Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Over-Reading the News: How LLM Forecasters Over-Update on Headlines, and What Survives Correcting It

Rudraneel Das

rudraneel93@gmail.com

Jadavpur University https://orcid.org/0009-0009-6173-0262

Research Article

Keywords: ArtificialIntelligence, LLM, Forecasting, AIResearch, MachineLearning, Research, Calibration, News

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11603009/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Over-Reading the News: How LLM Forecasters Over-Update on Headlines, and What Survives Correcting It

Rudraneel Das *rudraneel93@gmail.com*
Independent Researcher, India
ORCID: 0009-0009-6173-0262

Abstract

Forecasting language models are judged by accuracy, but how they use the news is rarely measured. We read the same questions with no evidence, with the real headlines and with unrelated placebo headlines, and, as controls, with one headline repeated and with headlines about a different event on the same topic. We use four open models from three families (7B to 120B parameters) and three proprietary frontier models. Open models over-update: headlines raise the mean forecast by 11–22 percentage points across four question sets, including a fresh set with registered predictions and independently written prediction-market questions, and by 15–21 points on 700 questions that were still open when read, a prospective test registered and externally timestamped in advance. Two label-free diagnostics derived from Bayesian updating flag the shift as excessive: treated as a prior and a posterior, the question-only and evidence forecasts should have equal means, and their movement should match the fall in uncertainty; neither holds. In the Qwen3 family the tested controls disfavour a volume-only account: repeating a headline lowers the push, paraphrases from other outlets add nothing to it, and same-topic headlines about other events produce a fifth of it (OLMo-2, by contrast, responds to same-topic coverage). The push needs headlines about the event, yet much of it is unrelated to the outcome (+0.20 on events that did not happen, +0.28 on those that did). In a limited test of three frontier models on 200 questions the push is smaller (6–10 points) and mostly tied to the outcome. Reading headlines one at a time and averaging halves the push without losing ranking skill. Auditing sixteen inference-time fixes and ten further rules on held-out questions, shrinking toward the base rate is the only family that passes our gate at the available power; with one consistent estimator, a measured gain beats a fixed halving clearly for Qwen3-32B and inconclusively for Qwen3-8B. Benchmarks built from headlines with open models therefore partly measure how far a model over-reads the news.

1 Introduction

Language models are increasingly used as probabilistic forecasters of real-world events, usually by reading recent news about each question (Halawi et al., 2024; Karger et al., 2025). Such systems are evaluated by their accuracy against outcomes. Accuracy alone cannot say *how* a forecaster uses the news: whether it moves its belief by as much as the news warrants, by too much, or in response to the mere fact that a topic is being covered. Human judges over-react to some kinds of information and judge an event likelier when instances come readily to mind (Tversky & Kahneman, 1973; Augenblick & Rabin, 2021). Whether language-model forecasters do the same matters for every system and benchmark built from headlines.

We answer this with a controlled design: the same model reads the same question with no evidence, with the real headlines, and with real but unrelated placebo headlines; further controls repeat a single headline and swap in headlines about a different event on the same topic. The answer is clear for the four open models we study, from three families and 7B to 120B parameters, and holds in weaker form for three proprietary

frontier models. Headlines about a question push forecasts toward YES by 11–22 percentage points. Two label-free diagnostics flag the movement as excessive: if the question-only and evidence forecasts are read as one forecaster's prior and posterior, their means should agree and their movement should match the fall in uncertainty, and neither holds. These are diagnostics under that reading, not proofs that the model violates Bayes' rule, since the two forecasts are separate samples. The controls narrow the cause. Under the tested conditions the push is not explained by how much coverage there is or by the topic: repeating one headline lowers it, and same-topic headlines about other events produce a fifth of it. The models respond to headlines about the event, but move much further than the outcome warrants, including on events that did not happen. The news nevertheless contains real information, which calibration recovers. These errors explain a second, broader finding. We tested every inference-time fix proposed for one production forecasting system, Kāla, over several weeks, on questions that played no part in its design, and screened further rules afterwards. Of 26 candidates only one family survives, a shrink toward the base rate. The over-update explains why: it creates a YES bias on questions built from headlines and over-dispersion on market questions, and the anchored shrink is the one simple correction that handles both.

Our contributions are:

1. **Over-updating on genuine news**: a registered, placebo-controlled measurement with genuine (non-adversarial) headlines and a question-only baseline. Headlines raise forecasts by 11–22 points on four question sets for Qwen3 at two sizes, by 19 points for OLMo-2 and by 12 points for gpt-oss-120b, including on prediction-market questions written independently of our pipeline. Two label-free diagnostics (equal means of prior and posterior; movement matching uncertainty reduction) flag the update as excessive under the reading of the two forecasts as prior and posterior. The push survives a neutral prompt, thinking mode and an explicit warning, and the news still adds information after calibration. In three proprietary frontier models the push is smaller and mostly tied to the outcome (Section 4, Figure 1).

2. **Content more than volume, in one family**: registered repetition, paraphrase and same-topic controls, replicated in Qwen3-32B; OLMo-2 responds to same-topic coverage and differs. Repeating a headline lowers the push, and same-topic headlines about other events give a fifth of it, rising with their similarity to the event. The push grows with the number of distinct headlines, and reading them one at a time halves it without losing ranking skill. Under the tested conditions the controls disfavour an account in which the volume or topic of coverage is itself taken as evidence; they do not isolate a general mechanism (Section 4.2).

3. **A correction without outcomes**: subtracting the evidence-induced shift, measured from paired reads alone, recovers about 70% of the attainable improvement where the question-only forecast is calibrated on average, and fails, as predicted, where it is not (Section 4.2).

4. **Two kinds of miscalibration**: on questions written from headlines the dominant error is a YES bias, on market and data-series questions it is over-dispersion; one anchored shrink corrects both (Section 4.1).

5. **A complete held-out audit** of sixteen inference-time fixes and ten further screened rules, with every failure, registration status, a taxonomy of four mechanisms behind the false positives, and multiplicity adjustment (Section 5, Appendix N).

6. **A registered replication at four times the scale** showing that miscalibration persists and that the larger model forecasts no better (Section 6).

7. **A theory that failed in public**: a proposition for the benefit of topic-level anchors, a quantitative prediction committed before the data existed, its failure on 1,557 new questions, a registered test of why, and the falsification of the resulting explanation on a fresh set (Appendices A and B).

8. **A label audit** of the automatically resolved questions, with every headline result re-estimated under audit-calibrated corrections (Section 7).

2 Related work

Evidence use by language models. Language models are sensitive to the presentation of their inputs, including the order of in-context examples, long contexts, answer options and ranked items (Lu et al., 2022; Liu et al., 2024; Pezeshkpour & Hruschka, 2024; Tang et al., 2024). Irrelevant context degrades their reasoning (Shi et al., 2023), and they readily adopt coherent external evidence that contradicts their own knowledge while favouring evidence that confirms it (Xie et al., 2024). For forecasting, Schoenegger et al. (2024) report an acquiescence effect: mean LLM forecasts well above 50% on questions that resolved evenly. Yuan et al. (2025) find that models revise forecasts in the right direction but by too little when given new information. Closest to our design, Lu & Zhang (2026) show that injecting a few LLM-written articles into the news corpus of retrieval-augmented forecasters built on open 7–8B models shifts probabilities by +0.13 to +0.22 beyond a neutral-article placebo, with effects growing with the number and query similarity of injected articles. Sarfati et al. (2026) find that internal activations are better calibrated than verbalised forecasts, and Turtel et al. (2026) separate bias, information and noise in LLM forecasts. Our contribution differs from this work in three ways: we study genuine, non-adversarial headlines in ordinary operation rather than fabricated ones; we compare against a question-only read of the same question, which locates the acquiescence effect in the evidence rather than the prior; we split the shift by outcome and test it against two label-free consequences of Bayesian updating (Augenblick & Rabin, 2021); and repetition and same-topic controls separate event content from the volume and topic of coverage. These controls argue against a pure availability account (Tversky & Kahneman, 1973) in which coverage itself is taken as evidence: the models respond to headlines about the event, but by far more than the outcome warrants. This differs from the under-updating reported by Yuan et al. (2025), whose models revised forecasts given information selected to be relevant to the outcome.

Forecasting, calibration and audits. ForecastBench (Karger et al., 2025) and Halawi et al. (2024) benchmark LLM forecasters against crowds; we instead hold the model fixed and audit the layer around it. Calibration of neural networks by temperature or logit scaling is standard (Guo et al., 2017; Zhao et al., 2021); shrinkage toward a base rate is the forecasting analogue of the attenuation familiar from human forecasting research (Tetlock & Gardner, 2015; Mellers et al., 2014). The rule that survives our audit is a constrained case of linear-in-log-odds recalibration (Turner et al., 2014), which is Platt scaling (Platt, 1999) applied to probabilities; the aggregation literature studies when forecasts should be made more or less extreme (Baron et al., 2014; Satopää et al., 2014), and non-parametric alternatives such as isotonic regression are standard in machine learning (Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005). Our finding that the right shrink depends on the question population connects to calibration within subgroups, studied as multicalibration (Hébert-Johnson et al., 2018) and in the fairness literature (Kleinberg et al., 2017), and to the degradation of calibration under distribution shift (Ovadia et al., 2019); topic-local anchors with partially pooled gains, which failed our external test, are a simple hierarchical form of group-wise recalibration. Our question is not whether recalibration helps but which constrained form survives held-out evaluation across question sets and model sizes, and how it compares with the many other fixes proposed for LLM forecasters. Closest to our concerns, Paleka et al. (2025) evaluate forecasters through logical consistency before resolution, and Paleka et al. (2026) catalogue evaluation pitfalls, notably temporal leakage through retrieval, misdated documents and unverified training cutoffs, which motivate our temporal checks (Section 3). A companion manuscript (Anonymous, 2026) analyses one failed fix, order ensembling, in detail; here we summarise it as one entry of the audit. Our contribution is empirical and methodological: which of these ideas survive held-out evaluation in a forecasting pipeline, and why the others do not.

3 Setup

Questions. Most comparisons use two benchmarks built with the same pipeline (cf. Karger et al., 2025): an LLM writes yes/no questions from news headlines about things not yet known at an as-of time; a search-grounded LLM later resolves them, and questions without cited resolution are dropped; each question's evidence is only headlines published before its as-of time. India-Bench covers Indian news (519 questions used); News-Bench Global covers world news (2,963 questions, base rate 0.22, median 15 evidence items, closing dates February 2025 to October 2026). Disjoint subsets form the evaluation sets: Stage 1 (1,516

Table 1: Model inventory. Revisions are Hugging Face commit hashes where recorded; "main" means the default branch at run time, not recorded as a hash. Replies are parsed by one rule (the last "Probability YES: p"); each n reported elsewhere counts questions with a parsed reply in every arm that table compares.

Model	Version	Inference	Used for	Question sets
Qwen3-8B	main (27–28 Oct 2026); `b968826d` for later reads	vLLM, fp16, 2×T4; later bf16, L40S	audit, coverage, controls, prospective	all
Qwen3-32B	`9216db57`	vLLM (unpinned), bf16, H100	scale replication, coverage	calibration, Stages 1, 4, 5, fresh
Qwen3-32B-AWQ	official 4-bit	vLLM, 2×T4	placebo	Stage 1
OLMo-2-7B-Instruct	`allenai/OLMo-2-1124-7B-Instruct`	vLLM, bf16, L40S	second family	Stage 1
gpt-oss-120b	reasoning effort low	vLLM, H100	larger reasoning model	Stage 1
Llama-3.1-8B-Instruct	unsloth mirror of Meta weights	vLLM, fp16 / bf16	order case study, prospective	order set, prospective
Qwen2.5-7B-Instruct	main	vLLM, fp16	order case study	order set
GPT-6.1-sol, Gemini-4-argon, Claude-Sonnet-5.5	provider defaults via API router	low reasoning effort, 700 tokens	frontier comparison	200 late fresh

questions from both benchmarks), Stage 4 and Stage 5 (800 untouched News-Bench questions each). Each set is split chronologically by closing date: the first 40% (dev) sets the base rate, the remaining 60% is the test set. The order-ensembling case study uses a separate set of 1,676 resolved market, macroeconomic (FRED) and Indian-news questions (Appendix P).

Models. Table 1 lists every model, its version, inference setting and use. All open models use temperature 0.7, top-p 0.8, top-k 20, a per-prompt seed, one fixed prompt asking for a final line "Probability YES: p" (Appendix E), thinking off unless stated. Some early questions close before a model's training cutoff, so the model may have seen the outcome; this affects both arms of every within-model comparison equally but could favour either model in comparisons between models.

Reads. The main forecast for each question is one read with its time-locked evidence. For the coverage analyses some sets were also read with no evidence (Stage 1 and the calibration set for both models; the external set for Qwen3-8B), and Stage 1 was read by Qwen3-8B with placebo evidence: the evidence list of a question from a different topic, all published before the question's as-of time, with the same number of items, chosen by a seeded random draw before any call. Content controls on 600 Stage 1 questions (Qwen3-8B) used the newest 1, 3 or 7 headlines; the newest headline listed 15 times, or the newest three cycled five times; and the headlines of a donor question from the same feed about a different event (no shared headline, low question-text overlap, all dated before the recipient's as-of time), chosen for the highest headline similarity.

Estimator. The primary estimator of a gain is Algorithm 1: on a calibration sample, one read per question, anchor z_b equal to the log-odds of the sample's YES rate, and g the one-parameter maximum-likelihood fit of $y \sim \sigma(z_b + g(z - z_b))$; at deployment the anchor is the target set's dev base rate. It is the rule a practitioner would deploy, and every headline claim about the measured gain uses it. The audit's originally registered gain of 0.37 was instead the coefficient on the news move in a logistic regression that also contained the question-only forecast. We report that registered analysis separately where it was registered (Table 5, row 6a; Section 6, H1–H2), because changing a registered estimator after the fact would hide the change; where the two disagree, the primary estimator decides the conclusion.

Temporal validity. Every test-split question has an as-of date after the public release of Qwen3 (29 April 2025), so no test outcome can be encoded in the model weights; only some dev-split questions, which set the base-rate anchor, are earlier (the earliest as-of date is 20 January 2025). Evidence is restricted to headlines first seen before the as-of time. In a separate sample of live questions about 14% were already decided when asked, and one of the 50 audited questions (Section 7) was; such questions remain in all arms.

Baseline and metric. The comparator is keep-half: $z^* = z_b + 0.5\,(z - z_b)$, where z is the model's log-odds forecast and z_b the log-odds of the dev base rate. We report the Brier score and paired differences (candidate minus comparator; negative favours the candidate). Probabilities are clipped to $[0.01, 0.99]$.

Uncertainty and gates. Confidence intervals come from a bootstrap that resamples groups of questions sharing a source and ISO week of closing (4,000 resamples), a proxy for event clusters; Section 5.2 checks the main comparisons, and the push, under event-level clusters built from question wording. A candidate is promoted only if its mean difference is at most -0.003 and its 95% upper bound is below zero. Two

Table 2: Bias versus over-dispersion (forward-in-time design; Brier on later questions). "Shift share": fraction of the anchored gain's improvement over raw forecasts obtained by the shift alone. Mean forecast, base rate and Platt slope are averaged over the training windows.

Set	Model	Mean forecast	Base rate	Platt slope	Raw	Shift only	Anchored gain	Shift share	Shift − anchored [95% CI]
News-Bench	8B	0.44	0.21	0.40	0.2509	0.2019	0.1934	0.85	+0.0084 [+0.0008, +0.0159]
News-Bench	32B	0.52	0.21	0.43	0.2803	0.2057	0.1970	0.90	+0.0087 [+0.0007, +0.0166]
External	8B	0.37	0.34	0.49	0.2417	0.2414	0.2220	0.02	+0.0194 [+0.0094, +0.0299]
External	32B	0.39	0.34	0.69	0.2216	0.2174	0.2096	0.35	+0.0078 [+0.0031, +0.0129]
Fresh (registered)	8B	0.40*	0.14	—	0.2278	0.1333	0.1230	0.90	—
Fresh (registered)	32B	0.44*	0.14	—	0.2512	0.1354	0.1226	0.90	—

*Over all fresh questions rather than averaged over training windows.

features of these intervals deserve mention. Paired differences between two smooth transformations of the same forecasts have small variance, so even small effects can be statistically clear; we therefore emphasise the size of each effect and its consistency across sets. And source-by-week groups are a proxy for events: questions about one event spread over several weeks or sources would make the intervals too narrow. The gate's minimum of 0.003 Brier (about 1.5% of a typical score) was fixed before the audit as the smallest improvement worth a second model call or a production change. We classify each comparison as *worse* (interval above zero), *survives* (passes the gate), *below threshold* (interval below zero but mean above -0.003), *inconclusive* (interval includes both a gate-sized gain and a harm), or *no detectable benefit* (interval excludes a gate-sized gain). We report raw intervals and, as a sensitivity analysis, Holm and Benjamini–Hochberg adjustments over the sixteen comparisons. Because Stage 1 was used exploratorily, Stages 4 and 5 form the primary pooled evidence and the three-set pool is a sensitivity analysis.

Registration status. We distinguish three levels honestly. V: protocol committed to version control before the analysis, with the commit order verifiable. L: protocol written in the lab log before the analysis but committed together with results, so the order rests on the log alone. E: exploratory or post-hoc analysis. Table 5 gives the level for each comparison.

4 Reading the news moves forecasts too far

This section first shows how the raw forecasts fail (Section 4.1) and then where the failure comes from (Section 4.2). Both use the forward-in-time design described in Appendix A: questions are ordered by closing date and every rule is fitted on earlier questions and scored on later ones.

4.1 Two kinds of miscalibration (exploratory)

The anchored shrink pulls forecasts toward the base rate, which corrects two different errors at once: forecasts that are too high (or low) on average, a *bias*, and forecasts that are spread too far from the base rate, *over-dispersion*. To separate them we compared, in the forward-in-time design of Appendix A, the raw forecasts, a pure log-odds shift $z + a$ that corrects only the bias, Platt scaling, and the anchored global gain, all fitted on earlier questions (Table 2; level E, computed after the external test).

The two populations fail differently. On News-Bench, whose questions an LLM wrote from headlines, both models forecast YES about twice as often as it happens (mean forecasts 0.44 and 0.52 against a base rate of 0.21), and a shift alone recovers 85–90% of the anchored gain's improvement. On the external questions, drawn from prediction markets and macroeconomic data series, the mean forecast is close to the base rate and the error is mostly over-dispersion (Platt slopes 0.49 and 0.69). "Over-reaction to the news" is therefore an accurate description only of the second population; on the first, the larger error is that the models treat the events in generated questions as far more likely than they are. The anchored shrink beats the shift in all four cases and matches Platt scaling, because it corrects both errors with one parameter; this, rather than a single model property, is why it survives where other fixes fail, and it explains part of why the right gain differs so much between populations. We registered a prediction for a fresh set built like News-Bench before any of its questions existed: the shift should recover more than half of the improvement and the

Table 3: Mean forecast with and without evidence (cluster-bootstrap 95% intervals). Placebo: real headlines taken from a question of a different topic, all published before the question's as-of time, with the same number of items. Log-odds: mean $\text{logit}(p_{\text{real}}) - \text{logit}(p_{\text{none}})$, which does not depend on the starting level; placebo minus none in log-odds is -0.79 $[-0.86, -0.71]$ for Qwen3-8B and $+0.08$ $[0.02, 0.15]$ for OLMo-2. n counts questions with parsed replies in the arms compared (for gpt-oss-120b, 1,507 have none and real reads; 1,506 also have a placebo read). The registered parser rejects ten 4-bit Qwen3-32B replies written as "**Probability YES:** p"; including them ($n = 1{,}513$) leaves every entry unchanged to three decimals.

Set	Model	Base rate	No evidence	Placebo	Real evidence	Real − none	Log-odds	AUC none / placebo / real
Stage 1 ($n = 1{,}515$)	8B	0.28	0.253	0.162	0.475	+0.222 [0.205, 0.240]	+1.28	0.56 / 0.57 / 0.64
Stage 1 ($n = 1{,}516$)	32B	0.28	0.338	—	0.542	+0.204 [0.188, 0.220]	+1.06	0.55 / — / 0.65
Stage 1, 4-bit ($n = 1{,}503$)	32B	0.28	0.329	0.280	0.531	+0.202 [0.185, 0.218]	+1.05	0.56 / 0.57 / 0.64
Stage 1 ($n = 1{,}492$)	OLMo-2-7B	0.28	0.326	0.339	0.520	+0.194 [0.178, 0.209]	+0.98	0.56 / 0.56 / 0.62
Stage 1 ($n = 1{,}507$)	gpt-oss-120b	0.28	0.267	0.236	0.382	+0.116 [0.102, 0.131]	+0.68	0.54 / 0.58 / 0.64
Stage 1, thinking on ($n = 597$)	8B	0.29	—	—	—	+0.183 [0.156, 0.210]	+0.98	0.56 / — / 0.65
Stage 1, warning prompt ($n = 600$)	8B	0.29	0.219	—	0.467	+0.248 [0.223, 0.273]	+1.36	0.51 / — / 0.63
Stage 1, neutral prompt ($n = 595$)	8B	0.29	0.365	—	0.572	+0.207 [0.172, 0.241]	+1.26	0.55 / — / 0.62
Calibration set ($n = 300$)	8B	0.20	0.238	—	0.452	+0.214 [0.181, 0.248]	+1.21	0.63 / — / 0.67
Calibration set ($n = 300$)	32B	0.20	0.355	—	0.518	+0.163 [0.126, 0.201]	+0.85	0.62 / — / 0.66
External ($n = 1{,}557$)	8B	0.36	0.256	—	0.392	+0.136 [0.117, 0.154]	+0.91	0.60 / — / 0.70
Fresh ($n = 952$)	8B	0.14	0.227	—	0.399	+0.172 [0.153, 0.191]	+1.11	0.58 / — / 0.67
Fresh ($n = 953$)	32B	0.14	0.327	—	0.441	+0.115 [0.095, 0.135]	+0.66	0.61 / — / 0.68

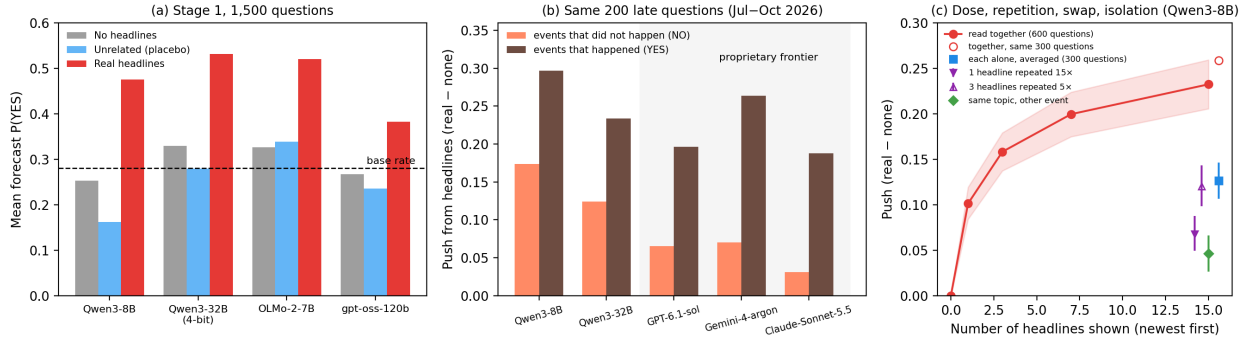


Figure 1: Headlines move forecasts too far. (a) Mean forecast with no headlines, with unrelated (placebo) headlines and with the real headlines, for four open models on Stage 1; dashed line: base rate. (b) The push from headlines split by outcome on the same 200 late questions (closing July–October 2026): the outcome-unrelated part (push on events that did not happen) is large for Qwen3 and small for three proprietary frontier models, while the outcome-related part is similar. (c) The push grows with the number of headlines read together (Qwen3-8B, 600 questions; band: 95% interval); reading each headline alone and averaging halves it (300 questions, with the joint read on the same questions shown as an open circle); repeating the newest one or three headlines up to 15 items gives less push than showing them once, and 14 same-topic headlines about another event give less than one headline about the event (Table 4).

mean forecast should exceed the base rate by more than 0.10. Both held for both models (shift share 0.90; mean forecasts 0.40 and 0.44 against a base rate of 0.14; Table 2). The fresh set's low base rate, partly a consequence of its stricter resolver, favours level corrections, so the size of the shift share there should not be over-read. Appendix N lists the other rules we screened.

4.2 Where the bias comes from: reading the news

Because some sets were also read without evidence, we can ask where the YES bias of Section 4.1 comes from. Every test below was registered before it was computed (Appendix Q); the external question-only reads and the placebo reads were generated after registration, and their analysis scripts were committed before the outputs existed.

The effect

Reading the news raises forecasts by 11–22 points in open models. In log-odds, which does not depend on where a forecast starts, the push is +0.66 to +1.28 across models and sets (Table 3), multiplying the odds of YES by about 1.9 to 3.6. Clustering questions by event rather than by source and week leaves the Stage 1 interval clear of zero (8B: $[0.199, 0.257]$). Without evidence, both models' mean forecast on Stage 1 is close to the base rate (0.25 and 0.34 against 0.28); with the headlines it rises to 0.48 and 0.54 (Table 3). The rise is nearly as large on questions that resolved NO (+0.20, +0.18) as on those that resolved YES (+0.28, +0.26): most of it is a shift that does not depend on the outcome, while a smaller part carries information (AUC rises from 0.56 to 0.64). The registered criteria for "the news induces the bias" were met for both models, and those for "the bias is already in the prior" were not. Across questions, the rise is unrelated to how many headlines a question happens to have (8B: +0.25 with 1–5 items, +0.22 with 11–15), but this observational comparison is confounded; when the number of headlines is manipulated directly (below), the push grows with it. On a fresh set of 952 questions built after these analyses with the same pipeline but a different question writer and resolver (Appendix B), the registered prediction of a rise above 0.10 held for both models (+0.172 and +0.115), again larger on YES than NO questions but substantial on both (+0.24/+0.16 and +0.19/+0.10).

Why the shift looks like an error, not just news. A rise on events that did not happen is not by itself irrational: a headline can rightly raise a 20% event to 40%, and the event can still fail. Three diagnostics that do not depend on the outcome split nevertheless flag the shift as excessive. They rest on an interpretive assumption: that the question-only forecast is the model's belief before the evidence, and the evidence forecast its belief after. Two separately sampled replies need not form one coherent belief state, the question-only reply need not be a calibrated prior, and one evidence bundle per question is a single draw rather than an average over possible evidence. The diagnostics therefore show behaviour inconsistent with a coherent, calibrated forecaster under this reading, not a direct violation of Bayes' rule. First, under it the expected belief after the evidence equals the belief before it (the law of iterated expectations); averaged over 1,500 questions, the mean forecast should barely move, yet it rises by 22 points. Second, the movement far exceeds the reduction in uncertainty that rational updating would produce (excess movement, below), and this test needs no labels at all. Third, the forecasts with evidence are miscalibrated in the large: a mean of 0.48 against a base rate of 0.28, or about 0.34 after the audit-calibrated label correction of Section 7, while the question-only mean (0.25) is close to it. Sampling noise accounts for little of the movement: with four reads per arm on 300 questions (per-read variance 0.007 without and 0.009 with evidence), excess movement computed from four-read means is 0.197 $[0.172, 0.224]$, against 0.207 from single reads. All three diagnostics point the same way.

It is not specific to generated questions. Questions written from headlines could inflate the effect, because the same news both selects the question and supplies its evidence. We had predicted that the rise would be small (below 0.10) on the external set, if it were an artefact of questions written from headlines. It was smaller, +0.136, but above that threshold, so the prediction failed. On the 1,150 prediction-market questions in that set (Polymarket, Manifold, Metaculus, Kalshi, INFER), which were written by others and independently of our headlines, the push is +0.153 $[0.133, 0.174]$ (exploratory breakdown). There it is more informative (rise +0.27 on YES and +0.10 on NO; AUC 0.57 to 0.72), it is negative only for macroeconomic series (−0.07), and it partly corrects a no-evidence forecast that is too low. Benchmark construction therefore changes how outcome-related the push is, but not whether it exists.

Unrelated news does not raise forecasts. To separate the presence of news from its content, the 8B model read each Stage 1 question with real headlines from a question on a different topic. If any news made the model say yes, the placebo should raise forecasts; instead it lowered them, to 0.16 (−0.091 $[-0.102, -0.080]$ against no evidence), with no ranking information (AUC 0.57). The registered content hypothesis was supported and the presence hypothesis rejected. The placebo shows that the push needs relevant headlines (and, for this model, that the absence of relevant headlines lowers the forecast); it cannot say whether the model responds to the topic being covered or to what the headlines say about the event. The next controls separate the two.

Table 4: Content controls (Qwen3-8B, 600 Stage 1 questions, base rate 0.29; cluster-bootstrap 95% intervals). Push: mean forecast minus the question-only mean (0.251). Repetition and swap arms were registered before any read (Appendix Q).

Headlines shown	Items	Push [95% CI]	AUC
None	0	—	0.53
Newest one	1	+0.101 [0.084, 0.119]	0.58
Newest one, repeated	15	+0.068 [0.049, 0.088]	0.58
Newest three	3	+0.158 [0.138, 0.179]	0.58
Newest three, repeated	15	+0.120 [0.098, 0.143]	0.63
Newest one + 14 paraphrases, other outlets	15	+0.098 [0.075, 0.121]	0.57
Same topic, other event	∼14	+0.047 [0.027, 0.066]	0.57
All real headlines	∼15	+0.232 [0.206, 0.260]	0.64

Content more than volume. The placebo and the dose test leave open several accounts: the model might respond to the amount of coverage, to repetition of one narrative, to topical salience, or to the event-specific content of the headlines. We registered two controls on 600 Stage 1 questions (Qwen3-8B; Table 4). *Repetition* holds the content fixed and multiplies the coverage. Listing the newest headline 15 times moves the forecast *less* than listing it once (push +0.068 against +0.101; difference −0.033 [−0.046, −0.021]), and cycling the newest three five times moves it less than listing them once (−0.038 [−0.051, −0.025]). *Same-topic swap* gives each question the headlines of another question from the same feed about a different event, chosen for the highest headline similarity, with no shared headline and all dated before the question's as-of time. It produces a push of only +0.047 [0.027, 0.066], a fifth of the real headlines' +0.232. That is less than a single headline about the event achieves, although about 14 swapped headlines are shown. The swap's push rises with how similar the donor's headlines are to the question's own (−0.017, +0.055 and +0.103 by similarity tercile), and so does its ranking skill (AUC 0.54, 0.56, 0.61). A hand check of 30 donors found 4 about the same event, all among the most similar, so the top tercile partly measures real information. Both registered hypotheses that would have supported a coverage account failed. Identical copies are not realistic coverage: a model may recognise and discount duplicates, which fifteen independent outlets would not trigger. The controls therefore make a volume-only account less plausible under the tested conditions, and point to distinct headlines about the event as the driver of the dose effect; they do not establish a general distinction between content and coverage. Two registered follow-ups sharpen and limit the picture. Making the corroboration realistic, the newest headline plus 14 paraphrases attributed to different outlets moves the forecast exactly as much as the headline alone (+0.098 against +0.101; difference −0.004 [−0.018, +0.012]) and more than 15 identical copies (+0.030 [0.016, 0.043]): apparent independent corroboration adds nothing beyond the one fact, and the drop with copies comes from recognising duplicates. Across models, repetition lowers the push in Qwen3-32B (−0.048) and OLMo-2 (−0.032), and the swap gives almost nothing in Qwen3-32B (5% of the real push) but more than half of it in OLMo-2 (53%). The content-over-volume pattern therefore holds in the Qwen3 family; OLMo-2 also responds to coverage of the topic.

How general it is, and what changes it

A second model family, and a neutral prompt. OLMo-2-7B, from a different developer and training corpus, shows a push of +0.194 [0.178, 0.209], again barely related to the outcome, and unrelated headlines leave it nearly unchanged; replacing the instruction to consider "how the evidence moves" the base rate with "Give your forecast" leaves Qwen3-8B's push almost intact (+0.207 against +0.232 on the same questions; details in Appendix K).

A larger reasoning model, thinking, warnings, and dose. Four further registered tests used new reads. gpt-oss-120b, a 120-billion-parameter reasoning model from a third developer, shows a smaller push (+0.116 [0.102, 0.131]; +0.68 in log-odds) that is more related to the outcome (+0.18 on YES against +0.09 on NO questions). For Qwen3-32B, run in its official 4-bit version, the push matches the full-precision run (+0.202) and unrelated headlines lower its forecast (−0.050). Thinking mode cuts Qwen3-8B's push only by about a fifth (+0.183 against +0.233), failing our registered criterion of a halving, and an explicit warning

that coverage is not evidence does not reduce it at all (+0.248 against +0.232). Showing only the newest 1, 3 or 7 headlines gives a push of +0.10, +0.16 and +0.20 against +0.23 with all of them (Figure 1c): the push grows with the number of distinct relevant headlines when it is varied directly, as Lu & Zhang (2026) report for fabricated articles, and ranking skill grows with it (AUC 0.57 to 0.64). Repeating the same headlines does not have this effect (above).

Proprietary frontier models. We then ran three current proprietary models, one from each major provider (GPT-6.1-sol, Gemini-4-argon and Claude-Sonnet-5.5, through a commercial API router), on the 200 latest fresh-set questions (closing July–October 2026, to limit the chance that outcomes were in their training data), with no, real and placebo headlines; the test was registered before any call, with a spending cap. All three show a push (+0.088 $[0.052, 0.123]$, +0.103 $[0.063, 0.144]$ and +0.058 $[0.030, 0.088]$) and no rise from unrelated headlines (−0.021, −0.022, −0.001), meeting both registered criteria. But their push is much more tied to the outcome (Figure 1b): on events that did not happen it is +0.066, +0.070 and +0.031 (all intervals above zero), against +0.174 and +0.124 for Qwen3-8B and -32B on the same questions, while on events that did happen it is similar (+0.20, +0.26, +0.19 against +0.30 and +0.23). For Claude-Sonnet-5.5 the evidence improves the raw Brier score (0.122 to 0.112); for the Qwen models on the same questions it worsens it. Their question-only forecasts already rank outcomes well (AUC 0.67, 0.69 and 0.79, against 0.56 and 0.60 for Qwen3), which may reflect better priors or some knowledge of outcomes; the outcome split is less affected by such leakage than the level of the forecasts, but it cannot rule it out.

Smaller in the frontier models we tested. Across the seven models the outcome-unrelated part of the push falls steeply from the smaller open models to the frontier models, from about 17 points (Qwen3-8B) through 9 (gpt-oss-120b) to 3–7 points for the frontier models, while the outcome-related part changes little. The frontier models' updates are therefore much more tied to the outcome. This is a limited comparison: 200 questions, three models, and question-only forecasts that may reflect knowledge of outcomes. We cannot attribute this to capability as such: size, training data, post-training and provider-side settings all differ between these models, and the frontier test used 200 questions. None of the models we tested is free of the effect, and within one model it grows with the number of distinct headlines about the event. For forecasting systems built on open models, which is how most published LLM forecasting pipelines and benchmarks are run, the effect is large; for frontier models it is a smaller bias that calibration can absorb.

Changing how the news is read. Because the push grows with the number of headlines read together, we registered two ways of changing the reading itself, on 300 Stage 1 questions with Qwen3-8B. Reading each headline on its own and averaging the resulting forecasts halves the push (+0.126 $[0.106, 0.146]$ against +0.258 $[0.220, 0.298]$ for the joint read) while keeping ranking skill (AUC 0.615 against 0.621), and its raw Brier score (0.217) is better than both the joint read (0.275) and the question-only read (0.241) and close to the base rate (about 0.211). Asking the model instead to label each headline YES, NO or NEUTRAL before forecasting makes matters worse: the push rises to +0.351 $[0.305, 0.395]$ and the raw Brier score to 0.327. The model follows its own labels (on the 57 questions it labels with no YES its forecast falls by 0.12), but its labels lean YES (1,411 YES against 664 NO). Isolated reading costs one call per headline, and Lu & Zhang (2026) show that an adversary who controls most of the retrieved articles can defeat it; for ordinary news, it is the one change to the reading that reduced the push without losing information.

A label-free test of rational updating. Augenblick & Rabin (2021) show that for a Bayesian, the expected squared movement of a belief equals the expected reduction in its uncertainty, $p(1-p)$; movement beyond that is excess. Treating the question-only and evidence reads as a belief before and after the evidence, this gives a test that needs no outcome labels. Every open model moves far more than its uncertainty falls; in fact uncertainty rises on average when the evidence is read (excess movement 0.175, 0.141, 0.140 and 0.111 for Qwen3-8B, Qwen3-32B, OLMo-2 and gpt-oss-120b on Stage 1, all intervals above zero). Sampling noise accounts for little of this: between two reads with identical evidence the excess is 0.019 (8B) and 0.022 (32B), leaving 0.156 $[0.143, 0.168]$ and 0.119 $[0.109, 0.129]$. On the 200 late questions the frontier models show less excess (GPT-6.1-sol 0.072, Gemini-4-argon 0.111, Claude-Sonnet-5.5 0.044) than Qwen3-8B (0.167) on the same questions, and their placebo reads show almost none (0.004, 0.011, 0.000). Excess movement thus orders the models the same way as the outcome-based split, without needing outcomes, which makes it usable for

monitoring a forecaster before its questions resolve. Two caveats apply: the two reads are separate samples rather than one agent's belief before and after the evidence, and for the frontier models we have no repeat reads with which to measure their noise floor, although their near-zero placebo excess suggests it is small.

What did not explain or reduce it. Masking the question's words in the headlines, listing every headline twice, showing the model its own prior, and eliciting per-headline likelihoods all failed their registered criteria (Appendix K).

Negation (exploratory). Two registered attempts to separate belief change from agreement used the complement of each question ("Will this fail to happen?" and, more naturally, "Will X avoid being charged. . . "). Both failed their manipulation checks, so neither test counts: even with correct natural complements (29 of 30 in a hand check), the model's forecasts for a question and its complement are nearly uncorrelated (-0.13; a coherent forecaster would be close to -1). The failure is informative in itself. Reading the same headlines, the model raises $P(X)$ by 0.23 and $P(\text{not } X)$ by 0.06, and on complements its forecasts with evidence lose all ranking skill (AUC 0.51, against 0.64 on the original questions). Its use of the evidence follows the wording of the question more than the event.

A prospective test. To test the push on questions whose outcomes did not yet exist, we added question-only reads to a separate prospective study: 700 open questions from a live forecasting feed, resolving between November and December 2026, whose evidence reads had already been made and externally timestamped. The new reads were registered before they were made, and their file was timestamped by an RFC 3161 authority while none of the 700 questions had resolved. Both registered label-free predictions held, for Qwen3-8B and for Llama-3.1-8B, a fourth model family: the push is $+0.205$ $[0.171, 0.241]$ and $+0.151$ $[0.115, 0.192]$, and excess movement is 0.150 $[0.121, 0.181]$ and 0.116 $[0.088, 0.150]$. The outcome-dependent predictions (push on questions that resolve NO, calibration in the large, and the primary estimator against keep-half) are registered and will be scored when the questions resolve.

What the push responds to. On Stage 1 the push rises with how closely the headlines match the question's wording, although that match does not predict the outcome there; on the external market questions the match does predict the outcome, but the push does not rise with it. Questions with a headline pointing toward NO still receive most of the push (about three quarters). The push itself needs no outcome labels, but statements about the bias do, and under the audit-calibrated label correction every label-dependent conclusion keeps its direction (Appendix K).

Consequences for correction

The news still carries real information. None of this means the news is useless to these models. After each read is calibrated with an anchored gain fitted on earlier questions, the read with evidence beats the question-only read on every set (-0.015 $[-0.023, -0.007]$ and -0.012 $[-0.019, -0.006]$ on Stage 1; -0.028 $[-0.043, -0.013]$ on the external set), and ranking skill rises with evidence everywhere (Table 3). The models extract information from the news but convert it into probabilities on a badly shifted and over-stretched scale.

A correction without outcomes, and its limit. If the bias is an evidence-induced shift, it can be measured without outcomes, from paired reads with and without evidence, and subtracted. In the forward design on Stage 1, subtracting the mean shift estimated on earlier questions beat raw forecasts by 0.046 $[0.026, 0.066]$ (8B) and 0.058 $[0.041, 0.076]$ (32B), recovering 99% and 89% of the improvement of an outcome-fitted shift and about 70% of the improvement of the anchored gain, which needs outcomes and remains better. On the fresh set, registered before its questions existed, it beat raw forecasts for both models (-0.067 $[-0.080, -0.053]$ and -0.054 $[-0.063, -0.045]$), although our stated condition for success (a question-only mean within 0.10 of the base rate) held only for the 8B model. On the external set the same correction failed ($+0.007$ $[-0.009, +0.022]$), as we had predicted before computing it: there the no-evidence forecast is itself too low, so removing the evidence-induced shift over-corrects. The correction is therefore useful exactly

Table 5: All comparisons. Δ = Brier(candidate) − Brier(comparator) on held-out questions; negative favours the candidate. Reg.: V verifiable registration, L logged, E exploratory (Section 3).

#	Candidate	Comparator	N	Δ [95% CI]	Reg.	Verdict
Evidence selection (Qwen3-8B, Stage 1 test)						
1	Relevance-filtered evidence	All evidence	909	$+0.0090$ $[-0.0028, +0.0205]^a$	V	inconclusive
2	Relevance-filtered evidence	Random subset, same size	909	-0.0004 $[-0.0054, +0.0044]$	V	no detectable benefit
3	Evidence with protected minimum	All evidence	909	$+0.0004$ $[-0.0083, +0.0093]$	V	inconclusive
Update rules (Qwen3-8B)						
4	Gain on move from no-news forecast	Keep-half	909	$+0.0127$ $[+0.0026, +0.0228]$	L	worse
5	Variance-scaled gain > 1	Keep-half	909	$+0.0191$ $[+0.0129, +0.0253]$	E	worse
6a	Anchored shrink, registered regression gain 0.37	Keep-half	1,869	-0.0046 $[-0.0076, -0.0016]^b$	L, E^e	survives
6b	**Anchored shrink, primary estimator (gain 0.23)**	Keep-half	960	$\mathbf{-0.0049}$ $[\mathbf{-0.0126}, \mathbf{+0.0028}]^f$	E	**inconclusive**
7	Lead-time-dependent gain	Gain 0.37	480	$+0.0018$ $[-0.0041, +0.0078]$	L	inconclusive
8	Two-channel (prior and news gains)	Gain 0.37	480	-0.0007 $[-0.0040, +0.0028]$	V	inconclusive
9	Disagreement-adaptive gain	Constant gain	909	$+0.0004$ $[+0.0000, +0.0008]$	V	no detectable benefit
Multiple reads (Qwen3-8B)						
10	Second read, repeated	One read	909	-0.0013 $[-0.0039, +0.0013]$	L	inconclusive
11	Second read, reversed order	One read	909	-0.0026 $[-0.0059, +0.0003]$	L	inconclusive
12	Reversed second read	Repeated second read	909	-0.0013 $[-0.0045, +0.0019]$	L	inconclusive
13	Reversed read added to gain 0.37	Gain 0.37 alone	480	-0.0025 $[-0.0060, +0.0008]$	L	inconclusive
Market fusion and output handling						
14	Fused market + model, weight fitted earlier	Market price alone	142	$+0.0005^c$	L	no benefit (no CI)
15	Drop exact 0/1 reads (Qwen2.5-7B)	Forward $\times 3$	1,676	-0.0007 $[-0.0062, +0.0045]$	V	inconclusive
16	Data series as a table (Qwen2.5-7B, FRED)	Chronological list	190	-0.017 $[-0.064, +0.012]^d$	V	inconclusive

^{a}97.5% CI (two co-primary contrasts). ^{b}Pooled over three disjoint sets (Table 14). ^{c}Chronological split, no CI computed. ^{d}Clustered by data series (20 series). ^{e}L for Stages 4 and 5, E for Stage 1 (Table 14). ^{f}Primary pool, Stages 4 and 5 (Table 15); for Qwen3-32B the same comparison is -0.0132 $[-0.0228, -0.0035]$. Both rows apply a shrink toward the base rate, as does the comparator; keep-half itself improves on raw reads by far more (Table 16). With Holm adjustment over the sixteen comparisons only rows 5 (worse, $p_{\text{Holm}} < 0.001$) and 6a ($p_{\text{Holm}} = 0.037$) remain significant; Benjamini–Hochberg gives the same two.

where the model's no-evidence forecast is calibrated on average, a condition that a few resolved questions can check.

5 What corrects it: a held-out audit of sixteen fixes

Table 5 lists all comparisons; Appendix L shows them graphically (Figure 6) together with the design of the audit (Figure 5). One family passes: shrinking toward the base rate, whether by the fixed halving that serves as comparator or by a measured gain; most others are inconclusive at the precision available, two show no detectable benefit, and two are significantly worse. A failure to clear the gate is not evidence that a fix is useless; Table 5 distinguishes the two. Two further post-hoc ideas, a selective revision gate and an adaptive-curvature shrink, changed Brier by $+0.00003$ and $+0.0004$ on Stage 5 and are not counted. The order-ensembling claim that motivated much of this work is treated in Appendix P.

5.1 The survivor: shrinking toward the base rate

The rule is $z^* = z_b + g\,(z - z_b)$, with z_b the base-rate anchor; keep-half is $g = 0.5$. Every shrink with g well below one turns the raw reads' negative Brier skill positive (Table 16); this is the large and robust effect. Whether a gain *measured* on resolved questions beats the fixed halving is a smaller question, and its answer depends on the model. With the primary estimator (Algorithm 1, 300 calibration questions) the gains are 0.23 for Qwen3-8B and 0.18 for Qwen3-32B. On the primary held-out sets the measured gain beats keep-half clearly for 32B (-0.0132 $[-0.0228, -0.0035]$) and only inconclusively for 8B (-0.0049 $[-0.0126, +0.0028]$; Section 5.2). We therefore claim that the models need shrinking toward the base rate far more than by one half, and that a measured gain is the better rule for 32B; for 8B a fixed halving cannot be ruled out as equally good.

The registered analysis. The audit registered a different estimator: a regression of outcomes on the base-rate log-odds and the model's news move on the same 300 questions, which gave 0.37 for Qwen3-8B (95% CI 0.14–0.65) and 0.58 for gpt-oss-120b. The value 0.37 beat keep-half on all three held-out sets (Table 14, Figure 7); the pooled estimate excludes zero and the three sets agree (Cochran's $Q = 1.47$, 2 df). On Stage

5 it also won in 200 of 200 random re-splits of dev and test. For gpt-oss the measured gain is close to 0.5, so keep-half is nearly optimal for that model; the gain differs between these two models, although Section 6 shows that a single calibration set of 300 questions cannot reliably separate gains that are closer together.

Why it works. Read noise explains only 37–40% of the shrinkage the models need; the rest is systematic over-scaling of the news, which is why one constant captures most of the improvement and per-question adaptive gains find nothing more (Appendix M).

Figure 8 shows where the improvement comes from: keep-half leaves forecasts above 0.3 systematically too high, and the smaller gain pulls them onto the diagonal without losing resolution at the low end.

5.2 The primary estimator, end to end

Table 15 gives the held-out comparison with Algorithm 1 for both models (level E, because the estimator was fixed after the registered analysis). It gives smaller gains than the registered regression, 0.23 for 8B and 0.18 for 32B, so it now orders the two models as their held-out optima do (Section 6). On the primary sets the 32B result strengthens (-0.0132), while the 8B result keeps its sign but loses significance. On Stage 1 both gains are worse than keep-half: that set mixes Indian questions, whose best gain is about 0.33, with global questions under one anchor, which is exactly the failure that population-local anchors address (Appendix A). Coarser clusters widen the intervals only slightly (8B: source–month $[-0.0123, +0.0032]$, topic–month $[-0.0123, +0.0032]$; 32B: source–month $[-0.0232, -0.0033]$, topic–month $[-0.0232, -0.0033]$). Clustering by event instead, joining questions whose wording is similar (TF-IDF cosine at least 0.4 or 0.6, alone or together with source–week), leaves both conclusions unchanged: the 8B interval at its widest is $[-0.0138, +0.0033]$ and the 32B upper bound stays below zero (-0.0021 at its widest; exploratory).

Calibration, not discrimination. For $0 < g < 1$ the shrink is a monotone transformation, so it changes only calibration: raw reads have Brier skill -0.21 and -0.34 against the base rate, and any reasonable shrink makes it positive, while AUC stays at 0.67 and 0.66 (Table 16).

5.3 Comparison with standard recalibration (E)

Platt scaling and isotonic regression fitted on the same calibration set are statistically indistinguishable from the measured shrink on average but swing widely between sets, because their intercepts carry the calibration set's base rate into sets whose base rates differ; temperature scaling, which has no base-rate anchor, is far worse (+0.04 Brier). What makes the shrink robust is the anchor (details in Appendix M, Table 19).

5.4 Four mechanisms behind the failures

1. Adaptive rules fit noise. Rules that let the gain depend on lead time (row 7), on the no-news prior separately from the news move (row 8), or on read-to-read disagreement (row 9) add parameters without adding signal. The two-channel rule confirmed our prediction that the news move is over-weighted more than prior knowledge (fitted 0.30 versus 0.39) yet gained nothing; disagreement between identical reads carries no information about how far to trust them.

2. Degenerate outputs and analysis choices create gains. Some apparent improvements come from a few exact 0 or 1 outputs, the clipping threshold that converts them to log-odds, and the choice of calibration. Appendix P summarises the clearest case: an apparent gain from shuffling evidence order that came almost entirely from exact 0/1 outputs on chronologically listed data series.

3. In-sample fusion weights do not travel. On 236 questions with market prices, an error-covariance weight fitted on all questions suggested that adding the model to the market helps. Fitted on the earliest 40% and applied to the later 60%, the weight was slightly negative and the fusion was worse than the market alone (row 14).

Table 6: Registered scale replication. Δ = Brier(candidate) − Brier(comparator), pooled over the three held-out sets unless stated (fixed effect over per-set cluster bootstraps).

	Hypothesis	Qwen3-8B	Qwen3-32B	Verdict (32B)
	Gain g on calibration set [95% CI]	0.37 [0.14, 0.65]	0.46 [0.21, 0.75]	
	Reliability $\hat{\lambda}$ / over-scaling $\hat{\kappa}$	0.77 / 0.48	0.71 / 0.65	
H1	Own gain vs keep-half	−0.0046 [−0.0077, −0.0016]	−0.0037 [−0.0048, −0.0026]	pass
H2	Own gain vs gain of the 8B model (0.37)	—	+0.0046 [0.0025, 0.0067]	8B gain better
H3	Over-scaling persists (κ upper bound < 1)	[0.18, 0.85]	[0.30, 1.11]	not shown
H4	Order vs repeat sensitivity, $\lvert\Delta z\rvert$	0.563 vs 0.399	0.516 vs 0.456	pass, smaller
	difference [95% CI]	+0.165 [0.111, 0.219]	+0.060 [0.012, 0.108]	
H5	Reversed second read (Stage 5)	−0.0024 [−0.0059, 0.0008]	−0.0008 [−0.0039, 0.0021]	fail
H6	Change in g / κ, 32B − 8B	+0.09 [−0.17, 0.36] / +0.17 [−0.18, 0.57]		no detectable trend
H7	8B+32B pair vs 32B alone	−0.0082 [−0.0108, −0.0055]		pass (see text)
H8	Pair vs 32B forward+reversed (Stage 5)	−0.0086 [−0.0137, −0.0037]		report only
E	32B vs 8B, each with own gain	+0.0089 [0.0040, 0.0138]		32B worse
E	Pair vs 8B alone	+0.0008 [−0.0015, 0.0031]		no gain

H2: own gain (0.46) minus the 8B gain (0.37), both applied to 32B; positive favours 0.37. E: exploratory, added after H7 was seen.

4. Fixes aimed at the wrong cause. Relevance filtering passed its own labelling gate (precision 0.93) yet did nothing for forecasts: a random subset of the same size performed identically (row 2), and dropping evidence mostly damps the news move, which shrinkage already does better. A gain above one (row 5) pushes forecasts the wrong way for a model that over-reacts. Formatting a data series as a table (row 16) left the degenerate behaviour largely in place because its trigger is the ordering of values, not their wording.

6 Does it hold at four times the scale?

Every comparison above uses 7–8B models. A reviewer's first question is whether the survivor, and the failures, are artefacts of small models. We ran Qwen3-32B, from the same family as Qwen3-8B, on exactly the same prompts: the 300-question calibration set (two reads with evidence, two without), the three held-out sets, and the repeated and reversed reads of Stage 1 (9,664 generations, one run, bf16 on a single H100, same sampling and parser). All 9,664 replies ended normally and 99.95% parsed. The protocol and its eight hypotheses (H1–H8) were committed before any call, with an amendment adding H6–H8 committed while the job was running and before any output was read. All questions had resolved before the run, so every result in this section is final.

The survivor replicates (registered estimator). The 32B model's gain, measured on the calibration set by the same regression that gave 0.37, is 0.46. Applied unchanged, it beats keep-half on each held-out set (−0.0026, −0.0046, −0.0044; all three intervals exclude zero) and in the pool (H1, Table 6). H2 tested whether the gain is a property of the model: if so, the 32B model's own gain should beat the 8B gain transplanted onto it. The opposite happened: 0.37 beat 0.46 by 0.0046 [0.0025, 0.0067]. The 32B model in fact needs more shrinkage, not less. What replicates is therefore the need to shrink far below 0.5 with a gain measured on resolved questions; that the two models need different gains is not shown. Before the run we had written that a gain within 0.08 of 0.5 could not beat keep-half by 0.003; it did, because the 32B model's held-out optimum is far below 0.46 (Figure 9), so even a modest step from 0.5 toward it pays.

Miscalibration does not disappear with scale. Raw 32B forecasts are more over-confident than 8B forecasts (raw Brier 0.278 against 0.249), and both models' held-out optimal gains are far below one (Appendix M).

Order sensitivity. Reversing the evidence moves the 32B model's log-odds by a third of the 8B model's excess over a plain re-run, and averaging a reversed read again does not improve accuracy (H4, H5).

The larger model is not the better forecaster. With each model shrunk by its own frozen regression gain, Qwen3-32B is worse than Qwen3-8B on every set (+0.0012, +0.0178, +0.0106; pooled +0.0089 [0.0040, 0.0138]); with the Algorithm 1 gains the gap shrinks to +0.0032 [+0.0011, +0.0053]. Part of this is the imprecise 32B gain, but not all: with each model's oracle gain chosen on the test set, the 8B model is better on Stages 4 and 5 (0.1725 vs 0.1793; 0.1574 vs 0.1601) and slightly worse on Stage 1 (0.2104 vs 0.2094), and its AUC is equal or higher on all three (0.638 vs 0.635, 0.673 vs 0.644, 0.719 vs 0.707). The difference in AUC is 0.009 [−0.014, 0.036] in favour of the 8B model (paired cluster bootstrap, 1,868 questions). The models also ran on different hardware and numerical formats (8B: fp16 on T4 GPUs; 32B: bf16 on one H100), which we cannot separate from model size. For news-based binary forecasting from headlines with this prompt, four times the parameters bought no accuracy; after correcting for label errors (Section 7) the larger model is indistinguishable from the smaller one (+0.0017, range over corrections [−0.0011, +0.0041]), so we claim only that it is not better.

The question population matters, and a theory of it failed. A gain measured on one set of questions transfers poorly to another: the right amount of shrinkage is a property of the model and the question population together. Exploratory analyses suggested anchoring on each topic's base rate; we derived a theory of when this helps, committed a quantitative prediction before the data existed, and tested it on 1,557 new questions. The rule and the theory both failed, and so did the explanation we registered for the failure on a fresh set. Appendix A gives the population analyses and Appendix B the registered test.

7 How good are the labels?

All questions in the main sets were resolved by a search-grounded LLM. To estimate how often it is wrong, we drew 200 questions at random from the pooled held-out test splits (in two batches of 50 and 150, seeds fixed before drawing) and checked each against contemporaneous web reporting. 187 could be verified; 13 could not (for example, a price on a date for which no record was found) and are excluded. Fourteen labels are clearly wrong and eight more are questionable (Appendix J). The clear-error rate is therefore 7.5% (14/187; Wilson 95% interval 4.5–12.2%), or 11.8% counting the questionable cases. The errors are strongly one-sided: twelve of the fourteen are events that happened but were labelled NO (a ceasefire agreement signed, a record broken, a lawsuit filed, a model released), so about 9.2% of NO labels should be YES (95% interval 5.3–15.3%) against 3.6% of YES labels that should be NO (1.0–12.1%). The pattern is consistent with a resolver that defaults to NO when it fails to find confirming evidence. Two of the erroneous questions were already decided when the question was asked, which echoes the separate estimate in Section 3.

Effect on the results. One-sided errors of this kind reward lower forecasts, and shrinkage toward a low base rate lowers most forecasts, so they could in principle manufacture part of the survivor. We therefore re-ran the whole pipeline under corrected labels: in each of 200 draws, observed NO labels were switched to YES and observed YES labels to NO at the audited rates, in the calibration set and every held-out set, and all gains, anchors and topic rules were refitted. Table 17 shows that the corrections barely move any result, at the audit's point estimates or at its upper bounds. Two effects offset each other: correcting labels raises every set's base rate, which raises the anchors, while the fitted gains fall slightly, which strengthens the case for shrinking far below one half. The uncertainty in the 8B results is sampling uncertainty (Table 15), not label error. An earlier analysis on 50 questions that corrected only NO labels, and used the regression gains, had suggested that the 8B result was sensitive to labels; with the larger audit, two-sided corrections and the deployment estimator, that sensitivity disappears.

Effect on other comparisons. Label errors shared by both arms cannot create the differences between evidence policies, read counts or update rules in Table 5, but they attenuate every effect and add noise.

8 Recommendations

For anyone building or evaluating an LLM forecaster that reads news:

1. Read a sample of questions with and without evidence, and with placebo evidence. The difference in mean forecast measures the push; report it alongside accuracy, since a Bayesian's mean forecast would not move. Do not expect a warning in the prompt or a thinking mode to remove it; in our tests the first did nothing and the second removed about a fifth of it. Reading each headline separately and averaging the forecasts halved it without losing ranking skill; asking the model to label each headline first made it worse. Deduplicating headlines is not a remedy: repetition lowered the push. Where the question-only forecast is calibrated on average, subtracting the push is a correction that needs no outcomes.

2. Report excess belief movement (Augenblick & Rabin, 2021) between question-only and evidence reads: it needs no outcomes and ranked our models in the same order as outcome-based measures.

3. When building benchmarks from headlines, report the mean forecast against the base rate: a large gap signals that the benchmark rewards discounting the news as much as forecasting skill.

4. Shrink forecasts toward a base rate taken from the population being forecast, with a log-odds gain of one half or less. Keep-half is a strong default; a gain measured on a few hundred resolved questions from that population, and refitted as questions resolve, was better for one of our two models and no worse for the other. Per-topic anchors did not help on new questions. Shrinking is the largest and most reliable improvement we found; it is calibration, so report it alongside discrimination (AUC) and a base-rate baseline.

5. Diagnose before correcting: compare the mean forecast with the base rate and fit a calibration slope. A large gap means a bias, a slope well below one means over-dispersion; the anchored shrink handles both, but other fixes usually target only one.

6. Check coherence under negation: ask a sample of questions in both directions, with and without evidence. In our tests the forecasts for a question and its complement were nearly unrelated, a failure that accuracy on one direction hides.

7. Do not assume a larger model forecasts better; compare sizes on the same held-out questions after shrinking each.

8. When a second read is available, report the reliability $\hat{\lambda}$ alongside the gain; it separates noise from systematic over-scaling and needs no outcomes.

9. Compare every new rule against that shrinkage, not against raw outputs, and on questions not used to design it.

10. Report the rate of exact 0/1 outputs, the clipping threshold and every calibration variant; check whether a gain survives without degenerate outputs.

11. Compare evidence filters against a random subset of the same size.

12. Fit fusion weights on earlier questions and test them on later ones.

13. Register quantitative predictions, not only directions, and check which parameters a prediction carries across populations.

14. Timestamp forecasts with an external authority before outcomes are known; version-control logs alone do not prove the order of protocol and result.

9 Limitations

Over-update. The findings come mainly from Qwen3 at two sizes, with OLMo-2-7B, gpt-oss-120b, thinking mode and two prompt variants as single-set controls; three proprietary frontier models were tested on 200 questions only, through an API router without control of provider-side settings, and their question-only forecasts rank outcomes well enough to suggest possible training-data leakage; the evidence is headlines.

The placebo was run on one set, for Qwen3-8B, Qwen3-32B (4-bit), OLMo-2 and gpt-oss-120b, which agree that unrelated news does not raise forecasts but disagree on whether it lowers them (it does for the Qwen3 models and, slightly, for gpt-oss-120b, not for OLMo-2); and the direction and coverage measures rely on simple word lists. The content controls were run for Qwen3-8B, with repetition and swap replicated in Qwen3-32B and OLMo-2 (where the swap result did not replicate) and the paraphrase control for Qwen3-8B only; the swap's donors were chosen automatically, and a hand check found about one in eight to concern the same event, which makes the swap's push an upper bound for content-free topical coverage. The two label-free tests treat two separately sampled reads as one belief before and after the evidence. The frontier comparison is not stratified by each model's training cutoff, which is not documented precisely enough to do so. Other families, prompts, full-text evidence or reasoning modes may behave differently. **Labels.** Questions were written and resolved by LLMs. Our 200-question audit finds 7.5% clear label errors, mostly one-sided; corrections calibrated by the audit leave the results unchanged (Section 7), but the audit was done by the authors, with web sources, not by blinded independent annotators with agreement statistics, and the fresh set's stricter resolver was not audited separately. About 14% of a sample of live questions were already decided when asked. **Registration.** Seven of the sixteen comparisons, and all eight scale-replication hypotheses, have verifiable order by version control (Appendix Q); the rest rely on our own log, and the recalibration, population and label analyses are exploratory. **Effect size.** The surviving effect is about 2% of the Brier score. It is consistent and cheap, but small next to the gap between these models and market prices where those exist. **Intervals.** Paired differences of smooth recalibrations have small variance, and source-by-week clusters approximate events; both can make intervals narrower than the true uncertainty. **Models, prompt and leakage.** We study open models from 7B to 32B, one family at two sizes, with one prompt template, thinking disabled and one sample per question; the finding that the larger model is worse may not hold under other prompts, with reasoning enabled, or for other families, and early questions may fall before the models' training cutoffs. **Estimator and exploratory status.** The registered gains came from a regression estimator; the primary estimator was fixed afterwards, and under it the measured gain's advantage over keep-half is inconclusive for Qwen3-8B. The population analyses, the recalibration comparison and the event-level clustering are exploratory re-analyses of already-scored questions. **Benchmark construction.** Most questions were written from headlines by an LLM, so newsworthy events are over-represented and question wording may echo the evidence; the market questions of the external set show the push is not confined to such questions, but its size and outcome-relatedness depend on how questions are made. **Prospective test.** The external test of Appendix B used questions new to these models and to every shrinkage analysis, but already resolved; its questions had been used, with a different model, in the order case study. Only the label-free part of the prospective test (Section 4.2) has been scored; its outcome-dependent predictions await resolution, and every other result was obtained on resolved questions. Given the evaluation pitfalls catalogued by Paleka et al. (2026), scoring those predictions is the most important remaining step. **Single system.** All tests come from one pipeline and one team; independent replication matters more than any single estimate.

10 Conclusion

Language-model forecasters that read the news over-update on it. For the four open models we study, from three families and 7B to 120B parameters, headlines about a question raise forecasts by 11–22 points, by far more than the reduction in their uncertainty would justify under a Bayesian reading, and much of the rise appears on events that did not happen. Under the tested conditions the push is not explained by the volume or topic of coverage: repeating a headline lowers it, and same-topic headlines about other events produce a fifth of it. The models respond to what the headlines say about the event and then move too far. Three proprietary frontier models show a smaller push that is mostly tied to the outcome. The news still carries real information, but it arrives on a shifted and over-stretched probability scale, which is why raw forecasts are worse than the base rate. This behaviour explains much of what we found when auditing inference-time fixes: of 26 candidates, only shrinking toward the base rate passes our gate at the available power, because it corrects both the YES bias on headline-built questions and the over-dispersion on market questions. A measured gain improves on a fixed halving clearly for one model and inconclusively for the other. Adaptive rules, extra reads, filters and a theory of topic-level anchors all failed on held-out or new questions, and a model four times larger forecast no better. The practical lessons are to measure the push

directly with question-only, placebo and content-control reads; to shrink before trusting any further gain of a few thousandths; and to read accuracy on benchmarks built from headlines as partly a measure of how far a model over-reads the news.

Data and code availability. Row-level predictions, stored raw replies, prompts, question identifiers, outcomes, protocols and analysis scripts are released; scripts reproduce every number in this paper. Question text and prices from platforms whose terms restrict redistribution (Metaculus, Polymarket) and news headlines are replaced by SHA-256 digests with retrieval identifiers. The anonymised package (questions, digested prompts, all Qwen3-8B, Qwen3-32B and OLMo-2 replies, parsed probabilities, the 200-question label audit, protocols and scripts) is archived at Zenodo, doi:10.5281/zenodo.23868135; standalone scripts reproduce Table 15 and the point estimates of Table 3 from the released files alone. The coverage data (external, fresh, placebo, OLMo-2, neutral-prompt, dose, repetition and same-topic swap reads, with the donor table) are included with headlines digested; for questions taken from prediction markets the question text is digested and only parsed probabilities are released.

References

Anonymous. Apparent order-ensembling gains in LLM forecasting come from degenerate outputs and calibration choices, not information. Companion manuscript under review, 2026.

Ned Augenblick and Matthew Rabin. Belief movement, uncertainty reduction, and rational updating. *The Quarterly Journal of Economics*, 136(2):933–985, 2021.

Jonathan Baron, Barbara A. Mellers, Philip E. Tetlock, Eric Stone, and Lyle H. Ungar. Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2):133–145, 2014.

Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, 2 edition, 2006.

Wayne A. Fuller. *Measurement Error Models*. Wiley, 1987.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

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

Ursula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1939–1948, 2018.

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 (ICLR)*, 2025.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS)*, 2017.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Yao Lu, Max Bartolo, Alireza Karimi, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8086–8098, 2022.

Yuan Lu and Yukuan Zhang. The corroboration illusion: When more news makes LLM forecasts less true. *arXiv preprint arXiv:2609.22246*, 2026.

Barbara A. Mellers, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S. E. Zheng, Lyle H. Ungar, Victor P. Fincher, M. M. Halaby, and Philip E. Tetlock. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115, 2014.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

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*, 2025.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters. In *International Conference on Learning Representations*, 2026.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2018–2032, 2024.

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

Raphaël Sarfati, Pratyush Ranjan Tiwari, Siddharth Boppana, Christopher J. Earls, Srikar Varadaraj, and Eric Ho. What LLM forecasters know but don't say: Probing internal representations for calibration and faithfulness. *arXiv preprint arXiv:2607.08046*, 2026.

Ville A. Satopää, Jonathan Baron, Dean P. Foster, Barbara A. Mellers, Philip E. Tetlock, and Lyle H. Ungar. Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356, 2014.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, and Philip E. Tetlock. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *arXiv preprint arXiv:2402.19379*, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 31210–31227, 2023.

Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

Brandon M. Turner, Mark Steyvers, Edgar C. Merkle, David V. Budescu, and Thomas S. Wallsten. Forecast aggregation via recalibration. *Machine Learning*, 95(3):261–289, 2014.

Benjamin Turtel, Paul Wilczewski, Kris Skotheim, Ville A. Satopää, and Philip E. Tetlock. How proper scoring rules shape LLM forecasting. *arXiv preprint arXiv:2608.28482*, 2026.

Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024.

Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. Do language models update their forecasts with new information? *arXiv preprint arXiv:2509.23936*, 2025.

Bianca Zadrozny and Charles Elkan. 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, 2002.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706, 2021.

A Where the gain comes from: the question population, not only the model (exploratory)

H2 and the factorisation check both point the same way: a gain measured on one set of questions transfers poorly to another. We therefore looked, after the registered analyses, at how the gain depends on the questions. Everything in this subsection is exploratory (E).

Cross-population transfer. Calibrating the gain on one population and deploying it on another shows that the result depends far more on the target population than on the calibration population, and that a moderately misestimated gain costs little as long as it is well below one half and the anchor is local (Figure 12; Appendix M).

Topic-local anchors, evaluated forward in time. We then evaluated a rule that anchors on each region–topic's base rate and partially pools its gain toward the global gain, strictly forward in time: all questions were ordered by closing date, and at four origins (the 50th to 80th date percentiles) the rule was fitted on earlier questions and scored on the next 10% (1,246 test questions in total; prior strengths 50 for anchors and 100 for gains, fixed before running). A single global gain is no better than keep-half in this design (8B -0.0004, 32B -0.0052, both intervals including zero). Topic anchors improve on it for both models (-0.0051 $[-0.0082, -0.0021]$ and -0.0057 $[-0.0091, -0.0026]$), and the full rule beats keep-half by 0.0068 (8B, $[-0.0139, +0.0001]$) and 0.0115 (32B, $[-0.0208, -0.0024]$; Figure 12c).

These analyses use questions the registered analyses had already scored, and the idea of local anchors arose from looking at subsets of the same data; the forward-in-time design prevents fitting on the future but not this selection. They are therefore exploratory, and Appendix B reports a registered test on new questions. Taken together with the transfer matrix they support a narrower conclusion than a model-specific gain: these models are miscalibrated throughout, and the right amount of shrinkage is a property of the model and the question population together. The local-anchor rule did not survive that test.

B A pre-registered external test, and a theory that failed it

The population analyses above were exploratory and suggested one concrete rule: topic-local anchors with partially pooled gains. Before running a new test, we wrote down a simple theory of why such anchors should help, derived a quantitative prediction from it, and committed both, together with the analysis script, before any output on the new questions existed. Both the rule and the theory failed.

The theory. Let questions fall into groups k (here, sources) with calibrated base log-odds a_k and grand mean $\bar{a}$, and let V_a be the variance of a_k across questions. Suppose the calibrated log-odds of a question is $a_k + \kappa s$ and the model reads $z = a_k + d_k + s + e$, where s is the question-specific signal, e is noise, both have mean zero within each group, and $d_k = -\gamma(a_k - \bar{a})$ describes how the model treats group differences ($\gamma > 0$: it partly ignores them; $\gamma < 0$: it exaggerates them). Shrinkage with gain $g \in [0, 1]$ toward an anchor c predicts $c + g(z - c)$.

Proposition 1. *The mean squared log-odds error of shrinkage toward the global anchor $\bar{a}$ minus that of shrinkage toward the local anchor a_k, at the same gain g, is*

$$\Delta(g, \gamma) = V_a\,(1 - g)\big[(1 - g) + 2g\gamma\big].$$

Local anchoring therefore never hurts when $\gamma \geq 0$, and it hurts if and only if $\gamma < -(1 - g)/(2g)$.

Proof. With the local anchor, the error is $e_L = g(d_k + s + e) - \kappa s$. With the global anchor it is $e_G = e_L + (1 - g)(\bar{a} - a_k)$. Hence $\mathbb{E}[e_G^2] - \mathbb{E}[e_L^2] = (1 - g)^2 V_a + 2(1 - g)\,\mathbb{E}[(\bar{a} - a_k)e_L]$. Because s and e have mean zero within groups, only $g\,d_k$ contributes to the cross term, and $\mathbb{E}[(\bar{a} - a_k)\,g\,d_k] = g\gamma V_a$. Collecting terms gives the result; the sign condition follows because $V_a(1 - g) \geq 0$. □

The proposition is elementary, but it makes the benefit of local anchors a measurable quantity: it scales with the between-group spread V_a and shrinks as the gain approaches one.

Table 7: Pre-registered external test (Brier differences, forward-in-time design; cluster-bootstrap 95% intervals). TL: topic-local anchors with pooled gains; GG: one global gain; KH: keep-half; raw: unshrunk forecasts. Rows above the rule are registered; the last row is exploratory.

Contrast	Qwen3-8B	Qwen3-32B	Verdict
C1: TL − KH	$+0.0003$ $[-0.0050, +0.0055]$	$+0.0019$ $[-0.0031, +0.0066]$	fails gate
C2: TL − GG	$+0.0006$ $[-0.0047, +0.0056]$	$+0.0009$ $[-0.0024, +0.0041]$	no benefit
predicted by Proposition 1	-0.0044	-0.0050	prediction fails
C3: GG − KH	-0.0003 $[-0.0010, +0.0003]$	$+0.0009$ $[-0.0034, +0.0054]$	no difference
C1, excluding FRED	$+0.0012$ $[-0.0023, +0.0047]$	$+0.0007$ $[-0.0036, +0.0051]$	fails gate
GG − raw (E)	-0.0194 $[-0.0313, -0.0079]$	-0.0113 $[-0.0191, -0.0040]$	shrinkage helps

The registered test. The new questions are the 1,676 resolved market, FRED and Indian-news questions of the order case study, restricted to as-of dates after 1 May 2025 (1,557 questions), which had never been read by either Qwen3 model nor used in any shrinkage analysis. Both models read each question once with evidence, in chronological order, with the template and settings of Section 3. We registered the forward-in-time design of Appendix A (origins at the 50th–80th date percentiles, the next 10% as test; about 620 test questions and 183 clusters per model) and three contrasts: C1, topic-local rule minus keep-half (gate as in Table 5); C2, topic-local rule minus a single global gain; C3, global gain minus keep-half. From Proposition 1 we registered a prediction for C2: measured from outcomes only, V_a is 0.277 on the new set against 0.318 on News-Bench, so the benefit should be about 0.87 times the News-Bench forward values, that is -0.0044 (8B) and -0.0050 (32B), counted as supported if each observed mean lay within a factor of two. The raw outputs were committed before the analysis script was run.

Result. Topic-local anchors give no benefit on the new questions for either model, and the observed C2 has the wrong sign (Table 7). We therefore withdraw the topic-local rule as a recommendation. What does replicate is the basic finding: shrinkage toward the base rate beats the raw forecasts for both models by margins similar to those on News-Bench, and the shrunk forecasts have positive skill over the base rate (single global gain: Brier skill 0.11 for 8B and 0.16 for 32B; raw AUC 0.68 and 0.73). This comparison was not one of the registered contrasts, so it is marked exploratory, but it is the same comparison that motivated the audit.

Why did the prediction fail? A registered mechanism test. Before computing any of the quantities involved, we registered two explanations. *Over-separation:* the model may already exaggerate group differences ($\gamma < 0$), which by Proposition 1 makes local anchors redundant or harmful; we estimate $\beta = 1 - \gamma$ as the slope of each source's mean raw log-odds on its base log-odds. *Drift:* source base rates may change over time, making anchors from earlier questions stale. Over-separation was not supported: β is larger on the new set than on News-Bench for both models (8B: 0.30 against 0.10; 32B: 0.92 against -0.03), but it stays below the threshold $1+(1-g)/(2g)$ at which the proposition predicts harm. Drift was partly supported. The correlation of source base rates between training and test windows is 0.73 on News-Bench for both models but only 0.28 on the new set, meeting the registered margin; on News-Bench, base rates are stable enough that anchors computed from only the most recent 30% of each training window are *worse* than full-window anchors ($+0.0023$ $[+0.0004, +0.0044]$ for 8B). But on the new set recent-window anchors do not rescue the rule (8B -0.0002 $[-0.0042, +0.0037]$; 32B $+0.0011$ $[-0.0028, +0.0051]$): its source base rates are unstable, and recent data are too few to track them.

What the failure teaches (exploratory). A second difference between the sets was not part of the registered prediction: the fitted global gain is far larger on the new questions (8B 0.49, 32B 0.66) than on News-Bench (0.23, 0.18), so the models over-react much less there. The registered prediction scaled only by V_a and assumed the gain carried over; but the proposition's factor $(1-g)[(1-g)+2g\gamma]$ falls steeply as g rises. Inserting each set's own fitted g and β, and calibrating the Brier scale on News-Bench, gives predicted C2 values of -0.0030 (8B) and -0.0008 (32B), both inside the observed intervals. This is an explanation found after seeing the data, not a confirmation, and it competes with the drift explanation above; separating

them requires another fresh set. Its lesson is general: a prediction derived from a correct theory can still fail if it carries over parameters, here the degree of over-reaction and the stability of base rates, that differ between populations.

A fresh test of the explanation, and its falsification. We built a fresh set of 2,940 world-news questions with the same pipeline, new search terms and, because our usual model service was unavailable, an open model (gpt-oss-120b) as question writer and as a stricter resolver that accepts an outcome only when a later headline states it; 1,163 questions resolved, 952 of them after the models' release, with a base rate of 0.14. Before any question existed we registered that, if the gain-shift explanation were right, topic anchors should help on this News-Bench-like set, where over-reaction is strong, and that the explanation would be falsified if the fitted gain were low (below 0.35) and topic anchors still failed for both models. The gain was low (0.16 and 0.14) and topic anchors gave no benefit (-0.0009 $[-0.0041, +0.0020]$ and -0.0007 $[-0.0038, +0.0023]$); the numeric predictions (-0.0065, -0.0068) fell outside both intervals. The explanation is falsified: the topic-anchor benefit on News-Bench was specific to that set. The same fresh set confirmed, as registered, the news push and the bias it creates (Section 4).

C Protocols for all comparisons

Each comparison was specified before its outcome was scored (levels V and L) or is marked exploratory (E). Unless stated, the model is Qwen3-8B (thinking off, temperature 0.7, top-p 0.8, top-k 20, one sample per prompt, per-prompt seed derived from the prompt key), the split is the chronological 40/60 dev/test split by closing date, the base rate is the dev-split YES rate (clamped to $[0.05, 0.95]$), probabilities are clipped to $[0.01, 0.99]$, and the gate is a mean paired difference of at most -0.003 Brier with the 95% upper bound below zero, from a bootstrap over source-by-ISO-week clusters (4,000 resamples, seed 13).

Rows 1–3: evidence selection. 1,516 questions (519 India-Bench, 997 News-Bench); test $n = 909$. Arms: A, all time-locked evidence; B, items a relevance classifier marks relevant (trained on forward/reversed agreement labels; held-out labelling precision 0.93, recall 0.51); C, a random subset of the same size as B per question; D, B with a protected minimum of five items. All arms use keep-half. Co-primary contrasts B−A and D−A at 97.5% confidence; B−C isolates selection from compression. Items kept: B 8.7%, C 8.7%, D 35.9%. Test Brier: A 0.2134, B 0.2224, C 0.2228, D 0.2138. Raw (uncalibrated) scores invert the picture (A 0.2695, B 0.2560, C 0.2525): fewer items mean smaller news moves, which damps over-reaction; a random subset does the same.

Row 4: gain on the move from the no-news forecast. Candidate $p = \sigma(z_0 + \alpha(z_1 - z_0))$ centred on the question-only read z_0, with α fitted on dev by regularised log loss toward 0.5. Fitted $\alpha = 0.394$; test Brier 0.2259 against keep-half 0.2132. Centring on the no-news read inherits that read's miscalibration.

Row 5: variance-scaled gain above one. A proposed "martingale correction" multiplies the gain by $\sqrt{1 + \pi \operatorname{Var}(\Delta z)/8}$ (here 1.444, giving gain 0.722). Exploratory test on the same split: 0.2323 against 0.2132. For a model that over-reacts, a larger gain moves forecasts the wrong way.

Row 6: model-specific gain. The gain is the coefficient on the news move $z_{R1} - z_{P1}$ in a logistic regression of outcomes on $[1, z_{P1}, z_{R1} - z_{P1}]$ over a separate set of 300 resolved News-Bench questions, each read twice with evidence (R_1, R_2) and twice without (P_1, P_2): 0.367 (bootstrap 95% CI 0.14–0.65). It was then applied, unchanged, to the three held-out sets of Table 14.

Row 7: lead-time-dependent gain. $\alpha(h) = \alpha^\star(0.60 + 0.40e^{-h/72})$ with h the hours from the as-of time to resolution, against the constant gain, on the Stage 4 test split.

Rows 8–9: two-channel and disagreement-adaptive gains. Row 8 fits separate gains on the prior and the news move on Stage 1 and scores them on the Stage 4 test split (fitted 0.30 on news, 0.39 on the prior). Row 9 shrinks harder when two forward reads of the same question disagree: $g_i = g_0/(1 + c\,|z_{A,i} - z_{F2,i}|)$,

with g_0 and c chosen on the Stage 1 dev split by grid search and compared on the Stage 1 test split with a constant gain chosen the same way.

Rows 10–13: multiple reads. Arms K (one forward read), SAME-2 (two forward reads averaged in log-odds), REV-2 (forward and reversed reads averaged), all with keep-half. Primary contrast REV-2−SAME-2 with superiority, equivalence (90% CI inside ±0.003) or inconclusive readings. Test Brier: K 0.2132, SAME-2 0.2119, REV-2 0.2106. Row 13 repeats the comparison with gain 0.37 on Stage 5.

Row 14: market fusion. 236 questions with visible market prices and matched reads from two larger models. A Bates–Granger weight on the model (0.134 when fitted on all questions) was refitted on the earliest 40% and applied to the later 60% (142 questions): weight slightly negative, fusion 0.0005 worse than the market alone.

Rows 15–16: output handling (Qwen2.5-7B). Row 15 drops exact 0/1 reads before averaging three forward reads. Row 16 presents FRED series as a table of (release date, value) rows rather than a chronological list; clustered by series (20 series). Zeros fell from 32% to 28% of decodings.

D Extensions of Proposition 2

Correlated noise across reads. If the noise of two reads has correlation ρ, then $\mathrm{Var}(x_1 - x_2) = 2\sigma^2(1-\rho)$, so the estimator $\hat{\sigma}^2 = \mathrm{Var}(x_1 - x_2)/2$ understates the noise by a factor $1-\rho$ and $\hat{\lambda}$ is biased upward. Every pair of reads used in this paper has independent sampling seeds: the seed is derived from a hash of the prompt key, and keys differ between the two reads. Identical outputs nevertheless occur (58% of repeat pairs for Qwen3-8B and 40% for Qwen3-32B on the calibration set) because sampling at temperature 0.7 with top-k 20 often returns the same coarse probability; this is low sampling noise, not seed dependence, and it is the quantity $\hat{\lambda}$ should measure. Dependence through the shared prompt is, by construction, part of the reproducible component s.

Logistic link. The proposition concerns the best linear predictor of the calibrated log-odds. Under a probit link, $\Pr(Y = 1 \mid s) = \Phi(z_b + \kappa s)$ with Gaussian s and e, the conditional law of s given x is Gaussian with mean λx and variance $\lambda\sigma^2$, so $\Pr(Y = 1 \mid x) = \Phi\big((z_b + \kappa\lambda x)/\sqrt{1 + \kappa^2\lambda\sigma^2}\big)$. The fitted slope on the link scale is then $\kappa\lambda/\sqrt{1 + \kappa^2\lambda\sigma^2}$, slightly below $\kappa\lambda$; with the logistic link the same holds approximately through $\mathrm{logit}(p) \approx 1.7\,\Phi^{-1}(p)$. For the reliabilities and over-scaling factors we estimate, the correction is small relative to the sampling uncertainty of the fitted gains.

Two noise sources. When the move is defined from a question-only read, $m = x_{\text{news}} - x_{\text{prior}}$, both reads contribute noise and $\lambda = 1 - (\sigma^2_{\text{news}} + \sigma^2_{\text{prior}})/\,\mathrm{Var}(m)$; this is the form used in Section G for the main parameterisation (0.77 for Qwen3-8B).

E Question pipeline and prompt

Questions are generated by an LLM from dated headlines, phrased as yes/no events with a closing date, and filtered for duplicates. A search-grounded LLM resolves each question after its closing date; questions without a cited resolution are dropped. Evidence for a question is the set of stored headlines published before its as-of time, de-duplicated by content hash. Every forecast in this paper uses the template below (forward order lists items oldest first; reversed order lists them newest first).

```
You are an expert forecaster.  It is {now}.  Question:  {question}
Known by:  {closing date}
Sources:
[1] {date} · {source} · {headline}
...
Think about the base rate and how the evidence moves it.  Write two or three plain
```

```
sentences (base rate, what the latest evidence says, what must still happen), then
end with exactly one line:
Probability YES: <number between 0 and 1>
```

The question-only arms use the same template with "(none)" as the source list. Replies are parsed by the last "Probability YES:" line; unparsable replies are treated as missing.

F Reproducibility and compute

All Qwen3-8B reads were generated with vLLM on free Kaggle GPUs (two T4s, fp16). The Qwen3-32B replication ran once on a single H100 (bf16) on a commercial cloud with a hard time limit. The Qwen2.5-7B regeneration ran on an L40S. Analysis needs only a laptop CPU: every table and figure is produced by scripts in the released package from stored row-level reads and outcomes. Protocol entries, their commit hashes and the analysis scripts are released together, so the verifiable (V) ordering can be checked from the version history.

G Factorising the gain, and a rule for new readers

The practical cost of the surviving rule is that each model's gain must be measured on a few hundred resolved questions. If the gain could be predicted from quantities that need no outcomes, a new reader, or a different number of reads, could be calibrated immediately. A standard measurement-error argument (Fuller, 1987; Carroll et al., 2006) suggests how.

Proposition 2 (Gain factorisation)**.** *Let $x = s + e$ be a read's log-odds move from the base rate, where s is the reproducible part of the move and e is read noise with $\mathbb{E}[e \mid s] = 0$ and $\mathrm{Var}(e) = \sigma^2$, and suppose the calibrated log-odds of the outcome is $z_b + \kappa s$ for a model-specific constant κ. Then the linear predictor $z_b + gx$ with the smallest mean squared error for $z_b + \kappa s$ has*

$$g^\star = \kappa\,\lambda, \qquad \lambda = \frac{\mathrm{Var}(s)}{\mathrm{Var}(s) + \sigma^2}.$$

If k reads with independent noise are averaged, $g_k^\star = \kappa\,\lambda_k$ with $\lambda_k = \mathrm{Var}(s)/(\mathrm{Var}(s) + \sigma^2/k)$, and λ is identified without outcomes from two reads x_1, x_2 by $\sigma^2 = \mathrm{Var}(x_1 - x_2)/2$ and $\mathrm{Var}(s) = \mathrm{Var}(x) - \sigma^2$.

Proof. Minimising $\mathbb{E}[(\kappa s - gx)^2]$ over g gives $g^\star = \kappa\,\mathrm{Cov}(s, x)/\,\mathrm{Var}(x)$. Because e has conditional mean zero given s, $\mathrm{Cov}(s, x) = \mathrm{Var}(s)$ and $\mathrm{Var}(x) = \mathrm{Var}(s) + \sigma^2$, which gives $g^\star = \kappa\lambda$. Averaging k reads leaves s unchanged and divides the noise variance by k. For two reads, $x_1 - x_2 = e_1 - e_2$ has variance $2\sigma^2$. □

The result is classical attenuation applied to forecasts; its use here is that κ (over-scaling) and λ (reliability) have different costs. λ needs only two reads per question, while κ needs outcomes. Two consequences are falsifiable. First, the gain for k reads is $g_1\lambda_k/\lambda_1$, with no new outcomes. Second, if κ is stable across models of a family, a new model's gain is $\kappa\hat{\lambda}_{\mathrm{new}}$. The proposition is a linear approximation: with a logistic link, the fitted slope also absorbs nonlinearity, and order-swap noise need not behave like repeat noise.

Decomposition. On the calibration set (Qwen3-8B, 300 questions, two repeated reads), defining the move from the set's base rate gives $\hat{\lambda}_1 = 0.87$, fitted $g_1 = 0.23$ and so $\hat{\kappa} = 0.26$; with the move defined from the model's own no-news forecast, as in the main analysis, $\hat{\lambda}_1 = 0.77$ and $g_1 = 0.37$, so $\hat{\kappa} = 0.48$. Either way $\hat{\kappa}$ is well below one: most of the shrinkage corrects systematic over-scaling, not noise. Within the calibration set the two-read gain predicted from the one-read fit, $\hat{\kappa}\hat{\lambda}_2 = 0.244$, matches the fitted 0.256 (in-sample).

Exploratory transfer check (E). Applying $\hat{\kappa} = 0.26$ to two held-out Qwen3-8B sets with forward and reversed reads, using order-swap noise for $\hat{\lambda}$, predicts gains of 0.233 (Stage 5) and 0.231 (Stage 3). The outcome-fitted gains are 0.313 $[0.233, 0.396]$ and 0.353 $[0.288, 0.417]$ (cluster bootstrap). The prediction lies at the edge of the interval on Stage 5 and below it on Stage 3 (Figure 2). The factorisation captures the

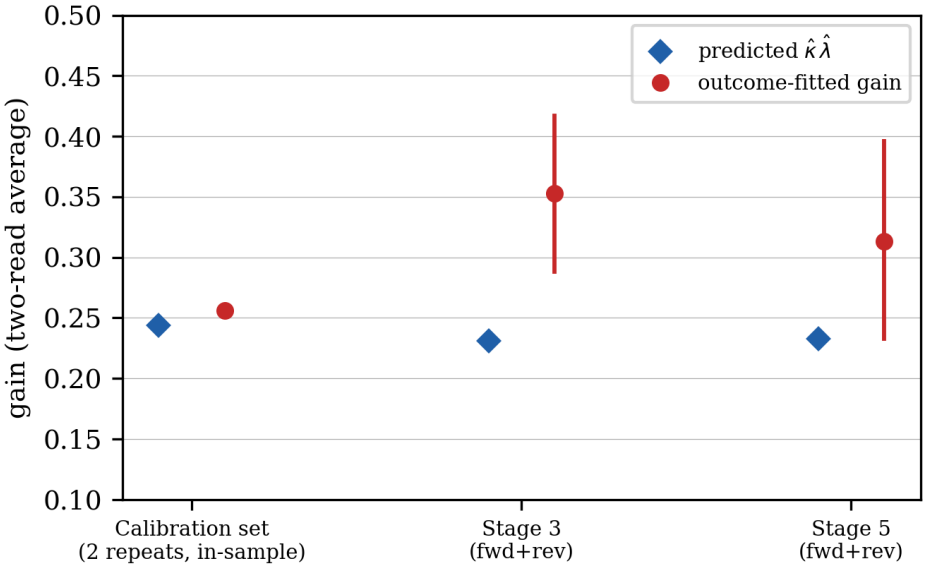


Figure 2: Predicted gain $\hat{\kappa}\hat{\lambda}$ (diamonds, no outcomes used beyond the calibration set) against the outcome-fitted gain with 95% cluster-bootstrap interval (circles), for the two-read average.

direction (gain well below one, slightly larger for averaged reads), but $\hat{\kappa}$ from the calibration set under-predicts the gain on these sets. Differences in base rate (0.20 on the calibration set, 0.21 on Stage 5, 0.28 on Stage 3) could explain part of the gap on Stage 3 but not on Stage 5; two other candidates are that order-swap noise is not exchangeable with repeat noise, and that the calibration-set estimate of κ is itself imprecise (the one-read gain's 95% interval spans 0.14–0.65 in the main parameterisation). Because these outcomes had been seen in earlier analyses, we treat this check as exploratory. The conclusion we draw is limited: reliability is cheap to measure and explains a minority of the shrinkage, but κ does not transfer across question sets well enough to replace measuring the gain on resolved questions.

H Scale replication: per-set results and robustness

Table 8: Per-set results for both models (forward read, test splits; registered regression gains 0.37 and 0.46). Brier scores; Δ = own gain − keep-half with 95% cluster-bootstrap interval; 0/1: share of exact 0 or 1 outputs.

Model	Set	N	Clusters	Raw	Keep-half	Own gain	Δ [95% CI]	0/1
Qwen3-8B	Stage 1	909	255	0.2690	0.2132	0.2105	-0.0027 $[-0.0077, +0.0021]$	0.3%
Qwen3-8B	Stage 4	480	218	0.2306	0.1769	0.1727	-0.0042 $[-0.0097, +0.0011]$	0.2%
Qwen3-8B	Stage 5	480	218	0.2303	0.1656	0.1583	-0.0073 $[-0.0128, -0.0020]$	0.4%
Qwen3-32B	Stage 1	910	255	0.2865	0.2141	0.2115	-0.0026 $[-0.0043, -0.0010]$	0.7%
Qwen3-32B	Stage 4	480	218	0.2813	0.1944	0.1899	-0.0046 $[-0.0066, -0.0026]$	0.8%
Qwen3-32B	Stage 5	480	218	0.2567	0.1733	0.1689	-0.0044 $[-0.0064, -0.0025]$	1.0%

Notes. The 32B run parsed 9,659 of 9,664 replies; unparsed replies are treated as missing. Stage 1 has 910 test questions for 32B and 909 for 8B because one 8B forward read did not parse. Exact 0/1 outputs are rare for both models on these sets (0.2–1.0%), so the degenerate-output mechanism of Section P does not drive any result here. Over the calibration set, the bootstrap difference in fitted gains between the models (H6) is $+0.09$ $[-0.17, 0.36]$; the calibration set is too small to detect a scale trend of plausible size.

Table 9: Robustness of the survivor (registered regression gains): pooled own gain − keep-half under alternative analysis choices (baseline: clip 0.01, dev 40%, dev base-rate anchor).

Variant	Qwen3-8B (gain 0.37)	Qwen3-32B (gain 0.46)
baseline	−0.0046 [−0.0077, −0.0016]	−0.0037 [−0.0048, −0.0026]
clip 0.001	−0.0046 [−0.0076, −0.0016]	−0.0037 [−0.0047, −0.0026]
clip 0.02	−0.0046 [−0.0076, −0.0016]	−0.0037 [−0.0047, −0.0026]
clip 0.05	−0.0045 [−0.0076, −0.0015]	−0.0036 [−0.0047, −0.0025]
dev 30%	−0.0047 [−0.0075, −0.0019]	−0.0039 [−0.0049, −0.0029]
dev 50%	−0.0047 [−0.0079, −0.0014]	−0.0038 [−0.0049, −0.0026]
anchor = calibration-set rate 0.197	−0.0041 [−0.0072, −0.0010]	−0.0036 [−0.0046, −0.0025]
anchor fixed 0.25	−0.0062 [−0.0089, −0.0034]	−0.0041 [−0.0051, −0.0031]

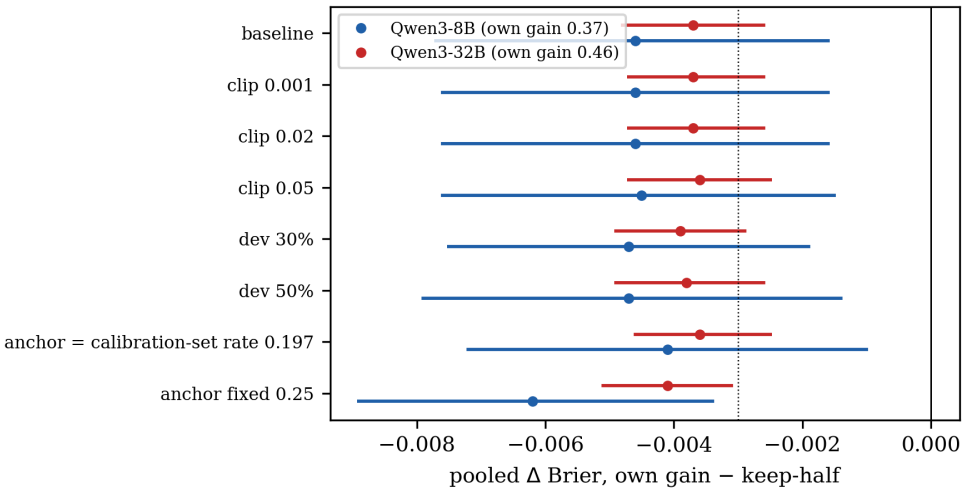


Figure 3: Robustness of the survivor (Table 9). Every variant keeps the 95% interval below zero for both models; the dotted line marks the −0.003 gate.

I Deployment recipe

The results suggest a simple, cheap procedure for any LLM forecaster. Figure 4 shows where it sits in a forecasting pipeline and Algorithm 1 states it. Only the base-rate anchor and the gain are learned; everything else is fixed. A diagnostic step measures the news push of Section 4.2 on a sample of paired reads; where resolved questions are not yet available and the question-only forecast is calibrated on average, subtracting that push is an interim correction. Topic-level anchors are not part of the recipe: they failed registered tests on two sets of new questions (Appendix B).

What not to add. On our evidence, none of the following improves a shrunk forecast enough to justify its cost: relevance filtering of evidence, a second read in another order, topic-level anchors, phrasing-class anchors, saturating or asymmetric gains, online or evidence-volume gains, a combination of question-only and evidence reads, gains that adapt to lead time or read disagreement, a larger model of the same family, or in-sample market-fusion weights. Each may still be worth testing in another system, but against the measured shrink and on held-out questions.

Table 10: Exploratory comparison of the two models and their pair (each shrunk by its own frozen gain; oracle gains chosen on the test split for reference only).

Set	N	8B	32B	Pair	Oracle g 8B	Oracle g 32B	AUC 8B	AUC 32B	AUC pair	corr(z)
Stage 1	909	0.2105	0.2117	0.2074	0.39	0.36	0.638	0.635	0.652	0.65
Stage 4	479	0.1726	0.1904	0.1775	0.34	0.22	0.673	0.644	0.673	0.66
Stage 5	480	0.1583	0.1689	0.1592	0.30	0.26	0.719	0.707	0.742	0.65

Table 11: Calibration-set size (exploratory): gain fitted on n random questions from the pooled held-out sets, scored on the rest (400 draws).

	Qwen3-8B		Qwen3-32B	
n	gain, median [90%]	P(beats keep-half)	gain, median [90%]	P(beats keep-half)
50	0.33 [0.03, 0.61]	0.60	0.28 [0.03, 0.51]	0.79
100	0.35 [0.18, 0.51]	0.81	0.31 [0.13, 0.47]	0.94
200	0.35 [0.23, 0.47]	0.91	0.32 [0.19, 0.42]	1.00
300	0.35 [0.25, 0.44]	0.98	0.31 [0.23, 0.39]	1.00
500	0.35 [0.29, 0.42]	0.98	0.31 [0.25, 0.37]	1.00
1000	0.35 [0.32, 0.40]	0.98	0.31 [0.27, 0.34]	1.00

J Label audit

Two hundred questions were drawn at random (two batches, fixed seeds) from the pooled held-out test splits and each label was checked against contemporaneous web reporting on 1 November 2026. Table 13 lists the questions judged wrong or questionable; the full list of 200 with verdicts and sources is released with the data.

Table 13: Audited labels judged wrong or questionable (Label: resolver label).

#	Question (abridged)	Label	Verdict	Evidence
6	Will "Dhurandhar 2" surpass the lifetime worldwide gross box office collection of "Baahubali 2: The Conclusion" by May 25, 2026?	N	wrong	Dhurandhar 2 passed Baahubali 2 worldwide 3 May 2026
11	Will a new formal ceasefire agreement between the United States and Iran be publicly announced and acknowledged by both…	N	wrong	US-Iran interim ceasefire signed 17 Jun 2026
23	Will Apple be reported as the leading smartphone vendor by market share (unit shipments) in China for calendar Q4 2025 by at…	N	wrong	Counterpoint: Apple #1 China Q4 2025 (published Jan 2026)
40	Will the year-over-year percentage change of the US Consumer Price Index for All Urban Consumers (CPI-U) for October 2025 be…	N	questionable	October 2025 CPI never published (shutdown); should have been voided
46	Will the Ministry of Electronics and Information Technology (MeitY) or another relevant Indian government body publicly…	Y	questionable	Broad semiconductor scheme announced; not startup-specific
63	Will the Government of India officially launch its "headline services output index" by September 14, 2026?	Y	questionable	Sub-sector services indices launched, no headline index
70	Will any US Senator or Representative formally introduce legislation or request a formal investigation into the Department of…	N	questionable	Senators wrote to the court; whether this is a formal investigation request is unclear
83	Will NASA officially announce the successful completion of all primary objectives of the Artemis II wet dress rehearsal by…	N	wrong	Artemis II WDR completed 19-20 Feb 2026, primary objectives met

#	Question (abridged)	Label	Verdict	Evidence
84	Will Stellantis or Tata Motors publicly announce the name of the specific Tata Motors platform that will be used for the...	Y	questionable	Platform widely reported but not officially named by the deadline
85	Will the Indian Space Research Organisation (ISRO) attempt an orbital launch using a Polar Satellite Launch Vehicle (PSLV)...	Y	wrong	No PSLV launch by 22 Jun 2026 (return-to-flight still pending in July)
93	Will Blue Origin's New Glenn rocket publicly launch its first NASA mission to Mars between October 27, 2025, and December 8,...	N	wrong	New Glenn launched NASA ESCAPADE to Mars 13 Nov 2025
95	Will the closing price of Brent Crude oil (ICE Futures Europe) be above $110.00 per barrel on any trading day between May 11,...	N	wrong	Brent closed above $110 mid-May 2026 (e.g. 15 May)
107	Will the CoinDesk Bitcoin Price Index (XBX) record a value of $57,000.00 USD or lower at any point between August 31, 2026,...	Y	wrong	BTC low $74.9k in Sep 2026, not <=$57k
116	Will the Indian Ministry of Health and Family Welfare issue new guidelines or regulations specifically restricting the...	Y	questionable	Guidelines issued in March, before the question was asked; whether new ones followed is unclear
118	Will at least one US state attorney general file a lawsuit against the US Environmental Protection Agency (EPA) by April 13,...	N	wrong	24-25 state AGs sued EPA 19 Mar 2026
134	Will the United States and Iran officially announce by April 27, 2026, that they have agreed to hold direct diplomatic talks...	N	wrong	US-Iran direct talks agreed and held in Islamabad 11 Apr 2026
137	Will Google officially announce the public deployment or availability of its new AI model for flood alerts in India, as...	Y	questionable	No dated announcement of a new flood model found
146	Will Nottingham Forest officially announce the appointment of a fourth different permanent manager for the 2025-2026 Premier...	N	questionable	A fourth permanent manager reported by one source only
160	Will OpenAI publicly announce the general availability of GPT-5.3 (or any version 5.x newer than 5.2) for broad public access...	N	wrong	GPT-5.3 Instant default for all ChatGPT 3 Mar 2026; GPT-5.4 6 Mar
163	Will the World Bank officially announce a decision to abandon its goal of devoting 45% of lending resources to climate change...	N	wrong	World Bank dropped 45% climate target late Jun 2026
174	Will the United States and Iran publicly announce an agreement for a new, formal ceasefire or peace deal, explicitly including...	N	wrong	US-Iran 14-point MoU incl. Hormuz signed 17-19 Jun 2026
183	Will Google officially announce the public release of the Gemini 3.1 Pro AI model by April 5, 2026?	N	wrong	Gemini 3.1 Pro publicly released (preview) 19 Feb 2026

K Further coverage analyses

Which claims depend on the labels. The push itself, the placebo effect, and the dependence of the push on coverage and on contradicting headlines compare reads of the same questions and need no outcome labels, so label errors cannot create them. Statements about the bias (mean forecast against base rate), the split by outcome, AUC and Brier scores do depend on the labels, and our audit (Section 7) found the resolver leaning toward NO. Under the audit-calibrated correction, Stage 1's base rate rises from 0.28 to about 0.34: the mean forecast with evidence (0.48) remains 0.14 above it, the 8B question-only mean (0.25) falls about 0.08 below it, and the push on questions that truly resolved NO changes only from +0.20 to about +0.19. The direction of every label-dependent conclusion survives, but "the question-only forecast is calibrated on average" holds only approximately. The fresh set used a different, stricter resolver whose error rate we did not audit, so its base rate (0.14) and bias figures should be read with more caution than its push.

Table 12: Topic-local shrinkage (exploratory): base rate and gain per region-topic fitted on the dev splits.

Region-topic	n_{dev}	Base rate	Gain 8B	Gain 32B
global-business	158	0.209	0.22	0.14
global-economy	288	0.240	0.28	0.16
global-politics	262	0.233	0.27	0.28
global-science	135	0.178	0.15	0.17
global-society	82	0.244	0.23	0.26
global-sports	152	0.250	0.23	0.16
global-technology	169	0.077	0.10	0.06

India-Bench topics had fewer than 40 dev questions each and keep the global anchor and frozen gain.

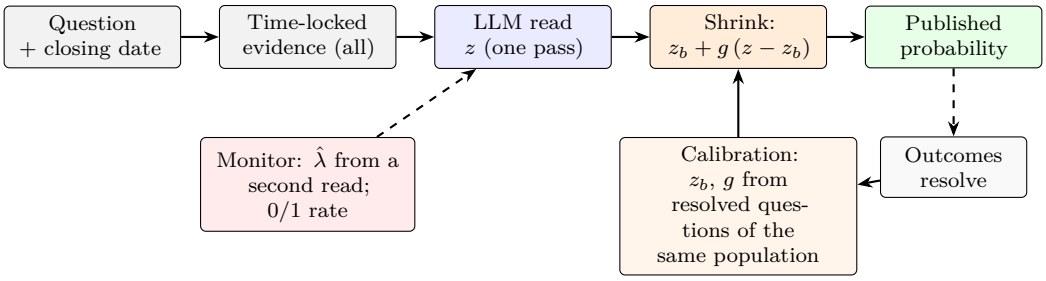


Figure 4: Where the measured shrink sits in a forecasting pipeline. Evidence is not filtered (rows 1–3); one read suffices (rows 10–13); the gain and anchor are refitted periodically from resolved questions.

The push is not matched to what coverage is worth. We measured coverage intensity as the share of a question's headlines that share at least two content words with the question (registered before computing, like every test in this section). On Stage 1 the push rises with coverage for both models (8B: from 0.18 in the lowest third of coverage to 0.27 in the highest; 32B: 0.16 to 0.24; slopes +0.17 $[0.11, 0.22]$ and +0.13 $[0.08, 0.19]$), although coverage does not predict the outcome there (logistic slope +0.13 $[-0.27, 0.53]$). On the external set the pattern is reversed: coverage predicts the outcome (logistic slope +1.09 $[0.75, 1.42]$), but the push does not rise with it (-0.03 $[-0.08, +0.02]$). We had registered that both patterns of the first kind would hold in all three set–model combinations, so the registered test failed; the reversal itself is exploratory. On the external set, where headline ages vary, fresh headlines push much harder than old ones (0.22 in the youngest third against 0.06 in the oldest).

Contradicting headlines barely reduce the push. Questions whose evidence contains at least one headline with a NO-direction word from a fixed list (rejected, delayed, failed, cancelled and similar) still receive a push of +0.19 $[0.17, 0.21]$ (8B) and +0.18 $[0.16, 0.20]$ (32B), against +0.25 and +0.23 for the remaining questions. Such words are rare (7% of headlines), the list is crude, and on the external set too few headlines contain them to form comparison groups, so this test could not be evaluated there.

Second model family and neutral prompt (detail). Two controls were registered after the main results. First, OLMo-2-7B-Instruct, an open model from a different developer and training corpus, read every Stage 1 question with no evidence, with the real headlines and with the same placebo headlines. Its push is +0.194 $[0.178, 0.209]$, as large as Qwen3's, and again barely related to the outcome (+0.22 on YES against +0.18 on NO questions); unrelated headlines leave its forecast nearly unchanged (+0.013 $[0.001, 0.024]$), so for this model irrelevant coverage neither raises nor, unlike for Qwen3-8B, lowers the forecast. Second, our template asks the model to consider "how the evidence moves" the base rate, which might itself invite an upward move. With that sentence replaced by "Give your forecast", Qwen3-8B's push on 595 Stage 1 questions is +0.207 $[0.172, 0.241]$, against +0.232 with the original prompt on the same questions (difference -0.026 $[-0.053, +0.002]$). The push is not an artefact of one model family or of the instruction to update.

Algorithm 1 Calibrate-and-shrink for an LLM forecaster

Require: model M; resolved calibration questions $\mathcal{C}$ from the target population; clip $\varepsilon = 0.01$

Diagnose: on a sample, read each question with and without evidence; record the mean push $\delta = \overline{z^{\text{ev}} - z^{\text{none}}}$ and the mean forecast against the base rate

Before outcomes exist (optional): publish $\sigma(z^{\text{ev}} - \delta)$ if the question-only mean is close to the base rate of a few resolved questions

for each question $i \in \mathcal{C}$ **do**

    $z_i \leftarrow \text{logit}(\text{clip}(M(\text{question}_i, \text{evidence}_i), \varepsilon))$

end for

$b \leftarrow$ YES rate of $\mathcal{C}$; $z_b \leftarrow \text{logit}(b)$

$g \leftarrow \arg\max_{g \in [0,1]} \sum_i \log \Pr(y_i \mid \sigma(z_b + g(z_i - z_b)))$ ▷ one-parameter logistic fit

Freeze (z_b, g); record the commit before using it

for each new question j **do**

    publish $p_j = \sigma\big(z_b + g\,(\text{logit}(\text{clip}(M(\cdot), \varepsilon)) - z_b)\big)$

end for

Every few hundred resolutions: refit on recent resolved questions; adopt the new (z_b, g) only if it beats the frozen one on later questions

Flag the model if exact 0/1 outputs exceed 5% in a category, or if the reliability $\hat{\lambda}$ from paired reads drifts

What did not explain or reduce the push (registered, round 6). Masking every headline word that also appears in the question leaves 83% of the push (+0.193), so it does not ride on lexical mention. Listing all headlines twice changes nothing (+0.005 [−0.009, +0.018]), and showing the model its own question-only forecast before the evidence lowers the push by only 6%. Asked headline by headline how likely each would be if the event happened and if it did not, the model judges most headlines only modestly diagnostic (median likelihood ratio 2.3; 28% at most 1.25); combining those judgements by Bayes' rule with its question-only forecast cuts the push to +0.097 and the raw Brier score from 0.275 to 0.241, but loses slightly more ranking skill (AUC 0.598 against 0.620) than our registered limit allowed. The model can say that a headline is weak evidence when asked about it alone, and still move far when reading headlines together.

L Additional audit figures and tables

The figures and tables below support Sections 5–7; they were moved here to keep the main text focused on the coverage effect.

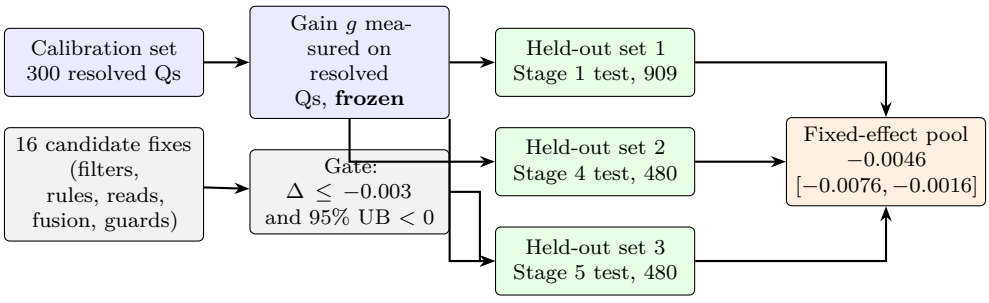


Figure 5: Design of the audit. The gain is estimated once on a separate calibration set and frozen; each candidate fix is scored against the gate on questions not used to design it; the surviving rule is then pooled over three disjoint held-out sets.

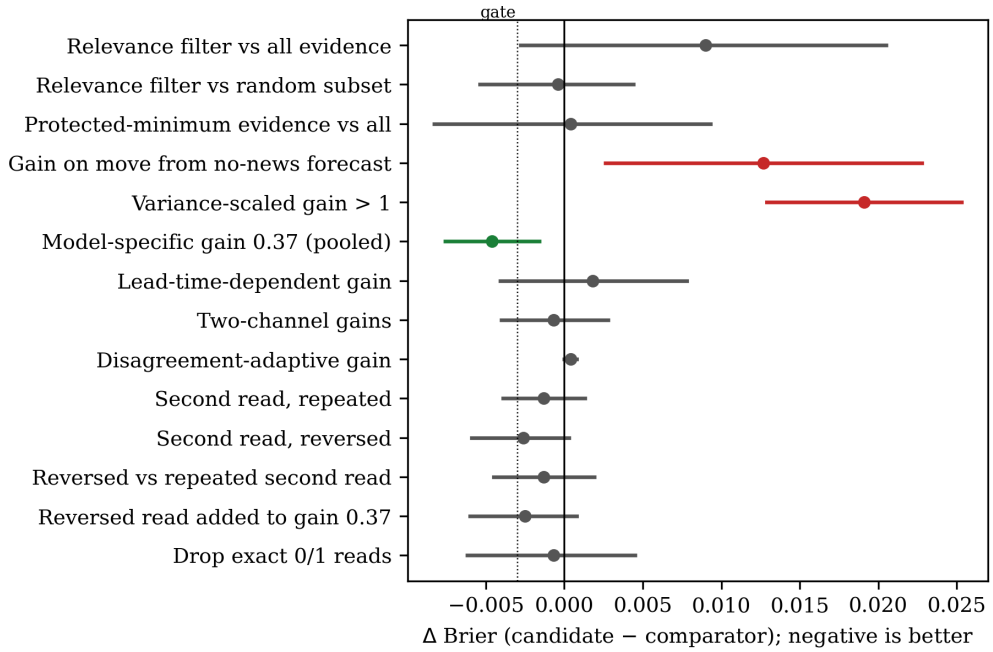


Figure 6: All comparisons with a confidence interval (Table 5; rows 14 and 16 omitted: no interval, or a different scale). Green: passes the gate (dotted line, −0.003, with the 95% upper bound below zero). Red: significantly worse than the comparator.

Table 14: Regression gain 0.37 (exactly 0.367) versus keep-half (0.5) on Qwen3-8B, forward read, on three disjoint held-out sets. The gain was estimated on a fourth set (300 News-Bench questions) before these tests.

Held-out set	N	Clusters	Δ [95% CI]	Reg.
Stage 1 test (India + News)	909	255	−0.0027 [−0.0077, +0.0021]	E
Stage 4 (News)	480	218	−0.0042 [−0.0097, +0.0011]	L
Stage 5 (News)	480	218	−0.0073 [−0.0128, −0.0020]	L
Primary pool (Stages 4 + 5)	960		−0.0058 [−0.0096, −0.0019]	
All three sets	1,869		−0.0046 [−0.0077, −0.0016]	

M Further audit detail

Standard recalibration methods. A fitted slope beating a fixed one is unsurprising, so the right comparison is with standard recalibrators fitted on the same calibration set (Table 19 in Appendix O; level E). On average, Platt scaling and isotonic regression are statistically indistinguishable from the measured shrink for both models, but neither beats keep-half with an interval excluding zero, and their per-set results swing widely: Platt scaling fitted on the calibration set is worse than keep-half on Stage 1 (0.2226 against 0.2132 for 8B) and much better on Stage 5 (0.1530 against 0.1656), because its intercept carries the calibration set's base rate (0.20) into sets whose base rates differ. Temperature scaling, which rescales log-odds around 0.5 with no base-rate anchor, is far worse than everything else (+0.04 Brier). Platt scaling fitted on each set's own dev split also fails to beat keep-half with an interval excluding zero. What makes the measured shrink robust is therefore the anchor: it keeps the base rate local to each set and learns only how far to move from it. This is consistent with the population analysis of Appendix A.

A caution on estimating the gain. Refitting the gain on each set's own small dev split was worse than using the value from the separate calibration set: on Stage 1 the dev-fitted gain was 0.19, which scored

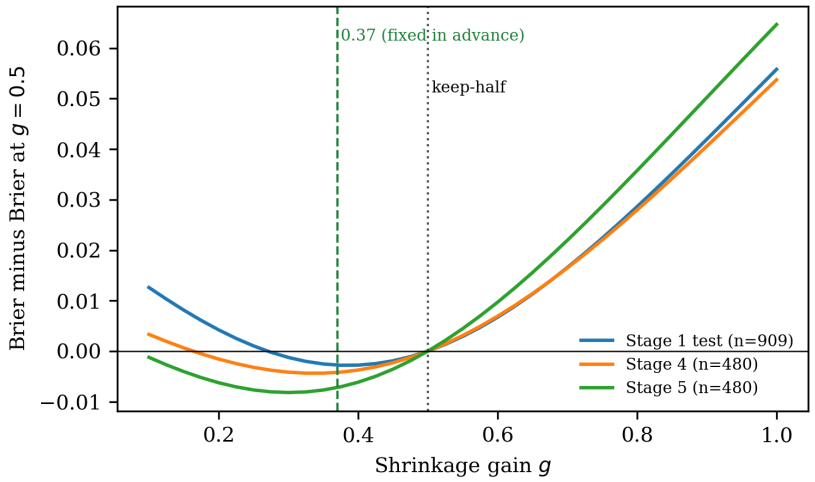


Figure 7: Test-set Brier score as a function of the shrinkage gain, relative to keep-half ($g = 0.5$), on the three disjoint held-out sets (Qwen3-8B, forward read). The gain 0.37 was estimated on a separate calibration set before these sets were scored; all three curves have their minimum between 0.30 and 0.40.

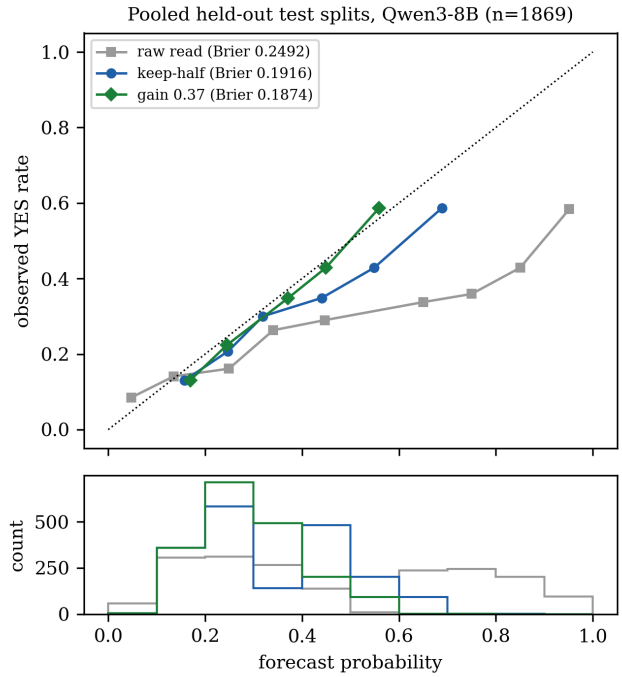


Figure 8: Reliability diagram on the pooled held-out test splits (Qwen3-8B, forward read; bins with at least 15 forecasts). The raw read is badly over-confident; keep-half still over-shoots above 0.3; the measured gain lies close to the diagonal. Bottom: forecast histograms.

0.2181 on test against 0.2104 for 0.37. Dev splits here are early in time and partly from a different source mix. The gain should be estimated on a reasonably large, representative calibration set and then frozen.

The calibration estimate is the weak link. For Qwen3-8B the frozen gain sits near each held-out optimum (0.30–0.39). For Qwen3-32B it does not: the held-out optima are 0.36, 0.22 and 0.26, while the

Table 15: Held-out Brier difference, Algorithm 1 gain minus keep-half (source–week clusters). Primary: Stages 4 and 5; all: three-set sensitivity pool.

Model	Gain [95% CI]	Stage 1	Stage 4	Stage 5	Primary pool (4+5)	All three
Qwen3-8B	$0.229\ [0.13, 0.33]$	$+0.0023$	-0.0026	-0.0072	$-0.0049\ [-0.0126, +0.0028]$	$-0.0022\ [-0.0083, +0.0039]$
Qwen3-32B	$0.184\ [0.11, 0.26]$	$+0.0028$	-0.0144	-0.0120	$-0.0132\ [-0.0228, -0.0035]$	$-0.0067\ [-0.0142, +0.0007]$

Table 16: Calibration and discrimination on the pooled held-out test splits. BSS: Brier skill relative to the dev base rate; Rel., Res.: reliability and resolution terms of the Murphy decomposition (10 bins; uncertainty 0.201); slope: logistic calibration slope.

	Qwen3-8B							Qwen3-32B			
Policy	Brier	BSS	Rel.	Res.	AUC	Slope	Log loss	Brier	BSS	AUC	Slope
Base rate only	0.2065	+0.000	0.0052	0.0000	—	—	0.607	0.2068	+0.000	—	—
Raw read	0.2492	-0.207	0.0631	0.0151	0.669	0.42	0.719	0.2775	-0.342	0.661	0.44
Keep-half	0.1916	+0.072	0.0055	0.0145	0.672	0.85	0.567	0.1986	+0.040	0.660	0.87
Algorithm 1 gain	0.1902	+0.079	0.0033	0.0126	0.672	1.86	0.564	0.1931	+0.066	0.660	2.38

calibration set gave 0.46 with an interval of 0.21–0.75. Three hundred questions locate the gain only roughly, and the 32B model's calibration-set behaviour differed from its behaviour on the larger sets. The rule is robust in sign but its size depends on a calibration set large enough to pin the gain down; Appendix A shows that the problem is the calibration population rather than its size.

A registered pass that does not mean what it seems. The pair of models (each shrunk by its own gain, log-odds averaged) beats the 32B model alone by 0.0082, passing H7 on all three sets. Our own taxonomy (Section 5.4) says to ask why. The exploratory comparisons explain it: the pair beats the 32B model because the 32B model is worse, and it is no better than the 8B model alone ($+0.0008\ [-0.0015, 0.0031]$). The two models' log-odds correlate at 0.65 and averaging them raises AUC slightly on two of three sets, but not enough to beat the better model's Brier. Adding a larger model to a smaller one is not a free improvement.

Calibration size is not the bottleneck. Drawing calibration subsets at random from the pooled held-out questions and scoring the fitted gain on the remaining questions, 300 questions locate the gain to a 90% range of 0.25–0.44 (8B) and 0.23–0.39 (32B), and the fitted gain beats keep-half on the remaining questions in 98% and 100% of draws (Figure 13a). The excess Brier over the oracle gain falls to about 0.0006 by 500 questions and barely improves after. The 32B model's frozen gain missed (0.46 against 0.31 on the pooled held-out questions) not because 300 is too few, but because the calibration questions were drawn from a different population than the test questions.

The gain tracks the subset's base rate. Fitting the gain separately by region and topic gives values from 0.0 (technology, 9% YES) through 0.2–0.3 (science, politics, business, sports, 20–28% YES) to 0.46–0.60 (economy, India, society, 35–41% YES), with nearly the same ordering for both models (Figure 13b). Part of this is an artefact of the anchor: when a subset's base rate differs from the anchor, a single gain must also absorb the mismatch. That suggests shrinking toward a local base rate rather than a global one.

Why it works. If each read is the true signal plus independent noise, the optimal gain equals the read's reliability, the classical attenuation for an error-prone regressor. A single read has reliability 0.83 for gpt-oss and 0.77 for Qwen3-8B, which would justify gains near 0.8. The fitted gains are much lower, so read noise explains only 37–40% of the needed shrinkage; the remaining 60–63% is systematic over-scaling of the news. The models move too far, consistently, which is why one constant captures most of the available improvement and why rules that try to adapt the gain per question find nothing more (rows 7–9). For Qwen3-32B the calibration set alone cannot establish this split ($\hat{\kappa}$'s interval reaches 1.1; Section 6), but its held-out optimal gains (0.22–0.36) lie far below its reliability (0.71), so the same conclusion holds on the larger sets.

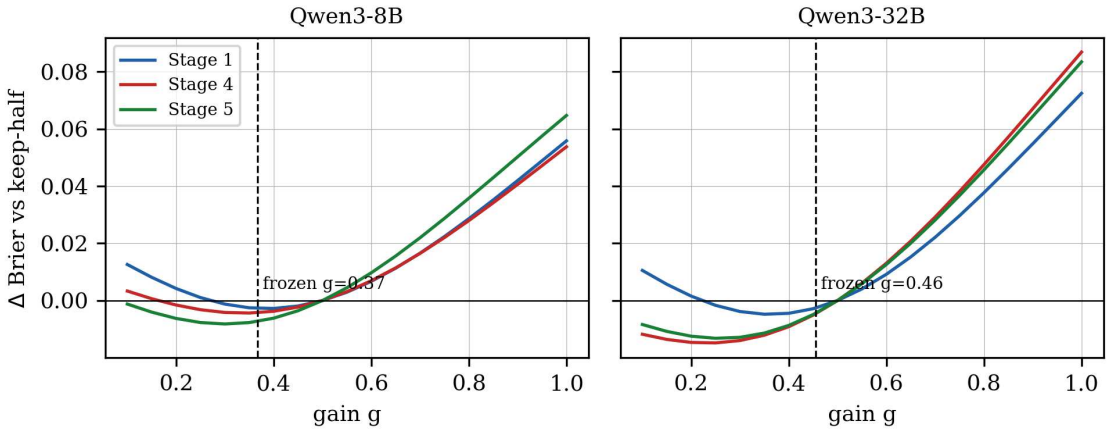


Figure 9: Test Brier relative to keep-half as a function of the gain, on each held-out set, for both models. Dashed: the gain frozen from the calibration set. For Qwen3-32B the held-out optimum (0.22–0.36) lies well below the frozen value; the frozen gain still improves on keep-half but leaves most of the available improvement unused.

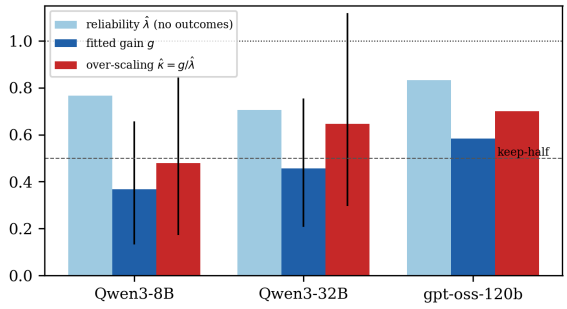


Figure 10: Reliability $\hat{\lambda}$ (no outcomes needed), fitted gain g and over-scaling $\hat{\kappa} = g/\hat{\lambda}$ on the calibration set, with 95% bootstrap intervals; gpt-oss-120b from an earlier analysis. Every model's gain is below its reliability, and well below one.

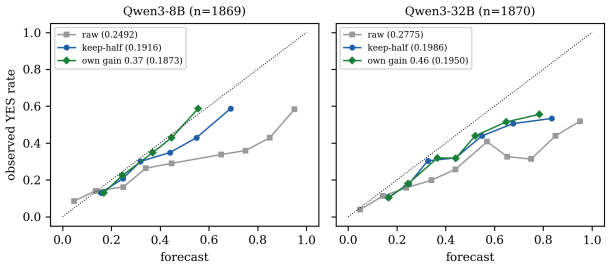


Figure 11: Reliability diagrams on the pooled held-out test splits (Brier in parentheses). The 32B model is more over-confident raw and remains under-shrunk at its frozen gain.

Calibration, not discrimination. For $0 < g < 1$ the shrink is a strictly increasing transformation of the forecast, so it cannot change ranking skill; it improves the Brier score only through calibration. Table 16 makes the decomposition explicit on the pooled held-out splits. Raw reads are worse than predicting the dev base rate for every question (Brier skill -0.21 and -0.34), with a calibration slope of about 0.43; any reasonable shrink turns this into positive skill. Keep-half leaves a slope just below one (forecasts still slightly too extreme), while the Algorithm 1 gain over-shrinks on the pooled splits (slope 1.9–2.4), because the pooled splits include Stage 1. Resolution falls slightly with stronger shrinkage, while reliability (miscalibration) falls much more. The models' ranking skill is modest (AUC 0.67 and 0.66) and is what any calibration rule has to work with.

Miscalibration does not disappear with scale. Raw 32B forecasts are more over-confident than 8B forecasts (raw Brier 0.278 against 0.249 pooled; Figure 11), and both models' held-out optimal gains are far below one. On the calibration set alone, however, the 32B model's over-scaling factor $\hat{\kappa} = 0.65$ has an interval reaching 1.11, so H3 is not shown there, and the change from 8B to 32B (H6) is not distinguishable from zero. Read noise is, if anything, larger for the bigger model ($\hat{\lambda} = 0.71$ against 0.77; Figure 10).

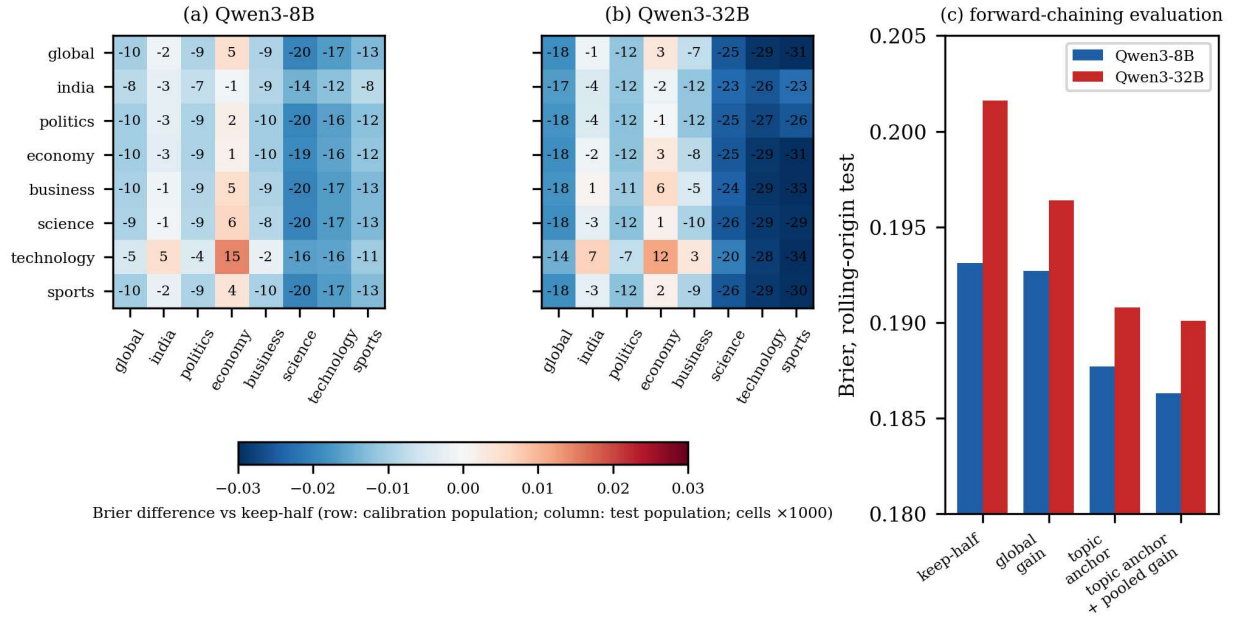


Figure 12: (a, b) Cross-population transfer: Brier difference against keep-half (×1000) when the gain is calibrated on the row population and deployed on the column population with its own anchor; blue favours the calibrated gain. (c) Forward-in-time evaluation of four rules on 1,246 test questions.

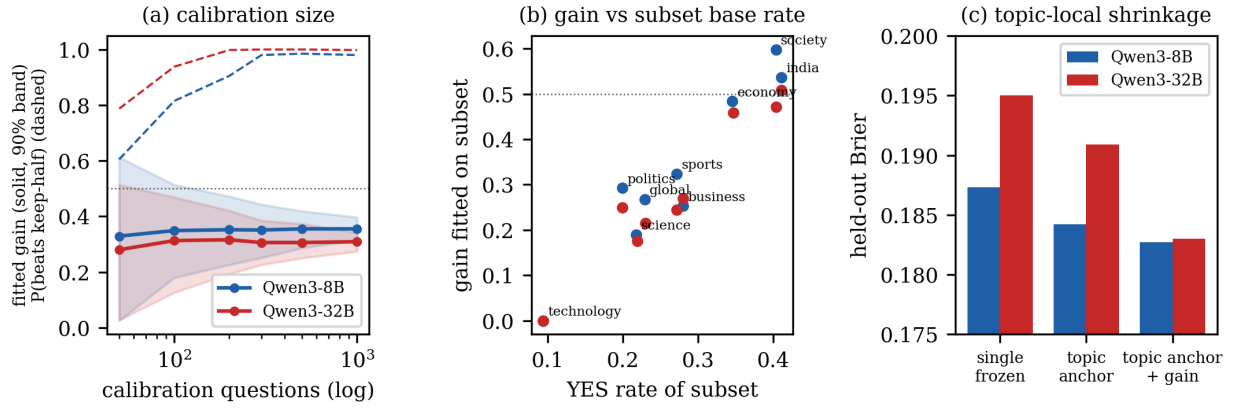


Figure 13: Exploratory analyses of the gain. (a) Gain fitted on random calibration subsets of the pooled held-out questions (median and 90% band) and the probability that it beats keep-half on the remaining questions. (b) Gain fitted per region or topic against that subset's YES rate. (c) In-sample-period comparison of the single frozen rule with topic anchors fitted on dev splits, superseded by the forward-in-time evaluation of Figure 12c.

Order sensitivity persists but shrinks; a second read still does not pay. Reversing the evidence moved the 32B model's log-odds by 0.516 on average against 0.456 for a plain re-run, a difference of 0.060 $[0.012, 0.108]$, about a third of the 8B model's excess (H4). Averaging a reversed second read again did not improve accuracy (H5).

Cross-population transfer. To test directly whether the right shrink is a property of the model or of the questions, we split the questions into eight populations (global and Indian news; six topics), divided each

Table 17: Main results under label correction calibrated by the 200-question audit. "Point": each observed NO switched to YES with probability 0.092 and each observed YES to NO with probability 0.036; "upper": the audit's upper 95% bounds (0.153, 0.121). All gains, anchors and topic rules are refitted in each of 200 draws; ranges show variation over corrections only, not sampling uncertainty. Observed values are question-weighted on questions read by both models and can differ slightly from the fixed-effect pools elsewhere.

Comparison (Brier difference)	Observed	Corrected (point) [range]	Corrected (upper) [range]
8B: Algorithm 1 gain vs keep-half, Stages 4+5	−0.0044	−0.0046 [−0.0077, −0.0006]	−0.0064 [−0.0115, −0.0009]
8B: same, all three sets	−0.0012	−0.0027 [−0.0058, +0.0012]	−0.0053 [−0.0092, −0.0004]
8B: topic-local forward rule vs keep-half	−0.0068	−0.0066 [−0.0102, −0.0036]	−0.0078 [−0.0121, −0.0038]
32B: Algorithm 1 gain vs keep-half, Stages 4+5	−0.0132	−0.0116 [−0.0160, −0.0063]	−0.0127 [−0.0192, −0.0069]
32B: same, all three sets	−0.0054	−0.0062 [−0.0098, −0.0021]	−0.0088 [−0.0138, −0.0032]
32B: topic-local forward rule vs keep-half	−0.0115	−0.0105 [−0.0143, −0.0061]	−0.0117 [−0.0166, −0.0067]
32B minus 8B, Algorithm 1 gains	+0.0028	+0.0017 [−0.0011, +0.0041]	+0.0010 [−0.0017, +0.0034]

chronologically into 40% calibration and 60% test, fitted the Algorithm 1 gain on each population's calibration part, and applied it to every population's test part with that population's own anchor (Figure 12a,b). Two patterns are clear. First, the result depends far more on where the rule is deployed than on where it was calibrated: rows of the matrices are nearly identical, columns are not. Gains calibrated anywhere improve on keep-half for most target populations, by roughly 0.01–0.03 Brier for science, technology and sports, but not for economy questions or Indian news, whose best gains are higher (8B: 0.42 and 0.33; 32B: 0.37 and 0.32) than those of the other topics (0.06–0.31). Second, the excess over the target's own best gain is small in most cells (median below 0.002), so a moderately misestimated gain costs little, as long as it is well below one half and the anchor is local.

N Further candidate rules screened after the audit

After the external test we screened ten further rules: the six in Table 18, anchors per phrasing class (threshold, deadline, other), a coverage-scaled version of the outcome-free correction of Section 4.2, a logistic combination of the question-only and evidence reads, and a label-free predictor of the gain (below). The six in the table were tested on News-Bench (forward pool, 1,245–1,246 test questions) and the external set (622–623), each registered before it was computed on these already-used data. A rule passed only if it beat the global anchored gain with a 95% upper bound below zero on both sets, for both models where applicable. None passed (Table 18). Phrasing-class anchors: -0.0011 $[-0.0029, +0.0007]$ and -0.0014 $[-0.0034, +0.0006]$ on News-Bench, $+0.0007$ and $+0.0004$ on the external set. Coverage-scaled correction against the constant one: -0.0003 $[-0.0043, +0.0034]$ and -0.0024 $[-0.0059, +0.0007]$. Two-read combination against the anchored gain: $+0.0035$ $[+0.0001, +0.0068]$, $+0.0032$ and $+0.0012$. Across all 57 comparisons with an interval computed on that day, Holm and Benjamini–Hochberg adjustment leave only the coverage-effect results of Section 4.2 and the value-of-news comparisons significant, together with the pooled-model comparison against the 32B model on News-Bench; no failed test becomes significant. We also asked whether a statistic of a population's raw forecasts alone (no outcomes) predicts its best gain, which would allow calibrating a new domain before any question resolves: across 16 populations per model no statistic did (best rank correlation 0.45, $p = 0.08$), and a gain predicted from forecast spread was no better than the median gain of the other populations (leave-one-population-out Brier 0.1920 against 0.1912 for 8B). A default gain borrowed from other populations recovered about two thirds of the attainable improvement over keep-half (0.1912 against 0.1968; oracle 0.1878).

The fitted exponent of the saturating rule is close to one for Qwen3-8B (0.91–1.09): over-reaction is linear in log-odds. Averaging the two models matches whichever model is better on a set (8B on News-Bench, 32B on the external set) without beating it.

Table 18: Screened rules (Brier difference against the global anchored gain; News-Bench / external set). All intervals include zero or favour the baseline unless shown.

Rule	Qwen3-8B	Qwen3-32B
Saturating shrink, $g\,\mathrm{sign}(m)\lvert m\rvert^{\alpha}$	$+0.0001$ / $+0.0001$	-0.0001 / $+0.0001$
Asymmetric gains for moves toward YES and NO	$+0.0015$ $[+0.0001, +0.0030]$ / $+0.0001$	$+0.0003$ / -0.0018 $[-0.0045, +0.0009]$
Online gain (exponential weights, half-life 300)	-0.0015 $[-0.0035, +0.0005]$ / -0.0004	-0.0013 $[-0.0037, +0.0009]$ / -0.0002
Gain scaled by evidence volume	-0.0004 / $+0.0002$	-0.0001 / $+0.0003$
Both models' log-odds averaged (vs 32B alone)	-0.0036 $[-0.0053, -0.0020]$ / $+0.0026$ $[-0.0043, +0.0094]$	
Both models' log-odds averaged (vs 8B alone)	0.0000 / -0.0095 $[-0.0172, -0.0016]$	
Disagreement-adaptive gain on the average	-0.0010 $[-0.0026, +0.0004]$ / $+0.0001$	

Table 19: Standard recalibration methods against keep-half and against the measured shrink, pooled over the three held-out sets (fixed effect over per-set cluster bootstraps). Methods marked "calibration set" are fitted on the same 300 questions as the regression gain, which is the comparator in the "vs measured gain" columns.

	Qwen3-8B		Qwen3-32B	
Method	vs keep-half	vs measured gain	vs keep-half	vs measured gain
Regression gain (0.37 / 0.46), dev anchor	-0.0046 $[-0.0077, -0.0016]$	—	-0.0037 $[-0.0048, -0.0026]$	—
Platt / linear-in-log-odds, calibration set	-0.0015 $[-0.0084, +0.0054]$	$+0.0034$ $[-0.0009, +0.0076]$	-0.0062 $[-0.0146, +0.0022]$	-0.0025 $[-0.0099, +0.0049]$
Isotonic regression, calibration set	$+0.0073$ $[+0.0003, +0.0142]$	$+0.0117$ $[+0.0056, +0.0179]$	-0.0069 $[-0.0145, +0.0007]$	-0.0031 $[-0.0098, +0.0035]$
Temperature (slope only, no anchor)	$+0.0408$ $[+0.0341, +0.0476]$	$+0.0456$ $[+0.0363, +0.0550]$	$+0.0471$ $[+0.0409, +0.0533]$	$+0.0510$ $[+0.0441, +0.0579]$
Platt, fitted on each set's dev split	-0.0028 $[-0.0098, +0.0042]$	$+0.0018$ $[-0.0024, +0.0059]$	-0.0068 $[-0.0153, +0.0016]$	-0.0031 $[-0.0106, +0.0043]$
Algorithm 1 gain, fitted on each set's dev split	-0.0022 $[-0.0085, +0.0041]$	$+0.0022$ $[-0.0010, +0.0054]$	-0.0059 $[-0.0137, +0.0020]$	-0.0022 $[-0.0089, +0.0046]$

O Standard recalibration methods: full results

P Case study: apparent gains from shuffling evidence order

The order-ensembling claim that motivated much of this work is analysed in detail in a companion manuscript (Anonymous, 2026); we summarise it because it illustrates the second failure mechanism. On 1,676 resolved questions forecast by Qwen2.5-7B-Instruct, three shuffled evidence orders beat three chronological readings by 0.0151 Brier as first analysed. The gain halves when exact 0/1 outputs are clipped at 0.01 instead of 10^{-6}, and 87% of what remains comes from the 12.8% of questions with any exact 0 or 1 output; without them it is 0.0023 $[-0.0004, 0.0052]$. Its largest source is macroeconomic data series presented in chronological order, where the model answers exactly 0.0 after the last move of the series (40% of decodings after a decline against 26% after a rise, with no ranking skill, AUC 0.48); shuffling merely restores the base rate. Regenerating all replies from the open weights, on two hardware platforms, reproduces these findings, and Llama-3.1-8B, which produces no excess of zeros, shows no gain. Order sensitivity itself is real: reversing the evidence moved Qwen3-8B's log-odds 41% more than a plain re-run (0.563 against 0.399). But it is a consistency problem rather than lost information, and averaging a reversed read improves accuracy by at most a few thousandths (rows 11–13; H5).

Q Registration provenance

Table 20 lists, for each group of comparisons, the commit that first contains its protocol and the commit that first contains its result. Commit timestamps are set by the committer and are not externally certified; the order of commits within the repository history is what supports level V. Level L means that protocol and result first appear in the same commit, so their order rests on the written log alone.

R A worked example

One News-Bench question from the Stage 5 test split illustrates the over-reaction that the survivor corrects. The question, written by the pipeline on 10 November 2025 from headlines published 7–9 November, asks:

Table 20: First commits containing each protocol and result (repository history; local time, UTC+5:30).

Comparisons	Protocol commit	Result commit	Note	Level
1–3 evidence selection	`22015b1`, `4ce232f`, `7a2d050` (27 Oct 20:46–21:28)	`986658a` (29 Oct 00:27)	amendments before any call	V
4 gain on no-news move	`986658a`	`986658a`	logged 27 Oct 23:00	L
5 variance-scaled gain	—	`986658a`	post hoc	E
6 gain 0.37 (Stages 4, 5)	`986658a`	`986658a`	logged 28 Oct	L
6 gain 0.37 (Stage 1 test)	—	`986658a`	post hoc	E
7 lead-time gain	`986658a`	`986658a`	logged 28 Oct	L
8–9 two-channel, disagreement	`43f40bd` (31 Oct 20:33)	`758e034` (31 Oct 20:34)	analysis script in protocol commit	V
10–13 multiple reads	`986658a`	`986658a`	logged 27 Oct	L
14 market fusion	`986658a`	`986658a`	logged 28 Oct	L
15–16 output handling	`7d0c853` (31 Oct 21:05)	`b18dbfd` (31 Oct 21:10)		V
H1–H5 scale replication	`7f100cc` (1 Nov 04:03)	this paper	before any 32B call	V
H6–H8 scale amendment	`88ed2f6` (1 Nov 04:16)	this paper	before any 32B output read	V
Gain factorisation check	`e2cd68d` (1 Nov 03:53)	this paper	retrospective sets: E	E
External test, topic anchors, theory prediction	`8be4c3c`, `70479a3`	`0cbe0b3`	raw outputs committed first (`9c43441`)	V
Mechanism test (over-separation, drift)	`e7f6cdc`	`2be2d9b`, corrected `4ab16fa`	data already seen†	V†
Screens 1–2 (six rules)	`2bfc41f`, `fbecbd6`	`bbed2c3`, `1c1d74e`	data already seen†	V†
News push, phrasing anchors	`b1cd492`	`638d14f`	data already seen†	V†
News push, external set (8B)	`638d14f`, script `ea1738d`	`5e88483`	reads generated after registration (`e6d4536`)	V
Outcome-free correction, dose	`27f7a08`	`1996ea7`	data already seen†	V†
Outcome-free correction, external set	`1996ea7`	`17c66de`	outcome predicted in advance	V
Placebo evidence (8B)	`27f7a08`, script `17c66de`	`fe2d709`	reads generated after registration (`799561a`)	V
Coverage intensity, recency	`1f62b83`	`9128bba`	data already seen†	V†
Contradicting headlines, value of news, two reads	`224ac2f`	`54c3a81`	data already seen†	V†
Fresh set (topic anchors, bias, push, outcome-free correction)	`1681338`, `4ab16fa`, `1c1d74e`, `638d14f`, script `8de7efe`	`779298c`	questions created after registration; outputs `3e1fbbe`	V
Second family (OLMo-2), neutral prompt	`8d78d10`, script `76eb3d3`	`9114195`	reads generated after registration (`ed1153f`)	V
gpt-oss-120b, thinking, warning prompt, dose	`810501d`, script `e54546c`	`bf4eb5c`	reads generated after registration (`a915047`)	V
Isolated reading, evidence tally	`923a347`, script `09114ca`	`f6c657e`	reads generated after registration (`c192363`)	V
Qwen3-32B placebo (4-bit, free GPUs)	`3fb2b34`	`71622c8`	reads generated after registration (`f6b896c`)	V
Proprietary frontier models (API router)	`f8b8fa7`, amended `568db44`, script before calls	`d6d6f65`	calls made after registration (`b3b2d26`); seed fix logged	V
Excess belief movement (label-free)	`bf63a35`	`5e047bb`	data already seen†	V†
Repetition and same-topic swap	`6f871b9`, amended `3400f55`, script `394bcfa`	`19bc12a`	reads generated after registration (`d5802c9`); donor rule amended before any read	V
Masking, dilution, anchoring, likelihoods, complements	`d3e4cd5`, script `2e2a859`	`19a462c`	reads generated after registration (`839192c`)	V
Paraphrases, other models, repeat reads, natural complements	`1031186`, amended before reads, script `313ce7f`	`f7d1606`	reads generated after registration (`7a407fe`)	V
Prospective question-only reads (700 open questions)	`9ec5d01`, script `acd2895`	`2ecfa35`	reads made and RFC 3161-stamped before any outcome; outcome tests pending	V

†Protocol committed before the quantity was computed, but on questions and reads already used in other analyses; the order is verifiable, the data were not new.

Will the Challenger, Gray & Christmas Job Cuts Report for November 2025, released in early December 2025, show a total number of announced US job cuts exceeding 160,000? The evidence consisted of 15 headlines from 13 outlets (headline text withheld; digests in the released package), most reporting that October 2025 layoffs had reached a 20-year high. The question resolved NO.

Qwen3-8B answered 0.85, reasoning that "job cuts have already surged" and that the trend "must continue at a rapid pace through November". Qwen3-32B answered 0.75 and quoted an annual range of job cuts as if it were a monthly base rate. Both readings extrapolate the latest month. With the Stage 5 dev base rate of 0.203, keep-half gives 0.546 (8B) and 0.467 (32B); each model's frozen gain gives 0.443 and 0.439. The Brier score of the 8B forecast falls from 0.723 raw to 0.298 under keep-half and 0.196 under its own gain. A single example proves nothing, but it shows the mechanism: the models move far with salient recent news, and a fixed pull toward the base rate removes much of the damage without needing to know which news was misleading.

S Datasheet for the question sets

Motivation. India-Bench and News-Bench were built to evaluate news-based forecasting with strict time-locking, in languages and regions under-represented in market-based benchmarks.

Composition. Binary questions with a closing date between February 2025 and October 2026; each has an as-of time, a list of headlines published before it (median 15), a resolution with cited sources, and a category. News-Bench Global has 2,963 questions (base rate 0.22); India-Bench contributes 519. Topic YES rates range from 0.09 (technology) to 0.41 (India).

Collection. Headlines come from public news feeds with publication and first-seen timestamps; an LLM drafts questions about events not yet known at the as-of time; duplicates and questions without a clear resolution criterion are dropped. A search-grounded LLM resolves each question after its closing date, and unresolvable questions are removed.

Known issues. Resolution is automatic and can be wrong, especially for questions whose wording is stricter than the event (for example, a deadline that passes days before the event occurs). A sample of live questions found about 14% already decided when asked. Both issues affect every arm of every comparison equally and cannot create a difference between arms, but they add noise and cap attainable accuracy.

Distribution. Question text written by the pipeline, outcomes, model replies and row-level predictions are released for research use. Headline text is replaced by SHA-256 digests with retrieval identifiers; third-party platform content with restrictive terms (Metaculus, Polymarket) is likewise digested.

T Reproducibility checklist

- **Models and versions.** `Qwen/Qwen3-8B` (main reads: Hugging Face revision `main` at run time, 27–28 October 2026, not recorded as a hash; vLLM v0.32.0, fp16, two T4 GPUs; later control reads on a commercial cloud GPU in bf16, from the cached snapshot `b968826d`, with vLLM 0.30.0 recorded for the repetition and swap reads) and `Qwen/Qwen3-32B` (snapshot `9216db57`, vLLM installed unpinned at run time on 1 November 2026, bf16, one H100), thinking off; Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct for the order case study; sampling temperature 0.7, top-p 0.8, top-k 20, maximum 700 new tokens, per-prompt seed from the first 8 hex digits of SHA-256 of the prompt key.
- **Prompts.** The exact template is in Appendix E. Released prompts contain question text and dates in full, but headline text is replaced by a SHA-256 digest with source name, date and URL; reconstructing the exact inputs requires retrieving the headlines from those sources, and the digests verify that a retrieved headline is the one used.
- **Two levels of reproduction.** From the released files alone, anyone can recompute every number in this paper from the stored predictions and replies (two standalone scripts check the main tables). Regenerating the inputs and replies requires re-retrieving headlines from the digests, the question-writer and resolver models, and, for the frontier models, an API router whose provider-side settings we could not fix; this second level is not guaranteed to reproduce the replies exactly, because sampled generation and inference software change.
- **Raw outputs.** Every reply used in this paper is stored and released, including the 9,664 Qwen3-32B replies.
- **Splits.** Chronological by closing date; first 40% dev, remaining 60% test; question identifiers per split are released.
- **Uncertainty.** Cluster bootstrap over source-by-ISO-week groups, 4,000 resamples, seed 13; fixed-effect pooling across sets with Cochran's Q.
- **Multiple comparisons.** We report every comparison with raw intervals. For the sixteen audit comparisons (Table 5) we also report Holm and Benjamini–Hochberg adjustments; the coverage hypotheses are reported unadjusted, each with its registered threshold. The surviving family also replicates on three disjoint question sets and at two model sizes, which guards against a chance finding more strongly than a correction on one set.
- **Compute.** Qwen3-8B on free Kaggle T4 GPUs (about 40 minutes per 4,700 generations); Qwen3-32B on one H100 for about 30 minutes including model download (under US$3); analysis on a laptop CPU.

U Broader impact

Calibrated forecasts help people decide under uncertainty, and over-confident ones mislead. The main practical result of this paper, that LLM forecasters are miscalibrated, by a YES bias or by over-dispersion depending on the questions, and that a measured shrink corrects much of both, reduces a specific harm: confident-sounding probabilities that overstate how likely news-driven events are. The audit format may also discourage a common failure in applied machine learning, reporting the one variant that worked. A risk is over-generalisation: our results come from seven models (four open, three proprietary), headline evidence and automatically resolved questions, the frontier results rest on 200 questions, and the size of the news push differs widely between models, so it must be measured for each system, not assumed.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- README.md
- CHANGESv2.md
- evidencepackageanonymized.zip