Forecasts as a Behavioral Probe of Language Models

Anonymous Authors[1]

Abstract

We use probabilistic forecasting as a black-box behavioral probe of latent beliefs in language models. Eight open-weight LLMs forecast 75 resolved binary questions across five issue axes, each paired with a polarity-inverted twin and elicited under pro/anti framing primes. The probe surfaces belief-like structure: forecasts share an axis-level lean across models, shift predictably under priming, and remain partially coherent under polarity inversion (median complement error 0.13). But the signal is partial. Only one of five axes survives a polarity-coding control, and the priming effect cannot be separated from compliance with assertive prompts. Forecasting is a usable probe for LLM priors, but apparent model "views" depend heavily on how questions are coded, inverted, and framed.

1. Introduction

Activation-level probes (Burns et al., 2023; Marks & Tegmark, 2023; Templeton et al., 2024; Park et al., 2024) are the standard tool for asking what language models "know," but they require white-box access, and most production models expose only sampling APIs. Apparent internal structure does not always correspond to coherent downstream behavior (Vafa et al., 2024).

Forecasting offers a black-box alternative. Probabilities on resolved propositions are comparable across models and admit tests of structure, coherence, and context sensitivity. Crucially, because the output is a scalar in $[0, 1]$, framing interventions and inversion checks yield directly measurable shifts in expressed belief rather than categorical labels that require interpretation. We construct a 75-question probe set grouped into five issue axes, each question carrying a polarity sign and a polarity-inverted twin. Eight open-weight models forecast each question under baseline, pro-frame, and anti-frame conditions, $N = 10$ samples per cell. We ask three things of the resulting model-by-question matrix:

1. Do forecasts exhibit axis-level structure beyond what polarity coding alone could produce, and do models differ from one another in stable ways?
2. Are forecasts coherent under polarity inversion?
3. Do framing primes shift forecasts predictably along the same axes?

Contributions. We release a 75-question forecast probe across five issue axes with polarity signs and polarity-inverted twins. Across eight open-weight models, the apparent shared axis-level lean is content-driven on only one of five axes (environmentalism, $\delta = +0.29$); on the others it is near-zero or reversed (econ $+0.16$, intl $+0.01$, tech -0.07, pol -0.20), and on political polarization forecasts are *negatively* associated with the assigned axis labels. A within-question permutation test detects between-model differentiation on two axes (international escalation, technological acceleration; Holm-adjusted $p < 0.005$), the same two where framing primes move forecasts most. Polarity-inverted forecasts are partially coherent (median $e = 0.13$ per cell), and framing primes shift forecasts in the expected direction on all 40 (model, axis) cells (caveats in §6).

2. Related Work

World models and their evaluation. Sequence models trained on restricted synthetic tasks can develop internal state representations that behave like world models (Li et al., 2023), but Vafa et al. (2024) show that apparent world-model competence often fails under coherence tests beyond next-token prediction.

Activation-level probing. Linear classifiers (Burns et al., 2023) and propositional probes (Marks & Tegmark, 2023) read beliefs from hidden states; sparse autoencoders decompose activations into features (Templeton et al., 2024); the linear representation hypothesis (Park et al., 2024) motivates much of this work. All of it requires white-box access.

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

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

Table 1. Issue axes for the 75 forecasting questions ($n = 15$ per axis). Polarity sign s: $+1$ if YES supports the axis, -1 if it contradicts it.

Issue axis	Claim	s (+/−)
Econ. Resilience	Economy absorbs shocks without failure	3 / 12
Environmentalism	Climate risks severe and accelerating	12 / 3
Tech Acceleration	AI/tech progress is rapid and transformative	14 / 1
Int'l Escalation	Great-power tensions intensifying	13 / 2
Pol. Polarization	US divisions deepen into institutional action	13 / 2

LLM forecasting. Retrieval-augmented systems can approach human aggregate accuracy on future-event questions (Halawi et al., 2024), and LLM ensembles can rival human crowds (Schoenegger et al., 2024). Live and rolling benchmarks treat forecasting on continually resolving questions as an evaluation of reasoning under uncertainty (Yang et al., 2026; Zeng et al., 2026). Existing work treats the probability output as a forecast to be scored for accuracy; we treat it as a quantitative measurement of belief, recording shifts under controlled interventions rather than accuracy alone.

3. Forecast Probe

The dataset contains 75 resolved binary questions, resolved March 2025–May 2026, grouped into five post-hoc *issue axes* (Table 1): 25 from Kalshi prediction markets[1] (5 per axis), 50 author-written (10 per axis). Each question carries a polarity sign $s \in \{+1, -1\}$ ($s = +1$ if YES supports the axis, $s = -1$ if YES contradicts it) and a polarity-inverted twin used for the negation-pair consistency check of Burns et al. (2023). Polarity signs are author-assigned; the coding is single-coder, a limitation discussed in Section 6. Construction details and the edits log are in Appendix A.

4. Elicitation and Metrics

Models and elicitation. We query eight open-weight models via OpenRouter; a ninth was excluded for an overall parse rate below the 90% retention threshold (Appendix B). Each retained model produces $N = 10$ independent samples per question per condition at temperature 1.0 from a zero-shot prompt that requests a probability in $[0, 1]$. Originals receive three conditions (baseline, pro-prime, anti-prime); inverses receive baseline only. The total is $8 \times (75 \times 3 + 75 \times 1) \times 10 = 24{,}000$ forecasts before parsing. Resolution dates are recorded for the 25 Kalshi questions; the 50 author-written questions resolve by end of 2025 against the criterion in the question text. The full model card (providers, release dates, claimed cutoffs) is Table 2 in Appendix B.

[1] https://kalshi.com

Polarity-adjusted score. For each question q with polarity s_q, we define the polarity-adjusted score

$$Z_{m,q} = \begin{cases} P_m(\text{YES} \mid q) & \text{if } s_q = +1, \\ 1 - P_m(\text{YES} \mid q) & \text{if } s_q = -1, \end{cases}$$

where P_m is the empirical mean over the ten samples for model m, and the polarity-adjusted outcome y_q^Z is y_q when $s_q = +1$ and $1 - y_q$ when $s_q = -1$. Values of $Z_{m,q}$ above 0.5 indicate the model leans toward the axis on that question regardless of phrasing; $y_q^Z = 1$ indicates the resolution agreed with the axis. Axis-level shared lean is summarized by the per-model signed error $E_{m,a} = \text{mean}_{q \in a}(Z_{m,q} - y_q^Z)$; bootstrap CIs and the formal rejection rule are in Appendix D.

Between-model differentiation. To test whether models differ in axis-specific ways, we permute the eight model labels within each question and compare the observed between-model variance of axis-level means to the resulting per-axis null ($n = 10{,}000$ permutations; Holm-adjusted across five axes; details in Appendix D).

Accuracy. Brier score $\text{Br}_{m,a} = \frac{1}{|a|} \sum_{q \in a} (P_m(\text{YES} \mid q) - y_q)^2$ against two climatology baselines: in-sample axis-mean and leave-one-out (LOO).

Framing interventions. For each (model, question) we run a *pro-prime* asserting the axis is true and an *anti-prime* asserting the opposite (Table 3, Appendix C); the framing is prepended to the same zero-shot prompt with all other parameters fixed. We report the per-axis pooled spread $\Delta_a = \text{mean}_{m,q \in a}(Z_{m,q}^{\text{pro}} - Z_{m,q}^{\text{anti}})$ and per-cell $\Delta_{m,a}$; paired-bootstrap CIs and the 40-cell sign check are in Appendix D.

Coherence under polarity inversion. For each pair (q, q^{-1}) we compute the coherence error $e_{m,q} = |\bar{P}_m(\text{YES} \mid q) + \bar{P}_m(\text{YES} \mid q^{-1}) - 1|$, where $\bar{P}_m$ is the empirical mean over the ten samples per cell.

5. Results

5.1. Forecasts Exhibit a Shared Axis-Level Lean

Figure 1 shows the polarity-adjusted mean score per (model, axis) cell. All eight models sit on the same side of 0.5 on four of five axes: leaning toward economic resilience and environmentalism, away from international escalation and political polarization. Technological acceleration is the exception.

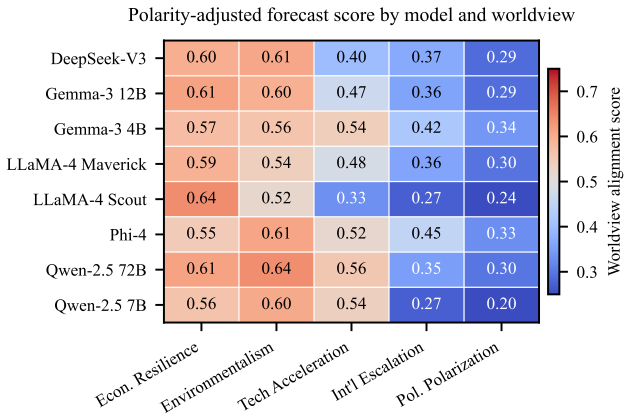

Figure 1. Polarity-adjusted mean score $\bar{Z}_{m,a}$ per (model, axis) cell. Red cells lean toward the axis ($\bar{Z} > 0.5$); blue cells lean against ($\bar{Z} < 0.5$). All eight models are on the same side of 0.5 on four of five axes (lean toward economic resilience and environmentalism, lean against international escalation and political polarization); technological acceleration is the exception. Comparison to polarity-adjusted outcomes ($\bar{y}^Z$ per axis: econ 0.67, env 0.27, int'l 0.47, pol 0.47, tech 0.53) is reported in Section 5.1.

The shared lean is quantified by the per-model signed errors $E_{m,a} = \bar{Z}_{m,a} - \bar{y}^Z_a$. At the point-estimate level, all eight $E_{m,a}$ are positive on environmentalism ($\bar{y}^Z = 0.27$, range $[+0.26, +0.37]$); all eight are negative on economic resilience ($\bar{y}^Z = 0.67$, range $[-0.12, -0.03]$), international escalation ($\bar{y}^Z = 0.47$, range $[-0.19, -0.01]$), and political polarization ($\bar{y}^Z = 0.47$, range $[-0.23, -0.12]$); on technological acceleration the per-model errors disagree in sign, three positive and five negative ($\bar{y}^Z = 0.53$, range $[-0.20, +0.02]$).

Question-bootstrap CIs ($n_b = 10{,}000$): the cross-model mean $\bar{E}_a$ excludes zero only on environmentalism (point $+0.32$, CI $[+0.14, +0.50]$); on the four other axes the cross-model CI includes zero (econ res $[-0.25, +0.09]$, int'l escal $[-0.30, +0.08]$, pol polarization $[-0.38, +0.04]$, tech accel $[-0.27, +0.16]$). Per-model, $8/8$ CIs exclude zero on environmentalism, $2/8$ on political polarization, $1/8$ on technological acceleration, $0/8$ on economic resilience and international escalation. Interpretation in Section 6.

5.2. Models Do Not Strongly Differentiate by Axis

To test whether models differ from one another in axis-specific ways, we ask: within each question, would it matter which model is credited with which forecast? We randomly reassign the eight forecasts on each question to the eight models, recompute each model's axis-level mean, and compare the observed between-model variance of those means to the $n = 10{,}000$-shuffle null. Two axes reject (Holm-adjusted across five): international escalation ($p_{\text{adj}} < 0.001$) and technological acceleration ($p_{\text{adj}} < 0.005$). On the other three, we cannot reject the null: the visible differences among models' axis-level means look like chance (Figure 4, Appendix E.2). The two axes with significant between-model variance are also the two with the largest prime-induced shifts (Section 5.4).

5.3. The Shared Lean Costs Accuracy

The shared lean is associated with an accuracy cost when it disagrees with outcomes. Because all eight models lean the same way on the same axes (§5.1), they share that cost. We score forecasts with Brier against two climatology baselines (lower is better): an in-sample climatology that forecasts each question with the axis mean outcome, and a leave-one-out (LOO) climatology that excludes the question being scored. LOO is the fairer baseline and is uniformly worse (higher) than in-sample by ≈ 0.03 per axis. Under LOO, all eight models beat baseline on economic resilience (model Briers 0.18–0.21 vs. LOO 0.28); 4 of 8 beat baseline on international escalation; 2 of 8 on technological acceleration; and only 1 of 8 on environmentalism and political polarization. The pattern is the §5.1 pattern: where the lean is largest in absolute terms (env at $E_{m,a} \in [+0.26, +0.37]$, pol at $[-0.23, -0.12]$), models uniformly fail to beat climatology; where the lean is smallest (econ at $[-0.12, -0.03]$), they uniformly beat it. Axis-level structure should therefore not be interpreted as forecasting accuracy. Full per-model, per-axis Brier in Appendix F.

5.4. Forecasts Are Sensitive to Framing Along the Same Axes

Each question is also forecast under two framing conditions: a *pro-prime* that prepends a sentence asserting the axis is true (e.g., for environmentalism, "Climate change is accelerating and its impacts are increasingly severe"), and an *anti-prime* that prepends a sentence asserting the opposite (Appendix C for all ten sentences). Figure 2 shows the resulting polarity-adjusted scores. The pro-prime shifts forecasts toward the axis and the anti-prime shifts them away in every one of the 40 (model, axis) cells (per-axis binomial $p = 0.0039$). Pooled spreads by axis: technological acceleration $\Delta = 0.36$, international escalation 0.21, economic resilience 0.18, environmentalism 0.17, political polarization 0.15. The two largest spreads are on the same two axes that reject the permutation null (Section 5.2); the smallest is on political polarization. The placebo-prime caveat that limits the reading of this result is in Section 6; a preliminary corrective-priming analysis is in Appendix G.

5.5. Polarity Inversion Is Partially Coherent

A coherent probabilistic belief about an event implies that probabilities on the question and its negation sum to one. Pooled across the $8 \times 75 = 600$ (model, question) cells,

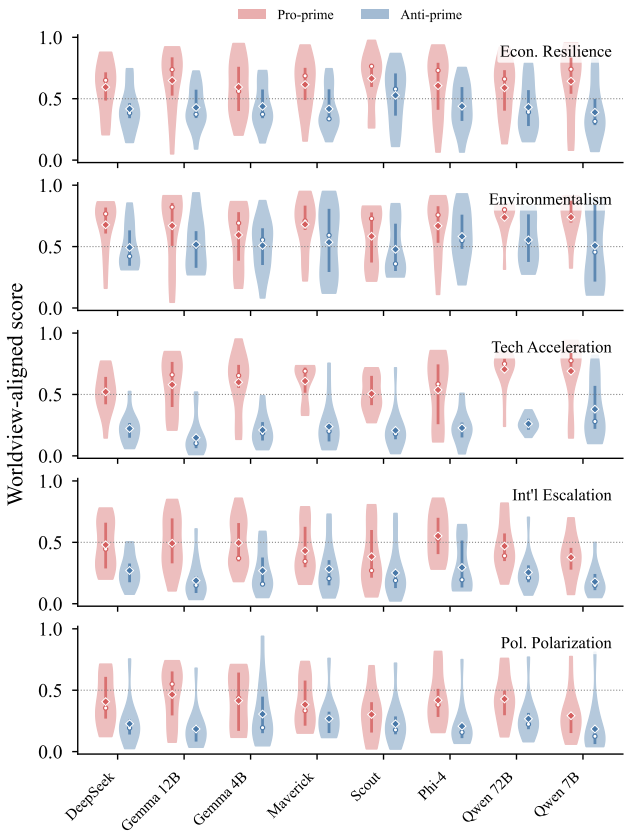

Figure 2. Polarity-adjusted scores under pro-prime (red) and anti-prime (blue) framings for each model and axis. Diamonds mark condition means across questions; dotted line marks 0.5. The pro-prime is above the anti-prime in every one of the 40 cells.

the median coherence error $e_{m,q}$ is 0.13 and 55% of cells satisfy $e_{m,q} \leq 0.15$; per-model medians range from 0.07 (DeepSeek-V3) to 0.19 (LLaMA-4 Scout), and per-axis medians are tightest on economic resilience (0.11) and loosest on technological acceleration (0.18). The full pooled distribution is in Figure 3 (Appendix E). Only 1 of 74 rated inverses falls in the pre-specified low-naturalness stratum, so we report this pooled distribution rather than stratifying (Appendix A).

A median coherence error of 0.13 is non-trivial. Models that cannot reliably treat q and q^{-1} as complementary on individual questions are unlikely to be expressing a stable axis-level belief; this weakens the reading of the shared axis-level lean (§5.1) as evidence of a coherent prior and is consistent with the polarity-coding artifact discussed in §6.

6. Discussion and Limitations

Content-driven discrimination is rare. The discrimination $\delta_a = \bar{P}(\text{YES} \mid s = +1, a) - \bar{P}(\text{YES} \mid s = -1, a)$ is content-driven only on environmentalism (+0.29). On the others it is small or reversed: economic resilience +0.16, international escalation +0.01, technological acceleration −0.07, political polarization −0.20. The apparent shared axis-level lean of §5.1 therefore does not license a "shared worldview" reading on four of five axes; on political polarization, forecasts are negatively associated with the assigned axis labels. A balanced-sample δ using the inverse twins (Section 5.5) would sharpen this check at the cost of assuming coherence we only observe partially.

Future work. *Logits over scalars.* The parsed probability discards most of the answer-position distribution; reading logits directly should sharpen the signal and remove format dependence. *Mechanistic grounding.* Pair axis-level behavioral signatures with activation probes (Burns et al., 2023; Marks & Tegmark, 2023) or SAE features (Templeton et al., 2024) to test whether the observed matrix structure corresponds to identifiable internal directions. *Scale via live questions.* An auto-generated stream of binary questions resolving after each model's cutoff removes leakage by construction and grows the probe by orders of magnitude. The most direct extensions are procedural: placebo primes, held-out axis definitions, independent polarity coding, and logits-based elicitation.

Reproducibility

We release the question bank (source URLs, polarity labels, inverse twins), prompts, parsing code, raw predictions, and analysis scripts. The model card (Table 2) includes exact OpenRouter strings, run dates, temperature, and top-p. Code and predictions are under MIT; Kalshi question text follows Kalshi's terms. Resolution dates for the 25 Kalshi questions combined with per-model cutoffs in Table 2 support a leakage-stratified rerun without re-elicitation.

Impact Statement

The framing-prime result, that pro- and anti-framed prompts move every (model, axis) cell in the asserted direction, indicates how readily forecasts can be steered by assertive context, with implications for downstream uses where models inform consequential probabilistic judgments. The accompanying methodological caution applies to any work eliciting LLM "views" on contested issues from fixed YES/NO framings without parallel inverse controls: reported leans can be substantially polarity-coding artifacts.

References

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *Proceedings of the 11th International Conference on*

Learning Representations (ICLR), 2023.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, 2023.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 39643–39666, 2024.

Schoenegger, P., Tuminauskas, I., Park, P. S., Bastos, R. V. S., and Tetlock, P. E. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45), 2024.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jermyn, A., Olsson, C., Olah, C., Henighan, T., et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Vafa, K., Chen, J. Y., Rambachan, A., Kleinberg, J., and Mullainathan, S. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Yang, Q., Mahns, S., Li, S., Gu, A., Wu, J., and Xu, H. LLM-as-a-prophet: Understanding predictive intelligence with prophet arena. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VpiHkMSPqI.

Zeng, Z., Liu, J., Chen, S., He, T., Liao, Y., Tian, Y., wangjinpeng.levi, Wang, Z., YangYang, Yin, L., Yin, M., Zhenwei, Z., Cai, T., Chen, X., Chen, Z., Chen, J., Du, Y., Gao, X., Guo, J., HU, L., Jiao, J., Li, X., Liu, J., nishuang, Wen, Z., Zhang, G., Zhang, K., , Blanchet, J., Qiu, X., Wang, M., and Huang, W. Futurex: An advanced live benchmark for LLM agents in future prediction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=z28PLIEj6l.

A. Dataset Construction

A.1. Sources, Curation, and Edits

The released dataset is 75 questions: 25 from Kalshi prediction markets (5 per axis) and 50 author-written (10 per axis); totals 15 per axis. The Kalshi resolutions span March 2025–May 2026 and rely on Kalshi's own oracle (which cites official statistics, government records, or recognized news outlets); the author-written questions resolve by end of 2025 against criteria specified in the question text. Construction was a two-pass curation followed by a small set of targeted edits.

The pilot dataset (`data/data2_signed.csv`, 86 rows) was trimmed to balance per-axis counts and remove near-duplicates: six economic resilience items dropped (five Kalshi CPI threshold variants and one custom CPI question), four technological acceleration items dropped (three Kalshi FLOP threshold variants and one Kalshi robotaxi question), one international escalation item dropped (a duplicate Strait-of-Hormuz horizon), and one environmentalism item dropped (Kalshi Florida hurricane) with one added (`KXHMONTH-26JAN`, hottest January). These are balancing and de-duplication decisions; dropped items are released alongside the retained set.

Reviewer-touched items. Four items received reviewer-level attention beyond a single-pass coding. We list them so the assignments can be audited.

- `KXICEERO-26MAR-APR01_Kalshi` (political polarization). In the pilot this question was tagged polarity-ambiguous and held at $s = 0$. For this release we (i) rephrase the question to drop pre-2025 context dependency for models with August 2024 cutoffs and (ii) commit $s = -1$, parallel to `CUSTOM-BUDGET-ONTIME-25`: Congress passing appropriations on a contested issue counts as institutional functioning, anti-axis. The rephrased text is "When will ICE removal operations be funded? (Before Apr 1, 2026)."
- `KXDEBTLEVEL-25_Kalshi` (economic resilience). $s = -1$. Reviewer decision: rising US debt indicates fiscal stress, against the economic-resilience axis.
- `KXAMERICAPROFIT-25_Kalshi` (economic resilience). $s = +1$. Reviewer decision: a monthly surplus indicates fiscal capacity, supporting the axis.
- `KXHURPATHGULFCOAST-25_Kalshi` (environmentalism). $s = +1$. Reviewer decision: a Gulf-Coast Cat-2+ landfall is a climate-supporting event regardless of the question's neutral phrasing.

Status of analyses. Forecasts are collected on all 75 questions; results in Section 5 use the full 75-question dataset. The rephrased `KXICEERO` item was re-elicited under all three conditions (baseline, pro-prime, anti-prime) plus the new inverse twin's baseline; the missing primed conditions for `KXHMONTH-26JAN` from the pilot were filled at the same time.

A.2. Inverse Twins

Each question is paired with a polarity-inverted twin whose YES outcome corresponds to the original's NO. Each twin is reviewed for naturalness on a 1–5 scale (high ≥ 4 indicates a clean logical complement; low ≤ 3 a forced reformulation). In the released set, 73 twins are rated ≥ 4, 1 is rated 3 (flagged `low_naturalness_3` in the released data), and the twin for the rephrased `KXICEERO` ("Will ICE removal operations remain unfunded through April 1, 2026?") is unrated. The low-naturalness stratum has only one item, so we report pooled coherence statistics in Section 5.5 rather than stratifying.

B. Elicitation Details

B.1. Baseline Prompt

Forecasts are elicited zero-shot with a single template, instantiated with the question text:

```
You are a probabilistic forecaster.  Briefly explain your reasoning in one
sentence, then on a new line output ONLY the probability as a number between 0
and 1.
Question:  {question_text}
Reasoning:
```

Table 2. Models analyzed. Release dates are public announcement dates. Cutoffs are provider-documented training-data cutoffs where reported. All models accessed via OpenRouter at temperature 1.0; exact OpenRouter model strings and run timestamps accompany the released data. Gemma-3 27B was queried but excluded because the zero-shot template returned malformed probability strings on more than half of its queries.

Model	Provider	Release	Cutoff
DeepSeek-V3	DeepSeek	Dec 2024	Jul 2024
Gemma-3 4B	Google	Mar 2025	Aug 2024
Gemma-3 12B	Google	Mar 2025	Aug 2024
LLaMA-4 Maverick	Meta	Apr 2025	Aug 2024
LLaMA-4 Scout	Meta	Apr 2025	Aug 2024
Phi-4	Microsoft	Dec 2024	Jun 2024
Qwen-2.5 7B	Alibaba	Sep 2024	not reported
Qwen-2.5 72B	Alibaba	Sep 2024	not reported

Temperature is fixed at 1.0 and the top-p default is left at the provider setting (released alongside the data). The model is sampled $N = 10$ times per (question, condition) cell. Each sample produces one forecast.

B.2. Priming Prompt

The pro- and anti-prime conditions prepend a single framing sentence to the same template, after the role declaration and before the reasoning instruction:

```
You are a probabilistic forecaster. {framing} Briefly explain your reasoning in
one sentence, then on a new line output ONLY the probability as a number between 0
and 1.
Question: {question_text}
Reasoning:
```

The full set of ten framing sentences (one pro and one anti per axis) is reproduced in Table 3.

B.3. Probability Parsing and Model Retention

A response is scanned for a probability literal using a four-step rule: (1) the last non-empty line is parsed as either a bare number in $[0, 1]$ or as the final well-formed probability literal on that line; (2) if that fails, all lines are scanned bottom-up for an explicit label of the form `Probability: X.XX`; (3) if that fails, scanned again for a bare probability literal on its own line; (4) as a last resort, a percentage on the last line is matched and divided by 100. Responses yielding no valid probability under any rule are dropped.

Retention rule. A model is retained if its parse rate averaged over baseline and primed conditions reaches 90%. Seven retained models clear 99% overall. Phi-4 is the next-lowest retained, at 93.3% baseline, 95.7% pro-prime, 98.1% anti-prime (95.7% overall); roughly 7% of its baseline cells fail to parse and are silently dropped. The excluded Gemma-3 27B was at 94.9% baseline, 82.4% pro-prime, 83.2% anti-prime (86.8% overall), the only model below the 90% threshold. The residual concern is that Phi-4's baseline parse rate is lower than its primed rates, so any baseline-vs-primed comparison for Phi-4 (notably Appendix **??**) compares a slightly filtered baseline subset to a fuller primed set.

C. Priming Framings

The framing sentences were authored manually to match in length (23–29 words; mean 26), declarative form, and topical specificity across pro and anti pairs within each axis. They are deliberately direct rather than subtle; a softer placebo-controlled variant is named in Section 6 as a planned follow-up. Each framing is prepended verbatim to the baseline prompt with no other change.

D. Statistical Tests

This section consolidates the inferential procedures referenced in Section 4.

Table 3. Priming framings for causal intervention. Each question receives the framing corresponding to its own worldview. The framing sentence is prepended to the forecasting prompt; all other instructions are identical to the baseline condition.

Worldview	Pro-Prime	Anti-Prime
Econ. Resilience	The global economy has proven remarkably resilient to recent shocks, with strong labor markets, robust consumer spending, and financial systems that have absorbed disruptions without systemic failure.	The global economy is increasingly fragile, with mounting debt levels, overextended financial markets, and structural vulnerabilities that make a severe recession or financial crisis increasingly likely.
Environmentalism	Climate change is accelerating faster than models predicted, with record-breaking temperatures, rapid ice loss, and increasingly severe extreme weather events becoming the new normal.	Environmental risks are frequently overstated by media and advocacy groups. Natural systems have shown remarkable resilience, and technological adaptation has consistently outpaced worst-case climate projections.
Tech Acceleration	AI and technology development is accelerating at an unprecedented pace, with breakthroughs in reasoning, robotics, and scientific discovery arriving faster than experts anticipated.	Recent AI progress has been largely incremental and overhyped. Fundamental technical barriers remain unsolved, and the pace of meaningful capability gains is slowing as low-hanging fruit is exhausted.
Int'l Escalation	International tensions are escalating across multiple fronts, with great-power competition intensifying, alliance structures fragmenting, and the risk of direct military confrontation rising significantly.	Despite media focus on conflicts, international institutions and diplomatic channels continue to function effectively. Historical patterns show that escalation risks are typically contained by mutual deterrence and economic interdependence.
Pol. Polarization	Political polarization in the United States has reached historically extreme levels, with institutional norms eroding, partisan action becoming more aggressive, and the likelihood of unprecedented political disruptions increasing.	Despite heated rhetoric, American political institutions have proven remarkably durable. Polarization in elite discourse overstates actual public divisions, and institutional guardrails continue to constrain extreme political actions.

D.1. Shared Axis-Level Lean (Bootstrap over Questions)

For each axis a and each model m we compute the signed error $E_{m,a} = \mathrm{mean}_{q \in a}(Z_{m,q} - y_q^Z)$. We resample questions within a with replacement $n_b = 10{,}000$ times and recompute $E_{m,a}^{(b)}$ on each resample. The 95% bootstrap CI is the central interval over $\{E_{m,a}^{(b)}\}$. The cross-model mean $\bar{E}_a = \mathrm{mean}_m E_{m,a}$ uses the same resampled question set so that the eight per-model intervals share their sampling variance. A shared lean is declared on axis a when all eight per-model CIs lie on the same side of zero and the $\bar{E}_a$ CI excludes zero.

D.2. Between-Model Differentiation (Label Permutation)

For each axis a we permute the eight model labels independently within each question. This preserves each question's empirical distribution of polarity-adjusted scores and breaks any stable model identity. For each of $n = 10{,}000$ permutations we recompute the axis-level mean per model and the between-model variance of those means. The observed variance is compared to this null. p-values across the five axes are Holm-adjusted; only adjusted $p < 0.05$ rejections are reported.

D.3. Priming Spread (Paired Bootstrap)

For each cell (m, a) we resample questions with replacement within a and recompute $\Delta_{m,a}^{(b)} = \mathrm{mean}_{q \in a}(Z_{m,q}^{\text{pro,(b)}} - Z_{m,q}^{\text{anti,(b)}})$ on $n_b = 10{,}000$ paired bootstrap samples. Per-axis pooled CIs are computed analogously on $\Delta_a^{(b)} = \mathrm{mean}_{m,q \in a}(Z_{m,q}^{\text{pro,(b)}} - Z_{m,q}^{\text{anti,(b)}})$. The sign test on $\{\Delta_{m,a}\}_{m,a}$ across the 40 cells is reported as a coarse directional check, not as the inferential test.

D.4. Coherence Error

For each pair (q, q^{-1}) and each model m we average $P_m(\text{YES} \mid q)$ and $P_m(\text{YES} \mid q^{-1})$ over the ten samples per cell, then compute $e_{m,q} = |\bar{P}_m(q) + \bar{P}_m(q^{-1}) - 1|$. This is $|\mathbb{E}[\cdot]|$, not $\mathbb{E}[|\cdot|]$; per-cell values are a downward-biased estimate of per-sample coherence error and should be read with that in mind. We report pooled medians and the 25th/75th percentiles,

Table 4. Brier scores by axis for the eight retained models (75-question dataset), sorted by overall score (lower is better). **Bold**: best model per axis. *Climatology (in-sample)* forecasts the axis mean outcome on every question. *Climatology (LOO)* forecasts each question with the axis mean computed on the other 14 questions; LOO is the fairer baseline.

Model	Econ.	Env.	Tech	Int'l	Pol.	Avg.
Maverick	0.181	**0.249**	0.269	**0.223**	0.281	0.241
Gemma 4B	0.211	0.319	**0.238**	0.267	**0.248**	0.257
Scout	**0.178**	0.272	0.295	0.265	0.297	0.262
DeepSeek	0.186	0.314	0.328	0.253	0.300	0.276
Phi-4	0.187	0.364	0.309	0.232	0.322	0.283
Gemma 12B	0.206	0.350	0.325	0.265	0.288	0.287
Qwen 72B	0.185	0.362	0.331	0.231	0.340	0.290
Qwen 7B	0.188	0.378	0.390	0.289	0.340	0.317
Climatology (in-sample)	0.240	0.222	0.240	0.222	0.222	0.229
Climatology (LOO)	0.276	0.255	0.276	0.255	0.255	0.263

per-model medians, and per-axis medians; the per-axis numbers are reported once in Section 5.5 and not duplicated here.

E. Supplementary Figures

E.1. Coherence Error Distribution

E.2. Permutation Null

E.3. Within-Question Variance

The pattern of tightening is axis-specific (Wilcoxon, baseline vs. primed). On environmentalism the pro-prime tightens within-question SD ($p < .001$); the anti-prime does not ($p = 0.08$). On political polarization the anti-prime tightens strongly ($p < .001$); the pro-prime does not ($p = 0.85$). On international escalation only the anti-prime tightens ($p = 0.008$). On technological acceleration both primes tighten ($p = 0.01$ pro, $p < .001$ anti). On economic resilience neither prime tightens significantly ($p = 0.26$ pro, $p = 0.41$ anti).

A uniform anchoring account predicts symmetric variance reduction under both primes, which is not what we see. The asymmetry is consistent with three accounts: framings resolving competition among latent priors, models complying with opinionated context, and anchoring to whichever prime aligns with the model's existing lean. The data here cannot distinguish among them.

The Phi-4 baseline SD is computed on a slightly filtered subset relative to its primed SD (Appendix B); the qualitative pattern survives if Phi-4 is dropped.

E.4. Priming Spread Heatmap

All eight models put technological acceleration first in axis-level priming spread. The last-place axis varies: political polarization is last for 4 of 8 models, environmentalism for 3, economic resilience for 1. Pooled axis ranking is in Section 5.4.

F. Brier Scores by Axis

G. Corrective Priming

We define each (model, axis) cell's *baseline lean* as the sign of $\bar{Z}^{\text{baseline}}_{m,a} - 0.5$, i.e. whether the model leans toward or away from the axis on average. The corrective condition applies the prime opposing that lean: if the model leans toward the axis ($\bar{Z} > 0.5$), the anti-prime is used; if it leans away ($\bar{Z} < 0.5$), the pro-prime. The lean (counts of pro vs anti among the eight models) is reported in the *Lean* column of Table 5.

The corrective prime improves Brier on only one of five axes: environmentalism ($0.326 \rightarrow 0.292$). On the other four it worsens Brier: economic resilience ($0.190 \rightarrow 0.231$), technological acceleration ($0.311 \rightarrow 0.333$), international escalation ($0.253 \rightarrow 0.262$), and political polarization ($0.302 \rightarrow 0.305$).

Table 5. Corrective priming Brier scores. *Lean*: models leaning pro/anti. **Bold**: best Brier.

Worldview	Lean	Base	Baseline	Corr.	Same
Econ. Resil.	8/0	0.240	**0.190**	0.231	0.202
Environ.	8/0	0.222	0.326	**0.292**	0.375
Tech Accel.	4/4	0.240	**0.311**	0.333	0.338
Int'l Escal.	0/8	0.222	**0.253**	0.262	0.282
Pol. Polar.	0/8	0.222	**0.302**	0.305	0.358

Reading the cases against $\bar{Z}_{\text{baseline}}$ and $\bar{y}^Z$:

- *env*: corrective pushes $\bar{Z}$ from 0.58 toward outcome 0.27, Brier improves.
- *econ res*: corrective pushes $\bar{Z}$ from 0.58 down, but outcome is at 0.67; the prime moves away from truth.
- *tech*: baseline $\bar{Z} = 0.50$ already straddles outcome 0.53, so the corrective overshoots.
- *int'l escal*: corrective points toward truth ($\bar{Z} = 0.34$ toward 0.47) but the small gain is offset by per-question variance.
- *pol*: corrective points toward truth ($\bar{Z} = 0.31$ toward 0.47) but the addition of one $s = -1$ outcome-NO question in this axis flattens the per-cell improvement, and Brier nudges up by 0.003.

Same-direction priming (applying the prime aligned with the baseline lean) worsens Brier in every cell. Single-sentence priming is therefore a brittle debiasing tool: debiasing requires knowing whether the baseline lean is wrong, not merely its direction. The selection-on-DV caveat that bounds this analysis is in Section 6.

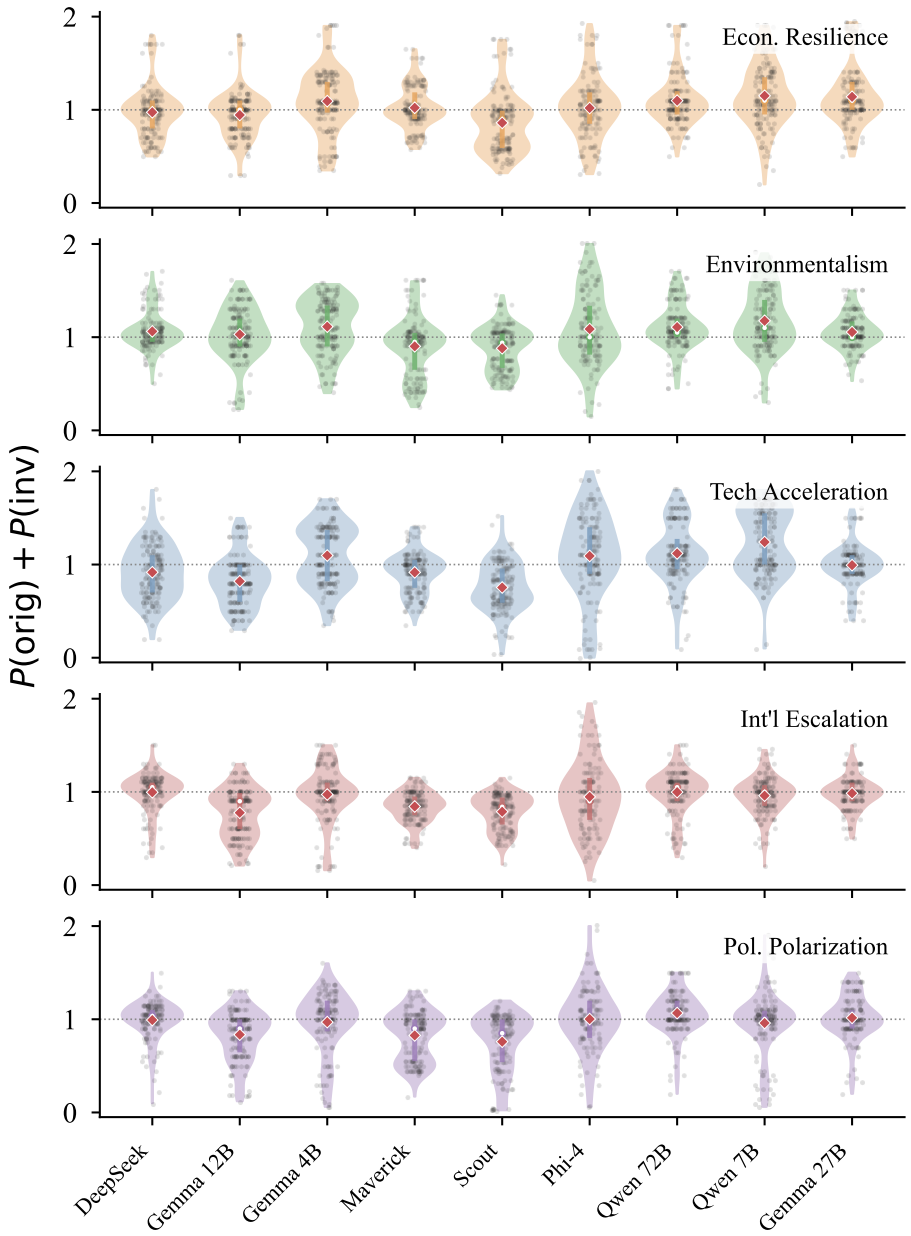

Figure 3. Coherence error $e_{m,q} = |\bar{P}_m(q) + \bar{P}_m(q^{-1}) - 1|$ per (model, question) cell. Sample means are taken first, then the absolute deviation from one. Pooled median 0.13 across 600 cells; pooled mean 0.18. Per-model medians range from 0.07 (DeepSeek-V3) to 0.19 (LLaMA-4 Scout).

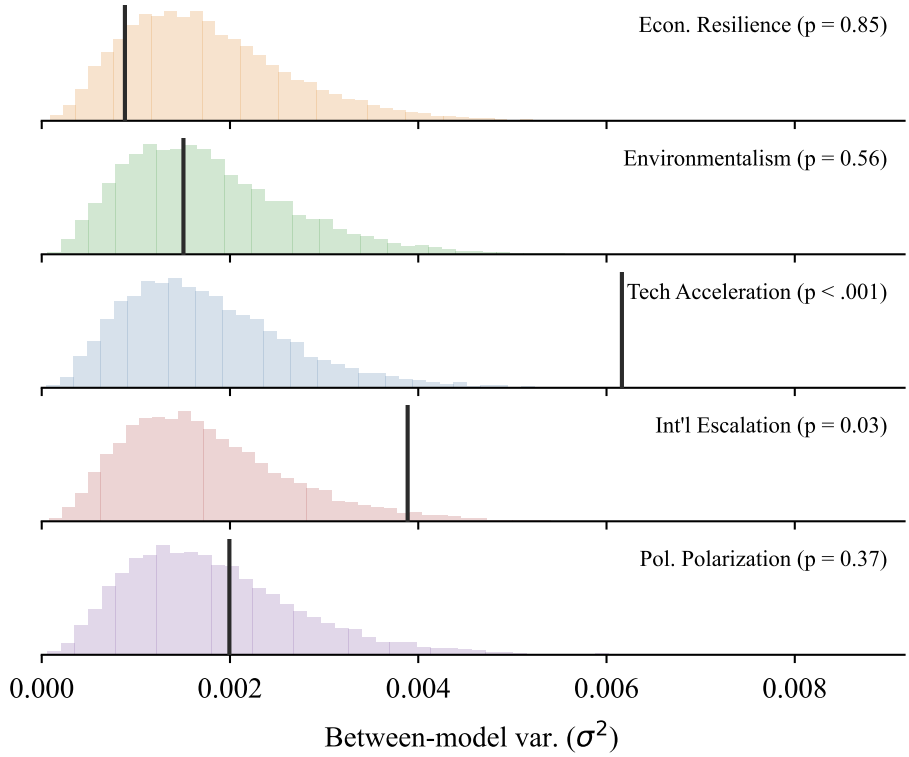

Figure 4. Per-axis permutation null for between-model variance. Histograms are the null distribution of axis-level between-model variance under $n = 10{,}000$ within-question reassignments of forecasts to models; vertical lines mark the observed variance. International escalation and technological acceleration sit in the far right tails ($p_{\text{adj}} < 0.001$ and $p_{\text{adj}} < 0.005$ after Holm correction); the other three observed variances fall inside their nulls.

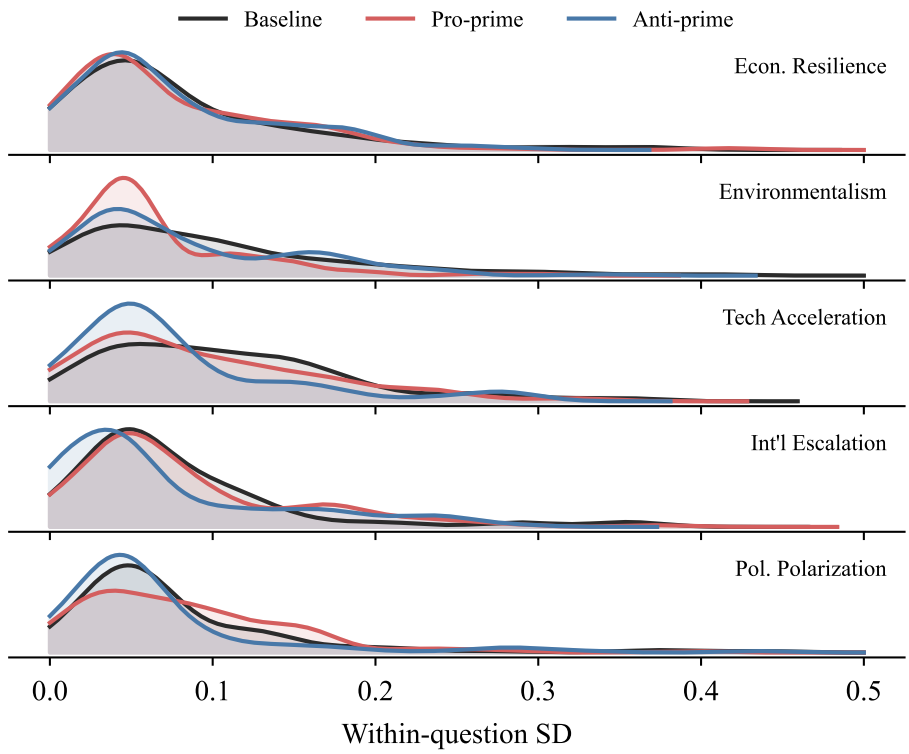

Figure 5. Within-question standard deviation across the ten samples per (model, question, condition), pooled across eight models. Baseline (black), pro-prime (red), anti-prime (blue). Both primes tighten the distribution relative to baseline on most axes.

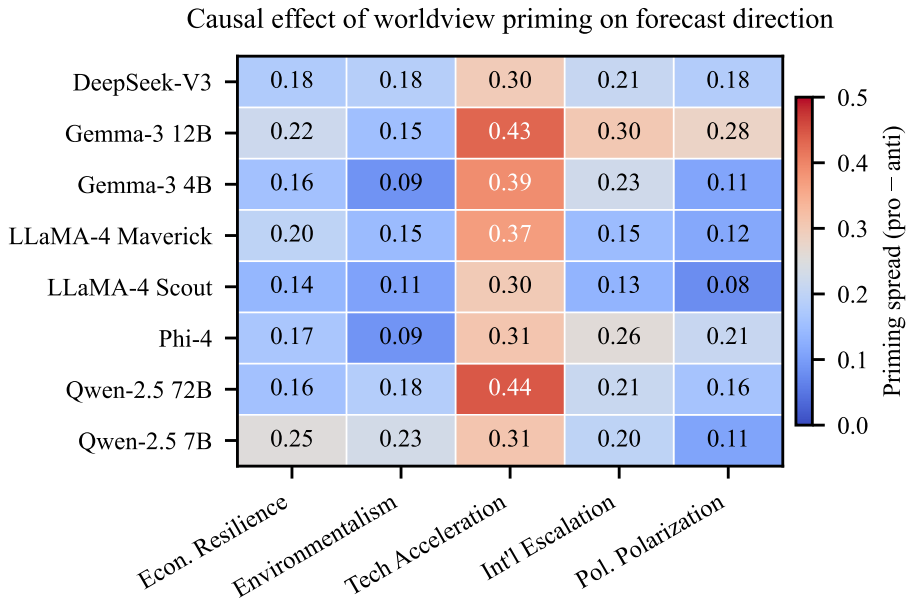

Figure 6. Per-cell priming spread $\Delta_{m,a}$ with rows indexing models and columns indexing axes. All 40 cells are positive (sign test, $p < .001$). Largest spreads on technological acceleration; smallest on political polarization.

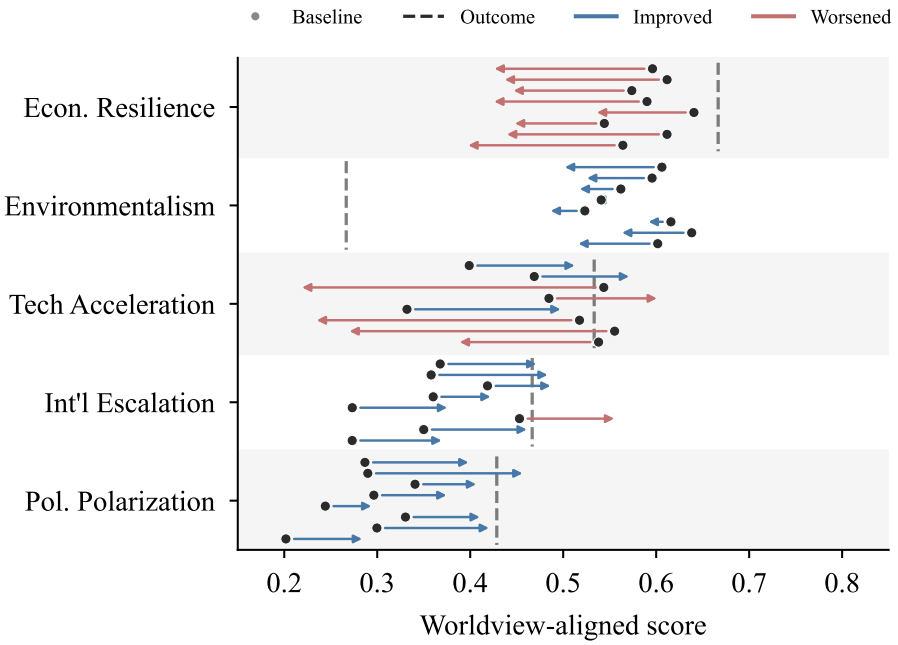

Figure 7. Corrective priming landscape. Each point is a (model, axis) cell. Black dots mark baseline polarity-adjusted means; arrows show the shift under the prime opposing the model's baseline lean; dashed line marks the axis outcome rate. Green arrows end closer to the outcome rate than they start; red arrows end further.