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:
```