Calibration as Capital, Refusal as Information:
Empirical Calibration of Frontier LLMs and Meta-Agents on Bitcoin Forecasting

Chris
Strategy Arena, independent research
contact@strategyarena.io

Dataset snapshot: 2026-05-17

Abstract

Large language models are increasingly deployed as probabilistic forecasters, yet their stated confidences are rarely checked against resolved outcomes. We report a live Bitcoin forecasting experiment covering 10,539 public forecast records and 8,367 verified binary forecasts from 9 providers and meta-agents. Each forecast states a confidence in [0,100] and is evaluated against a resolved market outcome for short-horizon Bitcoin questions. We compute reliability curves, Brier scores, expected calibration error, maximum calibration error, and NEUTRAL/refusal rates. The best calibrated provider in this snapshot is Table Ronde (Brier=0.220), while the worst is Gemini (Brier=0.398). The dominant failure modes are under-confidence, over-confidence, and confidence values that carry little ranking information. The full CSV dataset is bundled with this submission and publicly hosted under CC-BY 4.0. We argue that calibration monitoring is a cheap, auditable infrastructure layer for any multi-agent LLM system deployed on domains where ground truth resolves quickly.

Keywords: LLM evaluation; probabilistic forecasting; calibration; Brier score; reliability diagrams; agentic systems; financial time series; open data.

1 Introduction

Probability calibration is the property that events predicted with probability p occur in approximately a p fraction of cases. It is a classical idea in forecasting [Brier, 1950, Murphy, 1973, Gneiting and Raftery, 2007] and a renewed concern in modern machine learning [Guo et al., 2017]. Large language models (LLMs), however, are frequently used as components in agentic systems that ask them to state a confidence level, even though those stated confidences are rarely measured against outcomes in live production conditions.

Bitcoin forecasting is not used here as an investment claim. It is used as a clean calibration laboratory: outcomes resolve quickly, ground truth is observable, and repeated probabilistic statements can be collected at low cost. The central question is narrow: when an LLM or meta-agent states a confidence about a future binary event, what is the empirical hit rate?

The contribution of this paper is empirical and infrastructural. We publish a reproducible calibration snapshot, the code-level metric definitions, and the resulting failure modes. The dataset contains 10,539 public forecast records, of which 8,367 are verified binary forecasts after excluding 2,172 NEUTRAL answers from the binary Brier calculation.

Table 1: Provider-level calibration metrics from the public dataset snapshot bundled with this submission. Lower Brier and ECE are better.

Provider	Binary N	Neutral N	Accuracy	Brier	ECE	Claim → empirical
Table Ronde	1,068	408	66.8%	0.220	0.086	55% → 68%
GPT	1,006	422	71.8%	0.225	0.112	75% → 68%
Hydra	1,014	462	71.2%	0.245	0.216	55% → 75%
Grok	150	75	74.7%	0.250	0.197	55% → 75%
Claude	401	199	77.6%	0.250	0.226	55% → 78%
Meta	1,031	445	73.5%	0.261	0.234	55% → 75%
DeepSeek	1,401	0	46.3%	0.302	0.190	55% → 58%
Chimera	1,329	147	44.3%	0.340	0.343	65% → 38%
Gemini	967	14	34.6%	0.398	0.421	75% → 31%

2 Forecasting System and Dataset

Every prediction round asks multiple providers a fixed set of Bitcoin questions: direction over 4h/12h/24h horizons, volatility expansion, and magnitude exceedance. Each answer contains a categorical prediction (YES, NO, or NEUTRAL), a confidence in $[0, 100]$, and, once the horizon resolves, an observed outcome and correctness label.

The public CSV bundled with this submission is available at https://strategyarena.io/api/calibration/dataset.csv. It contains the fields `timestamp`, `provider`, `question`, `confidence`, `predicted`, `actual`, and `correct`. No private API key or login is required to reproduce the tables in this paper.

3 Methodology

For binary Brier-score computation, NEUTRAL answers are excluded. If an agent answered YES with confidence c, then $p_{yes} = c/100$. If it answered NO with confidence c, then $p_{yes} = 1 - c/100$. The resolved outcome is encoded as $y \in \{0, 1\}$. The Brier score is

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} (p_{yes,i} - y_i)^2.$$

A non-informative binary forecaster has a Brier score close to 0.25 when class balance is near even. Lower values indicate better probabilistic skill.

We also compute expected calibration error (ECE). Confidence values are bucketed in ten bins. In each bin b, let $\text{conf}(b)$ denote the average stated confidence and $\text{acc}(b)$ the empirical correctness rate. Then

$$\text{ECE} = \sum_{b} \frac{n_b}{N} \left| \text{acc}(b) - \text{conf}(b) \right|.$$

Maximum calibration error (MCE) is the largest absolute bin gap. We additionally report a "killer statistic" per provider: the most-populated confidence bin and its empirical hit rate.

4 Results

Figure 1 shows representative reliability curves. The diagonal is perfect calibration. Providers above the line are under-confident; providers below it are over-confident.

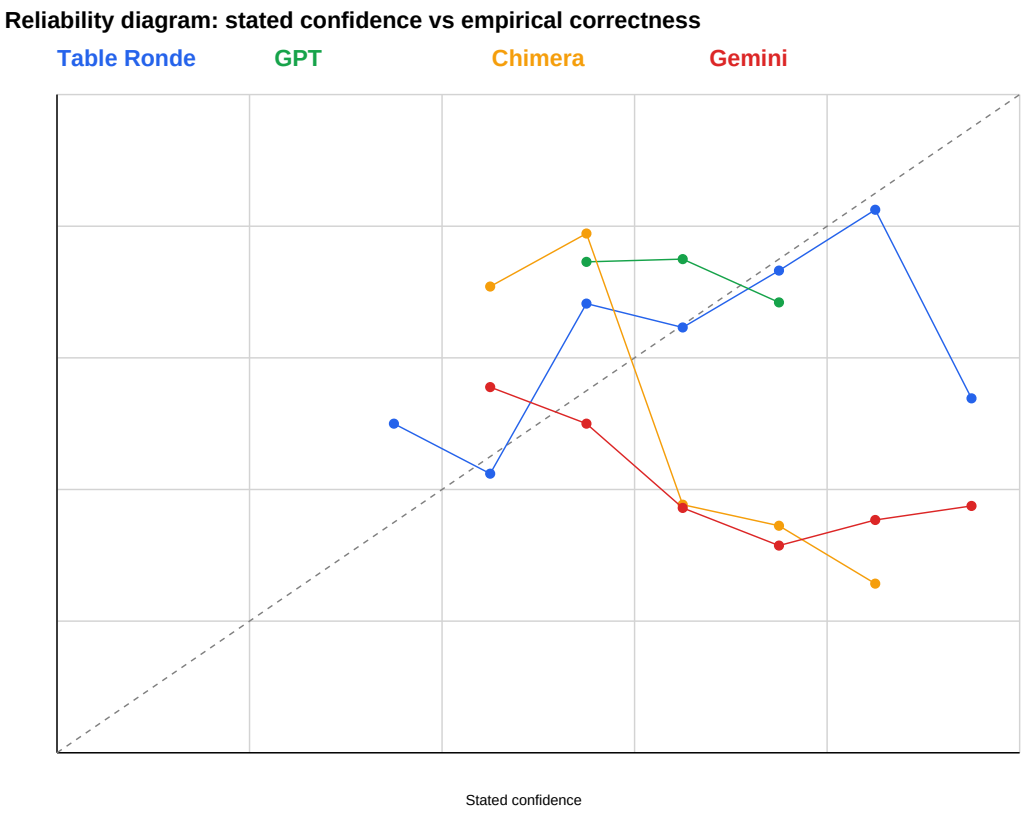

Figure 1: Reliability diagram for representative providers.

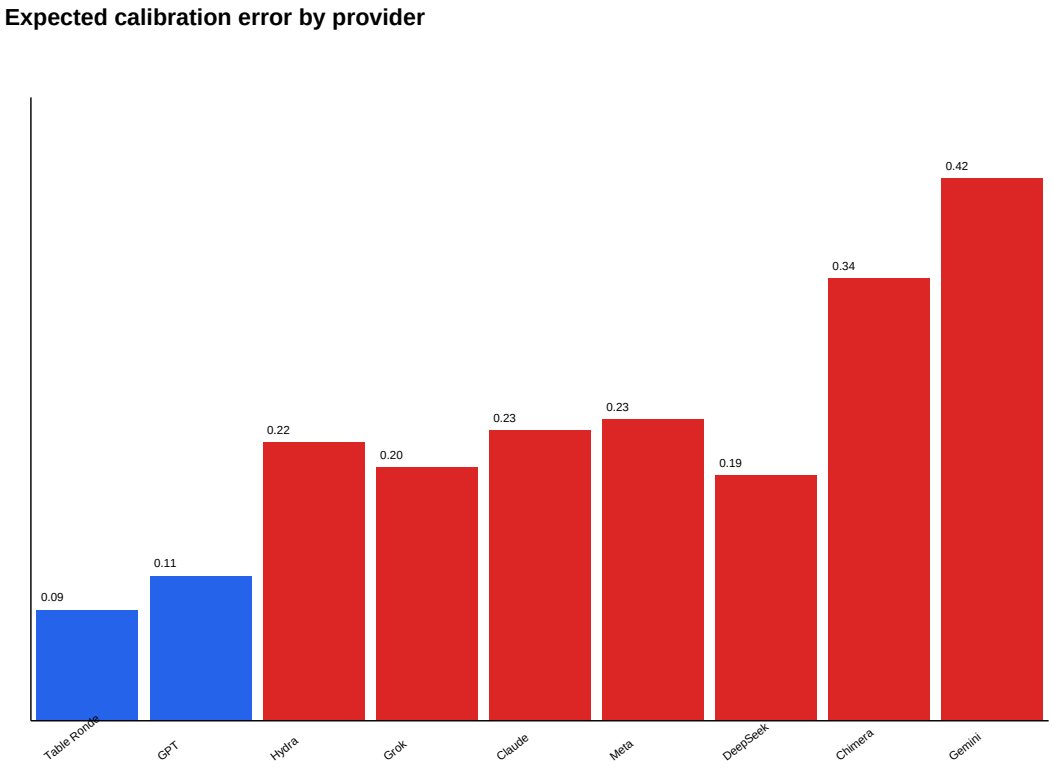

Figure 2: Expected calibration error by provider. Lower is better.

The results reveal three distinct failure modes. First, some providers are under-confident: their empirical accuracy exceeds their stated confidence. Second, some providers are over-confident: stated confidence is much higher than correctness. Third, for some agents confidence has weak ranking power, producing flat reliability curves across confidence bins.

5 Refusal and NEUTRAL Behavior

NEUTRAL answers are not discarded as uninteresting noise. They are an information channel about model hesitation and safety behavior. Figure 3 reports NEUTRAL rates by provider and question. High refusal rates may be a rational expression of uncertainty, but they also reduce actionability in downstream agentic systems.

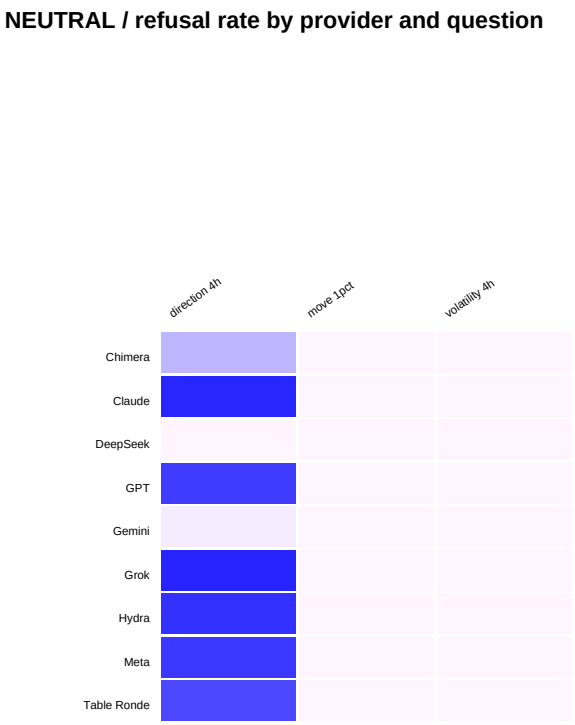

Figure 3: NEUTRAL/refusal rate by provider and forecast question. Darker cells indicate more NEUTRAL answers.

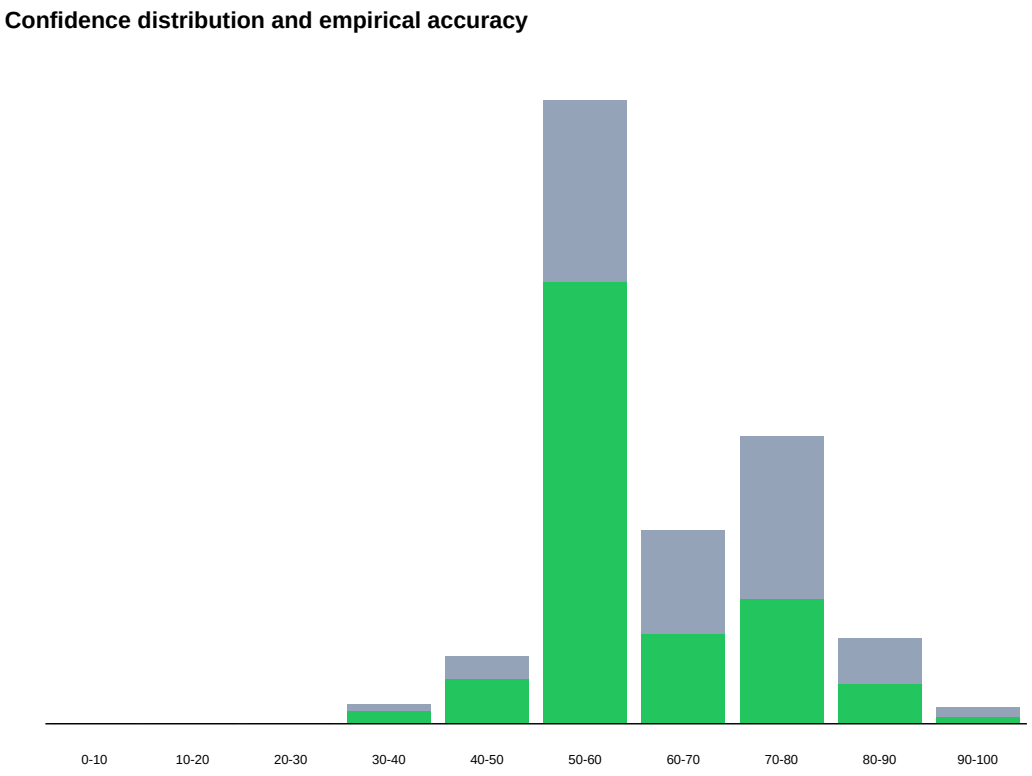

Figure 4: Distribution of stated confidence values. Green area shows the fraction correct in each bin.

6 Discussion and Related Work

Reliability diagrams and proper scoring rules are standard in meteorology and statistics [Brier, 1950, Murphy, 1973, Gneiting and Raftery, 2007]. Calibration failures in modern neural networks motivated post-hoc methods such as temperature scaling [Guo et al., 2017]. LLMs introduce a related but distinct problem: the model is often asked to state confidence in natural language or structured JSON, and that stated confidence may be affected by instruction tuning, safety training, refusal policies, and user-facing hedging norms [Ouyang et al., 2022, Bai et al., 2022].

Our hypothesis is that calibration should be treated as operational capital in multi-agent systems. A system that knows which agents are under-confident, over-confident, or prone to NEUTRAL answers can weight their votes more safely than a system that treats all stated confidences as literal probabilities.

7 Limitations

This study has several important limitations. First, all forecasts concern Bitcoin; transfer to equities, macroeconomic releases, or other crypto assets is unproven. Second, the observation window is non-stationary: market regimes change, and calibration can drift. Third, providers may

be correlated because LLMs share public training data and similar safety tuning. Fourth, a single scalar Brier score hides question-level heterogeneity. Finally, the experiment is paper/research infrastructure, not investment advice.

8 Open Data and Reproducibility

The dataset is public and the bundled CSV snapshot is included in `data/calibration_dataset.csv`. The public endpoint is:

`https://strategyarena.io/api/calibration/dataset.csv`

The analysis can be reproduced by computing the provider-level metrics described above. The bundle also includes `data/summary_metrics.json`, which records the exact snapshot counts and provider metrics used to generate the tables and figures.

9 Conclusion

The main result is not that LLMs can trade Bitcoin. The result is that live calibration measurement exposes meaningful differences between forecasters that would be invisible from accuracy alone. Calibration, refusal, and confidence ranking are measurable, cheap, and operationally useful. We recommend that institutions deploying multi-LLM forecasting systems maintain public or internal calibration ledgers rather than relying on unverified confidence claims.

A Verification Recipes

The current public provider summary can be inspected with:

```
curl -s https://strategyarena.io/api/calibration | jq '.providers'
```

The raw CSV used by this bundle can be downloaded with:

```
curl -s https://strategyarena.io/api/calibration/dataset.csv \
  -o calibration_dataset.csv
```

B Excluded Data

NEUTRAL answers are excluded from binary Brier and ECE calculations, but they are preserved in the released CSV and reported as refusal/hesitation behavior. Open forecasts whose horizon has not resolved are not present in the resolved CSV snapshot.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.