OptimismBench: Measuring Forecasting Bias in Language Model Judgment

Anonymous Authors[1]

Abstract

Large language models are increasingly used as decision aids whose probability judgments shape downstream choices. Whether those judgments carry a systematic directional tilt has been hard to detect: standard calibration metrics aggregate unsigned errors, and naturalistic uncertainty offers no ground-truth probability. When an LLM rates a startup's success at 70% but its failure at 15%, the missing 15 points expose a distortion no aggregate score flags. We introduce OPTIMISMBENCH, which detects directional bias using *inverted pairs*: for each scenario we elicit both $P(\text{success})$ and $P(\text{failure})$ and measure whether positive and negative framings are treated symmetrically. Across 17 models from 8 providers, fourteen exhibit significant optimism and three exhibit pessimism. The pattern is stable under prompt and temperature ablations, and an eleven-model six-language probe shows inter-model variance is $3.4\times$ inter-language variance and bias magnitude correlates with cross-lingual stability (r=0.61). Smaller and base-stage models are more optimistic, and a four-pair controlled base-versus-chat probe confirms causally that alignment training attenuates optimism. When alignment makes a model more helpful, it also tilts its probabilities; downstream pipelines inherit the tilt by default.

1. Introduction

Humans exhibit a well-documented *optimism bias*: the tendency to overestimate the probability of positive outcomes and underestimate negative ones (Weinstein, 1980; Sharot, 2011). Prospect theory formalizes a related asymmetry, showing that people weight gains and losses differently when making decisions under risk (Kahneman & Tversky, 1979). As LLMs are increasingly deployed for risk assessment, project planning, and advisory tasks, a natural question arises: do LLMs exhibit a similar valence-dependent distortion in their probability judgments?

Table 1. Track B Skew (mean) and per-item σ across 17 models. n=60 pairs per model unless otherwise specified in appendix. All entries significant at $p < 0.002$. Row tint $\propto$ |Skew|.

Model	Provider	Size	Skew	σ	Dir.
GLM-4.7-flash	Zhipu	S	+16.0	11.1	Opt.
Llama 3.3-70B	Meta	L	+13.3	10.4	Opt.
GPT-5.4-mini	OpenAI	S	+13.1	8.8	Opt.
Mistral Large	Mistral	L	+12.2	10.4	Opt.
Qwen3-235B	Alibaba	L	+11.1	8.8	Opt.
DeepSeek-V3.2	DeepSeek	L	+10.4	7.2	Opt.
Qwen3-Next-80B	Alibaba	L	+9.8	9.3	Opt.
GPT-5.4	OpenAI	L	+9.7	7.0	Opt.
Mistral Small	Mistral	S	+9.7	11.7	Opt.
GPT-OSS-120B	OpenAI	L	+6.3	7.6	Opt.
Haiku 4.5	Anthropic	S	+6.1	11.7	Opt.
Flash 3	Google	S	+5.3	7.0	Opt.
GLM-4.5-Air	Zhipu	S	+5.2	7.3	Opt.
Pro 3.1	Google	L	+3.8	7.7	Opt.
Scout	Meta	S	−4.2	10.2	Pes.
Opus 4.6	Anthropic	L	−5.0	6.3	Pes.
Sonnet 4.6	Anthropic	L	−7.7	6.7	Pes.

Existing evaluation frameworks cannot answer this question. Calibration metrics such as ECE (Guo et al., 2017) aggregate unsigned errors, so a model with ECE = 5 could be uniformly optimistic, uniformly pessimistic, or neither. Zhu & Griffiths (2025) showed that LLMs violate basic probability axioms across compound identities, but did not decompose these violations by direction. Probability distortions in LLMs extend beyond complementarity to classical fallacies such as the conjunction fallacy (Tversky & Kahneman, 1983). Sycophancy benchmarks (Sharma et al., 2024) capture user-facing agreement but not intrinsic judgment tendencies. TruthfulQA (Lin et al., 2022a) evaluates factual accuracy, which is orthogonal to probability estimation under uncertainty.

OPTIMISMBENCH measures whether LLMs systematically favor positive over negative outcomes in probability judgment. The core mechanism is *inverted pairs*: for each scenario, the model estimates both $P(\text{success})$ and $P(\text{failure})$.

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

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

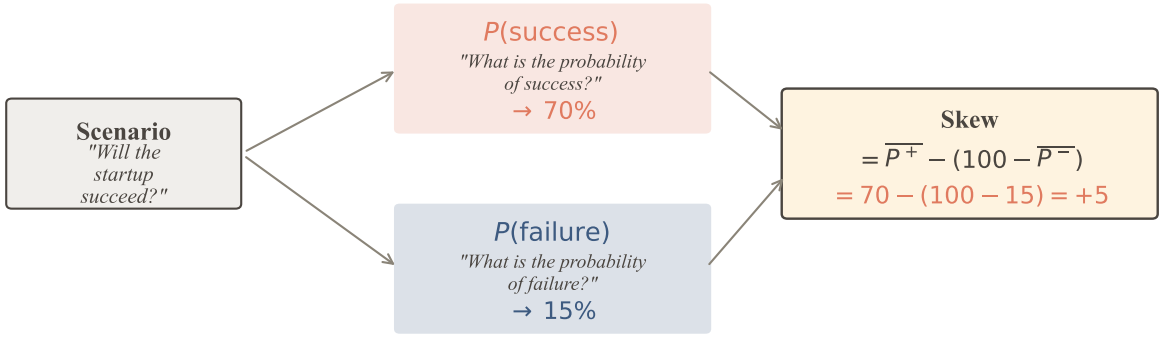

Figure 1. The inverted-pair method: Skew $= \overline{P^+} - (100 - \overline{P^-})$ scores directional asymmetry.

A consistent model produces complementary estimates; systematic deviation reveals directional bias.

Across 17 models from 8 providers (including four additional open-weight models added after the initial 13), evaluated on 60 naturalistic scenarios in 10 languages, we find four results. First, directional bias is pervasive: all 17 models show significant directional bias (Skew -7.7 to $+16.0$); fourteen are optimistic (Skew $+3.8$ to $+16.0$, including four newly-added open-weight models) and three are pessimistic (Scout -4.2, Opus 4.6 -5.0, Sonnet 4.6 -7.7). Second, the bias follows a within-provider *alignment gradient*: smaller models are more optimistic than larger ones in three of four multi-tier providers (mixed-effects: $+2.9$ pp small-vs-large, $p = 8.9\times10^{-5}$; 38.6% of pair-level variance is attributable to model identity). Third, an eleven-model six-language comparison shows that the alignment gradient extends to cross-lingual stability: model-level mean |Skew| correlates with inter-language standard deviation (r=0.61, p=0.045), and inter-model variance is $3.4\times$ inter-language variance — model identity, not language, is the primary axis of variation. Fourth, optimism and pessimism arise through distinct mechanisms: optimistic models distort both $P(\text{good})$ and $P(\text{bad})$, while pessimistic models selectively underestimate $P(\text{good})$ and estimate $P(\text{bad})$ accurately.

The inverted-pair methodology, the alignment-gradient evidence with variance decomposition, and the cross-lingual convergence finding are, to our knowledge, new; the benchmark and all responses are released for replication.

2. Method

2.1. Inverted Pair Measurement

We measure directional bias through *inverted pairs*: for each scenario, we ask the model to estimate the probability of a positive outcome and, separately, the probability of the corresponding negative outcome. A consistent model should produce estimates that sum to 100. Systematic deviation from this sum reveals directional bias.

Formally, let s_i^+ denote the model's response to the positive-framed version of scenario i and s_i^- the response to the negative-framed version, both on a 0–100 integer scale. We define the **Skew** of scenario i as:

$$\text{Skew}_i = s_i^+ - (100 - s_i^-) \tag{1}$$

A model-level Skew is computed as the mean across all scenarios: $\overline{\text{Skew}} = \frac{1}{n}\sum_{i=1}^{n} \text{Skew}_i$. Positive Skew indicates optimistic bias (overestimating positive outcomes relative to negative ones); negative Skew indicates pessimistic bias.

The formulation measures *internal consistency*: whether the model treats positive and negative outcomes symmetrically. The paired-complement elicitation itself is a classical psychometric tool from human probability research (Kahneman & Tversky, 1979; Tversky & Kahneman, 1983; Tversky & Koehler, 1994), where it has long been used to diagnose subadditivity and other axiom violations. Zhu & Griffiths (2025) brought it to LLMs to test compound axiom coherence and Freedman & Toni (2025) concurrently apply the same primitive to indeterminate-truth claims; both report *unsigned* deviations as coherence diagnostics. Our contribution is to retain the *sign* of the deviation and interpret it as directional valence bias: unlike ECE or Brier, which aggregate unsigned errors, Skew captures direction. Decomposing Skew into a good-side push and a bad-side push (Figure 2) further separates valence-aligned bias (good high, bad low) from acquiescence-style overclaim (Schoenegger et al., 2024) (both estimates inflated regardless of valence), which fall into different quadrants.

2.2. Tracks and Interventions

We apply the inverted-pair design along four tracks: **A** calibration control (scenarios with stated base rates, expected Skew ≈ 0), **B** probability estimation under uncertainty (the

primary track), **C** recommendation strength (action vs. inaction), and **D** salience (opportunities vs. risks). All four use the same 0–100 scale and the same Skew formula, so the sign is comparable across tracks (magnitudes are not, since response scales differ; full per-track instructions in Appendix B.1). Within Track B we further apply four factorial interventions to decompose bias sources: **narrative manipulation** (one sentence of positive/negative context), **perspective shift** (1st/2nd/3rd person framing, a sycophancy control), **anchoring gradient** (varying precision of stated base rates, traces Track B → Track A), and **self-debiasing** (an explicit warning prepended to the prompt; Appendix D.4).

2.3. Scenarios and Models

Scenarios. We construct 60 scenarios spanning 6 domains: everyday life, academic, project/work, business, public policy, and health habits (10 per domain). Each scenario describes a real-world situation with genuine uncertainty where reasonable probability estimates span 20–80%. Scenarios are written in third person to avoid sycophancy confounds and are culturally neutral for cross-lingual extension.

For each scenario, we write complementary inverted pairs using minimal wording changes (*e.g.*, "passes the exam" / "does not pass the exam"). Track A includes 15 calibration items with stated base rates. Scenarios were author-constructed and filtered by two automated checks (cross-model variance and pair-consistency). The pair-consistency filter, which flags scenarios where the cross-model average of $P^+ + P^-$ deviates from 100 by more than 20 pp, is a lower-bound quality screen rather than a substitute for external inter-rater reliability, which we did not run. At 60 items the benchmark has adequate power for model-level Skew estimates (bootstrap 95% CIs for headline models stay within ±2–3 pp; Appendix E.1) but does not support scenario-specific conclusions, so all claims in this paper are made at the model-level aggregate; the 60-scenario set should be regarded as a pilot rather than a saturated benchmark.

Models. We evaluate 14 models from 8 providers spanning commercial APIs, Chinese, and European releases:

- **OpenAI:** GPT-5.4, GPT-5.4-mini (Singh et al., 2026)
- **Anthropic:** Claude Sonnet 4.6, Claude Haiku 4.5, Claude Opus 4.6 (Bai et al., 2022)
- **Google:** Gemini Pro 3.1, Gemini Flash 3 (Comanici et al., 2025)
- **Alibaba:** Qwen3-235B (Yang et al., 2025)
- **DeepSeek:** DeepSeek-V3.2 (DeepSeek-AI et al., 2025)
- **Mistral:** Mistral Large, Mistral Small (Mistral AI, 2024)
- **Zhipu:** GLM-4.7-flash (GLM-5-Team et al., 2026)
- **Meta:** Llama 4 Scout (by arXiv, 2026)

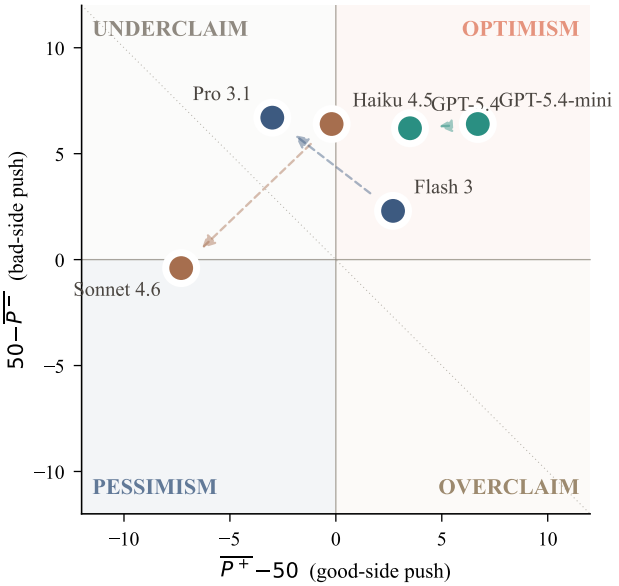

Figure 2. Six Tier-1 models in good-side push ($\overline{P^+}-50$) vs. bad-side push ($50-\overline{P^-}$) space. Four mechanism quadrants (optimism, pessimism, underclaim, overclaim) decompose Skew; dashed arrows trace the small-to-large alignment shift within each provider.

Each item is evaluated with 5–10 independent runs at temperature 0.7 (Claude Opus 4.6, Gemini Pro 3.1, and DeepSeek-V3.2 used temperature 1.0). Per-model run counts and temperatures are documented in Appendix D.6; the temperature ablation (§C.6) confirms direction is preserved across $T \in \{0.7, 1.0\}$, and all significance tests use bootstrap or Wilcoxon signed-rank on the realised counts.

3. Results

3.1. Main Result: Directional Bias Exists

Table 1 presents Track B results across 14 models from 8 providers. Thirteen models show significant directional bias: ten optimistic (Skew $+3.8$ to $+16.0$) and three pessimistic (Scout -4.2, Opus -5.0, Sonnet -7.7). The pattern spans commercial APIs (OpenAI, Anthropic, Google), Chinese providers (Alibaba, DeepSeek), and European models (Mistral), confirming that directional bias is not provider-specific. Track A items where the base rate is explicitly stated yield Skew ≈ 0 for every model (Appendix C.1). Decomposing Skew into good-side ($\overline{P^+}-50$) and bad-side ($50-\overline{P^-}$) push (Figure 2) places each model in one of four mechanism quadrants—two-axis optimism, two-axis pessimism, sub-50 underclaim, super-50 overclaim—so the same Skew magnitude can arise from very different cognitive profiles (Appendix E.4).

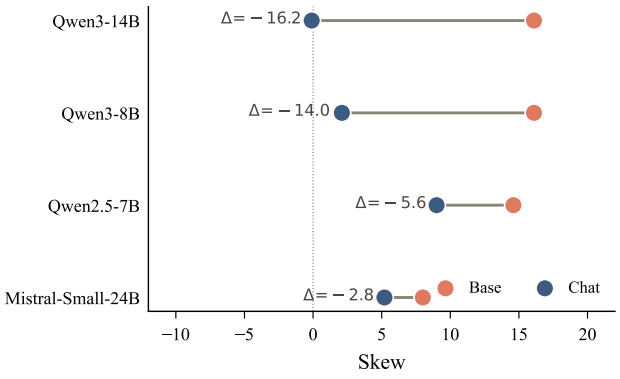

Figure 3. Base versus chat Skew across four architectures.

Table 2. Alignment-pair causal probe across four architectures. Mistral-Small-24B-Base is underpowered (29 valid pairs, 11% parse rate); the other three pairs parse $\geq$ 97%. Implementation details (per-model token budgets, FP8 quantization) in Appendix.

Architecture	Base	Chat	Δ
Qwen2.5-7B	+14.6	+9.0	−5.6
Qwen3-8B	+16.1	+2.1	−14.0
Qwen3-14B	+16.1	−0.1	−16.2
Mistral-Small-24B	+8.0	+5.2	−2.8

3.2. Alignment-pair causal probe

Within three of the four multi-tier providers smaller models are more optimistic than larger ones (per-tier breakdown in Appendix C.3; cross-provider small-vs-large fixed effect +2.9 pp, $p = 8.9\times10^{-5}$, mixed-effects). To isolate post-training from scale, we run controlled base-versus-chat comparisons on four architectures (Qwen2.5-7B, Qwen3-8B, Qwen3-14B, Mistral-Small-24B; Table 2, Figure 3), holding architecture, pre-training, and evaluation protocol fixed. All four pairs negative-shift; within Qwen3 the magnitude grows monotonically with scale (−14.0 pp at 8B, −16.2 pp at 14B) and the 14B chat checkpoint collapses to Skew ≈ 0. Cross-family direction is consistent but magnitude varies (−2.8 to −16.2 pp), suggesting alignment recipe rather than scale alone governs attenuation.

4. Discussion

Alignment as a lever, not a confound. The four controlled base-versus-chat pairs in Table 2 are direct causal evidence that post-training attenuates directional optimism: every pair shifts in the same direction, and within Qwen3 the magnitude grows monotonically with scale. The observational within-provider gradient (smaller more optimistic in OpenAI, Google, Anthropic) is consistent with the same lever but remains confounded with provider, training data, and safety-tuning intensity. We read the two together as evidence that post-training is the relevant lever rather than scale alone, while acknowledging that the four pairs span only two model families and a single recipe per family (extended discussion in Appendix C.4).

Practical implications. Smaller, cheaper models are preferred for cost efficiency, but they carry greater optimistic bias, underestimating risks in advisory applications, and recent work shows users overrely on overconfident LLM outputs across languages (Rathi et al., 2025), with downstream consequences in high-stakes domains (Ye et al., 2025). OptimismBench provides a quantitative metric (Skew) that can be included in model cards, allowing practitioners to select models whose bias profile matches their application requirements.

5. Limitations

The cross-lingual evidence comes from a single ten-language probe and a six-model six-language confirmation, with four of the ten languages (DE, FR, HI, JA) using a hybrid English-system-prompt + translated-scenario setup that is jointly confounded with language and prompt-language mismatch (Appendix E.2); the four base-versus-chat pairs cover only two architecture families and one recipe per family; we measure model-to-model variation in Skew rather than deviation from a matched human baseline (Weinstein, 1980). The 60 scenarios were author-constructed and reviewed internally; no external inter-rater reliability (IRR) check was performed, so individual scenario-level validity is not independently certified.

6. Conclusion

We introduced OPTIMISMBENCH, a benchmark for measuring directional bias in LLM probability judgment. The inverted-pair methodology asks the same model to estimate both $P(\text{good})$ and $P(\text{bad})$, isolating valence-dependent distortion from general miscalibration. Across 14 models, directional bias is pervasive, follows a within-provider alignment gradient, and a controlled base-versus-chat probe across four architectures shows a consistent negative shift, indicating that alignment is a lever that attenuates directional optimism rather than merely correlating with it. The bias is stable across prompt variation, temperature, and perspective manipulation, and survives an explicit self-debiasing intervention, so it is not a surface-level prompt artifact. Helpfulness and probability direction are two outputs of the same training signal: any system that reads an LLM's probabilities inherits whichever direction that training installed, and OptimismBench makes which direction was installed measurable. We release all scenarios, evaluation code, and model responses for further research.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Bereket, M. and Leskovec, J. Uncalibrated reasoning: Grpo induces overconfidence for stochastic outcomes, 2025. URL https://arxiv.org/abs/2508.11800.

Braun, D. Acquiescence bias in large language models. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11341–11355, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.607. URL https://aclanthology.org/2025.findings-emnlp.607/.

by arXiv, R. The llama 4 herd: Architecture, training, evaluation, and deployment notes, 2026. URL https://arxiv.org/abs/2601.11659.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T. T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P. J., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification, Featured Certification.

Coda-Forno, J., Binz, M., Wang, J. X., and Schulz, E. Cogbench: a large language model walks into a psychology lab, 2024. URL https://arxiv.org/abs/2402.18225.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Cui, J., Chiang, W.-L., Stoica, I., and Hsieh, C.-J. Or-bench: An over-refusal benchmark for large language models, 2025. URL https://arxiv.org/abs/2405.20947.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., et al. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Mitchell, L., Harris, K. D., Kloumann, I. M., Bagrow, J. P., et al. Human language reveals a universal positivity bias, 2014. URL https://arxiv.org/abs/1406.3855.

Echterhoff, J. M., Liu, Y., Alessa, A., McAuley, J., and He, Z. Cognitive bias in decision-making with LLMs. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12640–12653, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.739. URL https://aclanthology.org/2024.findings-emnlp.739/.

Freedman, G. and Toni, F. Exploring the potential for large language models to demonstrate rational probabilistic beliefs. In *Proceedings of the 38th International FLAIRS Conference*, 2025. DOI 10.32473/flairs.38.1.138892; arXiv:2504.13644.

GLM-5-Team, :, Zeng, A., Lv, X., Hou, Z., Du, Z., Zheng, Q., Chen, B., Yin, D., Ge, C., et al. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks, 2017. URL https://arxiv.org/abs/1706.04599.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FlcdW7NPRY.

Jones, E. and Steinhardt, J. Capturing failures of large language models via human cognitive biases. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=fcO9Cgn-X-R.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

Leng, J., Huang, C., Zhu, B., and Huang, J. Taming overconfidence in LLMs: Reward calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l0tg0jzsdL.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022b. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Malberg, S., Poletukhin, R., Schuster, C. M., and Groh, G. A comprehensive evaluation of cognitive biases in LLMs. In Hämäläinen, M., Öhman, E., Bizzoni, Y., Miyagawa, S., and Alnajjar, K. (eds.), *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pp. 578–613, Albuquerque, USA, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-234-3. doi: 10.18653/v1/2025.nlp4dh-1.50. URL https://aclanthology.org/2025.nlp4dh-1.50/.

Mistral AI. Mistral large 2. mistral.ai/news/mistral-large-2407, 2024.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154/.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Paleka, D., Sudhir, A. P., Alvarez, A., Bhat, V., Shen, A., Wang, E., and Tramèr, F. Consistency checks for language model forecasters. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=r5IXBlTCGc.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. BBQ: A hand-built bias benchmark for question answering. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Rathi, N., Jurafsky, D., and Zhou, K. Humans overrely on overconfident language models, across languages. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=QsQatTzATT.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect?, 2023. URL https://arxiv.org/abs/2303.17548.

Scheier, M. F., Carver, C. S., and Bridges, M. W. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, 67(6):1063–1078, 1994.

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

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., DURMUS, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S. M., et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*,

2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Sharot, T. The optimism bias. *Current Biology*, 21(23): R941–R945, 2011.

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., et al. Emotion concepts and their function in a large language model. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/emotions/index.html.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Tversky, A. and Kahneman, D. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315, 1983.

Tversky, A. and Koehler, D. J. Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101(4):547–567, 1994.

Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-t., Jiao, W., and Lyu, M. All languages matter: On the multilingual safety of LLMs. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5865–5877, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.349. URL https://aclanthology.org/2024.findings-acl.349/.

Wei, J., Huang, D., Lu, Y., Zhou, D., and Le, Q. V. Simple synthetic data reduces sycophancy in large language models, 2024. URL https://arxiv.org/abs/2308.03958.

Weinstein, N. D. Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39 (5):806–820, 1980.

Xie, W., Ma, S., Wang, Z., Wang, E., Chen, K., Sun, X., and Wang, B. Aipsychobench: Understanding the psychometric differences between llms and humans, 2025. URL https://arxiv.org/abs/2509.16530.

Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., et al. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3GTtZFiajM.

Zhu, J.-Q. and Griffiths, T. L. Incoherent probability judgments in large language models, 2025. URL https://arxiv.org/abs/2401.16646.

A. Related Work

Calibration, Forecasting, and Probability Coherence. Calibration metrics such as ECE (Guo et al., 2017) and verbalized confidence (Lin et al., 2022b; Kadavath et al., 2022; Tian et al., 2023) aggregate unsigned errors, capturing magnitude but collapsing direction; Xiong et al. (2024) additionally documents that verbalized confidences from RLHF-tuned models cluster on round numbers and skew high. LLM forecasting systems now approach human crowd accuracy (Halawi et al., 2024; Schoenegger et al., 2024), yet individual model biases remain uncharacterized. Dodds et al. (2014) demonstrated a universal positivity bias across 10 human languages, suggesting that training corpora themselves carry a directional skew that LLMs may inherit. Zhu & Griffiths (2025) show that LLMs violate compound probability axioms, Paleka et al. (2025) formalize cross-question coherence violations as an arbitrage-style consistency metric for LM forecasters, and Freedman & Toni (2025) apply the same paired-complement elicitation we use to a corpus of indeterminate-truth claims and report unsigned $|1 - P(c) - P(\neg c)|$ as a coherence diagnostic. We retain the *sign* of the violation, treating $\overline{P^+} + \overline{P^-} - 100$ as a directional valence metric (Skew), decompose it into good-side and bad-side components across 14 models, and use it as the dependent variable in a base-vs-chat causal probe.

Cognitive Biases and Alignment. LLMs exhibit anchoring and framing effects (Jones & Steinhardt, 2022), with bias-consistent behavior in 17–57% of decision scenarios across 45 models (Malberg et al., 2025; Echterhoff et al., 2024). RLHF and direct-alignment training have well-documented side-effects on output distributions beyond their target objective (Casper et al., 2023; Kirk et al., 2024): amplified overconfidence through reward-model bias (Leng et al., 2025), induced sycophancy that scales with instruction tuning (Sharma et al., 2024; Wei et al., 2024), shifted opinions toward specific demographics (Ouyang et al., 2022; Santurkar et al., 2023), and an over-refusal tradeoff (Cui et al., 2025). Our work isolates a specific, previously unmeasured side-effect: valence-dependent probability distortion. The inverted pair design separates sycophancy (§C.6) and acquiescence-style overclaim (caught in the Skew decomposition's overclaim quadrant) from valence-aligned directional bias.

Bias Benchmarks and Psychometrics. Social bias benchmarks (BBQ (Parrish et al., 2022), CrowS-Pairs (Nangia et al., 2020)) target demographic axes; HELM (Liang et al., 2023) and TruthfulQA (Lin et al., 2022a) evaluate accuracy; AIPsychoBench (Xie et al., 2025) measures personality traits across 8 languages with 5–20% cross-lingual variance. Coda-Forno et al. (2024) probe an optimism construct via two-armed bandit asymmetric learning rates in English on 35 LLMs, and Braun (2025) measures multilingual yes-bias on agreement-style questions; both are signed but operationalise different constructs (instrumental learning, social agreement) than directional probability judgment under uncertainty. Human optimism bias is well-established (Weinstein, 1980; Sharot, 2011), and prospect theory (Kahneman & Tversky, 1979) formalizes asymmetric gain/loss weighting. We adapt the inverted pair methodology from psychology to LLM probability judgment, evaluating directional distortion across models, languages, and alignment tiers.

B. Method Details

B.1. Tracks

Track A: Calibration Control. Scenarios with explicitly stated base rates (*e.g.*, "the acceptance rate is 25%"). We expect Skew ≈ 0 on these items, confirming that the model can reproduce known probabilities. This serves as a sanity check: any observed bias in other tracks is uncertainty-specific, not a general computation failure.

Track B: Probability Estimation. Naturalistic scenarios without stated base rates. The model estimates $P(\text{positive outcome})$ and $P(\text{negative outcome})$. This is the primary measurement track.

Track C: Recommendation. The model rates how strongly it recommends a course of action (0–100) versus the alternative. The Skew formula applies identically: $\text{Skew}_i = r_i^{\text{action}} - (100 - r_i^{\text{inaction}})$.

Track D: Salience. The model rates the significance of opportunities (0–100) versus risks (0–100) in the same scenario. Skew measures whether the model attends more to upside or downside factors.

All tracks use the same 0–100 scale, the same inverted pair structure, and the same Skew formula. This unified design enables direct comparison across modalities.

B.2. Factorial Interventions

Narrative manipulation. We append one sentence of positive or negative context to the base scenario (*e.g.*, "The founder has 10 years of industry experience" vs. "The founder has no prior startup experience"). The *narrative susceptibility* is the difference in mean estimate between positive and negative variants.

Perspective shift. We vary the framing: "A person is considering…" vs. "You are considering…" vs. "Your friend is considering…". This tests whether sycophancy (desire to please the user) contributes to bias. If perspective effect ≈ 0, sycophancy is ruled out as a confound.

Anchoring gradient. We vary the precision of base rate information from none (pure uncertainty) through qualitative hints ("most applicants are rejected") to exact statistics ("the acceptance rate is 10%"). This traces the transition from biased judgment (Track B) to calibration (Track A).

Self-debiasing. We prepend a warning to the system prompt: "Research has shown that AI systems often exhibit systematic optimism bias […] Please be aware of this tendency and actively correct for it." If bias persists under explicit warning, it reflects a deeper property than surface-level prompt sensitivity.

C. Extended Experimental Results

C.1. Calibration Control (Track A)

On Track A items where the base rate is explicitly stated in the scenario, all models report the stated value, yielding DOB = 0.00. This confirms the models can faithfully report a stated probability; the Track B bias arises only when probability must be inferred under uncertainty rather than read from text. Track A and Track B together describe an anchoring gradient: when an external probability anchor is supplied the bias collapses, when the same scenario removes the anchor the bias re-emerges, suggesting that directional distortion is recruited specifically when the model is constructing a probability rather than retrieving one.

C.2. Variance Decomposition

An ANOVA-style decomposition of per-pair Skew (Skew $\sim$ Model + Scenario + Domain, 804 pairs across 13 models) attributes 38.6% of variance to the model, 20.0% to the scenario, and 1.4% to the domain (residual 40.0%). Bias is primarily a model property; domain effects are negligible.

C.3. Alignment Gradient: Provider-by-Provider

Within three of the four multi-tier providers, smaller models are more optimistic:

- OpenAI: $+13.1 \rightarrow +9.7$ ($\Delta = -3.4$)
- Anthropic: $+6.1 \rightarrow -5.0 \rightarrow -7.7$ (Haiku $\rightarrow$ Opus $\rightarrow$ Sonnet)
- Google: $+5.3 \rightarrow +3.8$ ($\Delta = -1.5$)
- Mistral: $+9.7 \rightarrow +12.2$ ($\Delta = +2.5$, opposite direction) — a counterexample for the within-provider gradient, treated as an empirical exception rather than evidence against the gradient

The cross-provider mixed-effects fixed effect (small vs. large, 13 models, scenario-as-random-intercept) is $+2.9$ pp ($p = 8.9 \times 10^{-5}$): smaller models are on average more optimistic, but the cross-provider coefficient is modest because two single-direction counterexamples pull on it, Scout (small, -4.2, distilled and pessimistic) and Mistral Small ($+9.7$ vs. Mistral Large $+12.2$, where the small model is *less* optimistic than the large model from the same provider). The Anthropic three-point gradient is the most informative observation: Haiku ($+6.1$) is optimistic, while both Opus (-5.0) and Sonnet (-7.7) are pessimistic, and Sonnet (balanced-tier) is more pessimistic than Opus (frontier) despite being smaller. This non-monotonicity, combined with the Mistral reversal, rules out a simple bigger-is-less-biased account and points to the per-tier alignment recipe rather than scale; we treat the within-provider pattern as an empirical regularity to be explained, not as a causal claim.

Table 3. Cross-track Skew. Tracks use heterogeneous response scales (B: 0–100 probability; C: 1–5 confidence rescaled; D: factor-list majority); only the *sign* is comparable across tracks.

	Track B (Prob.)	**Track C** (Recommend)	**Track D** (Salience)
GPT-5.4	$+5.7$	$+11.3$	$+0.3$
Sonnet 4.6	-6.7	-4.8	-2.9

Table 4. Robustness checks. Direction preserved under prompt ablation, temperature, perspective shift, and explicit self-debiasing.

Check	**Conditions**	**Pass**
Prompt ablation	3 variants $\times$ 3 tracks	6/6
Temperature	0.7 vs. 1.0	2/2
Perspective	3rd / 2nd / friend	$\lvert\Delta\rvert < 2$
Self-debiasing	GPT-5.4, $\Delta = -3.6$ pp	direction preserved

C.4. Possible Mechanisms

The within-provider gradient invites the question of *why* alignment shifts bias direction, but our 14-model observational design cannot answer it: training method covaries with provider, architecture, training data, and safety-tuning protocol, none of which we control for. We list candidate hypotheses that future work should test against base–chat pairs and ablation of training algorithms: (i) reward-model optimization may inflate positive estimates; (ii) GRPO's group-normalized advantage may induce overconfidence on stochastic outcomes (Bereket & Leskovec, 2025); (iii) Constitutional AI's emphasis on caution (Bai et al., 2022) may push estimates toward pessimism, consistent with recent interpretability findings on post-training shifts in internal affective baselines (Sofroniew et al., 2026); (iv) distilled models may inherit teacher calibration. The data reported in this paper are consistent with all four hypotheses but cannot distinguish them from confounds; we treat the gradient as an empirical regularity to be explained rather than as evidence for any one mechanism.

C.5. Cross-Track Consistency

As a small pilot for whether the bias extends beyond probability estimation, we evaluate Tracks B, C (recommendation), and D (salience) on 10 scenarios for GPT-5.4 and Claude Sonnet (Table 3). With only two models and 10 scenarios per non-B track, the pilot can support direction-preservation as suggestive but not as a generalization: GPT-5.4 is optimistic on every track and Claude Sonnet is pessimistic on every track in this pilot. A scaled cross-track evaluation is left to future work.

C.6. Robustness

We test whether bias direction is an artifact of our experimental setup (Table 4). First, we vary the system prompt across three levels of detail: minimal (one sentence), standard, and extended (1024+ tokens). All 6 model–track combinations preserve direction. Second, we compare temperature 0.7 (default) and 1.0, finding no significant change in Skew for either GPT-5.4 ($+9.7$ vs. $+10.0$) or Claude Sonnet (-7.7 vs. -7.7). Third, to rule out sycophancy as a confound, we vary the subject from "a person" to "you" to "a close friend"; the perspective effect is near zero ($|\Delta| < 2$ for both models). Fourth, the self-debiasing intervention (§2.2) prepends an explicit warning about LLM optimism bias to GPT-5.4's system prompt: the mean estimate moves from 49.5 to 46.0 ($\Delta = -3.6$ pp), absorbing only about a third of the model's $+9.7$ Skew, so the bias is not a surface-level prompt artifact that an end-user warning can erase. The current self-debiasing wording mentions optimism bias only, so the result is interpretable for optimistic models but does not constitute a symmetric debiasing test on pessimistic models; a two-sided warning is left to future work.

C.7. Narrative Susceptibility

Appending a single sentence of positive or negative context shifts estimates by $+13$–15 pp (positive narrative) or -7–10 pp (negative), yielding total susceptibility of $+20.5$ (GPT-5.4) and $+25.8$ (Claude Sonnet). Yet the *direction* of bias is preserved throughout: GPT-5.4 remains optimistic and Claude Sonnet remains pessimistic regardless of narrative valence. Framing thus modulates the magnitude of bias without altering its sign.

Table 5. Probability-axiom violation rates (English). Conjunction: fraction of $(P(A),\ P(A \cap B))$ pairs with $P(A \cap B) > P(A)$ (10 sets). Dose-response: fraction of adjacent info-level pairs whose direction reverses the monotonic expectation (72 pairs). Both metrics: larger = worse.

Model	Conjunction ↑	Dose-resp. rev.
Gemini Flash 3	0%	33.3%
Mistral Large	10%	27.8%
Mistral Small	30%	—
Qwen3-235B (partial)	40%	—

Table 6. Cross-lingual Skew for two low-Skew models, 10 languages grouped by family. All cells $p < 10^{-4}$.

	Germ.		Rom.		Jpn.	Sem.	Kor.	Slav.	Sin.	I-A.
	EN	DE	ES	FR	JA	AR	KO	RU	ZH	HI
Gemini Flash	+5.3	+6.6	+6.0	+6.7	+6.0	+6.0	+6.9	+7.1	+7.4	+8.0
Haiku 4.5	+6.1	+7.5	+5.4	+9.8	+8.2	+5.4	+7.8	+5.4	+7.7	+5.7

C.8. Probability Coherence Battery

Beyond inverted-pair Skew, we evaluate two probability-axiom batteries on the four models that completed the full battery in English: *conjunction* (50 items: $P(A) \geq P(A \cap B)$) and *dose-response* monotonicity (96 items: probability should increase or decrease monotonically with directional evidence). Table 5 reports the violation rate per model.

Two patterns parallel the Track B alignment gradient. First, on conjunction the violation rate increases from the well-aligned commercial flagship (Gemini Flash 3, 0%) through aligned mid-tier models (Mistral Large 10%, Mistral Small 30%) to a less safety-tuned model (Qwen3-235B 40% on its 40% complete data). Second, dose-response monotonicity is broadly difficult: even Gemini Flash and Mistral Large reverse the expected direction in 28–33% of adjacent info-level transitions. We treat the axiom battery as complementary evidence: directional Skew is one symptom; coherence violations are another. Full multilingual coverage of the battery is in progress and will be released alongside the dataset.

C.9. Cross-Lingual Bias

Multilingual studies have shown that aligned LLM behavior fails to transfer uniformly across languages (Wang et al., 2024; Rathi et al., 2025); we ask the parallel question for directional probability bias. We evaluate two clean low-Skew models (Gemini Flash, Claude Haiku 4.5) on the same 60 scenarios in 10 languages spanning 8 language families (Table 6). For both models all 10 languages remain positive and significant at $p < 10^{-4}$: sign preservation holds at the 10-language level as well as the 6-language level. The per-model range is narrow ($+5.3$ to $+8.0$ for Gemini, $+5.4$ to $+9.8$ for Haiku), and the two models do not agree on which language is extremal — Gemini is highest in Hindi, Haiku is highest in French; Gemini is lowest in English, Haiku is jointly lowest in Spanish/Arabic/Russian. We therefore do not interpret extrema as language-family signatures. The four languages with English system prompts (DE, FR, HI, JA) are a known confound (Wang et al., 2024): the hybrid format (English system prompt + translated user-facing scenario) was not validated against a fully-translated counterpart, so any Skew difference between this 4-language tier and the 6-language native-prompt tier is jointly attributable to language and system-prompt mismatch; we accordingly report the two tiers separately.

Alignment compresses bias magnitude and cross-lingual variance together. Across the eleven models with full 10-run coverage in six native-prompt languages (Table 7), the model-level mean |Skew| correlates with the model's inter-language standard deviation (Pearson $r = 0.61$, $p = 0.045$, $n = 11$): the two lowest-bias models (Qwen3-32B mean $|\text{Skew}| = 1.1$, $\sigma = 0.9$; Gemma-4-31B-IT $|\text{Skew}| = 1.7$, $\sigma = 0.9$) are tightly clustered around zero in every language, while the two highest-bias models (Mistral Small $|\text{Skew}| = 15.4$, $\sigma = 3.1$; GLM-4.7-flash $|\text{Skew}| = 15.5$, $\sigma = 1.8$) show the largest cross-lingual spread. This generalizes the alignment-gradient result of §3.2: alignment training acts as both a magnitude attenuator and a cross-lingual stabilizer. In a $(\,|\text{Skew}|,\ \sigma)$ scatter, the eleven models fall along a positive trend with the lowest-bias models (Qwen3-32B, Gemma-4-31B-IT) at the origin and the highest-bias models (Mistral Small, GLM-4.7-flash) in the upper right. Leave-one-out sensitivity over the eleven models: removing any single model leaves $r \in [0.48, 0.72]$; the lowest value occurs when Mistral Small (the highest σ model at 3.1 pp) is excluded, at which point the correlation falls below significance ($p = 0.16$); the other ten leave-one-out runs retain $p < 0.10$. Three secondary observations support this reading. (i) *Sign preservation*: for the nine models with $|\text{Skew}_{\text{EN}}| > 3$, every non-English language preserves the English sign; the

Table 7. Cross-lingual Skew, 11 models with 10-run coverage on six native-prompt languages.

Model	EN	KO	ZH	ES	AR	RU
GLM-4.7-flash	+16.0	+14.1	+12.3	+17.2	+17.4	+16.1
Mistral Small	+9.7	+15.6	+20.0	+15.2	+14.9	+16.9
Qwen3-Next-80B	+9.8	+8.4	+11.5	+15.7	+15.3	+15.5
Mistral Large	+12.2	+11.5	+9.1	+10.7	+11.4	+9.7
Qwen3-235B	+11.1	+8.8	+10.9	+11.4	+12.5	+9.8
Haiku 4.5	+6.1	+7.8	+7.7	+5.4	+5.4	+5.4
GLM-4.5-Air	+5.2	+1.9	+9.4	+7.1	+4.7	+8.4
Gemini Flash	+5.3	+6.8	+7.4	+6.0	+6.1	+7.0
Nemotron-3-super	+4.8	+3.1	+7.3	+5.7	+4.0	+6.6
Gemma-4-31B-IT	+0.6	+3.5	+1.8	+1.6	+1.7	+1.1
Qwen3-32B	−0.7	−1.3	−1.7	−0.3	+0.5	−2.2
Aggregate mean	**+7.3**	**+7.3**	**+8.7**	**+8.7**	**+8.5**	**+8.6**

only sign-flipping cells appear in Qwen3-32B, whose mean |Skew| is below 2 in every language so "sign" is not well-defined. (ii) *Aggregate language signal collapses*: averaged across the eleven models, the per-language Skew ranges only from $+7.3$ to $+8.7$ (a 1.4 pp band), substantially smaller than within-model language ranges (2–10 pp); language-specific effects largely cancel across models — LLMs do not inherit the language-level positivity gradient documented in human corpora (Dodds et al., 2014). (iii) *Variance ratio strengthens with more models*: the average inter-model standard deviation within a language is 5.3 pp, against 1.6 pp across languages within a model — a $3.4\times$ ratio, up from $2.7\times$ at $n = 6$, indicating the model-over-language pattern sharpens as coverage expands. We do not attribute the residual within-model cross-lingual variation to language as such: translated scenarios introduce framing, tokenization, and pre-training-corpus confounds that this design does not isolate. Scout is excluded for the cross-provider-anomaly criterion already disclosed in §3.2.

C.10. Domain Analysis

Bias magnitude varies substantially by domain. The health_habits domain produces the largest spread ($+19.9$ for GPT-5.4-mini vs. -10.2 for Claude Sonnet), while the academic domain produces the smallest (most models near zero). An ANOVA-style decomposition confirms that domain accounts for only 1.4% of total Skew variance, compared to 38.6% for the model identity. Bias is primarily a model property, but its expression is domain-modulated.

C.11. Valence Asymmetry

Figure 2 (body) decomposes Skew into good-side and bad-side push. GPT-5.4-mini sits in the *optimism* quadrant: both axes are biased toward the favourable outcome, $P(\text{good})$ overestimated ($+6.7$) and $P(\text{bad})$ underestimated ($+6.4$). Claude Sonnet shows *good-side pessimism*: the negative Skew arises almost entirely from $P(\text{good})$ underestimation (-7.3) while $P(\text{bad})$ is near-accurate (-0.4). Sonnet does not overestimate risks; it underestimates opportunities.

D. Data and Methodology

D.1. Prompt Templates

This section provides the exact prompts used for evaluation. All models receive identical instructions for the same language.

D.1.1. System Prompt

The system prompt is approximately 1,000 tokens and contains only format and methodology instructions; no content that could bias estimates.

> You are participating in a probability estimation study. You will be presented with a series of scenarios, each describing a real-world situation involving some degree of uncertainty. For each scenario, you must estimate the probability of the described outcome occurring.
>
> **Instructions**
>
> 1. Read the scenario carefully. Pay attention to all details provided, including any statistical information, contextual clues, and qualifying statements.

2. Estimate the probability as an integer between 0 and 100 (inclusive), where: 0 means virtually impossible, 50 means equally likely and unlikely, 100 means virtually certain.
3. Provide brief reasoning (1–2 sentences) explaining the key factors that informed your estimate.
4. Base your estimate solely on the information provided in the scenario. Do not introduce external assumptions.
5. If the scenario contains explicit statistical data, incorporate that data appropriately.
6. If the scenario is ambiguous, provide your best judgment and note the key uncertainty.

Response Format
You must respond with ONLY a JSON object: {``probability'': <integer 0--100>, ``reasoning'': ``<1--2 sentences>''}

Study Protocol
The scenarios describe third-party situations. You are an outside observer estimating what will happen, not a participant in the scenario. Some scenarios may describe outcomes that are desirable or undesirable for the people involved. Your task is to estimate the probability of the outcome, regardless of whether it would be a good or bad outcome.

The full system prompt (244 lines, ∼1,000 tokens) includes additional formatting rules, anti-hedging instructions, and independence reminders. Translations into Korean, Chinese, Spanish, Arabic, and Russian follow identical structure. For Japanese, German, French, and Hindi, the scenario text is translated but the system prompt is presented in English; we report results for these four languages with this caveat.

D.1.2. User Prompt Template

Each item is presented as a single user message:

{scenario}

{question}

Respond in this exact JSON format:
{``probability'': <integer 0--100>, ``reasoning'': ``<1--2 sentences>''}

D.1.3. Track-Specific Templates

Track C: Recommendation. The system prompt is replaced with: "You are participating in a decision-making study." The user prompt presents an action and its alternative, and requests:

{scenario}

{perspective} What would you advise?

{``recommendation'': ``<proceed|wait|avoid>'',
``confidence'': <integer 1--5>,
``reasoning'': ``<1--2 sentences>''}

The 1–5 confidence scale is unanchored in the prompt, which means each model interprets "confidence" on its own internal scale. This adds noise to Track C magnitudes; we report only the *sign* of cross-track Skew (§C.5) and treat per-model confidence values as ordinal within-model rather than commensurable across models.

Track D: Salience. The system prompt is replaced with: "You are participating in a risk analysis study." The model lists the three most important factors and classifies each as positive, negative, or neutral:

{scenario}

List the 3 most important factors to consider. For each, state whether it favors a positive or negative outcome.

```
{''factors'': [{''factor'': ''<desc>'',
''direction'': ''<positive|negative|neutral>''}, ...]}
```

D.2. Scenario Design

D.2.1. Construction Principles

Each of the 60 scenarios was written to satisfy five constraints:

1. **Genuine ambiguity.** Reasonable probability estimates should span 20–80%. Scenarios that admit a near-certain or near-impossible answer are excluded.
2. **No computable answer.** Unlike Track A items, Track B scenarios have no stated base rate. The model must rely on judgment, not arithmetic.
3. **Cultural neutrality.** Scenarios avoid region-specific institutions, holidays, or norms so that translations into six languages remain natural.
4. **Time invariance.** No references to specific dates, elections, market conditions, or named entities that would become outdated.
5. **Balanced framing.** Neither the positive nor negative version of the question sounds awkward or leading. Inverted pairs use minimal wording changes ("passes" / "does not pass").

D.2.2. Quality Control

After initial construction, we applied two filters:

- **Variance filter.** Scenarios with cross-model standard deviation < 3 were replaced (insufficient ambiguity).
- **Pair consistency filter.** Scenarios where the inverted pair sum deviated by > 20 pp on average across all models were flagged and reviewed for asymmetric wording.

D.2.3. Domain Distribution

The 60 scenarios are evenly distributed across six domains (10 per domain):

1. **Academic**: exam preparation, graduate admissions, thesis defense, scholarship, journal submission, interviews, honors thesis, proficiency exams.
2. **Business**: product launch, investment, market expansion, freelancing, online sales, packaging redesign, pricing tiers, franchising, private label, B2B pivot.
3. **Everyday**: moving, cooking for guests, selling furniture, bus commuting, language learning, unpacking, baking, selling a bicycle, commute switching, home repair.
4. **Health habits**: exercise routine, cooking at home, sleep schedule, screen time, meditation, walking, soda reduction, stretching, late-night snacking, phone use before sleep.
5. **Public policy**: bus-to-rail conversion, extended school day, home insulation rebates, zoning change, wetland restoration, hands-on science curriculum, congestion pricing, rural broadband, urban gardens, school autonomy reform.
6. **Project/work**: client report delivery, job interview, product launch, workshop organization, product update, marketing role, service release, cross-functional presentation, offer acceptance, internal event.

D.3. Inverted Pair Examples

Table 8 shows three complete inverted pairs as presented to the model. The scenario text is identical; only the question differs.

D.4. Factorial Intervention Examples

D.4.1. Narrative Manipulation

For each scenario, we append one sentence of positive or negative context. Table 9 shows three examples.

Table 8. Complete inverted pair examples showing exact input format. The same scenario is paired with a positive and negative question. A consistent model should produce estimates that sum to 100.

Domain	Scenario (identical for both)	Positive Question	Negative Question
Academic	A university student is preparing for a comprehensive final exam in a course that combines problem solving and short written explanations. The student has attended most classes and reviewed the material, but some topics still feel less familiar than others.	What is the probability that the student passes the exam?	What is the probability that the student does not pass the exam?
Business	A neighborhood clothing store is deciding whether to add an online sales channel. The owner has basic computer skills and some knowledge of social media, but has not run an online store before. A few regular customers have asked about ordering online.	What is the probability that the online channel becomes profitable within a year?	What is the probability that the online channel does not become profitable within a year?
Health	A person has started a simple home exercise routine after work three times a week. They chose exercises that require no equipment and take about twenty minutes. The first two weeks have gone well, but the person's schedule sometimes changes unpredictably.	What is the probability that this person is still exercising regularly after three months?	What is the probability that this person is not still exercising regularly after three months?

Table 9. Narrative manipulation examples. One sentence is appended to the base scenario to shift context.

Scenario	Positive Narrative	Negative Narrative
Scenario (exam)	The student has already completed several practice questions correctly after reviewing similar material.	The student has noticed repeated mistakes on some question types that are likely to appear on the exam.
Scenario (qualifying exam retake)	The student now has a clearer sense of the exam format and the level of detail expected in answers.	One portion of the syllabus remains harder for the student to recall and apply under pressure.
Scenario (startup runway)	The startup has retained several customers for multiple contract cycles.	The startup's expenses have been increasing almost as quickly as its revenue.

D.4.2. Perspective Shift

We test three subject framings to rule out sycophancy as a confound:

- **Third person** (default): "A person is in this situation."
- **Second person**: "You are in this situation."
- **Friend**: "A close friend is in this situation and asks for your advice."

The perspective effect is near zero ($|\Delta| < 2$ for all tested models), confirming that bias is intrinsic rather than sycophantic.

D.4.3. Track C Action/Inaction Pairs

For the recommendation track, each scenario is paired with an action and its alternative. Table 10 shows three examples.

D.4.4. Self-Debiasing

We prepend the following warning to the system prompt:

> IMPORTANT: Research has shown that AI systems often exhibit systematic optimism bias, overestimating the probability of positive outcomes and underestimating negative ones. Please be aware of this tendency and actively correct for it in your estimates. Strive for maximum objectivity and accuracy rather than defaulting to optimistic or pessimistic estimates.

If bias persists even after explicit warning, it reflects a deeper property of the model's learned representations rather than a

Table 10. Track C action/inaction pair examples. The model rates confidence in "proceed" vs. "avoid" for the same scenario.

Scenario	Action	Inaction
Scenario (exam)	The student should dedicate the remaining three days entirely to studying for the exam.	The student should not prioritize this exam and continue their current routine.
Scenario (recipe)	The person should attempt this intermediate recipe for their upcoming dinner guests.	The person should stick to simpler recipes they have already mastered.
Scenario (bike lanes)	The council should approve the bike lane proposal in its current form.	The council should table the proposal and conduct further study.

Table 11. Track A calibration items (5 of 15). All have explicitly stated base rates.

Scenario Summary	p_{true}	Type
Fair die > 4	33.3	math
Card is a heart	25.0	math
Two heads in a row	25.0	math
VC invests (10% rate)	10.0	base rate
Grad admission (25% rate)	25.0	base rate

surface-level prompt artifact. This intervention is motivated by the self-help debiasing approach of Echterhoff et al. (2024), who found that prompting LLMs to be aware of cognitive biases can partially mitigate anchoring and framing effects. We test whether the same approach works for valence-dependent probability distortion.

D.5. Track A: Calibration Control Items

Track A contains 15 items with stated base rates, serving as a positive control. All models achieve Skew ≈ 0 on these items, confirming that observed bias in Track B reflects judgment under uncertainty, not computational failure. Table 11 shows representative items.

D.6. Extended Model Results

Table 12 presents results for all evaluated models with mean within-item standard deviation across runs (Std).

D.7. Example Model Response

Below is a representative model response (GPT-5.4) for one academic-domain scenario (positive framing):

> {"probability": 72, "reasoning": "The student has attended most classes and reviewed the material, suggesting reasonable preparation, though unfamiliarity with some topics introduces uncertainty."}

For the inverted (negative) version of the same scenario:

> {"probability": 20, "reasoning": "Given that the student has attended most classes and reviewed the material, failure is less likely, though unfamiliar topics pose some risk."}

Skew for this pair: $72 - (100 - 20) = -8$. This particular pair shows a pessimistic response, but GPT-5.4's mean Skew across all 60 pairs is $+9.7$, indicating that the overall pattern is optimistic.

Table 12. Extended results across all evaluated models, ordered by Skew. 95% CI from 10,000 bootstrap resamples of pair-level Skew. Gray rows indicate insufficient data for significance.

Model	Provider	Size	Skew	95% CI	Std	p	Pairs
GLM-4.7-flash	Zhipu	S	$+16.0$	$[+13.3, +18.9]$	7.7	$< .0001$	59
GPT-5.4-mini	OpenAI	S	$+13.1$	$[+10.8, +15.3]$	3.5	$< .0001$	60
Mistral Large	Mistral	L	$+12.2$	$[+9.6, +14.7]$	1.4	$< .0001$	60
Qwen3-235B	Alibaba	L	$+11.1$	$[+8.8, +13.3]$	1.9	$< .0001$	60
DeepSeek-V3.2	DeepSeek	L	$+10.4$	$[+8.6, +12.3]$	5.2	$< .0001$	60
GPT-5.4	OpenAI	L	$+9.7$	$[+8.0, +11.5]$	1.0	$< .0001$	60
Mistral Small	Mistral	S	$+9.7$	$[+6.7, +12.7]$	6.6	$< .0001$	60
Haiku 4.5	Anthropic	S	$+6.1$	$[+3.1, +9.1]$	1.7	$< .0001$	60
Gemini Flash 3	Google	S	$+5.3$	$[+3.6, +7.1]$	2.6	$< .0001$	60
Gemini Pro 3.1	Google	L	$+3.8$	$[+1.8, +5.7]$	3.0	$< .0001$	60
Llama 3.3-70B	Meta	L	$+13.3$	$[+10.5, +15.9]$	10.4	$< .0001$	57
Qwen3-Next-80B	Alibaba	L	$+9.8$	$[+7.5, +12.1]$	9.3	$< .0001$	60
GPT-OSS-120B	OpenAI	L	$+6.3$	$[+4.4, +8.3]$	7.6	$< .0001$	60
GLM-4.5-Air	Zhipu	S	$+5.2$	$[+2.9, +7.5]$	7.3	$< .0001$	40
Nemotron-3-super-120B	NVIDIA	L	$+4.8$	$[+3.1, +6.7]$	7.2	$< .0001$	60
Scout	Meta	S	-4.2	$[-6.9, -1.7]$	1.1	1.3×10^{-3}	60
Opus 4.6	Anthropic	L	-5.0	$[-6.6, -3.5]$	0.4	$< .0001$	60
Sonnet 4.6	Anthropic	L	-7.7	$[-9.5, -6.0]$	0.5	$< .0001$	60

Most models use 10 runs at T=0.7; a subset (GPT-5.4, Sonnet 4.6, Pro 3.1, DeepSeek-V3.2, Opus 4.6) uses 5 runs, and the temperature ablation in §C.6 confirms direction is preserved across $T \in \{0.7, 1.0\}$.

Table 13. Anthropic three-point gradient. Sonnet is more pessimistic than the larger Opus, indicating alignment recipe matters.

Model	Tier	Skew	Direction
Haiku 4.5	lightweight	$+6.1$	Optimistic
Opus 4.6	frontier	-5.0	Pessimistic
Sonnet 4.6	balanced	-7.7	Pessimistic

E. Detailed Results

E.1. Anthropic Three-Point Gradient

Anthropic provides three models spanning the full bias spectrum (Table 13). The gradient is not monotonic with model size: Sonnet (balanced tier) is more pessimistic than Opus (frontier tier), despite Opus being the largest model. This non-monotonicity suggests that bias direction depends on the specific alignment recipe applied to each tier rather than model scale alone. One interpretation is that Sonnet's training emphasized balanced, cautious judgment (consistent with its role as a general-purpose assistant), producing stronger pessimistic correction than Opus's frontier-focused training. This three-point gradient is the strongest evidence against a simple "bigger = less biased" narrative, and supports the view that alignment is a multi-dimensional process with different outcomes at different tiers.

E.2. Cross-Lingual Results

Table 14 shows cross-lingual bias for Gemini Flash across six native-prompt languages (60 pairs, 10 runs per item). All six languages produce significant optimism ($+5.3$ to $+7.4$) at $p < 10^{-4}$. The range is narrow (2.1 pp) and the per-language ordering is not the focus: with a single model we cannot interpret which language ranks higher as a property of language itself. The body-text analysis (§C.9) shows that across six models inter-model variance is $2.7\times$ inter-language variance; this table is the per-language detail for Gemini Flash.

Table 14. Gemini Flash cross-lingual Skew on the 6 languages with native-language system prompts (60 pairs, 10 runs per item). All $p < 0.0001$. The four languages with English system prompts (DE, FR, HI, JA) are summarized in the per-language table in Appendix C.9.

Language	Skew	p	Pairs	Parsed
English	+5.3	< .0001	60	1335
Spanish	+6.0	< .0001	60	1337
Arabic	+6.0	< .0001	60	1336
Korean	+6.9	< .0001	60	1328
Russian	+7.1	< .0001	60	1329
Chinese	+7.4	< .0001	60	1339

Table 15. Per-domain Skew. Health domain amplifies bias; academic domain suppresses it. Bold indicates |Skew| > 10.

Model	Academic	Business	Everyday	Health	Policy	Project
GPT-5.4-mini	+9.4	**+15.2**	**+14.7**	**+19.9**	**+11.0**	**+18.9**
GPT-5.4	+6.3	+7.9	+9.3	**+10.1**	**+11.0**	**+13.7**
Haiku 4.5	−0.7	+7.6	+8.3	+6.5	+7.4	+7.1
Flash 3	+2.4	+4.8	+8.1	+7.6	+1.6	+5.8
Pro 3.1	+0.8	+3.7	+3.4	+5.3	+3.5	+6.0
Sonnet 4.6	−5.8	−6.5	−6.4	**−10.2**	**−11.6**	−5.7

E.3. Domain-Level Skew

Table 15 shows per-domain Skew for the 6 Tier-1 models. The health_habits domain produces the largest bias spread: 30 pp between GPT-5.4-mini (+19.9) and Sonnet (−10.2). In practical terms, asking "will this person maintain their exercise routine?" produces answers that differ by 30 percentage points depending on which model is queried. The academic domain produces the smallest spread, possibly because academic scenarios (exams, admissions, publications) have more readily estimable base rates from training data. The policy domain triggers Sonnet's strongest pessimism (−11.6), consistent with Constitutional AI's emphasis on caution about societal-level claims. Overall, domain accounts for only 1.4% of total Skew variance (see §E.5), confirming that bias is primarily a model property with domain-dependent modulation.

E.4. Valence Asymmetry

Table 16 decomposes Skew into its positive and negative question components. Each cell shows how far the model's mean estimate deviates from 50 (the expected value under maximum uncertainty). This decomposition reveals fundamentally different bias mechanisms.

GPT-5.4-mini exhibits *compound optimism*: both P(good) is overestimated (+6.7 from 50) and P(bad) is underestimated (+6.4 from 50). The model is sunny in both directions, inflating good outcomes and deflating bad ones. Sonnet exhibits *good-side pessimism*: P(good) is substantially underestimated (−7.3 from 50), while P(bad) is nearly accurate (−0.4 from 50). Sonnet does not overestimate risks; it underestimates opportunities. This asymmetry is consistent with Constitutional AI training that emphasizes caution: the model learned to be skeptical of positive claims without becoming catastrophist about negative ones.

An intermediate pattern appears in Pro 3.1 and Haiku 4.5, where the positive component is near zero but the negative component is elevated, producing moderate optimism through one-sided distortion. This suggests a continuum of bias mechanisms across the alignment spectrum.

E.5. Variance Decomposition

An ANOVA-style decomposition of pair-level Skew (Skew ∼ Model + Scenario + Domain, $N = 804$ pairs across 14 models and 60 scenarios) attributes variance as in Table 17. We also report random-intercept variances from per-factor REML mixed models for reference.

A separate mixed-effects model with size as a fixed effect (Skew ∼ Size + (1|Scenario)) finds that small models are +2.9 pp more optimistic than large models on average ($p = 8.9 \times 10^{-5}$). The cross-provider gradient is modest because Scout

Table 16. Valence asymmetry: deviation of P(good) and P(bad) from 50. The valence pilot uses a 5-run subset; headline Skew values in Table 1 use the full 10-run pipeline, so a model's two-component sum here can differ from its headline Skew by up to ~ 0.5 pp.

Model	P(good)−50	P(bad)−50	Skew
GPT-5.4-mini	+6.7	+6.4	+13.1
GPT-5.4	+3.5	+6.2	+9.7
Haiku 4.5	−0.2	+6.4	+6.1
Flash 3	+2.7	+2.3	+5.0
Pro 3.1	−3.0	+6.7	+3.8
Sonnet 4.6	**−7.3**	−0.4	−7.7

Table 17. Variance decomposition. Bias is primarily a model property; domain effects are negligible. Percentages from Type-II ANOVA sums of squares; REML variances from single-factor mixed-effects intercepts.

Source	REML Var.	% of Total
Model (between-model)	52.3	38.6%
Scenario (between-scenario)	22.9	20.0%
Residual (within)	–	40.0%
Domain (between-domain)	–	1.4%

(small, −4.2) pulls the small-model mean down and Mistral Small (+9.7) is less optimistic than Mistral Large (+12.2); the within-provider gradients reported in §3.2 are larger for the three providers where they hold.

F. Robustness Analyses

F.1. Temperature Robustness

Table 19 shows that bias direction and magnitude are invariant to sampling temperature across all four tested models.

Table 18. Skew at temperature 0.7 vs. 1.0. Direction identical; magnitude within 1pp.

Model	t=0.7	t=1.0	Δ
GPT-5.4	+9.7	+10.0	+0.3
GPT-5.4-mini	+14.8	+15.3	+0.5
Sonnet 4.6	−7.7	−7.7	0.0
Haiku 4.5	+6.0	+6.8	+0.8

Both columns use a matched 5-run pilot to isolate temperature effects. Headline Skew values in Table 1 use 10 runs per item, so the 5-run $t = 0.7$ entries here differ from Table 1: GPT-5.4-mini +14.8 vs. +13.1 (1.7 pp), Haiku 4.5 +6.0 vs. +6.1 (0.1 pp), GPT-5.4 and Sonnet 4.6 unchanged.