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.