Can LLMs Predict the Future? A Brier Score Analysis of Prediction Markets

Yuanbo Li, Zekun Li*& Xiaoyan Cong*
Brown University
Providence, RI 02912, USA
yuanbo_li@brown.edu

Abstract

We study whether model upgrades improve probability estimates for prediction-market questions. Our **Resolved Market Forecasting (RMF)** benchmark contains 3,000 resolved binary questions across nine domains, on which we evaluate six Claude and Qwen model variants using a question-only, zero-shot protocol. We assess Brier scores relative to an empirical base-rate predictor, examine their Murphy decomposition, and compare models through paired differences, with results stratified by event category and timing relative to training cutoffs. In the reported post-cutoff stratum, the four Claude models achieve Brier scores of 0.183–0.192, improving on the base-rate reference by 0.024–0.033. Qwen 32B does not significantly outperform that reference, although its paired Brier is 0.024 lower than that of the 7B checkpoint. The evaluated Claude version and tier upgrades yield no significant improvement. Within-model differences across event categories exceed the observed differences among Claude variants. These results show why forecasting scores should be interpreted alongside simple probability baselines and question composition: under this protocol, newer versions or higher model tiers do not consistently produce more accurate probabilities.

1 Introduction

An event forecast communicates both an expected outcome and a degree of uncertainty. Two models assigning probabilities of 0.55 and 0.95 to the same event make the same binary prediction, but express substantially different confidence. Evaluating such outputs requires a score that retains the probability estimate and a reference that gives the resulting error a meaningful scale. As LLMs are increasingly studied as probabilistic forecasters (Halawi et al., 2024; Schoenegger et al., 2024; Karger et al., 2025), these choices become central to interpreting their performance.

Brier scores and base-rate comparisons. We use the Brier score, the mean squared difference between a forecast probability and the observed binary outcome (Brier, 1950). A low score alone does not establish how much information a model extracts from an individual question. When outcomes are imbalanced, a constant forecast based on the marginal event frequency can already perform well. We therefore compare model predictions with an empirical base-rate reference within each evaluation stratum. This comparison asks whether the model improves on a forecast that ignores the content of every question.

Aggregate error also combines several properties of a forecast and its evaluation set. The Murphy decomposition separates reliability, resolution, and uncertainty (Murphy, 1973): whether stated probabilities align with observed frequencies, whether forecasts distinguish groups with different outcome rates, and how balanced the outcomes are overall. These components help interpret differences across event categories, where a change in raw Brier

*Equal contribution.

need not represent a change in calibration. Our analysis uses this established framework to describe both the magnitude and the composition of forecasting error.

A shared question-only protocol. We construct **Resolved Market Forecasting (RMF)** from public Polymarket records, drawing a stratified sample of 3,000 binary questions from more than 600,000 resolved markets. The records span nine domains and provide settlement outcomes, resolution dates, and category labels (Sec. 3). Each model receives the market question and returns a probability, with no retrieval or additional market context. This fixed input protocol lets us examine what the evaluated models produce from internalized knowledge. We will release the RMF data and evaluation code upon acceptance.

Comparing models, questions, and time periods. Our model selection supports two types of comparison: the 7B and 32B Qwen2.5 checkpoints differ in size, while the Claude Sonnet and Opus variants provide version and tier comparisons. For each pair, we compute the Brier difference on questions answered by both models and report a bootstrap confidence interval. Pairing keeps the question set fixed when assessing a model change; category breakdowns then show where aggregate comparisons conceal variation across topics. The empirical base-rate reference and paired model comparisons answer complementary questions: whether a model adds predictive information, and whether a different model improves the score on the same questions.

The resolved records also permit retrospective comparisons across the temporal strata specified in Sec. 3. We examine how scores and their components vary with event timing relative to training cutoffs, alongside category composition, market salience, and agreement between models. The model identifiers, cutoff assumptions, and sample definitions are provided in the setup and appendices.

The study contributes a question-only benchmark and an analysis that links aggregate Brier scores to baseline performance, paired model effects, and variation across evaluated events. Section 4.1 addresses whether models improve on the base rate and whether model upgrades help; Section 4.2 examines category differences; and Section 4.3 studies the temporal patterns.

2 Related Work

LLMs as probabilistic forecasters. Prior work evaluates LLMs as probabilistic forecasters under a range of information and aggregation settings. Halawi et al. (2024) train retrieval-augmented systems that approach crowd-level calibration on questions from Metaculus, Good Judgment Open, INFER, Polymarket, and Manifold. Schoenegger et al. (2024) show that an ensemble of 12 LLMs rivals a human crowd on 31 binary Metaculus questions. Without retrieval, single LLMs are near chance (Schoenegger & Park, 2023); expert forecasters beat frontier LLMs on FORECASTBENCH (Karger et al., 2025; Alur et al., 2025); and frontier models are systematically overconfident on 300 Kalshi questions, mostly worse than the base rate (Nel, 2025). Fine-tuning on Polymarket (Turtel et al., 2025) and scaling to frontier models (Lu, 2025) narrow but do not close the gap to human superforecasters. Temporal information access remains a concern when interpreting forecasting performance (Paleka et al., 2025a;b). LLMs memorize portions of their training data (Carlini et al., 2023), and pre-/post-cutoff contamination detection is now standard (Golchin & Surdeanu, 2024; Roberts et al., 2023; Shi et al., 2024). Lopez-Lira et al. (2025) formalize this as a non-identification problem for economic forecasting, Magar & Schwartz (2022) distinguish memorization from downstream exploitation, and Li et al. (2026) show that prompted "forgetting" of post-cutoff knowledge fails to replicate genuine temporal ignorance. We examine forecasting performance jointly across training-cutoff strata, model variants, and event categories.

Prediction markets as ground truth. Prediction markets aggregate information through trader incentives (Wolfers & Zitzewitz, 2004; Berg et al., 2008), and trained forecasters outperform population baselines on such tasks (Mellers et al., 2014; Tetlock, 2005); both moti-

vate using crowd prices and human forecasters as benchmarks for LLMs. Prior work compares against Metaculus participants (Schoenegger et al., 2024) or expert superforecasters (Karger et al., 2025; Mellers et al., 2015), both trained populations; Polymarket aggregates capital from a self-selected global pool of traders, a more representative deployment target. Markets recover binary outcomes across domains (Dreber et al., 2015) and often match or beat polls on identical questions (Atanasov et al., 2017), subject to crowd-wisdom conditions (Davis-Stober et al., 2014) and price-interpretation caveats under belief heterogeneity (Manski, 2006).

3 Setup

RMF. RMF draws from the public Polymarket record of resolved binary markets between 2021 and 2026, spanning nine domains: crypto, entertainment, finance, geopolitics, politics, science/tech, sports, weather, and a residual *other* category. Polymarket resolutions are determined by financial settlement, which filters out ambiguous, unverifiable, or trivially-known questions and forces precise resolution criteria; the resulting ground truth is both large in scale and clean enough for stratified analysis at the domain level.

Models. We evaluate six models in three families, with two variants per family, so that each family supports a within-family scaling contrast (older→newer for Claude, smaller→larger for Qwen): Claude Sonnet (**4.5**, **4.6**), Claude Opus (**4.1**, **4.6**) (Anthropic, 2024), and Qwen2.5 (**7B-Instruct** in bf16, **32B-Instruct-AWQ** in int4) (Qwen Team, 2024). The two Claude variants in each family share a tier (Sonnet at \$3 / \$15 per million input/output tokens, Opus at \$15 / \$75) but differ in training cutoff; Qwen variants share a cutoff (October 2023) but differ by roughly 4.5× in parameter count. Anthropic does not publicly disclose Claude parameter counts; full per-model release dates, vendor-published training cutoffs, and per-cutoff sample sizes are in App. A.

Table 1: Models evaluated on the n=3,000 sample. "n pre" and "n post" use the paper's shared boundary (pre $\leq$ 2024-12-31, post $\geq$ 2025-06-30). Cutoff dates for Sonnet 4.5 and Opus 4.1 are best-effort estimates from the corresponding release notes (Anthropic does not always publish an exact cutoff for intermediate model versions). For the Qwen2.5 checkpoints, both buckets fall after Qwen's own October 2023 cutoff and are therefore both strictly post-cutoff for those models; we still report counts under the paper's shared boundary so the rows are directly comparable.

Model	API / HF ID	Released	Vendor cutoff	n pre	n post
Sonnet 4.5	`claude-sonnet-4-5-20250929`	2025-09	2025-07	1,498	1,502
Sonnet 4.6	`claude-sonnet-4-6`	2025-12	2026-01	1,142	1,152
Opus 4.1	`claude-opus-4-1-20250805`	2025-08	2025-03	1,498	1,502
Opus 4.6	`claude-opus-4-6`	2025-11	2025-08	1,498	1,501
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`	2024-09	∼2023-10	1,498	1,502
Qwen2.5-32B-AWQ	`Qwen/Qwen2.5-32B-Instruct-AWQ`	2024-09	∼2023-10	1,498	1,502

Training-data cutoffs. Vendor-published training cutoffs vary across the six models (App. A); the earliest is Qwen2.5 (October 2023) and the latest is Sonnet 4.6 (January 2026). To keep *post-cutoff* strictly out-of-distribution for all four Claude models simultaneously, we use a single conservative boundary calibrated to the earliest Claude cutoff. A market is labeled *pre-cutoff* if it resolved on or before December 31, 2024, *post-cutoff* if it resolved on or after June 30, 2025, and excluded as ambiguous otherwise; the six-month exclusion zone serves as a buffer against soft-cutoff effects. Both Qwen buckets fall strictly after Qwen's own October 2023 cutoff and so are post-cutoff for Qwen regardless of which Claude-calibrated bucket they land in (see App. A).

Protocol and metrics. Each model receives only the market question and outputs a probability $p \in [0, 1]$ (zero-shot, no retrieval, no context). We evaluate on the same n=3,000 stratified markets across all six models (∼1,500 per period); per-model valid sample sizes

Table 2: Post-cutoff Brier across all six models. Δ_{BR}: base-rate Brier − model Brier (positive = model beats base rate). * = base rate outside the model's 95% CI. All four Claude models clear the base rate by 0.02–0.03 Brier; Qwen 32B is indistinguishable from it; Qwen 7B is significantly worse.

Model	n	Brier	Base rate	Δ_{BR}
Sonnet 4.5	1,502	.186	.216	+.030*
Sonnet 4.6	1,152	.189	.217	+.028*
Opus 4.1	1,502	.183	.216	+.033*
Opus 4.6	1,501	.192	.216	+.024*
Qwen 7B	1,502	.234	.216	−.018*
Qwen 32B	1,502	.210	.216	+.006

range from 1,152 to 1,502 post-cutoff after parse failures (App. A). We report the Brier score $\text{BS} = \frac{1}{N}\sum_i (p_i - o_i)^2$ (lower is better) with bootstrap 95% CIs (10K resamples), and the Murphy decomposition $\text{BS} = \text{REL} - \text{RES} + \text{UNC}$ (Murphy, 1973); full per-condition decomposition in App. C. For pairwise comparison between two models we use the *paired Brier difference* $\Delta_{A-B} = \frac{1}{|I|}\sum_{i\in I}\left[(p_i^A - o_i)^2 - (p_i^B - o_i)^2\right]$ over the set I of markets where both models made a prediction, with a 95% bootstrap CI over markets; a CI that contains zero indicates the two models are statistically indistinguishable on the matched set. Full procedure in App. D.

Baseline. We benchmark every model against the *empirical base-rate predictor*, which assigns the marginal $P(\text{Yes})$ in each cutoff stratum as a constant prediction to every market in that stratum; we use the empirical rate rather than 0.5 because Polymarket yes-rates are not balanced. A model *beats* the base rate when its Brier is strictly lower and its 95% bootstrap CI excludes the baseline's Brier (the * marker in Table 2); we report the matched effect size as $\Delta_{\text{BR}} = \text{Brier}_{\text{base}} - \text{Brier}_{\text{model}}$.

4 Results

We answer Q1 (do LLMs predict?) and Q2 (does scaling help?) in §4.1, identify the dominant axis (event category) in §4.2, and characterize the cutoff as a moderator in §4.3.

4.1 Frontier prediction barely beats base rate; scaling helps only below the frontier

Q1: Can LLMs predict the future well? Barely. All four Claude models converge to a narrow Brier band around 0.19 (Table 2), beating the base rate by only 0.024–0.033: significant at n>1,000 per stratum, practically marginal. The two Qwen models trail: 32B is indistinguishable from the base rate, and 7B is significantly *worse* than it.

Q2: Does scaling capacity help? Only below the frontier. Table 3 reports five paired contrasts: three within-family scaling tests (Sonnet 4.5→4.6, Opus 4.1→4.6, Qwen 7B→32B) and two cross-family tests (Sonnet vs. Opus within each generation). Only one is significant in the expected direction: Qwen 7B→32B. Inside the Claude lineup, none of the four contrasts improves significantly, and Opus 4.1→4.6 is significantly *negative*. Capacity helps when a model is weak (lifting Qwen from below the base rate to indistinguishable from it), but once a model reaches the Claude convergence band near 0.19, neither version nor tier moves the needle. The Opus pair has few markets in the memorization-gap window (App. B), so its negative paired difference reflects post-cutoff prediction, not memorization.

4.2 Category: matters far more than capacity

Event category affects post-cutoff Brier substantially more than capacity does within the frontier. Within-model Brier spreads across categories are several times larger than the largest within-Claude paired contrast (§4.1) and the Qwen size effect (Table 4; full per-cell

Table 3: Paired Brier difference Δ (later minus earlier; negative = later is better) on post-cutoff markets. * = 95% bootstrap CI excludes zero. Only Qwen 7B→32B improves significantly; within-Claude contrasts are null or significantly negative. Full CIs in App. D.

Contrast	Δ
Within-family version (older → newer)	
Sonnet 4.5 → 4.6	+.004
Opus 4.1 → 4.6	+.009*
Within-family size (smaller → larger)	
Qwen 7B → 32B	−.024*
Cross-family capacity (Sonnet → Opus)	
Within 4.6 generation	+.004
Within 4.5/4.1 generation	−.003

Table 4: Within-model post-cutoff Brier spread across categories ($n \geq 20$ per cell). The spread is several times the largest paired capacity effect from §4.1; best and worst categories differ across models. Full per-cell numbers in App. E.

Model	Best cat	Brier	Worst cat	Brier
Sonnet 4.5	weather	.126	sports	.195
Sonnet 4.6	geopolitics	.138	sports	.205
Opus 4.1	geopolitics	.102	sports	.205
Opus 4.6	weather	.129	geopolitics	.229
Qwen 7B	geopolitics	.149	sci/tech	.274
Qwen 32B	geopolitics	.143	other	.225

numbers in App. E). Best and worst categories also differ across models, so no universal LLM-friendly domain exists.

Table 5: Post-cutoff Brier by event category and model. Per-category n in parentheses (varies slightly for Sonnet 4.6 because of parse failures). Italic cells ($n < 20$) are directional only.

Category	Sonnet 4.5	Sonnet 4.6	Opus 4.1	Opus 4.6	Qwen 7B	Qwen 32B
crypto	.176 (178)	.170 (136)	.164 (178)	.165 (177)	.203 (178)	.192 (178)
entertainment	.150 *(14)*	.036 *(10)*	.150 *(14)*	.160 *(14)*	.199 *(14)*	.219 *(14)*
finance	.197 *(14)*	.187 *(11)*	.101 *(14)*	.218 *(14)*	.193 *(14)*	.202 *(14)*
geopolitics	.147 (34)	.138 (28)	.102 (34)	.229 (34)	.149 (34)	.143 (34)
other	.198 (745)	.200 (566)	.198 (745)	.202 (745)	.248 (745)	.225 (745)
politics	.183 (64)	.165 (48)	.179 (64)	.219 (64)	.220 (64)	.205 (64)
science/tech	.178 (95)	.190 (71)	.159 (95)	.170 (95)	.274 (95)	.179 (95)
sports	.195 (243)	.205 (191)	.205 (243)	.208 (243)	.229 (243)	.216 (243)
weather	.126 (115)	.166 (91)	.123 (115)	.129 (115)	.210 (115)	.171 (115)

4.3 Cutoff: a moderator of category, not a main effect

If models genuinely forecasted, the cutoff axis would be irrelevant. If models purely recalled, crossing the cutoff would cause a uniform Brier collapse. We find neither: pre- and post-cutoff Brier are within 0.03 for every Claude model, with no sharp memorization cliff as the Opus buffer varies from 0 to 360 days (App. G, App. F). As a main effect, the cutoff buys almost nothing.

The cutoff signal does not disappear; it concentrates inside the category axis. Stratifying each model's markets into log-volume tertiles as a proxy for salience (App. H), the salience gap turns from non-positive pre-cutoff to positive post-cutoff for every Claude model and for Qwen 32B: prominent events are easy when memorizable and hard when novel. A second signature is in cross-model agreement: mean per-event squared-error correlation

Table 6: Brier marginals across all six models, both periods. Δ_{BR}: base-rate minus model Brier on post-cutoff (positive = model beats base rate). * = base rate outside the model's post-cutoff 95% CI.

Model	n post	Pre	Post	Δ_{BR}
Sonnet 4.5	1,502	.196 [.184, .208]	.186 [.176, .196]	$+.030^*$
Sonnet 4.6	1,152	.220 [.205, .236]	.189 [.178, .201]	$+.028^*$
Opus 4.1	1,502	.211 [.198, .224]	.183 [.173, .193]	$+.033^*$
Opus 4.6	1,501	.211 [.198, .225]	.192 [.181, .203]	$+.024^*$
Qwen 7B	1,502	.264 [.251, .277]	.234 [.222, .246]	$-.018^*$
Qwen 32B	1,502	.231 [.220, .243]	.210 [.200, .220]	$+.006$

rises from 0.42 pre-cutoff to 0.62 post-cutoff, with the lift concentrated inside the frontier Claude lineup (App. I); the frontier converges on the same hedged predictions. The cutoff matters precisely where memorization could matter, and nowhere else.

5 Conclusion

We introduce **RMF**, a benchmark for evaluating LLM probability forecasts, using each model's training cutoff to separate real prediction from recall. Across six models in three families, three findings emerge: frontier LLMs barely beat the empirical base-rate predictor; capacity scaling helps only below the frontier and stops once a model reaches the Claude convergence band; and event category drives more variance in post-cutoff Brier than capacity does. Better reasoning benchmarks do not produce better frontier forecasters.

Limitations

Three limitations. *(i) No ceiling baseline.* We benchmark only against the empirical base rate, not against Polymarket crowd prices or expert forecasters; the gap to what is achievable is unknown. *(ii) Zero-shot, no-retrieval protocol.* Predictions use internalized knowledge alone; allowing internet access would likely shrink the gap to the base rate and may change the within-frontier scaling story. *(iii) Time-since-cutoff decay.* We treat all post-cutoff markets as equivalent, but markets resolving weeks vs. months past a cutoff differ; we do not decompose Brier by time-since-cutoff.

Ethical Considerations

Probabilistic LLM forecasters carry a deployment risk: users tend to trust calibrated-looking numerical outputs more than the underlying reliability warrants, especially in financial and policy settings where a confident probability can drive concrete decisions. We release RMF so that deployers can directly measure per-model reliability against ground truth instead of relying on aggregate benchmark scores or vendor claims. On the data side, RMF uses only public post-resolution Polymarket metadata (question, outcome, resolution date, category); we access no trader identities, order-book data, or live prices, and contribute no order flow to Polymarket itself. Frontier model versions are deprecated by vendors over time; we list exact model identifiers and vendor-published training cutoffs in App. A to support replication while the evaluated versions remain available.

Acknowledgements

We used Anthropic's Claude Opus 4.7 for writing assistance and prose editing on this manuscript; all research design, experiments, analyses, and findings are the authors'.

References

Rohan Alur, Bradly C. Stadie, Daniel Kang, Ryan Chen, Matt McManus, Michael Rickert, Tyler Lee, Michael Federici, Richard Zhu, Dennis Fogerty, Hayley Williamson, Nina Lozinski, Aaron Linsky, and Jasjeet S. Sekhon. AIA forecaster: Technical report. *arXiv preprint arXiv:2511.07678*, 2025.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Model card, 2024. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3):691–706, 2017.

Joyce E. Berg, Forrest D. Nelson, and Thomas A. Rietz. Prediction market accuracy in the long run. *International Journal of Forecasting*, 24(2):285–300, 2008.

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

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023.

Clintin P. Davis-Stober, David V. Budescu, Jason Dana, and Stephen B. Broomell. When is a crowd wise? *Decision*, 1(2):79–101, 2014.

Anna Dreber, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112 (50):15343–15347, 2015.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations*, 2024.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations*, 2025.

Zehan Li, Yuxuan Wang, Ali El Lahib, Ying-Jieh Xia, and Xinyu Pi. Simulated ignorance fails: A systematic study of LLM behaviors on forecasting problems before model knowledge cutoff. *arXiv preprint arXiv:2601.13717*, 2026.

Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust LLMs' economic forecasts? *arXiv preprint arXiv:2504.14765*, 2025.

Janna Lu. Evaluating LLMs on real-world forecasting against expert forecasters. *arXiv preprint arXiv:2507.04562*, 2025.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 157–165, 2022.

Charles F. Manski. Interpreting the predictions of prediction markets. *Economics Letters*, 91 (3):425–429, 2006.

Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott, Don Moore, Pavel Atanasov, Samuel A Swift, Terry Murray, Eric Stone, and Philip E Tetlock. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115, 2014.

Barbara Mellers, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip Tetlock. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3):267–281, 2015.

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

Lukas Nel. Do large language models know what they don't know? KalshiBench: A new benchmark for evaluating epistemic calibration via prediction markets. *arXiv preprint arXiv:2512.16030*, 2025.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters. *arXiv preprint arXiv:2506.00723*, 2025a.

Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency checks for language model forecasters. In *International Conference on Learning Representations (ICLR)*, 2025b.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. Data contamination through the lens of time. *arXiv preprint arXiv:2310.10628*, 2023.

Philipp Schoenegger and Peter S Park. Large language model prediction capabilities: Evidence from a real-world forecasting tournament. *arXiv preprint arXiv:2310.13014*, 2023.

Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, Rafael Valdece Sousa Bastos, and Philip E Tetlock. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528, 2024. doi: 10.1126/sciadv.adp1528.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Philip E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2005.

Benjamin Turtel, Danny Franklin, and Philipp Schoenegger. LLMs can teach themselves to better predict the future. *arXiv preprint arXiv:2502.05253*, 2025.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18 (2):107–126, 2004.

A Models Evaluated on RMF

Table 1 lists every model evaluated in this study, with vendor-published training cutoffs, release dates, and per-cutoff n=3,000 sample sizes. Each family contains two variants, supporting a within-family scaling contrast: older→newer for the two Claude families (Sonnet, Opus) and smaller→larger for Qwen2.5. The Qwen2.5-7B-Instruct checkpoint runs in bf16; the 32B-Instruct-AWQ checkpoint is quantized to int4 via AWQ for single consumer GPU inference.

B Pairwise Cutoff Coverage for Version-Robustness Pairs

The version-robustness checks in §4.1 compare two pairs of Claude models that share an architecture family but differ in training-data cutoff: Sonnet 4.5 (2025-07) vs. Sonnet 4.6 (2026-01), and Opus 4.1 (2025-03) vs. Opus 4.6 (2025-08). Table 7 splits the n=3,000 sample by where each market's resolution date falls relative to both members of a pair: (i) *both pre*,

in training for both members and so a paired retrieval comparison; (ii) *split*, in training for the newer member only, the window in which any memorization-gap effect should appear; and (iii) *both post*, out of training for both members and therefore a paired comparison on genuine prediction. Counts are over the markets where both members returned a parseable prediction (the intersection on which paired Brier contrasts can be computed).

Table 7: Pairwise cutoff coverage on the n=3,000 sample, restricted to markets jointly parseable by both models in the pair. "Split" is the memorization-gap window: only the newer member could plausibly have ingested these outcomes during training. The Opus pair has very few split markets (n=51) because the n=3,000 sample was constructed under the shared 2024-12-31 / 2025-06-30 boundary and the Opus 4.1 to Opus 4.6 gap (2025-03 to 2025-08) falls almost entirely inside the excluded buffer; consequently, the Opus pair primarily contrasts both-pre and both-post performance, not the memorization-gap window.

Pair (older vs. newer)	Both pre	Split	Both post	Total
Sonnet 4.5 (cutoff 2025-07) vs. Sonnet 4.6 (cutoff 2026-01)	1,150	480	664	2,294
Opus 4.1 (cutoff 2025-03) vs. Opus 4.6 (cutoff 2025-08)	1,498	51	1,450	2,999

C Murphy Brier Decomposition

Table 8 presents the full Murphy decomposition for Opus predictions. The Brier score of a single model decomposes as BS = REL − RES + UNC, where reliability (REL; lower is better) measures calibration error, resolution (RES; higher is better) measures discriminative ability, and uncertainty (UNC) is the irreducible base-rate variance.

D Paired Brier Difference

For two models A and B, the paired Brier difference on the intersection I of markets where both made a prediction is

$$\Delta_{A-B} = \frac{1}{|I|} \sum_{i \in I} \left[(p_i^A - o_i)^2 - (p_i^B - o_i)^2 \right]$$

We report a 95% bootstrap CI by resampling markets in I with replacement (10K resamples) and taking the 2.5th and 97.5th percentiles of the recomputed Δ. We use the paired difference rather than the difference of aggregate Briers because pairing removes the variance contributed by question difficulty: even when two models share an aggregate Brier they may disagree on which questions are easy, and the paired statistic isolates the model-driven contribution from that question-driven contribution. We restrict to the intersection rather than imputing missing predictions because parse failures are not random across models, and treating them as either zero, the model's mean prediction, or a uniform 0.5 would each bias Δ in a different direction. The sign convention used throughout the paper is $A - B$ for "A minus B" contrasts (so $\Delta < 0$ when B has lower mean squared error and is therefore the better predictor). Table 9 reports the full 95% CI for every paired contrast in Table 3.

E Post-Cutoff Brier by Category and Model

Table 5 reports post-cutoff Brier by event category for all six models on the n=3,000 sample. Per-cell n is identical across models within a category (the sample is shared), with small variations from parse failures on Sonnet 4.6. Cells with $n < 20$ (entertainment, finance) are shown in italics as directional only. Within a single column (a model), the spread across categories is several times the spread across columns within a single row (a category), supporting the category-over-capacity claim in §4.2.

Table 8: Murphy decomposition of Opus Brier score on the n=3,000 sample. Resolution drops on novel events (less discrimination) but reliability improves (better calibration).

Component	Pre	Post	Δ
Reliability ↓	.023 [.016, .034]	.012 [.007, .021]	−.011
Resolution ↑	.048 [.040, .061]	.034 [.027, .045]	−.014
Uncertainty	.233	.215	−.018
Brier ↓	.207	.193	−.014

Table 9: Paired Brier difference with 95% bootstrap CI for the five contrasts in Table 3. n is the intersection size.

Contrast	n	Δ	[95% CI]
Sonnet 4.5 → 4.6	1,152	+.004	[−.004, +.011]
Opus 4.1 → 4.6	1,501	+.009*	[+.001, +.017]
Qwen 7B → 32B	1,502	−.024*	[−.035, −.014]
Sonnet 4.6 vs. Opus 4.6	1,152	+.004	[−.005, +.014]
Sonnet 4.5 vs. Opus 4.1	1,502	−.003	[−.011, +.004]

F Temporal Ablation

To test whether performance depends on the exact cutoff buffer, we sweep the buffer width from 0 to 360 days. Post-cutoff Brier remains in $[0.180, 0.195]$ across all buffer widths, and the pre/post gap stays small and negative throughout. There is no sharp "memorization cliff"—consistent with a gradual rather than step-function boundary between memorized and novel events.

G Pre/Post Brier Marginals Across All Six Models

Table 6 reports pre- and post-cutoff Brier with bootstrap 95% CIs (10K resamples) and the post-cutoff base-rate predictor for each model. All four Claude models significantly beat the base rate post-cutoff; Qwen 32B is indistinguishable from it; Qwen 7B is significantly worse.

H Salience-Tertile Breakdown

We split each model's markets into log-volume tertiles (low / mid / high) and recompute Brier within each cutoff condition (Table 10). The salience gap (high − low Brier) is non-positive pre-cutoff for all four Claude models and turns positive post-cutoff for each, so Δ_{gap} isolates a memorization-on-prominent-events effect: prominent markets are relatively easy when memorizable and relatively hard when novel. Qwen 32B shows the same pattern; Qwen 7B does not ($\Delta_{\text{gap}} = -0.022$), consistent with 7B being too weak to memorize even prominent events.

I Per-Event Cross-Model Agreement

Restricting to the intersection where all six models made a prediction (n=1,142 pre-cutoff, n=1,152 post-cutoff), Table 11 reports the mean per-event squared-error correlation across pairs of models, by pair group. The overall mean rises from 0.42 pre-cutoff to 0.62 post-cutoff: novel-event errors are shared across models, consistent with a shared base-rate hedging strategy. The lift is largest within the frontier Claude lineup (mean 0.53 → 0.75); the four Claude models converge tightly on the same predictions for events none of them have seen. Within Qwen the lift is much smaller (0.51 → 0.56), and cross-family Claude×Qwen pairs lie in between (0.32 → 0.53).

Table 10: Brier by log-volume salience tertile, cutoff, and model.

	low	mid	high	gap
Sonnet 4.5				
pre	.217	.176	.190	−.026
post	.195	.177	.214	+.019
Δ_{gap}				+.045
Sonnet 4.6				
pre	.221	.226	.216	−.005
post	.202	.171	.220	+.018
Δ_{gap}				+.023
Opus 4.1				
pre	.220	.199	.214	−.006
post	.196	.168	.209	+.013
Δ_{gap}				+.019
Opus 4.6				
pre	.222	.206	.209	−.014
post	.207	.182	.213	+.005
Δ_{gap}				+.019
Qwen 7B				
pre	.251	.261	.273	+.022
post	.243	.222	.244	+.000
Δ_{gap}				−.022
Qwen 32B				
pre	.247	.217	.229	−.018
post	.215	.207	.228	+.014
Δ_{gap}				+.032

Table 11: Mean per-event squared-error correlation by pair group, on the all-six-model intersection. Within-Claude correlations rise sharply post-cutoff; within-Qwen does not.

Group	n pairs	Pre	Post
Within Claude (4 models)	6	.53	.75
Within Qwen (2 models)	1	.51	.56
Cross-family (Claude × Qwen)	8	.32	.53
Overall	15	.42	.62

J Artifact Licenses

We will release the RMF data under CC-BY 4.0 and the evaluation code under the MIT License. The LLMs we evaluate are used under their respective vendor terms: Claude models via the Anthropic API under Anthropic's usage policies (Anthropic, 2024); Qwen2.5 7B-Instruct and 32B-Instruct-AWQ under the licenses distributed with their respective Hugging Face releases (Qwen Team, 2024). The underlying Polymarket data is public post-resolution market metadata; we accessed no non-public data and contributed no order flow to the platform. All uses are consistent with the artifacts' intended purposes: the model APIs and open-weight releases are distributed for research and downstream evaluation, the Polymarket record is public historical market metadata, and RMF is released for non-commercial research use in LLM forecasting evaluation.