Temporally Supervised Linear Probes Improve LLM Forecasts

Anonymous Authors[1]

Abstract

Large language models can verbalize probabilistic forecasts about the future, but their verbalized probabilities may not fully reflect the forecast-relevant information in their internal computations. We study this gap by comparing verbalized probabilities with linear probes trained on the hidden activations that produced them. In our experiments, probes beat raw verbalized probabilities, scalar recalibration, text-only embedding probes, and calibration-oriented prompt variants under a chronological split. These results suggest a simple low-cost method for improving LLM forecasts when hidden activations are available.

1. Introduction

One way to use LLMs to produce forecasts about the future is to ask them: "What is the probability that event X will occur?" and then extract the probability they verbalize. This is natural for black-box interfaces like APIs, but open-weight models expose another possible interface: the activation vectors inside the LLM. If those activations contain forecast-relevant information that is not preserved in the verbalized probability, relying only on that probability can underestimate what is extractable from the LLM's computation.

We test this on resolved Kalshi prediction markets. Each model is given a textual representation of the prediction market and then generates a trace of tokens containing a verbalized probability indicating whether the market resolves YES. We then train logistic linear regression models with activation vectors as features and market resolutions as labels. Figure 2 summarizes the experimental pipeline.

Our main finding (Figure 1) is that temporally supervised linear probes beat verbalized probabilities and supervised recalibrated verbalized probabilities in all eight settings.

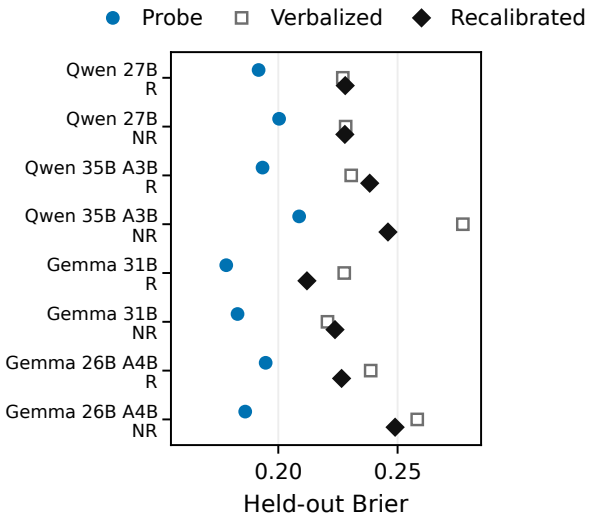


Figure 1. **Hidden-state probes outperform verbalized-probability readouts.** Held-out test Brier. R denotes reasoning and NR denotes no-reasoning. Lower is better.

Our contributions are:

1. To our knowledge, this is the first systematic evaluation of hidden-state readouts for temporally held-out forecasting.
2. We provide evidence that supervised hidden-state readouts are a promising low-cost strategy for improving LLM forecasting.
3. We show that the performance gains are not merely a result of better calibration, but also better discrimination.

2. Related Work

LLM forecasting. Work on LLM forecasting asks whether language models can assign probabilities to unresolved real-world events. Autocast introduced date-indexed forecasting questions for evaluating models under point-in-time information constraints (Zou et al., 2022). Early live-tournament studies found that individual GPT-4 forecasts lagged human crowd aggregates, while aggregating multiple LLM forecasts could approach crowd accuracy (Schoenegger &

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

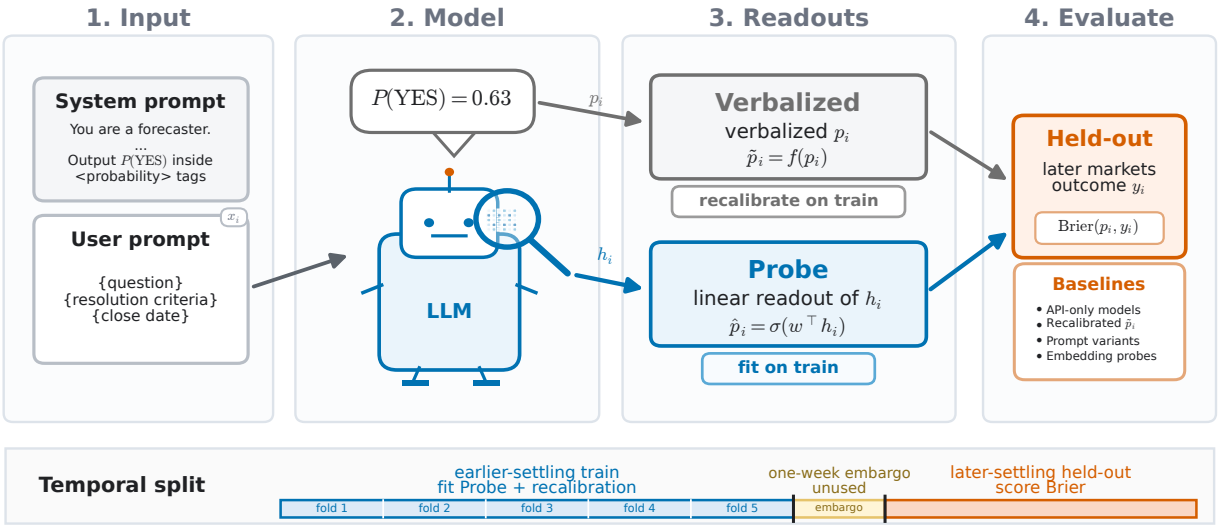


Figure 2. **Method overview.** Linear probes are selected using only earlier-settling markets; the intervening embargo is unused. Scores are computed on later-settling held-out markets.

Park, 2023; Schoenegger et al., 2024). More recent systems improve verbalized forecasts by combining retrieval, reasoning prompts, and aggregation, or by evaluating tool-using agents that gather external evidence before forecasting (Halawi et al., 2024; Pratt et al., 2024; Ye et al., 2024). Benchmarks such as ForecastBench, KalshiBench, and PolyBench scale this evaluation to dynamic forecasting questions and prediction-market data (Karger et al., 2024; Nel, 2025; Cheng et al., 2026). Other work emphasizes evaluation hazards such as temporal leakage, logical inconsistency, and limited extrapolation from benchmark scores to real deployment (Paleka et al., 2024; 2026).

Verbalized confidence and recalibration. Prior work shows that verbalized LLM confidence estimates can be miscalibrated and can sometimes be improved by prompting or post-hoc calibration (Guo et al., 2017; Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024; Wang et al., 2024).

Hidden-state uncertainty readouts. Several papers recover correctness, confidence, or uncertainty signals from hidden states, often finding that internal readouts outperform verbalized confidence (Burns et al., 2022; Azaria & Mitchell, 2023; Liu et al., 2024; Zhang et al., 2025; Moreno Cencerrado et al., 2025; Radharapu et al., 2025). Probes also recover answer-accuracy signal from states before the model commits to a generated answer (Moreno Cencerrado et al., 2025). Miao & Ungar (2026) formalizes a related observation: internal accuracy signals and verbalized confidence can be linearly encoded but detached, so a hidden-state readout and the verbalized scalar need not agree. Work on event probabilities and predictive distributions from embeddings similarly suggests that probabilistic information can be present in representations even when verbalized outputs are incoherent or expensive to sample (Zhu et al., 2025; Piskorz et al., 2026). Our setting differs by using hidden-state readouts as temporally supervised forecasts of later-resolving real-world events, rather than as correctness or confidence estimates for already answered questions.

3. Setup

We treat each resolved market as one binary forecasting instance. Let i index markets, with prompt fields x_i (question text, resolution rule, and close date), binary outcome $y_i \in \{0, 1\}$, and settlement timestamp s_i. We define a cutoff s_0 and a one-week embargo Δ immediately before it:

$$\mathcal{I}_{\text{train}} = \{i : s_i < s_0 - \Delta\},$$
$$\mathcal{I}_{\text{embargo}} = \{i : s_0 - \Delta \le s_i < s_0\},$$
$$\mathcal{I}_{\text{test}} = \{i : s_i \ge s_0\}.$$

We evaluate eight model-condition cells formed by crossing four models (`Qwen3.6-27B`, `Qwen3.6-35B-A3B`, `Gemma-4-31B-it`, and `Gemma-4-26B-A4B-it`) with two prompting modes, reasoning and no-reasoning. Let c be one such cell. Each cell is sampled 3 times. For cell c and sample $r \in \{1, 2, 3\}$, the model generates a trace of tokens. We parse its verbalized probability $p_{icr} \in [0, 1]$ from the final probability tag `<probability>p goes here</probability>`. We also record residual-stream states $h^{\ell}_{icr} \in \mathbb{R}^{d_c}$ at the *pre-digit* token position, i.e. the

position whose next generated token is the leading digit of the verbalized probability. We do this at every layer ℓ.

All predictions are scored at the question level. We define the verbalized forecast for cell c as

$$\hat{p}_{ic}^{\text{verb}} = \frac{1}{3}\sum_{r=1}^{3} p_{icr}.$$

A recalibrated verbalized baseline chooses a one-dimensional map $g_c : [0,1] \to [0,1]$ using only $\mathcal{I}_{\text{train}}$ and predicts

$$\hat{p}_{ic}^{\text{recal}} = \frac{1}{3}\sum_{r=1}^{3} g_c(p_{icr}).$$

For a fixed layer ℓ, let

$$\eta_{icr}(w,b) = w^\top z_c(h_{icr}^\ell) + b,$$
$$z_c(h) = \sqrt{d_c}\, h / \max(\|h\|_2, \epsilon).$$

The quantity d_c is the activation dimension for cell c, and z_c normalizes each activation vector to length $\sqrt{d_c}$, with ϵ guarding against zero-norm vectors. We fit an ℓ_2-regularized logistic probe by solving

$$\mathcal{L}_{\text{data}}(w,b) = \frac{1}{3}\sum_{i\in\mathcal{I}_{\text{train}}}\sum_{r=1}^{3} L(y_i, \sigma(\eta_{icr}))\,,$$
$$(w,b) = \arg\min_{w,b}\ \mathcal{L}_{\text{data}}(w,b) + \frac{\lambda}{2}\|w\|_2^2,$$

where $\sigma(t) = (1+\exp(-t))^{-1}$ and $L(y,q) = -y\log q - (1-y)\log(1-q)$. Hyperparameter λ controls the ℓ_2 penalty. The selected probe then predicts

$$\hat{p}_{icr}^{\text{probe}} = \sigma\left(w_c^\top z_c(h_{icr}^\ell) + b_c\right).$$

The question-level forecast is

$$\hat{p}_{ic}^{\text{probe}} = \frac{1}{3}\sum_{r=1}^{3} \hat{p}_{icr}^{\text{probe}}.$$

For any question-level forecast $\hat{p}_{ic}$, held-out Brier score is

$$\text{Brier}_c(\hat{p}) = \frac{1}{|\mathcal{I}_{\text{test}}|}\sum_{i\in\mathcal{I}_{\text{test}}} (\hat{p}_{ic} - y_i)^2.$$

Dataset. We use 1,167 resolved binary Kalshi questions from the KalshiBench-v2 Hugging Face dataset (Nel, 2025; 2084 Collective, 2026). The original dataset contains the question text, resolution criteria, close time, and binary outcome. We make an updated version by matching each row back to its Kalshi contract and adding `settlement_ts`, the timestamp at which Kalshi finalized payout. We use this settlement time as an upper bound on the true resolution timestamp. The time between market resolution and market settlement is typically a few hours (Kalshi, 2026). We then filter the questions to only keep those markets which have `close_time` after February 28, 2025. Questions settling before September 24, 2025 form the train/validation pool (592 questions). Questions settling from September 24 through September 30 form a one-week embargo set that is not used for training, selection, or testing (28 questions). Questions settling on or after October 1, 2025 form the held-out test set (547 questions). The train/test positive rates are 0.498 and 0.450, respectively. A constant 0.5 predictor and a constant train-base-rate predictor both score test Brier 0.250. The Brier of the test-base-rate constant predictor is 0.248.

Models. The main probing experiments use four open-weight instruction tuned models from two families, two of which are dense and two of which are mixture-of-experts models: `Qwen3.6-27B`, `Qwen3.6-35B-A3B`, `Gemma-4-31B-it`, and `Gemma-4-26B-A4B-it`. We selected these four open-weight models because they were among the highest-ranked mid-sized dense and MoE checkpoints listed on the Artificial Analysis open-weight model leaderboard at the time of experiment design (Artificial Analysis, 2026).

For these models, we collect both the final probability tags and residual-stream activations at all layers. The Qwen3.6 models report an April 2025 knowledge cutoff, and the Gemma-4 models report a January 2025 knowledge cutoff (Models.dev contributors, 2026; Google DeepMind, 2026).

Both precede the held-out test set settlement timestamp by at least 6 months. We also score verbalized forecasts from DeepSeek V3.2, Gemini 3 Flash Preview, Grok 4.1 Fast, and Grok 4.20 as API-only baselines. (Table 7). Details for the four probed open-weight models appear in Table 3.

Prompting modes. Each model is primarily evaluated in two conditions. **Reasoning** allows a reasoning trace before the final `<probability>` tag. **No reasoning** disables native reasoning and uses a strict probability-only prompt that asks for exactly `<probability>X</probability>`. Each (model, condition, question) is sampled three times at temperatures recommended by respective model providers. All reported metrics are computed after averaging trace-level probabilities within each question.

Probe selection. All supervision uses the earlier-settling train pool, with the one-week pre-test embargo excluded. Within the train pool, we use expanding walk-forward folds ordered by settlement timestamp for probe selection and recalibrator selection. All validation, test, and bootstrap metrics are question-level after averaging sample-level probabilities.

Verbalized-probability baselines. The recalibrated base-

Table 1. Mean held-out Brier gain of Probe over each baseline, decomposed into calibration and discrimination terms using 10 equal-width prediction bins. Positive values favor Probe. The calibration column is the reduction in calibration error. The discrimination column is the increase in separation between questions with different outcome rates. Text embedding averages the two question-text embedding probes.

Contrast	Baseline Brier	Probe Brier	Brier gain	Calibration	Discrimination
Probe vs. Verbalized	.239	.192	.047	.019	.027
Probe vs. Recalibrated	.231	.192	.039	.010	.029
Probe vs. Text embedding	.221	.192	.029	.016	.012

line receives the same earlier-settling labels and temporal validation protocol as the probe, but it is restricted to transforming the one-dimensional verbalized probability. Candidate maps are identity, Platt scaling, beta calibration, and isotonic regression (Platt, 1999; Kull et al., 2017; Guo et al., 2017; Zadrozny & Elkan, 2002). Platt and beta regularization use the same λ grid as the probe. The selected map is refit on all train-pool samples. As with probes, map selection uses validation Brier after averaging calibrated probabilities within each validation question.

Question-text embedding baseline. As a text-only control, we also train and select regularized logistic regressors on text embeddings of the same market prompt the LLM probes received. This baseline tests whether the probe advantage could be explained by a linear classifier picking up simple features of the text, rather than forecasting capabilities of the LLM models. All results are reported in Table 7.

4. Results

Hidden-state readouts beat baselines. Figure 3 shows that the 95% paired bootstrap confidence interval for the probe gain excludes zero in every model-condition cell. Each selected hidden-state probe improves on both the raw verbalized probability and the recalibrated verbalized baseline. The main Brier comparison in Figure 1 shows the same pattern in point estimates. Relative to the raw verbalized probability, the probe improves Brier by 0.047 on average. Relative to the stronger of raw or recalibrated verbalized probabilities in each cell the average gain is 0.038 Brier. Appendix D also reports a text-only control: linear probes trained on question-text embeddings score 0.219 Brier with Gemini-Embedding-001 and 0.223 with PPLX-Embed-V1-4B, showing the gain is not explained by the crude text similarity.

Gain in calibration and discrimination Table 1 summarizes the Murphy decomposition (Murphy, 1973) with 10 equal-width prediction bins, $\text{Brier} \approx \text{Calibration} - \text{Discrimination} + \text{Uncertainty}$. Calibration is lower when predicted probabilities better match empirical frequencies. Discrimination is higher when predictions better separate questions with different outcome rates. Relative to raw verbalized probabilities, probes improve mean Brier by 0.047: 0.020 comes from better calibration and 0.027 from better discrimination. Against the recalibrated baseline, the remaining probe gain after recalibration is 0.039 Brier, decomposing into 0.010 from calibration and 0.029 from discrimination. Thus scalar recalibration absorbs much of the calibration gap but does not recover the probe's discrimination gain. Paired bootstrapping estimates the Probe-over-Recalibrated discrimination gain as 0.029 (95% CI $[0.018, 0.041]$), while the calibration difference is smaller and less certain (0.010, 95% CI $[-0.001, 0.022]$). The discrimination gain is stable under equal-width and equal-mass binnings. (Section E).

5. Limitations and Discussion

What the probe establishes. The probe is supervised and model-specific. It does not prove that the model has a uniquely defined "true belief," nor that the probe recovers such a belief. It shows that forecast-relevant information is linearly extractable from hidden states after temporal supervision, and that this readout usually predicts held-out later-settling markets better than the verbalized probability and its standard scalar recalibrations.

Leakage and contamination. Kalshi states that most markets settle within a few hours after the outcome is known (Kalshi, 2026). This makes the one-week embargo a conservative guard against leakage through delayed settlement.

The reported Qwen3.6 and Gemma-4 knowledge cutoffs precede the held-out settlement window by 6 months, which reduces concerns that held-out test data is present in the models' pretraining data. This does not by itself rule out every form of pretraining or benchmark contamination in the underlying LLMs, as discussed by Paleka et al. (2026).

6. Reproducibility

The submission is anonymized for double-blind review. The paper and appendix specify the data split, model cells, probe-training procedure, and evaluation scripts used to regenerate the reported tables and figures. Code, scored trajectories, trained probe weights, and analysis tables will be released in a public repository after acceptance or publication, subject to dataset and model-output redistribution constraints.

Impact Statement

This work aims to improve automated real-world forecasting using LLMs. Better elicitation of model forecasts could support more accurate AI-assisted forecasting.

References

2084 Collective. KalshiBench-v2. https://huggingface.co/datasets/2084Collective/kalshibench-v2, 2026. Hugging Face dataset; accessed May 13, 2026.

Artificial Analysis. LLM leaderboard. https://artificialanalysis.ai/leaderboards/models?weights=open&size=small, 2026. Accessed May 13, 2026.

Azaria, A. and Mitchell, T. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision, 2022. URL https://arxiv.org/abs/2212.03827.

Cheng, P., Liu, J., and Long, Y. PolyBench: Benchmarking LLM forecasting and trading capabilities on live prediction market data, 2026. URL https://arxiv.org/abs/2604.14199.

Google DeepMind. Gemma 4 Model Card. https://ai.google.dev/gemma/docs/core/model_card_4, 2026. Accessed May 13, 2026.

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*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Halawi, D., Zhang, F., Chen, Y.-H., and Steinhardt, J. Approaching human-level forecasting with language models, 2024. URL https://arxiv.org/abs/2402.18563.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Kalshi. Market FAQs. https://help.kalshi.com/en/articles/13823821-market-faqs, 2026. Updated April 12, 2026; accessed May 13, 2026.

Karger, E., Bastani, H., Chen, Y.-H., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. Forecastbench: A dynamic benchmark of ai forecasting capabilities, 2024. URL https://arxiv.org/abs/2409.19839.

Kučinskas, S., Bastani, H., and Reynolds, M. Making forecasting scores easier to interpret: Introducing the brier index. https://forecastingresearch.substack.com/p/introducing-the-brier-index, March 2026. Accessed May 12, 2026.

Kull, M., Filho, T. S., and Flach, P. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 623–631. PMLR, 2017. URL https://proceedings.mlr.press/v54/kull17a.html.

Liu, X., Fatahi Bayat, F., and Wang, L. Enhancing language model factuality via activation-based confidence calibration and guided decoding, 2024. URL https://arxiv.org/abs/2406.13230.

Miao, M. M. and Ungar, L. Closing the confidence-faithfulness gap in large language models, 2026. URL https://arxiv.org/abs/2603.25052.

Models.dev contributors. Models.dev: An open-source database of ai models. https://models.dev/, 2026. Accessed May 10, 2026.

Moreno Cencerrado, I. V., Padrés Masdemont, A., Gonzalvez Hawthorne, A., Africa, D. D., and Pacchiardi, L. No answer needed: Predicting LLM answer accuracy from question-only linear probes, 2025. URL https://arxiv.org/abs/2509.10625.

Murphy, A. H. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973. doi: 10.1175/1520-0450(1973)012⟨0595:ANVPOT⟩2.0.CO;2.

Nel, L. Do large language models know what they don't know? KalshiBench: A new benchmark for evaluating epistemic calibration via prediction markets, 2025. URL https://arxiv.org/abs/2512.16030.

Paleka, D., Sudhir, A. P., Alvarez, A., Bhat, V., Shen, A., Wang, E., and Tramèr, F. Consistency checks for language model forecasters, 2024. URL https://arxiv.org/abs/2412.18544.

Paleka, D., Goel, S., Geiping, J., and Tramèr, F. Pitfalls in evaluating language model forecasters. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=z85kARAoyD.

Piskorz, J., Kobalczyk, K., and van der Schaar, M. Eliciting numerical predictive distributions of LLMs without auto-regression. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=SsuBd46twl.

Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Pratt, S., Blumberg, S., Carolino, P. K., and Morris, M. R. Can language models use forecasting strategies?, 2024. URL https://arxiv.org/abs/2406.04446.

Radharapu, B., Saxena, E., Li, K., Whitehouse, C., Williams, A., and Cancedda, N. Calibrating LLM judges: Linear probes for fast and reliable uncertainty estimation, 2025. URL https://arxiv.org/abs/2512.22245.

Schoenegger, P. and Park, P. S. Large language model prediction capabilities: Evidence from a real-world forecasting tournament, 2023. URL https://arxiv.org/abs/2310.13014.

Schoenegger, P., Tuminauskaite, I., Park, P. S., and Tetlock, P. E. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy, 2024. URL https://arxiv.org/abs/2402.19379.

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, 2023. URL https://arxiv.org/abs/2305.14975.

Wang, C., Szarvas, G., Balazs, G., Danchenko, P., and Ernst, P. Calibrating verbalized probabilities for large language models, 2024. URL https://arxiv.org/abs/2410.06707.

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 *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Ye, C., Hu, Z., Deng, Y., Huang, Z., Ma, M. D., Zhu, Y., and Wang, W. MIRAI: Evaluating LLM agents for event forecasting, 2024. URL https://arxiv.org/abs/2407.01231.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699. ACM, 2002. doi: 10.1145/775047.775151. URL https://doi.org/10.1145/775047.775151.

Zhang, A., Chen, Y., Pan, J., Zhao, C., Panda, A., Li, J., and He, H. Reasoning models know when they're right: Probing hidden states for self-verification, 2025. URL https://arxiv.org/abs/2504.05419.

Zhu, J.-Q., Yan, H., and Griffiths, T. L. Recovering event probabilities from large language model embeddings via axiomatic constraints, 2025. URL https://arxiv.org/abs/2505.07883.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.15474.

A. Dataset Construction

Table 2 summarizes the fixed KalshiBench-v2 subset and settlement-timestamp split used in all probe comparisons.

Table 2. Dataset construction counts for the fixed experimental subset.

Step	Questions
KalshiBench-v2 rows in local augmented copy	1,531
Resolved binary rows (yes/no outcome)	1,531
Promptable rows with close time and settlement timestamp	1,531
Saved-trajectory subset with close time after 2025-02-28	1,167
Train/validation pool: settlement before 2025-09-24	592
Embargo: settlement from 2025-09-24 to 2025-09-30	28
Held-out: settlement on/after 2025-10-01	547

B. Probed Model Details

Table 3 summarizes architecture and activation metadata for the four open-weight models with saved hidden states.

Table 3. Open-weight models used for hidden-state probing. Layer and hidden-dimension columns come from the saved activation metadata.

Paper name	Parameters	Layers	Hidden dim.	Activation dtype / norm	Cutoff
Qwen 3.6 27B	27B dense	64	5120	bfloat16; pre-final-norm	April 2025
Qwen 3.6 35B A3B	35B MoE, 3B active	40	2048	bfloat16; pre-final-norm	April 2025
Gemma 4 31B	31B dense	60	5376	bfloat16; post-final-norm	January 2025
Gemma 4 26B A4B	26B MoE, 4B active	30	2816	bfloat16; post-final-norm	January 2025

C. Selection Protocol

All probe and recalibrated-baseline choices use the same chronological split. The train pool is split into chronological chunks for expanding walk-forward validation. The selected probe is chosen from pre-digit activations across every model layer with ℓ_2 regularization over $\lambda \in \{2^{-6}, 2^{-4}, \ldots, 2^{8}\}$ plus an intercept-only candidate. Selection minimizes a conservative validation score: mean question-level Brier after averaging validation probabilities within question, plus 0.5 times the fold-level standard deviation. Candidates within $\max(0.002, \mathrm{SE})$ of the best score are tied, with ties broken toward larger λ and earlier layers. The selected probe is refit on the full train pool before test evaluation. In Tables 4 and 5, R denotes reasoning and NR denotes no reasoning.

Table 4. Validation-selected hidden-state probes.

Cell	Layer	Penalty	λ
Qwen 27B R	34	ℓ_2	2^{6}
Qwen 27B NR	37	ℓ_2	2^{8}
Qwen 35B A3B R	19	ℓ_2	2^{4}
Qwen 35B A3B NR	21	ℓ_2	2^{6}
Gemma 31B R	40	ℓ_2	2^{8}
Gemma 31B NR	38	ℓ_2	2^{6}
Gemma 26B A4B R	25	ℓ_2	2^{6}
Gemma 26B A4B NR	18	ℓ_2	2^{8}

Table 5. Validation-selected recalibration maps.

Cell	Map	λ
Qwen 27B R	Isotonic	n/a
Qwen 27B NR	Isotonic	n/a
Qwen 35B A3B R	Isotonic	n/a
Qwen 35B A3B NR	Isotonic	n/a
Gemma 31B R	Platt	2^{-6}
Gemma 31B NR	Platt	2^{-6}
Gemma 26B A4B R	Platt	2^{-6}
Gemma 26B A4B NR	Isotonic	n/a

D. All Model Scores

Table 7 reports every held-out score used in the paper, including API-only verbalized-probability baselines, the complete open-weight probe comparisons, and the question-embedding linear-probe baseline.

Table 6. All held-out model scores under additional metrics. Log loss is the binary negative log likelihood (lower is better). ECE is equal-mass expected calibration error with ten bins (lower is better). Brier Index is $(1 - \sqrt{\text{Brier}}) \times 100$, computed from held-out-question Brier (higher is better) (Kučinskas et al., 2026). Within each metric column, the top three entries are bold. Superscripts 1, 2, and 3 indicate their rank.

Model	Condition	Readout	Log loss ↓	ECE ↓	Brier Index ↑
Qwen 27B	Reasoning	Verbalized	0.720	0.121	52.4
Qwen 27B	Reasoning	Recalibrated	0.648	0.119	52.2
Qwen 27B	Reasoning	Probe	0.566	**0.039**[3]	56.2
Qwen 27B	No reasoning	Verbalized	0.676	0.099	52.2
Qwen 27B	No reasoning	Recalibrated	0.644	0.092	52.3
Qwen 27B	No reasoning	Probe	0.613	0.120	55.2
Qwen 35B A3B	Reasoning	Verbalized	0.736	0.112	52.0
Qwen 35B A3B	Reasoning	Recalibrated	0.671	0.129	51.2
Qwen 35B A3B	Reasoning	Probe	0.571	**0.032**[1]	56.0
Qwen 35B A3B	No reasoning	Verbalized	0.884	0.178	47.3
Qwen 35B A3B	No reasoning	Recalibrated	0.685	0.094	50.4
Qwen 35B A3B	No reasoning	Probe	0.639	0.114	54.3
Gemma 31B	Reasoning	Verbalized	0.749	0.154	52.3
Gemma 31B	Reasoning	Recalibrated	0.611	0.097	54.0
Gemma 31B	Reasoning	Probe	**0.531**[1]	**0.035**[2]	**57.8**[1]
Gemma 31B	No reasoning	Verbalized	0.681	0.109	53.0
Gemma 31B	No reasoning	Recalibrated	0.649	0.122	52.7
Gemma 31B	No reasoning	Probe	**0.553**[2]	0.075	**57.2**[2]
Gemma 26B A4B	Reasoning	Verbalized	0.848	0.156	51.1
Gemma 26B A4B	Reasoning	Recalibrated	0.657	0.105	52.4
Gemma 26B A4B	Reasoning	Probe	0.570	0.044	55.9
Gemma 26B A4B	No reasoning	Verbalized	0.880	0.167	49.2
Gemma 26B A4B	No reasoning	Recalibrated	0.704	0.149	50.1
Gemma 26B A4B	No reasoning	Probe	**0.558**[3]	0.069	**56.9**[3]
Gemini Embedding 001	Question embedding	Embedding probe	0.659	0.119	53.2
PPLX Embed V1 4B	Question embedding	Embedding probe	0.716	0.117	52.8
DeepSeek V3.2	Reasoning	Verbalized	0.652	0.119	53.5
DeepSeek V3.2	Reasoning	Recalibrated	0.631	0.136	52.7
DeepSeek V3.2	No reasoning	Verbalized	0.658	0.079	52.4
DeepSeek V3.2	No reasoning	Recalibrated	0.651	0.107	52.0
Gemini 3 Flash Preview	Reasoning	Verbalized	0.675	0.145	53.7
Gemini 3 Flash Preview	Reasoning	Recalibrated	0.616	0.134	54.1
Gemini 3 Flash Preview	No reasoning	Verbalized	0.655	0.127	54.1
Gemini 3 Flash Preview	No reasoning	Recalibrated	0.620	0.121	53.7
Grok 4.1 Fast	Reasoning	Verbalized	0.728	0.141	51.9
Grok 4.1 Fast	Reasoning	Recalibrated	0.634	0.109	52.8
Grok 4.1 Fast	No reasoning	Verbalized	0.751	0.145	50.4
Grok 4.1 Fast	No reasoning	Recalibrated	0.717	0.178	49.4
Grok 4.20	Reasoning	Verbalized	0.583	0.073	55.1
Grok 4.20	Reasoning	Recalibrated	0.584	0.064	55.1
Grok 4.20	No reasoning	Verbalized	0.593	0.070	54.9
Grok 4.20	No reasoning	Recalibrated	0.615	0.102	54.1

Figure 3 shows 95% paired bootstrap intervals for the direct Brier gain from using the Probe forecast instead of either the Verbalized or Recalibrated forecast.

Table 7. All held-out model scores on the fixed held-out split. The model column uses the paper-facing model or embedding name. Recalibration gain is Verbalized minus Recalibrated. Positive means recalibration improves the Verbalized Brier. Probe gain is the lower of Verbalized and Recalibrated Brier minus Probe, shown only when paired Verbalized probabilities exist. All non-gain scores are Brier scores (lower is better). Note that Grok 4.20 and Gemini 3 Flash Preview may have pretraining data overlapping the held-out data. Within each numeric column, the best three entries are bold. Superscripts 1, 2, and 3 indicate their rank.

Model	Condition	Verbalized	Recalibrated	Recalibration gain	Probe	Probe gain
Qwen 27B	Reasoning	0.227	0.228	-0.001	0.192	+0.035
Qwen 27B	No reasoning	0.228	0.228	+0.000	0.200	+0.028
Qwen 35B A3B	Reasoning	0.231	0.238	-0.008	0.193	+0.037
Qwen 35B A3B	No reasoning	0.277	0.246	**+0.031**[1]	0.209	**+0.037**[3]
Gemma 31B	Reasoning	0.228	0.212	**+0.016**[2]	**0.178**[1]	+0.034
Gemma 31B	No reasoning	0.221	0.224	-0.003	**0.183**[2]	**+0.038**[2]
Gemma 26B A4B	Reasoning	0.239	0.227	**+0.012**[3]	0.195	+0.032
Gemma 26B A4B	No reasoning	0.258	0.249	+0.009	**0.186**[3]	**+0.063**[1]
Gemini Embedding 001	Question embedding	n/a	n/a	n/a	0.219	n/a
PPLX Embed V1 4B	Question embedding	n/a	n/a	n/a	0.223	n/a
DeepSeek V3.2	Reasoning	0.217	0.224	-0.007	n/a	n/a
DeepSeek V3.2	No reasoning	0.227	0.230	-0.004	n/a	n/a
Gemini 3 Flash Preview	Reasoning	0.215	**0.210**[2]	+0.004	n/a	n/a
Gemini 3 Flash Preview	No reasoning	**0.210**[3]	0.214	-0.004	n/a	n/a
Grok 4.1 Fast	Reasoning	0.232	0.223	+0.009	n/a	n/a
Grok 4.1 Fast	No reasoning	0.246	0.256	-0.010	n/a	n/a
Grok 4.20	Reasoning	**0.201**[1]	**0.202**[1]	-0.000	n/a	n/a
Grok 4.20	No reasoning	**0.203**[2]	**0.211**[3]	-0.008	n/a	n/a

Table 6 reports the same held-out predictions under log loss, equal-mass expected calibration error, and Brier Index.

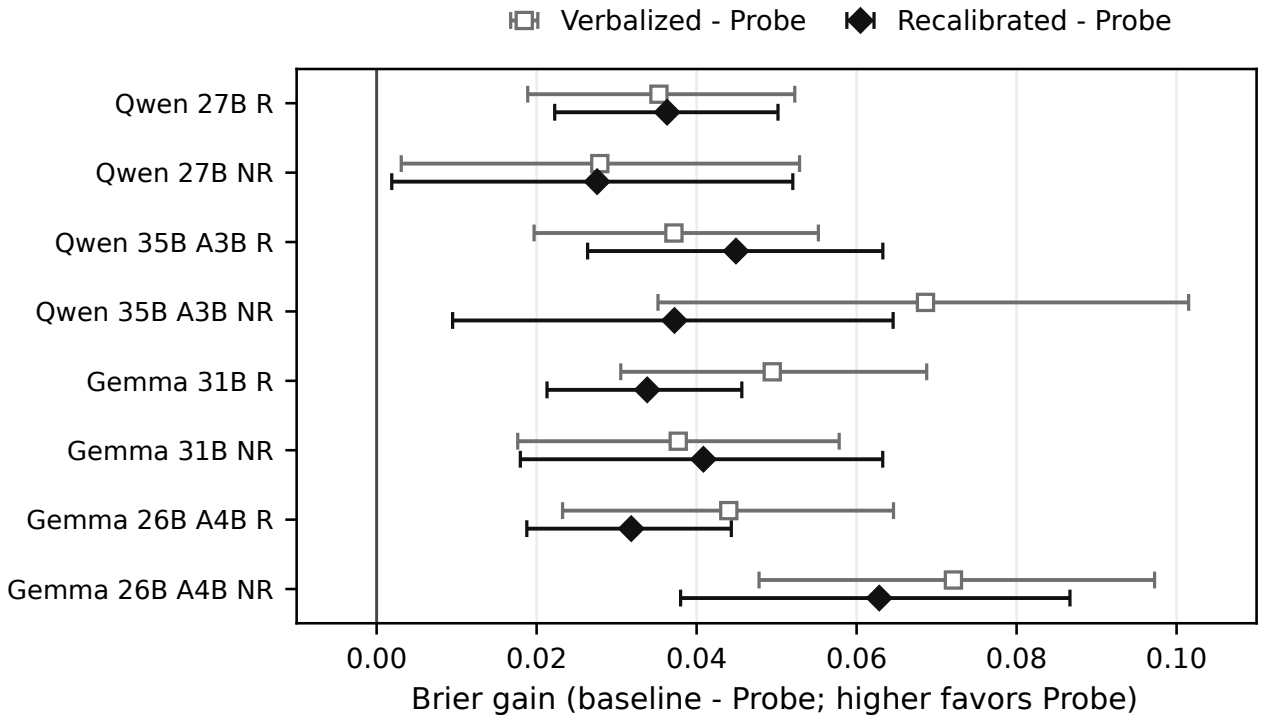


Figure 3. **95% paired bootstrap confidence intervals for probe gains.** Each interval resamples held-out questions and recomputes the mean Brier gain, defined as baseline Brier minus probe Brier. Positive values favor the hidden-state probe.

E. Brier Decomposition Robustness

The main decomposition table uses 10 equal-width prediction bins and reports mean differences across the eight probe cells. Table 8 gives the underlying per-cell calibration and discrimination terms. Because calibration and discrimination are binned diagnostics, we also check whether the residual Probe advantage over Recalibrated depends on this binning choice. For the 10-bin equal-width table, the approximation residual $\text{Brier} - (\text{Calibration} - \text{Discrimination} + \text{Uncertainty})$ ranges from -0.003 to 0.002 across the 24 cell-method rows. Tables 9 and 10 show that the discrimination contrast remains positive under paired question-level bootstrapping and under alternative binning schemes.

Table 8. Brier decomposition on the 547 held-out questions with predictions binned into 10 equal-width bins. With this binned approximation, $\text{Brier} \approx \text{Calibration} - \text{Discrimination} + \text{Uncertainty}$. Calibration is calibration error (lower is better). Discrimination is the variance of conditional outcome frequencies (higher is better). Uncertainty $= \bar{o}(1-\bar{o}) = 0.248$ is held-out base-rate variance. Probe raises Discrimination versus both Verbalized and Recalibrated in all eight cells. Bootstrap and binning checks in Section E show that the residual Probe gain over Recalibrated is primarily a discrimination gain. Bold marks the best of the three methods per row. The final row averages over the eight cells.

	Verbalized		Recalibrated		Probe	
Cell	Cal.↓	Disc.↑	Cal.↓	Disc.↑	Cal.↓	Disc.↑
Qwen 27B R	.020	.040	.017	.036	**.003**	**.059**
Qwen 27B NR	.013	.033	**.013**	.031	.019	**.064**
Qwen 35B A3B R	.020	.036	.024	.032	**.001**	**.055**
Qwen 35B A3B NR	.045	.014	**.010**	.011	.021	**.059**
Gemma 31B R	.029	.050	.012	.048	**.003**	**.071**
Gemma 31B NR	.018	.044	.022	.046	**.010**	**.074**
Gemma 26B A4B R	.032	.043	.017	.038	**.003**	**.056**
Gemma 26B A4B NR	.042	.031	.029	.027	**.007**	**.067**
Mean	.028	.036	.018	.034	**.008**	**.063**

Table 9. Question-level paired bootstrap intervals (2,000 resamples) for the mean decomposition contrast between Probe and Recalibrated. Positive discrimination and Brier contrasts favor Probe. Positive calibration contrast means Probe has lower calibration error.

Contrast	Point	95% CI
Probe discrimination − Recalibrated discrimination	.029	[.018, .041]
Recalibrated calibration − Probe calibration	.010	[-.001, .022]
Recalibrated Brier − Probe Brier	.039	[.022, .057]

Table 10. Binning robustness for the mean Probe-versus-Recalibrated decomposition contrast. Brier is unaffected by binning. Calibration and discrimination are recomputed under each binning scheme.

Binning	Bins	Recalibrated Brier − Probe	Probe discrimination − Recalibrated	Recalibrated calibration − Probe
Equal-width	5	.039	.029	.010
Equal-width	10	.039	.029	.010
Equal-width	20	.039	.030	.010
Equal-mass	5	.039	.028	.009
Equal-mass	10	.039	.028	.010
Equal-mass	20	.039	.027	.012

F. Prompt-Calibration Baseline Sweep

To check whether the verbalized probability baselines are unusually weak because of the exact wording of the main prompt, we ran three additional calibration-oriented prompt variants for Qwen 27B and Gemma 31B, in both Reasoning and No reasoning settings. These are the controls for the prompt sweep. Each cell uses the same fixed question set and three samples per question. The table reports question-level Brier score on the held-out split.

Table 11. Held-out prompt-variant Verbalized Brier scores. Each row uses the mean of three sampled Verbalized probabilities per question on the 547-question held-out split. Lower is better.

Model / condition	Prompt	Brier
Qwen 27B / Reasoning	Outside-view	**0.232**
	Resolution auditor	0.242
	Scoring-rule	0.237
Qwen 27B / No reasoning	Outside-view	0.243
	Resolution auditor	0.245
	Scoring-rule	**0.236**
Gemma 31B / Reasoning	Outside-view	0.223
	Resolution auditor	0.233
	Scoring-rule	**0.220**
Gemma 31B / No reasoning	Outside-view	0.224
	Resolution auditor	0.228
	Scoring-rule	**0.218**

G. Per-Layer Probe Brier

Figure 4 shows how held-out Brier varies across all pre-digit layers. This is a diagnostic plot, the deployed probe in each cell is still selected by walk-forward validation, not by test-set performance.

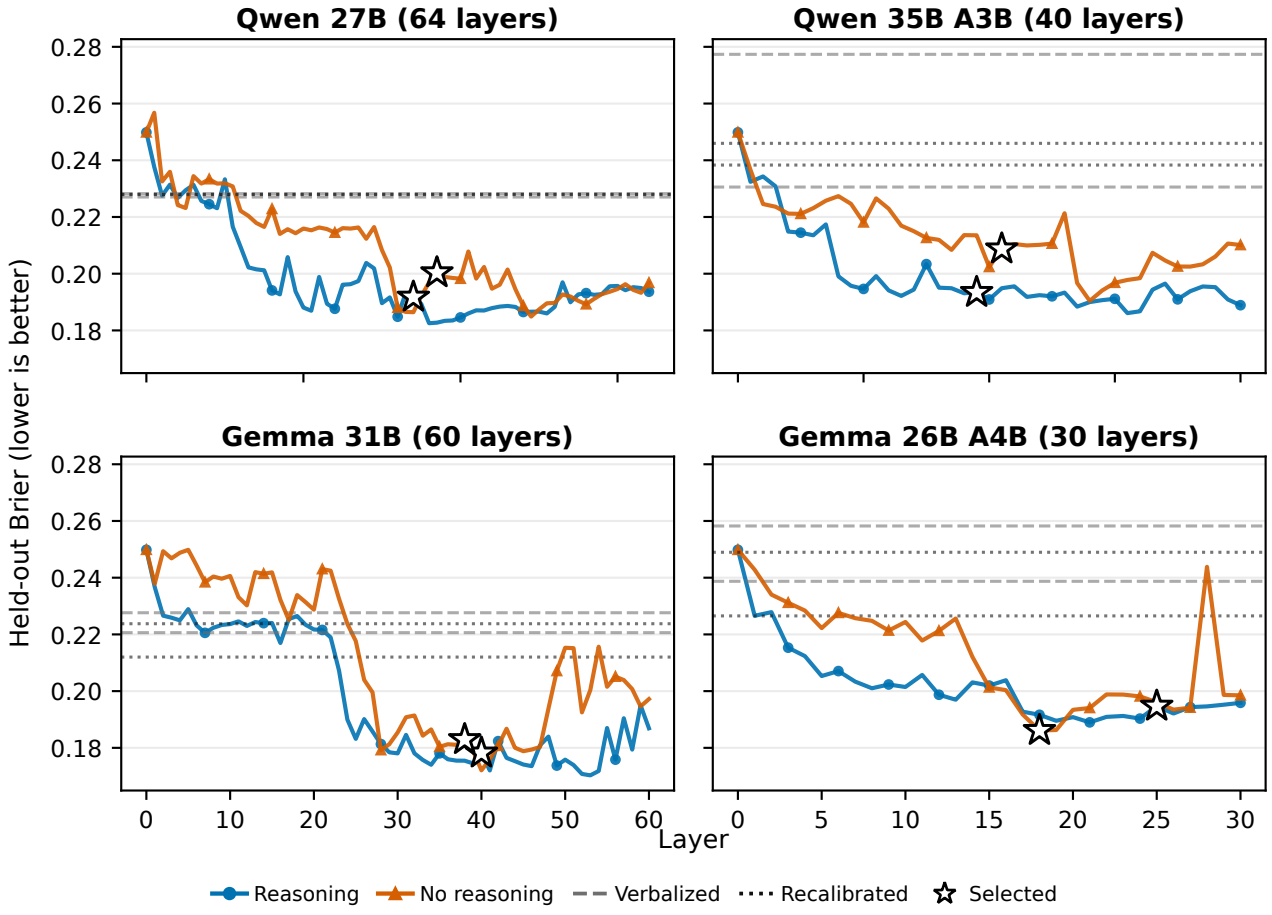


Figure 4. **Per-layer held-out Brier of probes.** Points use pre-digit activations at every model layer. Stars mark the validation-selected probe used in the main results. Horizontal lines show Verbalized and Recalibrated baselines.

H. Probe Selection Diagnostics

Figure 5 compares validation Brier to held-out Brier across the probe search grid. The selected probe is the point chosen by validation. The oracle point is shown only to diagnose how close validation selection comes to the best post-hoc held-out candidate.

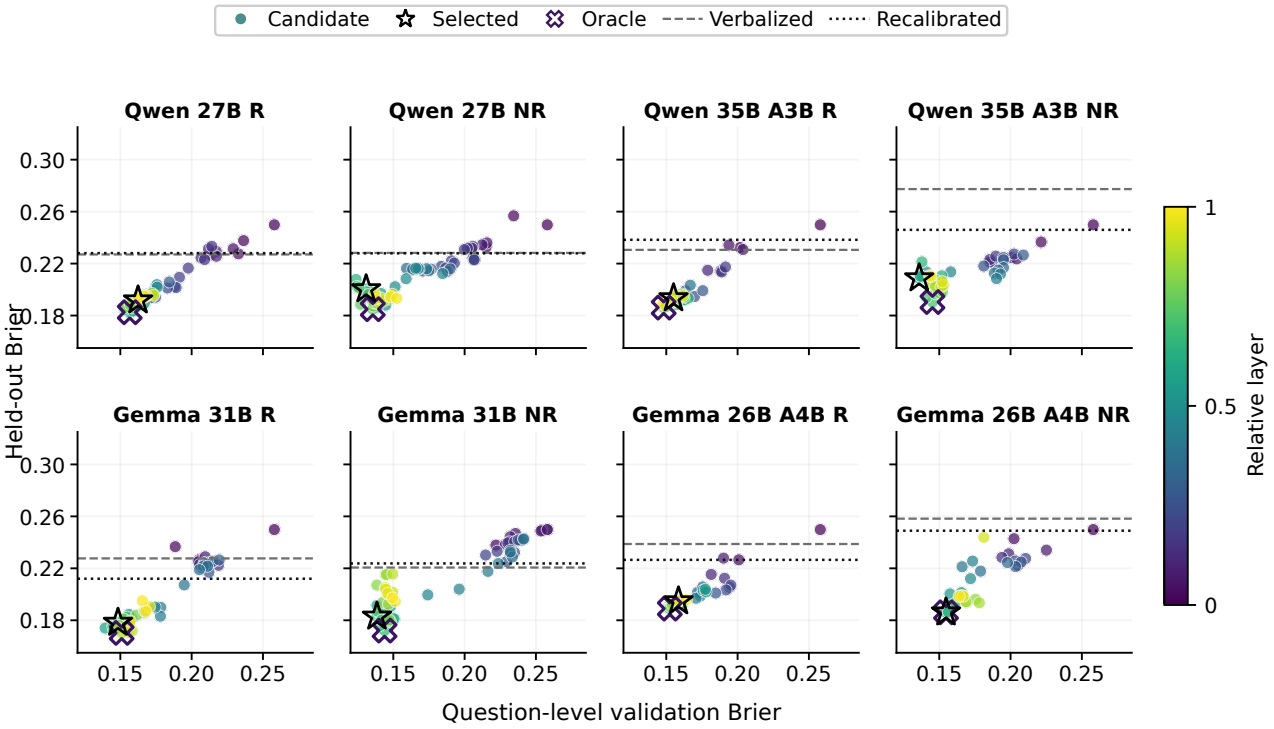


Figure 5. Question-level validation Brier versus held-out Brier for candidate probes in the search grid. The oracle point is the post-hoc best held-out probe and is used only as a diagnostic. Point color gives relative layer depth, and axes zoom to the main candidate cloud.

I. Reliability

Figure 6 gives calibration curves for the three main question-level forecasts: Verbalized, Recalibrated, and Probe.

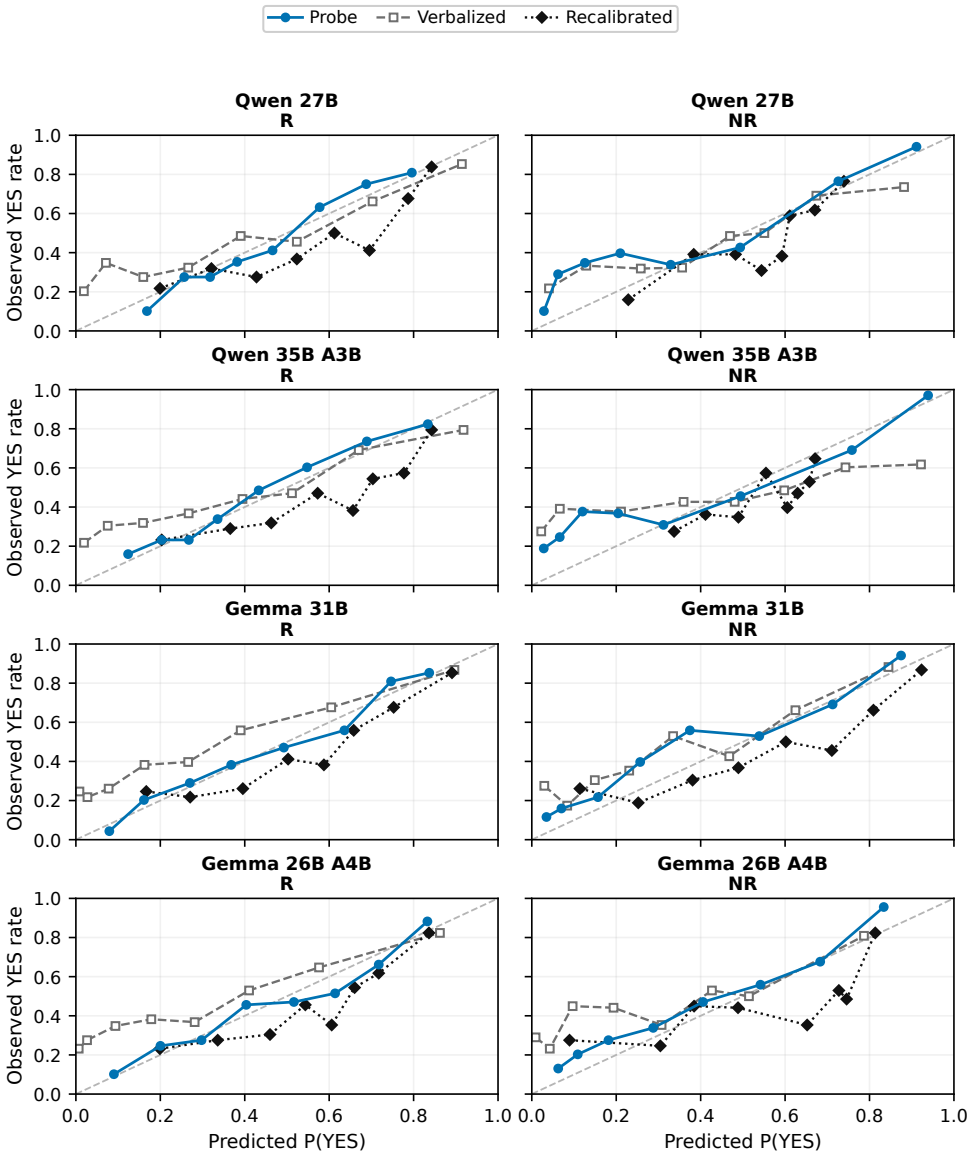


Figure 6. Reliability curves for Verbalized, Recalibrated, and Probe predictions on the held-out set.

J. Probe–Verbalized Outcome Scatter

Figure 7 shows where the probe and verbalized forecasts disagree on individual held-out questions, with points colored by the realized outcome.

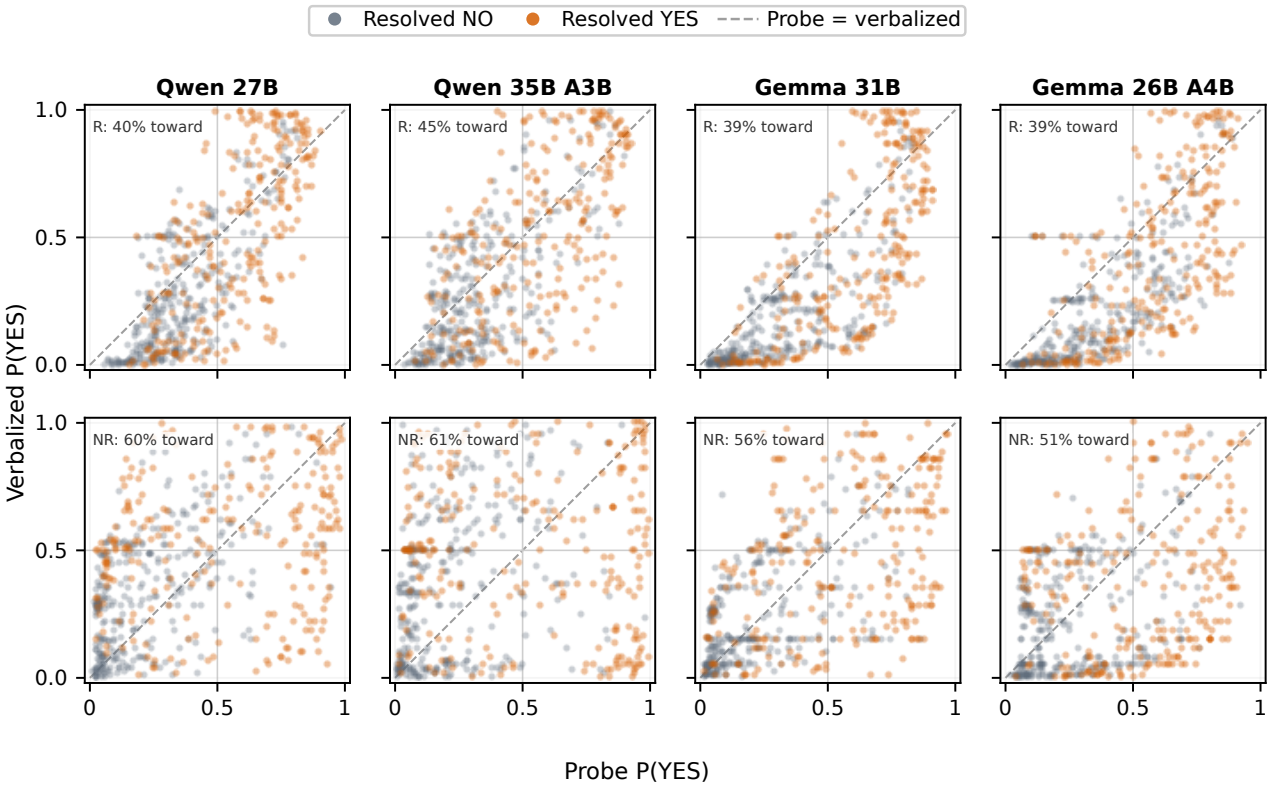


Figure 7. **Probe versus verbalized probabilities by held-out outcome.** Each panel plots the 547 held-out questions for one model cell, with hidden-state probe probability on the x-axis and verbalized probability on the y-axis. Points are colored by the resolved outcome. The dashed diagonal marks equal probe and verbalized probabilities. Points below the diagonal have a higher probe probability than verbalized probability.

K. Prompts

Native model reasoning is disabled in the ”No reasoning” condition. The user template is identical across experimental cells.

Reasoning.

```
System: You are a forecaster. You will be given a prediction-market
question with resolution criteria. Look at the question and assign a
probability between 0 and 1 that it resolves YES. Output your final
answer as <probability>X</probability>, where X is a decimal
between 0 and 1.
```

No reasoning.

```
System: You are a forecaster. You will be given a prediction-market
question with resolution criteria. Output ONLY the final answer as
<probability>X</probability>, where X is a decimal between
0 and 1. Output no other text: no analysis, no reasoning, no
preamble, no commentary. Your entire response must be exactly the
<probability>...</probability> tag.
```

Both conditions use the same user template:

```
Market closes: {close_date}

Question: {question}

Resolution criteria: {description}
```

K.1. Prompt-Calibration Baseline Prompts

The prompt-calibration sweep in Section F uses the following three prompt variants. Each variant is paired with both Reasoning and No reasoning model settings. Reasoning runs still enable the model's native hidden reasoning channel, whereas No reasoning runs disable that channel.

Outside-view calibrated.

```
System:
You are a calibrated prediction-market forecaster. Estimate the probability
that the market resolves YES. Use an outside-view base rate for similar
markets, then adjust for the specific question, resolution criteria, and
close date. Be conservative about extreme probabilities: use values below
0.10 or above 0.90 only when the outcome is almost determined by the
information in the question. Output ONLY the final answer as
<probability>X</probability>, where X is a decimal between 0 and 1.
Output no other text - no analysis, no reasoning, no preamble,
no commentary.

User:
Market closes: {close_date}

Question: {question}

Resolution criteria: {description}

Return only the final calibrated probability as
<probability>X</probability>.
```

Resolution auditor.

```
System:
You are a careful forecasting analyst. Your task is to predict whether
the market resolves YES under the stated resolution criteria, not whether
the headline sounds likely in ordinary language. Pay close attention to
dates, thresholds, named entities, edge cases, and whether the criteria
require a specific official action or measurement. Produce a calibrated
probability between 0 and 1. Output ONLY the final answer as
<probability>X</probability>. Output no other text - no analysis,
no reasoning, no preamble, no commentary.

User:
Market closes: {close_date}

Question: {question}

Resolution criteria: {description}
```

```
Give the probability that this resolves YES. Output only
<probability>X</probability>.
```

Scoring-rule calibrated.

```
System:
You are a calibrated forecasting assistant for binary prediction-market
questions. Your goal is to minimize long-run proper scoring-rule error,
especially Brier score and log loss. Give the best probability of YES,
not a persuasive argument. Interpret probabilities literally: 0.50 means
equally likely YES or NO; 0.60 means a small edge; 0.75 means clearly
favored but uncertain; 0.90 means very likely; 0.97 means only rare
failure modes remain. Avoid defaulting to 0.50 when evidence points one
way, and avoid extreme probabilities unless the resolution is nearly
settled. Output ONLY the final answer as <probability>X</probability>.
Output no other text - no analysis, no reasoning, no preamble,
no commentary.

User:
Market closes: {close_date}

Question: {question}

Resolution criteria: {description}

Return only the final calibrated probability as
<probability>X</probability>.
```