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`