no predictions in that bin. All models become increasingly overconfident at higher confidence levels.

	Claude Opus 4.5				**DeepSeek-V3.2**				**GPT-5.2-XHigh**			
Bin	Conf	Acc	N	Gap	Conf	Acc	N	Gap	Conf	Acc	N	Gap
0.0-0.1	.054	.194	36	-.14	.048	.200	10	-.15	.030	.000	1	+.03
0.1-0.2	.151	.188	32	-.04	.175	.250	8	-.08	—	—	0	—
0.2-0.3	.248	.333	42	-.09	.250	.000	4	+.25	—	—	0	—
0.3-0.4	.359	.355	31	+.00	.344	.333	9	+.01	—	—	0	—
0.4-0.5	.439	.333	36	+.11	.418	.545	11	-.13	—	—	0	—
0.5-0.6	.566	.353	34	+.21	.575	.365	63	+.21	.573	.429	42	+.14
0.6-0.7	.641	.724	29	-.08	.673	.463	67	+.21	.661	.480	50	+.18
0.7-0.8	.751	.542	24	+.21	.747	.400	30	+.35	.751	.488	41	+.26
0.8-0.9	.854	.688	16	+.17	.831	.517	58	+.31	.835	.387	62	+.45
0.9-1.0	.946	.700	20	**+.25**	.937	.308	39	**+.63**	.959	.337	104	**+.62**

	Qwen3-235B-Thinking				**Kimi-K2**			
Bin	Conf	Acc	N	Gap	Conf	Acc	N	Gap
0.0-0.1	.039	.356	73	-.32	.047	.263	38	-.22
0.1-0.2	.153	.316	19	-.16	.141	.312	16	-.17
0.2-0.3	.262	.400	5	-.14	.249	.111	9	+.14
0.3-0.4	.341	.357	14	-.02	.314	.600	5	-.29
0.4-0.5	.442	.500	6	-.06	.465	.000	2	+.47
0.5-0.6	.556	.455	22	+.10	.570	.477	44	+.09
0.6-0.7	.664	.439	41	+.23	.668	.458	48	+.21
0.7-0.8	.756	.310	29	+.45	.750	.484	31	+.27
0.8-0.9	.846	.469	49	+.38	.849	.447	38	+.40
0.9-1.0	.941	.462	39	**+.48**	.948	.377	61	**+.57**

GPT-5.2-XHigh exhibits the most severe miscalibration. The model rarely expresses low confidence (only 1 prediction below 50%), concentrating 104 predictions (35% of total) in the 90-100% bin where accuracy is merely 33.7%—worse than chance. This represents a catastrophic +62.2% calibration gap.

DeepSeek-V3.2 shows a similar pattern to GPT-5.2, with a +63.0% gap in the highest confidence bin. When DeepSeek expresses 90%+ confidence, it is correct only 30.8% of the time.

Reasoning Models (Qwen3, Kimi-K2) both show substantial overconfidence at high confidence levels (+47.9% and +57.1% gaps respectively), despite their "thinking" architectures. Extended reasoning does not translate to better uncertainty awareness.

Summary: High-Confidence Performance. Table 7 summarizes performance in the critical 90-100% confidence bin, where models claim near-certainty:

Table 7: Performance in the 90-100% confidence bin. A well-calibrated model should achieve ∼95% accuracy when expressing 95% average confidence. All models fall catastrophically short.

Model	**Avg Conf**	**Actual Acc**	**Gap**	**N**
Claude Opus 4.5	94.6%	70.0%	+24.6%	20
DeepSeek-V3.2	93.7%	30.8%	+62.9%	39
GPT-5.2-XHigh	95.9%	33.7%	+62.2%	104
Qwen3-235B	94.1%	46.2%	+47.9%	39
Kimi-K2	94.8%	37.7%	+57.1%	61

5.4 Category Breakdown

Category analysis reveals domain-dependent performance. Models perform well on Entertainment, Sports, and Elections—domains with substantial training data—but struggle with Crypto and Science/Technology, suggesting calibration degrades in domains with higher inherent uncertainty or less training exposure.

Table 8: Performance by category for Claude Opus 4.5 (best overall). Performance varies substantially across domains, with Social (100% accuracy) and Entertainment (78.7%) being strongest, while Science (0% on 1 question) and Crypto (36.4%) are weakest.

Category	Acc	Brier	Category	Acc	Brier
Social (n=3)	100.0%	0.011	Crypto (n=11)	36.4%	0.240
Entertainment (n=47)	78.7%	0.187	Mentions (n=19)	52.6%	0.357
Climate (n=9)	77.8%	0.229	World (n=6)	50.0%	0.262
Sports (n=83)	75.9%	0.193	Economics (n=4)	50.0%	0.326
Elections (n=24)	75.0%	0.172	Sci/Tech (n=1)	0.0%	0.608
Financials (n=8)	75.0%	0.203			

5.5 Cost-Performance Analysis

Table 9: Cost-performance tradeoffs. More expensive models are not necessarily better calibrated. GPT-5.2-XHigh costs 2.6× more than Claude but shows 3× worse calibration.

Model	Cost (USD)	Tokens	Acc	ECE
DeepSeek-V3.2	$0.36	304K	64.3%	0.284
Kimi-K2	$0.94	624K	67.1%	0.298
Qwen3-235B	$1.19	594K	65.7%	0.297
Claude Opus 4.5	$11.63	224K	69.3%	0.120
GPT-5.2-XHigh	$30.32	2.07M	65.3%	0.395

Cost does not predict calibration quality. DeepSeek-V3.2 achieves comparable accuracy to GPT-5.2-XHigh at 1/84th the cost with substantially better calibration. This suggests calibration improvements require architectural or training innovations rather than simply more compute.

6 Analysis and Discussion

6.1 Why Are Models Overconfident?

We hypothesize several contributing factors. Notably, our prompt explicitly instructs models to "be calibrated: if you're 70% confident, you should be correct about 70% of the time on similar questions." Despite this direct instruction, all models exhibit substantial miscalibration, suggesting the problem runs deeper than prompt engineering.

Training Objective Misalignment. Standard language modeling objectives reward correct predictions without penalizing miscalibrated confidence. Models learn to maximize probability of correct tokens, not to appropriately quantify uncertainty.

RLHF Pressure for Confidence. Human feedback in RLHF may inadvertently reward confident-sounding responses over appropriately hedged ones. Users may rate uncertain responses as less helpful, creating pressure toward overconfidence.

Hindsight Leakage. Even with temporal filtering, models may have indirect signals about future events through patterns learned during training (e.g., seasonal trends, recurring events). This could inflate confidence without improving accuracy.

6.2 Why Does Reasoning Hurt Calibration?

The finding that GPT-5.2-XHigh shows worse calibration than simpler models is counterintuitive but may reflect:

Confirmation Bias in Extended Reasoning. Longer reasoning chains may reinforce initial hypotheses rather than genuinely updating on evidence. The model generates arguments supporting its prediction, increasing confidence without corresponding accuracy gains.

Verbosity Without Epistemic Humility. Extended reasoning produces more text but not necessarily better uncertainty quantification. The model may be optimized for persuasive reasoning rather than calibrated forecasting.

6.3 Implications for Deployment

Our findings have direct implications for LLM deployment:

1. **Don't trust high-confidence predictions.** When models express 90%+ confidence, expect 20-30% error rates, not $<$10%.
2. **More reasoning $\neq$ better calibration.** Extended reasoning modes may actually decrease reliability.
3. **Post-hoc calibration is necessary.** Temperature scaling or Platt scaling should be applied before using model confidences for decision-making.
4. **Domain matters.** Calibration varies substantially by category; validate on domain-specific data.

6.4 Comparison to Human Forecasters

For context, human superforecasters typically achieve Brier scores of 0.15-0.20 on similar prediction market questions [Tetlock & Gardner, 2015]. Claude Opus 4.5's Brier score of 0.227 is approaching but not matching expert human performance. Critically, superforecasters exhibit much better calibration (ECE $\approx$ 0.03-0.05), suggesting LLMs have particular deficits in uncertainty quantification rather than raw forecasting ability.

7 Limitations

Dataset Scope. Our evaluation uses 300 questions sampled from the full 1,531-question KalshiBench dataset. While this exceeds the 200-question evaluation used in ForecastBench [Karger et al., 2024], category-level analysis (especially for rare categories) has high variance. Some categories contain only 1-4 questions in our sample.

Temporal Constraints. Temporal filtering ensures validity but limits dataset size. Questions must resolve after all model cutoffs, reducing the available pool substantially.

Binary Outcomes Only. We evaluate only yes/no markets. Multi-outcome prediction markets and continuous forecasts present different calibration challenges not addressed here.

Prompt Sensitivity. Model calibration may be sensitive to prompt wording. We use a standardized prompt but do not exhaustively explore prompt variations.

Confidence Elicitation. Self-reported confidence (0-100) may not reflect internal probability estimates. Alternative elicitation methods (betting, proper scoring rule incentives) might yield different results.

8 Conclusion

We introduced KalshiBench, a benchmark for evaluating LLM epistemic calibration using temporally-filtered prediction market questions with verified real-world outcomes. Our evaluation of five frontier models reveals:

- **Universal overconfidence**: All models show substantial calibration errors (ECE 0.12-0.40).
- **Base-rate failures**: Only one model achieves positive Brier Skill Score.
- **Reasoning paradox**: Extended reasoning worsens rather than improves calibration.
- **Calibration-accuracy decoupling**: Models with similar accuracy show 3$\times$ variation in calibration.

These findings highlight epistemic calibration as a distinct capability—separate from accuracy—that current training approaches fail to adequately develop. Future work should explore calibration-aware training objectives, explicit uncertainty modeling architectures, and integration with human forecasting expertise.

Broader Impact. Improved LLM calibration is essential for safe deployment in high-stakes domains. Our work provides tools and baselines for measuring progress. Conversely, publication of calibration failures could be misused to manipulate users who overweight model confidence; we encourage deployment of properly calibrated systems.

Reproducibility. The full KalshiBench dataset (1,531 questions) is available at `https://huggingface.co/datasets/2084Collective/kalshibench-v2`. Our evaluation uses a 300-question sample with random seed 42. Code and evaluation scripts are open-sourced at `https://github.com/2084collective/kalshibench`.

References

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., Nelson, F. D., et al. The promise of prediction markets. *Science*, 320(5878):877–878, 2008.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330. PMLR, 2017.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. ForecastBench: A dynamic benchmark of AI forecasting capabilities. *arXiv preprint arXiv:2409.19839*, 2024.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 15682–15694, 2021.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Tetlock, P. E. and Gardner, D. *Superforecasting: The Art and Science of Prediction*. Crown Publishers, New York, 2015.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5433–5442. Association for Computational Linguistics, 2023.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

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

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., and Hendrycks, D. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A Extended Results

A.1 Full Confusion Matrices

Table 10: Confusion matrices for all models. TP=True Positive, FP=False Positive, FN=False Negative, TN=True Negative.

Model	TP	FP	FN	TN
Claude Opus 4.5	69	40	52	139
GPT-5.2-XHigh	43	26	78	153
DeepSeek-V3.2	55	41	66	138
Qwen3-235B	45	27	76	152
Kimi-K2	51	30	66	145

A.2 Full Reliability Diagram Data

Table 11 provides complete reliability diagram statistics including average confidence, accuracy, sample count, and calibration gap for each bin and model.

Table 11: Extended reliability diagram data showing average confidence within each bin.

	Claude Opus 4.5				**DeepSeek-V3.2**			
Bin	Conf	Acc	N	Gap	Conf	Acc	N	Gap
0.0-0.1	0.054	0.194	36	-0.141	0.048	0.200	10	-0.152
0.1-0.2	0.151	0.188	32	-0.037	0.175	0.250	8	-0.075
0.2-0.3	0.248	0.333	42	-0.085	0.250	0.000	4	+0.250
0.3-0.4	0.359	0.355	31	+0.004	0.344	0.333	9	+0.011
0.4-0.5	0.439	0.333	36	+0.106	0.418	0.545	11	-0.127
0.5-0.6	0.566	0.353	34	+0.213	0.575	0.365	63	+0.210
0.6-0.7	0.641	0.724	29	-0.083	0.673	0.463	67	+0.211
0.7-0.8	0.751	0.542	24	+0.210	0.747	0.400	30	+0.347
0.8-0.9	0.854	0.688	16	+0.167	0.831	0.517	58	+0.313
0.9-1.0	0.946	0.700	20	+0.246	0.937	0.308	39	+0.630
	GPT-5.2-XHigh				**Qwen3-235B-Thinking**			
Bin	Conf	Acc	N	Gap	Conf	Acc	N	Gap
0.0-0.1	0.030	0.000	1	+0.030	0.039	0.356	73	-0.317
0.1-0.2	—	—	0	—	0.153	0.316	19	-0.163
0.2-0.3	—	—	0	—	0.262	0.400	5	-0.138
0.3-0.4	—	—	0	—	0.341	0.357	14	-0.016
0.4-0.5	—	—	0	—	0.442	0.500	6	-0.058
0.5-0.6	0.573	0.429	42	+0.144	0.556	0.455	22	+0.101
0.6-0.7	0.661	0.480	50	+0.181	0.664	0.439	41	+0.225
0.7-0.8	0.751	0.488	41	+0.263	0.756	0.310	29	+0.446
0.8-0.9	0.835	0.387	62	+0.448	0.846	0.469	49	+0.376
0.9-1.0	0.959	0.337	104	+0.622	0.941	0.462	39	+0.479
	Kimi-K2							
Bin	Conf	Acc	N	Gap				
0.0-0.1	0.047	0.263	38	-0.216				
0.1-0.2	0.141	0.312	16	-0.172				
0.2-0.3	0.249	0.111	9	+0.138				
0.3-0.4	0.314	0.600	5	-0.286				
0.4-0.5	0.465	0.000	2	+0.465				
0.5-0.6	0.570	0.477	44	+0.093				
0.6-0.7	0.668	0.458	48	+0.210				
0.7-0.8	0.750	0.484	31	+0.266				
0.8-0.9	0.849	0.447	38	+0.402				
0.9-1.0	0.948	0.377	61	+0.570				

B Prompt Template

The exact system prompt used for all model evaluations:

```
SYSTEM PROMPT:
```

```
You are an expert forecaster evaluating prediction market questions.
Given a question and its description, predict whether the outcome
will be "yes" or "no".

You must respond in this exact format:
<think>
[Your reasoning about the prediction, considering base rates,
relevant factors, and uncertainty]
</think>
<answer>[yes or no]</answer>
<confidence>[a number from 0 to 100 representing your confidence
that the answer is "yes"]</confidence>

Be calibrated: if you're 70% confident, you should be correct
about 70% of the time on similar questions.

USER PROMPT:
Question: {question}

Description: {description}
```

The explicit calibration instruction ("Be calibrated: if you're 70% confident, you should be correct about 70% of the time") makes the observed miscalibration particularly notable—models fail to achieve calibration even when directly instructed to do so.

C Dataset Creation Details

The KalshiBench dataset was created through the following pipeline:

1. **Raw data collection**: Query Kalshi API for all resolved binary contracts.
2. **Temporal filtering**: Retain only contracts resolving after October 1, 2025.
3. **Deduplication**: Limit to 2 questions per series_ticker to reduce redundancy while preserving within-series diversity.
4. **Quality filtering**: Remove contracts with ambiguous resolution criteria or missing ground truth.
5. **Schema standardization**: Map to unified schema with fields: id, question, description, category, close_time, ground_truth.

The final dataset contains 300 questions across 13 categories, with category entropy of 3.01 bits (maximum possible: 3.70 bits), indicating reasonable diversity.