Presentation Robustness for LLM Forecasters

Anonymous Authors[1]

Abstract

Language models are increasingly used as probabilistic forecasters for real-world events. A basic reliability question is whether equivalent descriptions of the same event and evidence yield similar probabilities. We study this through presentation robustness: for each binary forecasting market, we hold the target, outcome, and source identities fixed while changing either source-summary wording or question phrasing. On 200 resolved Prophet Arena target markets and four LLMs, equivalent presentations frequently change forecasts, including side flips across the 0.5 decision boundary. These changes predict forecast error, separate useful stability from uninformative uncertainty, and reveal failures hidden by single-prompt evaluation. Source-summary rewrites and question rephrasings expose complementary failure modes, showing that robustness to one wording change does not imply robustness to the other. Prompt averaging helps when alternate wordings move the model toward a strong reference forecast and can hurt when they move away. Our results establish presentation robustness as a practical evaluation axis for LLM forecasting, alongside accuracy and calibration.

1. Introduction

Open-domain forecasting is an attractive testbed for language models because predictions can later be scored objectively. Recent work asks whether language models can compete with human or market-based forecasters on real-world questions (Halawi et al., 2024; Yang et al., 2025). This raises a robustness question: should a forecast depend on superficial changes in how the same task is phrased?

We study *presentation robustness* for LLM forecasters: how much a forecast changes when the same forecasting task is written in an equivalent form. We keep the target market and outcome fixed, and vary wording in two ways. In a *source-summary* test, we preserve source identities and rewrite only the source summaries. In a *question-rephrasing* test, we omit sources and vary only the wording around the same event and target market. The resulting forecasts measure how far probabilities move and whether that movement predicts reliability.

The amount a forecast moves is informative. It distinguishes useful stability from forecasts that merely stay near 0.5, and it separates errors caused by source wording from errors caused by question wording. Averaging equivalent prompts helps most when it moves the model toward the market midpoint and hurts when it moves away.

Concretely, we evaluate four LLMs on 200 resolved target-market questions and collect several forecasts for each model-question pair. This reveals information that a single probability hides: sensitivity to source wording, sensitivity to question wording, stability with or without confidence, and movement toward or away from a stronger reference forecast. The point is not that every wording change is harmful, but that equivalent wordings expose uncertainty about the prompt itself.

2. Related Work and Positioning

Forecasting evaluations usually emphasize accuracy, calibration, and aggregation. Brier score is a standard proper score for binary probabilistic forecasts (Brier, 1950; Gneiting & Raftery, 2007); calibration work asks whether stated confidence matches empirical frequency (Guo et al., 2017). Prediction markets provide a strong comparison point because prices often summarize dispersed information (Wolfers & Zitzewitz, 2004). Recent LLM forecasting systems add retrieval, reasoning, and aggregation around language models and compare against human or market forecasts (Halawi et al., 2024; Yang et al., 2025). Work on prompt sensitivity has shown that LLM evaluations can change substantially under prompt-format or instruction-paraphrase choices, motivating reports over multiple prompts rather than a single template (Sclar et al., 2024; Mizrahi et al., 2024).

Our question is narrower: conditional on a fixed target market and fixed evidence, does the model give approximately the same probability under equivalent presentations? Unlike

[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.

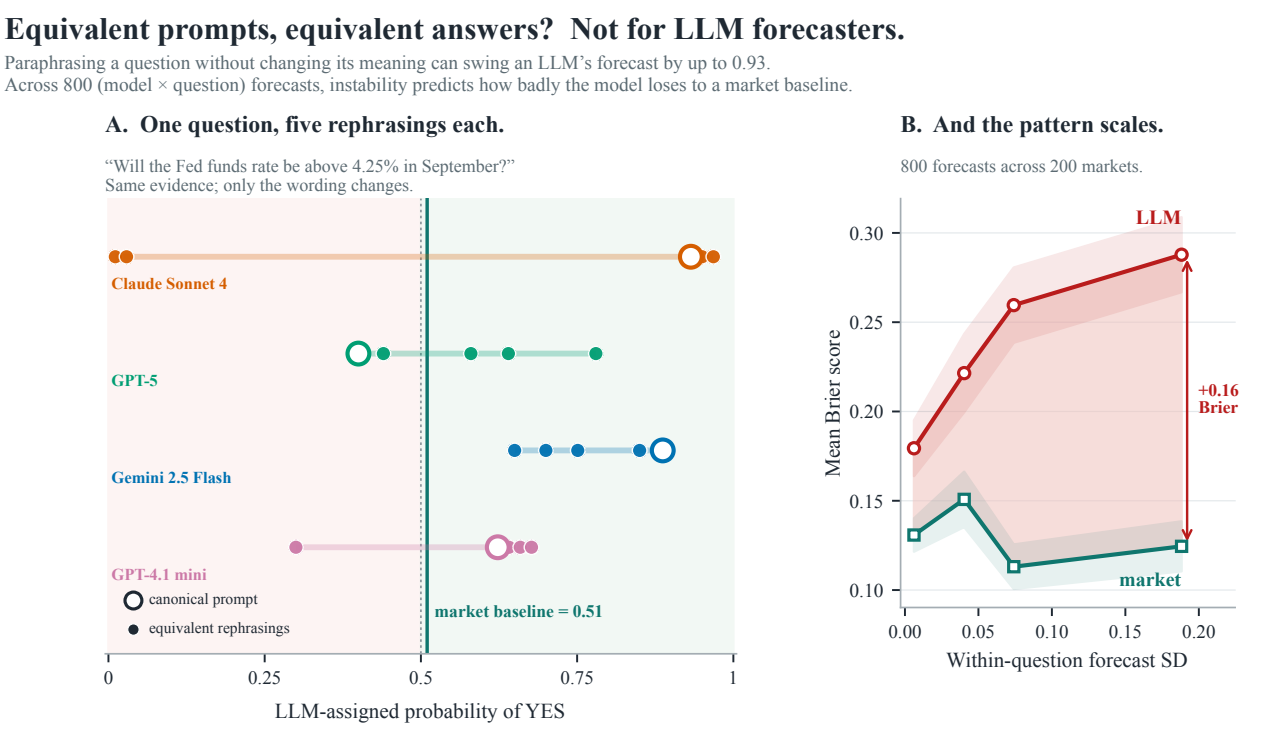


Figure 1. Equivalent prompts can produce sharply different forecasts for the same resolved market. Left: a real Fed funds target where only the prompt wording changes. Right: higher within-question rephrase instability corresponds to worse LLM Brier score relative to the market baseline.

ordinary accuracy or calibration evaluation, the output is a set of probabilities for the same resolved event. Compared with adversarial prompt-robustness benchmarks (Zhu et al., 2024), our variants preserve the target and available information.

3. Experimental Setup

Data. We use 200 resolved binary target-market questions sampled from the Prophet Arena processed subset. Each target has a canonical event title, target label, source context, snapshot time, market midpoint probability, and binary outcome. The sample resolves YES for 77 targets and NO for 123; a constant forecast at the empirical YES rate has Brier score 0.237, well above the market midpoint's 0.130. Splitting multi-market events into yes/no targets avoids conflating sibling markets.

Models. We evaluate four forecasters: GPT-5, GPT-4.1-mini, Claude Sonnet 4, and Gemini 2.5 Flash. For each model and target market we collect forecasts under two families of wording tests.

Wording tests. The source-summary test uses three source-context versions: the original ranked summaries and two rewritten versions generated to preserve the same evidence. The same rewrites are used for all forecasters; Appendix B summarizes generation and validation. The question-rephrasing test uses five source-free phrasings: one canonical question and four deterministic rephrasings. In both cases, the target event and YES/NO orientation are fixed.

Metrics. For each model-question row, let p_0 be the canonical forecast and $\bar{p}$ the mean over equivalent presentations. We compute Brier score $(p-y)^2$, variant standard deviation, range, side flips across 0.5, movement toward the market midpoint, and sharpness, which here means average distance from uncertainty, $|p-0.5|$. Binned analyses use variant standard deviation as instability and $|\bar{p}-0.5|$ as confidence. For Figure 2, stable/unstable and confident/uncertain are median splits within each wording family: source-summary thresholds are SD 0.017 and confidence 0.315; question-rephrase thresholds are SD 0.040 and confidence 0.286. For Figure 3a, high/low source and question fragility are within-model median splits on the corresponding SD, with "high" strictly above the median. We report pooled results over 800 model-question rows unless otherwise specified.

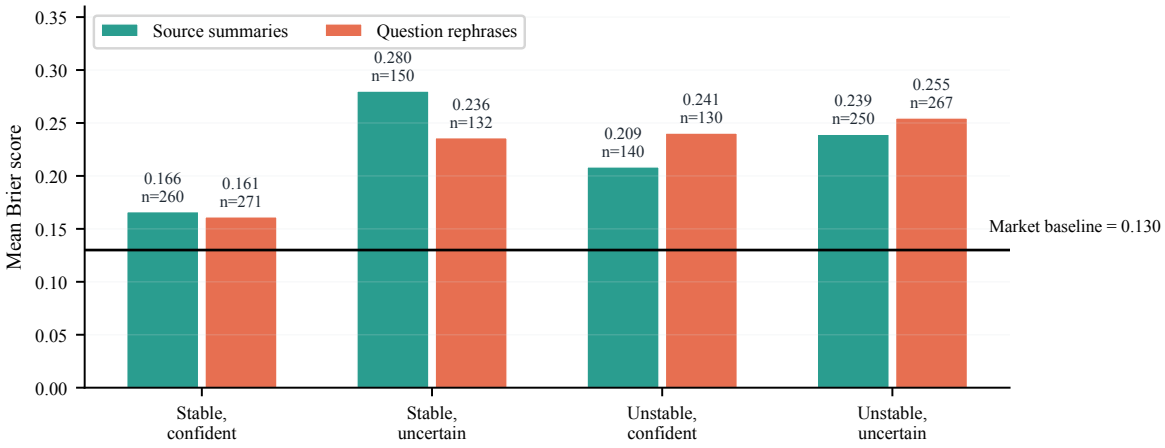
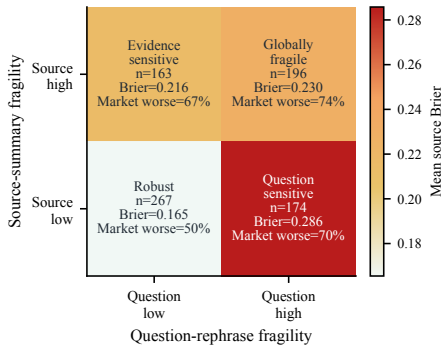
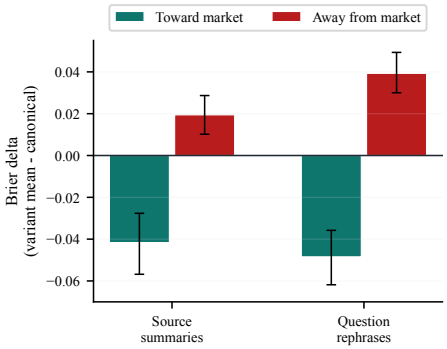


Figure 2. Forecast reliability depends on both stability and confidence. Stable, confident rows are the best LLM group; stable, uncertain rows can be much worse. The market midpoint remains substantially stronger overall.



(a) Source and question changes.



(b) When averaging helps.

Figure 3. Wording changes reveal structure. Left: source-summary and question-rephrase changes define distinct groups. Right: prompt averaging helps when it moves forecasts toward the market midpoint and hurts when it moves away.

4. Results

4.1. Stability Requires Confidence

Presentation stability is most useful when the model is also confident in the right direction. We group rows by variant movement and distance from 0.5, then compare mean Brier scores. Figure 2 shows that stable, confident rows are the best LLM group: source-summary forecasts have mean Brier 0.166, and question-rephrase forecasts have mean Brier 0.161. Stable, uncertain source-summary rows have mean Brier 0.280, worse than both unstable source groups.

A near-0.5 forecast can be stable for unhelpful reasons: a model may repeatedly return uncertainty even when useful evidence is available. The market midpoint remains the strongest comparison forecast in our study, with mean Brier 0.130.

4.2. Source Wording and Question Wording Fail Differently

The two sets of wording changes reveal different problems: some forecasts move under question rephrasing but not source rewrites, while others show the reverse. Figure 3a organizes these cases along the two axes. Rows stable under both tests have the lowest source-summary Brier score (0.165). Rows stable under source rewrites and unstable under question rephrasing are much worse (0.286), and rows unstable under both tests underperform the market most often: 74% have source-summary Brier worse than the market midpoint. The axes only weakly agree, so a single paraphrase test misses failures.

4.3. Averaging Helps When It Moves Toward the Market

Prompt averaging modestly improves average Brier score: source-summary averaging improves from 0.225 to 0.218,

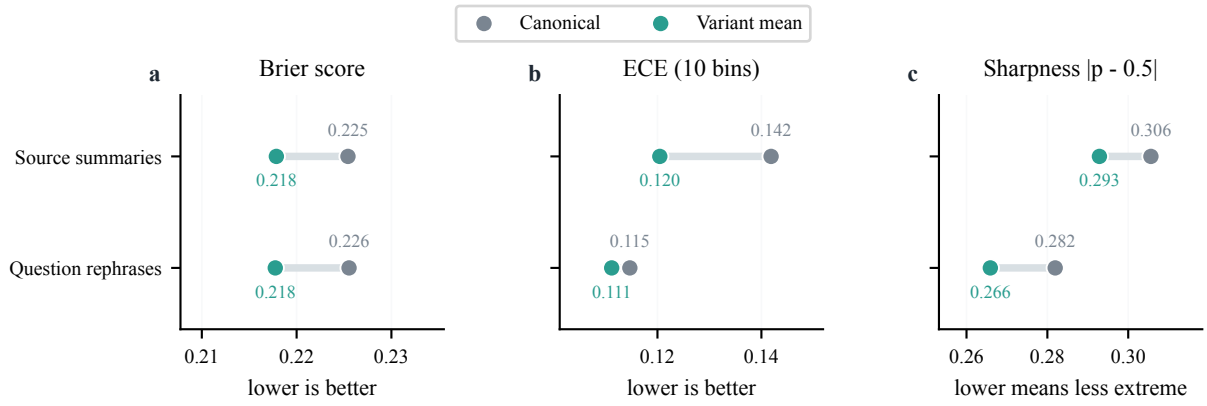


Figure 4. Prompt averaging improves Brier score modestly and moves forecasts toward 0.5. Source-summary averaging also reduces 10-bin calibration error, while question-rephrase averaging mainly reduces sharpness.

and question-rephrase averaging from 0.226 to 0.218. The gain is directional. When the variant mean moves the canonical forecast toward the market midpoint, Brier improves by 0.042 for source summaries and 0.048 for question rephrasings; when it moves away, Brier worsens by 0.019 and 0.039. The net gain is smaller because the rows are split across these cases: source-summary means move toward the market in 33% of rows, away in 32%, and tie in 36%; question-rephrase means move toward in 43%, away in 33%, and tie in 23%.

This makes prompt averaging useful as a correction for unusually bad canonical phrasings, but not as a uniformly safer forecast. Prompt averaging should therefore be paired with calibration or market-aware checks.

4.4. Averaging Moves Forecasts Toward 0.5

Prompt averaging also changes how extreme the forecasts are. Figure 4 shows that averaging lowers Brier score for both wording tests by about 0.008. For source summaries, it also lowers 10-bin ECE from 0.142 to 0.120; for question rephrasings, the ECE change is smaller. The most consistent effect is movement toward 0.5: mean sharpness falls from 0.306 to 0.293 for source summaries and from 0.282 to 0.266 for question rephrasings.

4.5. Concrete Prompt-Level Examples

The aggregate patterns correspond to concrete prompt-level failures. For an August 2025 CPI market with target "Above 2.4%," Claude Sonnet 4 assigns 0.75 to the canonical question; four equivalent rephrasings yield 0.15, 0.15, 0.15, and 0.00. For a September Fed funds market with target "Above 4.25%," Claude alternates between 0.95 and 0.02 under question rephrasing, while GPT-4.1-mini shifts from 0.05 on original source summaries to 0.96 on one source rewrite. Appendix D gives exact prompt text.

5. Discussion

Equivalent-presentation robustness gives a direct way to inspect LLM forecasting systems. A single forecast probability hides whether the model's answer is stable under harmless reformulations. Reporting the set of probabilities from equivalent presentations shows uncertainty about the prompt itself, identifies side flips, and separates errors caused by source wording from errors caused by question wording.

Overall, stability helps when the model is confident in the right direction. Stable uncertainty can still be a bad forecast. Some unstable rows improve after averaging because the alternate wordings move the forecast toward the market.

6. Limitations

The source rewrites are model-generated and validated, though some automatic checks flag possible imperfections; we therefore analyze source-summary and question-rephrasing results separately and expose exact example prompts in Appendix D. The market midpoint is a strong reference forecast, and many settings lack a market forecast. Finally, our sample covers 200 resolved target markets, so the grouping in Figure 3a should be tested on larger and more diverse forecasting streams.

7. Conclusion

LLM forecasters change under equivalent presentations. These changes are informative: they reveal when forecasts are stable, when they depend on source or question wording, and when prompt averaging is likely to help. Presentation robustness should therefore be part of evaluating LLM forecasting systems, alongside accuracy and calibration.

References

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

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.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models, 2024.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., and Stanovsky, G. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024.

Wolfers, J. and Zitzewitz, E. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004. doi: 10.1257/0895330041371321.

Yang, Q., Mahns, S., Li, S., Gu, A., Wu, J., and Xu, H. LLM-as-a-Prophet: Understanding predictive intelligence with prophet arena, 2025.

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Zhang, Y., Gong, N. Z., and Xie, X. PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts, 2024.

A. How the Wording Tests Are Built

Target-market unit. The unit of analysis is a resolved binary target market. If an event contains several sibling outcomes, each yes/no target is scored separately against its own market midpoint and resolution. This matters because paraphrase effects can otherwise be confounded with accidental changes in target orientation.

Source-summary variants. The source-summary test keeps the target market, source identities, and YES/NO orientation fixed. It compares the original ranked source summaries with two rewritten summary sets that are intended to preserve the same evidence while changing the wording. We analyze these separately from question rephrases because source rewrites can introduce small evidence-preservation errors; question rephrases are source-free.

Question-rephrase variants. The question-rephrasing test removes source context and varies only the surrounding phrasing of the same event and target market. Each model receives one canonical question and four deterministic rephrasings. A large movement or side flip here means the model's probability depends on how the same binary question is asked.

Row-level statistics. For each model-target row, the canonical forecast is p_0, and $\bar{p}$ is the mean over the equivalent presentations. We measure how much the variants move using their standard deviation and range. Side flips are crossings of the 0.5 decision boundary among equivalent presentations. We use Brier score $(p - y)^2$ for a binary outcome $y \in \{0, 1\}$ and compare LLM forecasts to the market midpoint.

Binning thresholds. The stability-confidence plot uses global median splits within each wording family. Source-summary rows are stable at SD ≤ 0.017 and confident at $|\bar{p} - 0.5| \geq 0.315$; question-rephrase rows are stable at SD ≤ 0.040 and confident at $|\bar{p} - 0.5| \geq 0.286$. The source-versus-question matrix instead uses model-specific SD medians to avoid letting one model dominate the high-fragility groups. The source/question SD medians are 0.000/0.000 for Claude Sonnet 4, 0.0236/0.0641 for Gemini 2.5 Flash, 0.0221/0.0490 for GPT-4.1-mini, and 0.0216/0.0399 for GPT-5; values strictly above the relevant median are labeled high fragility.

B. Source-Summary Rewrite Details

The source-summary rewrites were generated once, before forecasting, with OpenAI 'gpt-4.1-mini' at temperature 0. The rewrite prompt asked the model to preserve the factual claims, numbers, dates, entities, thresholds, and direction of evidence while producing two wording variants for each target market. The same rewrites were used for all forecasters. We ran automatic preservation checks and a separate 'gpt-4.1-mini' side-by-side judge; 88/400 rewrites passed all automatic checks and 395/400 were approved by the judge. We keep the automatic-check flags in the artifacts rather than filtering the main analysis, so the source-summary results should be read with this rewrite-quality caveat in mind.

C. Additional Figures

Figure 5 shows how the two families of equivalent presentations are generated. Figure 6 shows that side flips are more severe than ordinary probability movement. Averaging can rescue some bad canonical prompts when alternate variants move the forecast in the right direction.

D. Auditable Prompt-Level Examples

The main paper reports concrete failures because they help readers check the aggregate story. Tables 1 and 2 show the prompt text behind two of those failures.

CPI target. For the August 2025 CPI market with target "Above 2.4%," Claude Sonnet 4 moves from 0.75 on the canonical question to $0.15, 0.15, 0.15$, and 0.00 on four equivalent question rephrasings. The same model is stable at 0.85 across source-summary rewrites.

Fed funds target. For the September Fed funds market with target "Above 4.25%," Claude Sonnet 4 alternates between high and near-zero probabilities, including 0.95 and 0.02, under equivalent question rephrasing. GPT-4.1-mini shifts from 0.05 on original source summaries to 0.96 on one source rewrite.

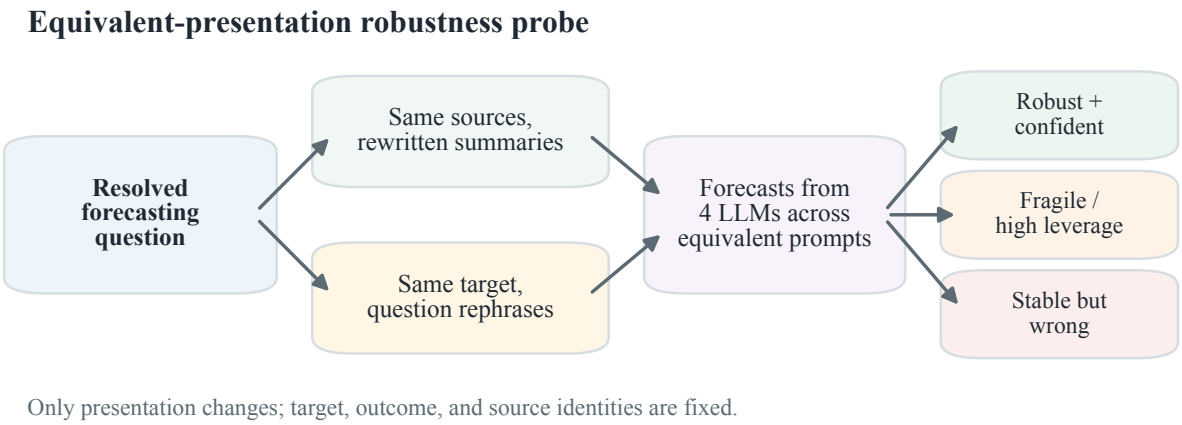


Figure 5. Equivalent-presentation test design. The target market, outcome, and orientation are fixed while either source-summary wording or question phrasing is varied.

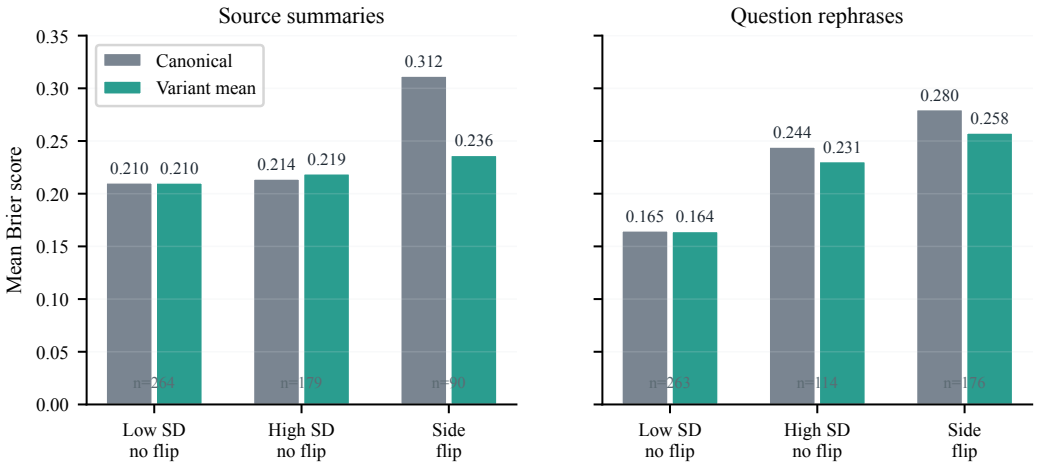


Figure 6. Side-flip behavior and prompt averaging. Side flips mark severe prompt dependence; variant means can improve over bad canonical prompts when the alternate presentations move the forecast in the right direction.

Market	Variant	p_{yes}	Exact question text
CPI above 2.4%	canonical	0.750	For the target market 'Above 2.4%', what is the probability that it resolves YES under the event 'Inflation in August 2025? (CPI YoY)'?
CPI above 2.4%	rephrase 01	0.150	For the event 'Inflation in August 2025? (CPI YoY)', estimate the probability that the target market 'Above 2.4%' resolves YES.
CPI above 2.4%	rephrase 02	0.150	What chance do you assign to the target market 'Above 2.4%' resolving YES for the event 'Inflation in August 2025? (CPI YoY)'?
CPI above 2.4%	rephrase 03	0.150	Give a calibrated YES probability for the target market 'Above 2.4%' in the event 'Inflation in August 2025? (CPI YoY)'.
CPI above 2.4%	rephrase 04	0.000	How likely is it that, for the event 'Inflation in August 2025? (CPI YoY)', the target market 'Above 2.4%' resolves YES?
Fed funds above 4.25%	canonical	0.950	For the target market 'Above 4.25%', what is the probability that it resolves YES under the event 'Fed funds rate in September?'?
Fed funds above 4.25%	rephrase 01	0.020	For the event 'Fed funds rate in September?', estimate the probability that the target market 'Above 4.25%' resolves YES.
Fed funds above 4.25%	rephrase 02	0.950	What chance do you assign to the target market 'Above 4.25%' resolving YES for the event 'Fed funds rate in September?'?
Fed funds above 4.25%	rephrase 03	0.950	Give a calibrated YES probability for the target market 'Above 4.25%' in the event 'Fed funds rate in September?'.
Fed funds above 4.25%	rephrase 04	0.020	How likely is it that, for the event 'Fed funds rate in September?', the target market 'Above 4.25%' resolves YES?

Table 1. Exact question rephrases for two side-flip examples. In each block, the event and target market are fixed and only the surrounding question wording changes.

Source	Original summary	Rewrite B
Fed preview	This article, published on September 16, reports that markets see a roughly 96% probability of a 25-basis-point (0.25%) rate cut at the Federal Reserve's September meeting, with only a small chance (4%) of a larger 50-basis-point move, providing clear, near-term market expectations relevant to predicting the Fed funds rate outcome.	On September 16, this article reports that markets price in roughly a 96% likelihood of a 25-basis-point (0.25%) rate cut at the Federal Reserve's September meeting, with a minor 4% chance of a 50-basis-point move, indicating clear short-term market expectations for the Fed funds rate.
Market impact analysis	Published today, this analysis emphasizes that markets are pricing in a high probability of a 25-basis-point cut at the September meeting, and discusses how Fed funds futures reflect these expectations, offering data-driven insights into probable rate outcomes.	This analysis, published today, stresses that markets are pricing in a strong probability of a 25-basis-point cut at the September meeting, with Fed funds futures reflecting these expectations and providing data-driven insights on likely rate outcomes.
Barron's meeting preview	Barron's outlines key considerations ahead of the Sept 16–17 Fed meeting, including inflation, labor market, and internal Fed dynamics–offers insight into what could sway the Fed's decision beyond consensus expectations.	Barron's highlights five key points to watch at the September 16–17 Fed meeting, including inflation trends, labor market status, and internal Fed factors, offering perspective on what might influence the Fed's decision beyond consensus views.

Table 2. Exact source-summary rewrite for the Fed funds source-side failure. For this same target, GPT-4.1-mini assigns $p_{\text{yes}} = 0.05$ with the original summaries and $p_{\text{yes}} = 0.96$ with Rewrite B.