Elicitation Format Drives Divergent LLM Geopolitical Forecasts

Suhas Hariharan [1] **George Ghetiu** [1] **Ari Weiler-Ofek** [1] **Hao Jie Pe** [1] **Tatsan Kantasit** [1] **Michal Bravansky** [1 2] **Raphael Tang** [1 3]

Abstract

Large language models are approaching expert-level performance on geopolitical forecasting tasks, but a broad literature on LLM behavior shows that model outputs can shift under minor prompt perturbations. Whether matched geopolitical forecasts are similarly unstable under benign changes in elicitation remains underexplored. We study that question in a closed-book setting using Claude, GPT-OSS, and Qwen models and matched country-index forecasting tasks that hold the country, index, and horizon fixed while varying question form. A closed-book ForecastBench control confirms that the models are competent forecasters. Yet on governance targets, binary questions produce much larger US-sphere versus China-sphere gaps than matched numerical forecasts of the same country–index pairs. A Human Freedom Index comparison shows a smaller cross-bloc gap on matched economic sub-indices, suggesting that the binary amplification is concentrated in politically evaluative concepts rather than country forecasting in general. Trilingual reruns reveal additional but less uniform instability, and mirrored improve/decline prompts do not support a simple yes-saying explanation. We therefore argue that evaluations of LLM geopolitical forecasting should report robustness to elicitation alongside resolved-event accuracy, especially for politically evaluative targets.

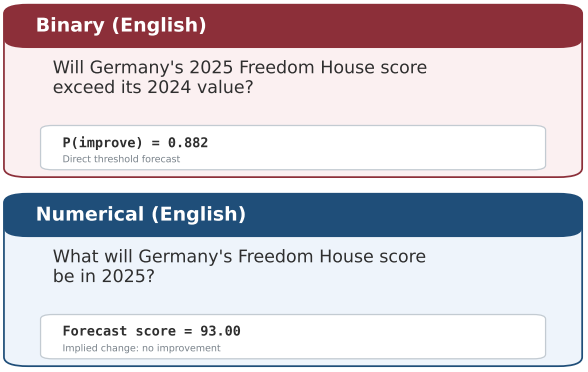

Figure 1. **Matched elicitation states.** We hold country, index, horizon, previous value, and model fixed, varying only whether the forecast is elicited as a binary threshold probability or a numerical point estimate. In the Qwen3 32B example shown, Germany's Freedom House score was 93 in 2024 and resolved to 95 in 2025.

1. Introduction

In geopolitical analysis, hindsight is cheap. Once an event has occurred, almost any outcome can be made to look predictable in retrospect. This is a familiar hindsight problem in judgment research (Fischhoff, 1975). Probabilistic forecasting is harder because probabilities must be assigned before events resolve and then checked against what actually happened.

In the Tetlock tradition, repeated forecasting tournaments made political judgment empirically tractable by asking many forecasters the same unresolved questions and comparing performance after resolution (Tetlock, 2005; Mellers et al., 2014; Tetlock & Mellers, 2014; Mellers et al., 2015; Tetlock & Gardner, 2015). They also made it possible to test whether updating, training, and aggregation improved judgment. That setup is especially useful in geopolitics, where evidence is often noisy, reference classes are weak, and feedback is slow (Mellers et al., 2024).

Recent work reports meaningful LLM forecasting performance from models alone, human–model teams, and tool-supported systems (Halawi et al., 2024; Schoenegger et al., 2024a;b; Ye et al., 2024). ForecastBench sharpened those comparisons by introducing rolling test sets with explicit cutoff logic to reduce contamination (Karger et al., 2024). But those evaluations mainly ask whether a model can forecast accurately at all. They say much less about whether the same forecast remains stable under benign changes in elicitation.

[1]University College London [2]METR [3]Microsoft. Correspondence to: Suhas Hariharan <hariharansuhas@gmail.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

This concern is especially salient for country-index forecasting. Unlike many event questions, which resolve on discrete occurrences or settled facts, country-index targets ask a model to predict year-over-year movements in comparative governance ratings. Work on global performance indicators has emphasized that such indices do not merely record reality but also frame issues, standardize comparison across countries, and can shape discourse and policy (Merry, 2011; Kelley & Simmons, 2019). Annual indices such as Freedom House, the Corruption Perceptions Index, the World Press Freedom Index, and the Human Freedom Index therefore offer a useful setting for studying elicitation robustness on politically evaluative targets (Freedom House, 2025; Transparency International, 2025; Reporters Without Borders, 2025; Cato Institute and Fraser Institute, 2024; 2025), since the same country–index–year target can be posed either as a binary threshold question or as a numerical point forecast.

In this paper, we evaluate matched country-index forecasts in a closed-book setting, where the clearest result is a large binary–numerical dissociation on governance targets. Binary questions produce substantially larger US-sphere versus China-sphere gaps than matched numerical forecasts of the same country–index pairs, while cross-language reruns reveal further instability, though less uniformly across models, and mirrored improve/decline prompts do not support a simple yes-saying account. For politically evaluative targets, then, forecast quality cannot be summarized by accuracy and calibration alone, but also depends on whether a forecast survives benign changes in elicitation. We therefore argue that robustness to elicitation should be treated as a core evaluation criterion for forecasting systems in politically sensitive domains.

2. Related Work

Recent work on forecasting has focused on how well humans and human–AI systems predict resolved events. The superforecasting literature established repeated, scored geopolitical forecasting as a useful setting for studying judgment under uncertainty (Tetlock, 2005; Mellers et al., 2014; Tetlock & Mellers, 2014; Mellers et al., 2015; Tetlock & Gardner, 2015). Recent LLM work extends that agenda to resolved event benchmarks, retrieval-heavy forecasting agents, and human–AI workflows (Zou et al., 2022; Halawi et al., 2024; Karger et al., 2024; Ye et al., 2024; Schoenegger et al., 2024a;b; Jeen et al., 2026; Yang et al., 2025). Recent evidence also suggests that forecasting performance of LLMs depends on what general topic is being forecast and how the task is posed (Karkar & Chopra, 2025). We examine a more target-controlled version of that question by holding the country–index target fixed and varying only elicitation.

This setup also connects to work on framing effects, prompt sensitivity, and evaluation-format bias. Classical work in judgment and decision-making showed that semantically equivalent choices can elicit different judgments under different frames (Tversky & Kahneman, 1981). Recent LLM work documents analogous sensitivity to wording, conversational stance, and evaluation order (Perez et al., 2022; Wei et al., 2023; Razavi et al., 2025; Zheng et al., 2024; Wang et al., 2024; Shi et al., 2025). We bring that invariance question into probabilistic forecasting rather than multiple-choice selection or preference judging.

The multilingual and political-bias literature is equally relevant. Prior studies show that multilingual models do not always preserve evaluative judgments across languages and may express political or cultural biases in language-dependent ways (Qi et al., 2023; Naous et al., 2024; Bang et al., 2024; Shin et al., 2024; Zhou & Zhang, 2024; Nie et al., 2024). A separate line of work studies confidence, uncertainty, and calibration (Kadavath et al., 2022; Kapoor et al., 2024; Hager et al., 2025). Such biases can also extend into the chain-of-thought of LLMs, which is not always a faithful account of model reasoning (Turpin et al., 2023; Madsen et al., 2024; Karvonen & Marks, 2025). Together these literatures motivate our distinction between calibration error and forecast robustness. A model can be acceptably calibrated on average while still being unstable under matched prompt variants.

3. Experimental Setup

Forecasting task. Each example fixes country, index, forecast year, and prior value. We ask either a binary threshold question, "Will the score in year $t+1$ exceed its year-t value?", or a numerical point estimate, "What will the score in year $t+1$ be?" Matched comparisons binarize numerical forecasts by whether they exceed the prior value, so only elicitation changes. We highlight that the two are not probability-equivalent. The binary form is natively probabilistic, while the numerical form yields a direction only after thresholding, which discards the uncertainty around the point estimate. The comparison is thus directional, asking whether the conditions agree on the direction of change rather than on a shared probability. Still, we explore more probability-equivalent formulations in Appendix H, where recomputing the numerical side as the sample fraction above the prior value preserves the divergence and supports the same conclusion.

All evaluations are closed book, with no retrieval, web search, or tools. Because model timing differs, the ForecastBench competence control uses timing-aware subsets with slice-specific base-rate baselines (Appendix D). Unless noted, we sample each question five times at temperature 0.7 and take the median over valid extractions, following

self-consistency-style aggregation (Wang et al., 2023). Appendix F gives prompts and Appendix E defines metrics.

Data and models. Countries are grouped into US-sphere, China-sphere, Russia-sphere, and non-aligned blocs from public geopolitical-alignment data (Appendix C). Governance targets come from Freedom House's *Freedom in the World*, Transparency International's Corruption Perceptions Index, and RSF's World Press Freedom Index; HFI provides matched governance and economic sub-indices for the domain-specificity analysis. Most paired comparisons focus on the US–China contrast, with Russia-sphere and non-aligned countries retained for multilingual and ancillary checks. The model pool is Claude Haiku 4.5 (Anthropic, 2025), GPT-OSS 20B and 120B (OpenAI, 2025), and Qwen3 32B (Qwen Team, 2025a). English framing, ForecastBench, trilingual, and acquiescence analyses use all four models; HFI uses the three-model subset in Appendix A. Appendices A–C list coverage, protocols, timing, country membership, and sources.

Metrics. Our main directional estimand is the US-sphere minus China-sphere forecast difference, the mean predicted improvement for US-sphere countries minus the corresponding China-sphere mean. Predicted improvement is the binary probability of exceeding the current score, or the implied direction after thresholding a numerical point forecast. We measure resolved-event accuracy with Brier score (Brier, 1950; Gneiting et al., 2007), and use supporting robustness/control quantities for matched variants, especially binary–numerical contradiction rate and the mirrored improve/decline diagnostic, $\mathrm{AS} = P(\mathrm{improve}) + P(\mathrm{decline}) - 1$.

4. Results

4.1. Binary framing amplifies US–China governance gaps

For the same country–index–year target, the binary version asks whether the next score will exceed the current one, while the numerical version asks for the next score directly. In both versions, the prompt provides the previous year's observed index value as the reference point. Figure 2 reports the mean predicted improvement for US-sphere countries minus the corresponding mean for China-sphere countries on matched governance questions, with model-level values in Appendix G. Positive values mean relatively more predicted improvement for the US sphere, and negative values mean the reverse.

Under binary threshold prompting, the forecast difference is positive in every model, from 0.131 for Claude Haiku 4.5 to 0.436 for Qwen3 32B. Under numerical prompting, the same comparative pattern becomes much weaker, and

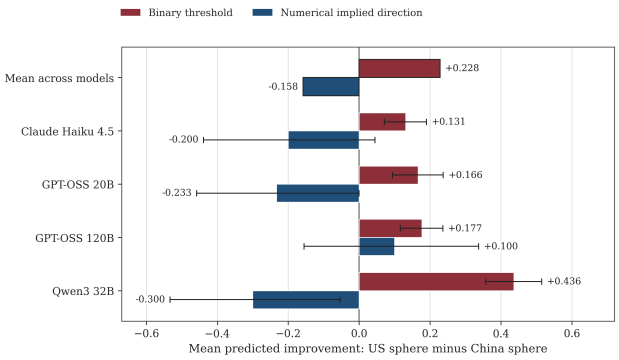

Figure 2. **Binary and numerical questions yield different cross-bloc governance forecasts.** Bars show US–China mean predicted-improvement differences on matched governance questions; error bars are 95% CIs.

three of the four models reverse sign from US-favoring to China-favoring. Averaged across the four models, the forecast difference falls from 0.228 under binary prompting to -0.158 under numerical prompting. The within-model swing has bootstrap 95% confidence intervals that exclude zero for Claude Haiku 4.5, GPT-OSS 20B, and Qwen3 32B; GPT-OSS 120B moves in the same direction with a wider interval. Asking for a threshold judgment rather than a point estimate can therefore change which bloc is assigned relatively more predicted improvement. Additional checks point in the same direction, as score-swap prompts and retrospective HFI variants preserve a substantial US-sphere minus China-sphere gap (Appendix K), making it unlikely that the dissociation is driven solely by visible score anchors or unresolved-target timing.

4.2. Binary prompts recast forecasts as reform-plausibility judgments

To understand why binary and numerical prompts behave differently, we ran an analysis using LLM-as-judge of 284 usable English governance traces from Claude Haiku 4.5 and Qwen3 32B (Appendices F and K give the exact protocol). A separate judge model analyzed and labeled each trace for (1) primary reasoning mode, (2) explicit reform assessment, and (3) anchoring on the previous score. The results allow us to differentiate counterfactual reasoning, which asks whether improvement is plausibly achievable, from statistical reasoning, which projects forward from the previous score or recent trajectory. In binary mode, 48% of traces were judged primarily counterfactual, compared with 7% in numerical mode, while strong anchoring remained common in both framings. The contrast is even more apparent on China- and Russia-sphere cases, where 72% of binary traces were judged primarily counterfactual, compared with 18% of the corresponding numerical traces. As such, the binary threshold appears to recast the forecast as a judgment

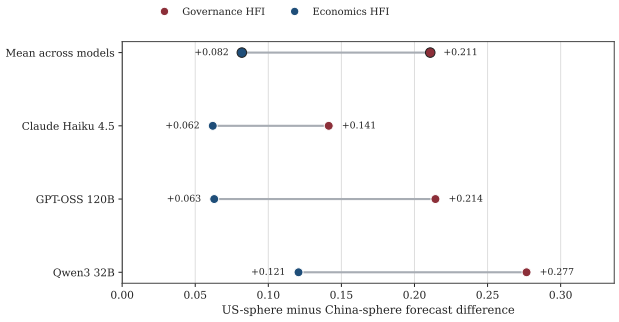

Figure 3. **The binary amplification is concentrated in governance concepts.** HFI binary gaps are larger for governance sub-indices than for matched economics sub-indices.

about whether reform is plausible rather than a projection of the next value, which admits evaluative priors about which countries can credibly improve.

4.3. The amplification is concentrated in governance targets

We next ask whether the binary gap is specific to politically evaluative targets. The Human Freedom Index provides a matched within-source comparison because its governance and economics sub-indices share the same source family, scale, and country inventory. Across the three-model HFI subset, the mean US-versus-China forecast difference under binary prompting is 0.211 for governance sub-indices and 0.082 for matched economic sub-indices, with model-level values in Appendix I. Figure 3 shows that the binary effect is concentrated in politically evaluative governance concepts such as expression, association, and rule of law. The same models are much less separated on matched economic concepts such as trade, sound money, and size of government. We report this HFI domain comparison for binary prompting only, as we did not observe this effect in the numerical-framing slice.

We therefore do not read this difference as a generic country-preference effect. Holding the HFI source family, countries, and scale fixed, the gap is much larger for governance than for economic sub-indices, which points to something more specific where binary prompting is especially sensitive to politically evaluative targets, rather than to country-index forecasting in general.

4.4. Prompt language shifts the gap less uniformly

We next hold the governance target fixed and translate the same matched binary governance questions into Chinese and Russian, with localized prompt details in Appendix F. Similar multilingual drift has been observed in prior work on political and evaluative judgments in LLMs (Qi et al., 2023; Naous et al., 2024; Zhou & Zhang, 2024). On this governance-only rerun slice, Chinese usually narrows the US-versus-China forecast gap, mainly by lowering predicted improvement for US-sphere countries rather than by raising it for China-sphere countries.

For Qwen3 32B, the gap falls from 0.436 in English to 0.211 in Chinese because the US-sphere mean drops by 0.240 while the China-sphere mean changes little. On the matched binary Freedom House question asking whether Germany's 2025 score will exceed its 2024 value of 93, Qwen3 32B falls from 0.882 in English to 0.150 in Chinese. Claude Haiku 4.5 moves in the same direction but more mildly. GPT-OSS is mixed internally. GPT-OSS 20B is essentially unchanged, whereas GPT-OSS 120B narrows the gap by combining a modest US-sphere decline with a small China-sphere increase. Appendix J reports the Chinese and Russian within-bloc shifts. These shifts are less uniform than the binary–numerical effect, but they reinforce the broader point that forecast outputs depend on how the forecasting problem is presented.

4.5. The gap is not explained by forecasting failure or simple yes-saying

Before interpreting the country-index results as an elicitation effect, we rule out two simpler explanations. First, on timing-aware closed-book ForecastBench controls, three of the four models beat their slice-specific base-rate baselines (Appendix D), so the country-index instability cannot be dismissed as generic forecasting failure. Because model timing differs, this is a within-slice competence check rather than a cross-model ranking. Second, mirrored improve/decline prompts do not support a simple yes-saying account. If the binary gap reflected indiscriminate agreement, models should assign high probability to both "improve" and "decline" variants of the same threshold question; instead, mirrored scores are negative on average across the four-model set (Appendix K).

5. Discussion and Conclusion

Matched elicitation changes can alter geopolitical forecasts even when targets are fixed. On identical country–index–year questions, binary and numerical prompts often yield different cross-bloc conclusions, and the HFI comparison localizes the strongest binary instability to politically evaluative governance targets rather than country forecasting in general. Binary and numerical elicitation may therefore not be surface forms of the same forecast. On governance targets, numerical prompts favor incremental projection, whereas binary prompts elicit a thresholded judgment about whether reform is plausible. Our trace analysis is consistent with that account, but the evidence is suggestive as chain-of-thought might not be always a faithful record of how the answer was produced (Turpin et al., 2023; Madsen et al.,

2024).

Several limitations qualify the claim. The country-index targets are politically evaluative benchmarks; we do not treat the indices as neutral ground truth, but hold the indexed target fixed and study how forecasts move when presentation changes. The binary–numerical comparison is directional rather than probability-equivalent, because binary probabilities are compared with implied directions from numerical point estimates. Appendix H recomputes the numerical side as the fraction of valid numerical samples above the previous score, and the mean binary-minus-numerical swing remains positive at 0.328. Mechanism evidence is also indirect, relying on English traces from a model subset and an external judge model rather than full-scale human annotation.

Holding a forecasting target fixed is therefore insufficient to guarantee stability. Accuracy and calibration remain necessary, but for politically sensitive forecasting they are incomplete without robustness checks across binary, numerical, and translated variants. As a proof of concept for future work, we trained a small Qwen3-4B model to align binary outputs with probabilities implied by its own numerical forecasts (Appendix L). While this does not fully resolve the inconsistency, we believe similar consistency interventions may reduce binary–numerical inconsistency across larger models and domains.

References

Aiyar, S., Ohnsorge, F., and Yilmazkuday, H. Horizontal and vertical connector countries in a geoeconomically fragmenting world. Discussion Paper 19352, Centre for Economic Policy Research, 2024. URL https://cepr.org/publications/dp19352.

Amazon Web Services. Anthropic claude haiku 4.5 model card, 2025a. URL https://docs.aws.amazon.com/bedrock/latest/userguide/model-card-claude-haiku-4-5.html. Accessed May 12, 2026.

Amazon Web Services. Qwen3 32b model card, 2025b. URL https://docs.aws.amazon.com/bedrock/latest/userguide/model-card-qwen-qwen3-32b.html. Accessed May 12, 2026.

Anthropic. System card: Haiku 4.5, October 2025. URL https://www.anthropic.com/claude-haiku-4-5-system-card. Reports February 2025 public-internet cutoff.

Bailey, M. A., Strezhnev, A., and Voeten, E. Estimating dynamic state preferences from United Nations voting data. *Journal of Conflict Resolution*, 61(2):430–456, 2017. doi: 10.1177/0022002715595700.

Bang, Y., Chen, D., Lee, N., and Fung, P. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11142–11159. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.600. URL https://aclanthology.org/2024.acl-long.600/.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Cato Institute and Fraser Institute. Human freedom index 2024, 2024. URL https://www.cato.org/sites/cato.org/files/2024-12/human-freedom-index-2024.pdf. Accessed April 8, 2026.

Cato Institute and Fraser Institute. Human freedom index 2025, 2025. URL https://www.cato.org/human-freedom-index/2025. Accessed May 12, 2026.

Fischhoff, B. Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3):288–299, 1975. doi: 10.1037/0096-1523.1.3.288.

Freedom House. Freedom in the world 2025, 2025. URL https://freedomhouse.org/report/freedom-world. Public report page dated February 2025; accessed April 8, 2026.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x. URL https://academic.oup.com/jrsssb/article-abstract/69/2/243/7109375.

Hager, S., Mueller, D., Duh, K., and Andrews, N. Uncertainty distillation: Teaching language models to express semantic confidence, 2025. URL https://arxiv.org/abs/2503.14749.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models, 2024. URL https://arxiv.org/abs/2402.18563.

Jeen, S., Aitchison, M., and Mantic. Training LLMs to predict world events. *Thinking Machines Lab: News*, March 2026. URL

https://thinkingmachines.ai/news/training-llms-to-predict-world-events/. Guest post with Mantic, published March 19, 2026.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Kapoor, S., Gruver, N., Roberts, M., Collins, K., Pal, A., Bhatt, U., Weller, A., Dooley, S., Goldblum, M., and Wilson, A. G. Large language models must be taught to know what they don't know, 2024. URL https://arxiv.org/abs/2406.08391.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. Forecastbench: A dynamic benchmark of ai forecasting capabilities, 2024. URL https://arxiv.org/abs/2409.19839.

Karkar, C. and Chopra, P. Future is unevenly distributed: Forecasting ability of llms depends on what we're asking, 2025. URL https://arxiv.org/abs/2511.18394.

Karvonen, A. and Marks, S. Robustly improving llm fairness in realistic settings via interpretability. 2025. URL https://arxiv.org/abs/2506.10922.

Kelley, J. G. and Simmons, B. A. Introduction: The power of global performance indicators. *International Organization*, 73(3):491–510, 2019. doi: 10.1017/S0020818319000146.

Lab, T. M. Tinker, 2025. URL https://thinkingmachines.ai/tinker/.

Madsen, A., Chandar, S., and Reddy, S. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 295–337. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.19. URL https://aclanthology.org/2024.findings-acl.19/.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., and Tetlock, P. E. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115, 2014. doi: 10.1177/0956797614524255. URL https://pubmed.ncbi.nlm.nih.gov/24659192/.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., and Tetlock, P. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10 (3):267–281, 2015. doi: 10.1177/1745691615577794. URL https://journals.sagepub.com/doi/10.1177/1745691615577794.

Mellers, B. A., McCoy, J. P., Lu, L., and Tetlock, P. E. Human and algorithmic predictions in geopolitical forecasting: Quantifying uncertainty in hard-to-quantify domains. *Perspectives on Psychological Science*, 19(5): 711–721, 2024. doi: 10.1177/17456916231185339. URL https://journals.sagepub.com/doi/10.1177/17456916231185339.

Merry, S. E. Measuring the world: Indicators, human rights, and global governance. *Current Anthropology*, 52(S3): S83–S95, 2011. doi: 10.1086/657241.

Naous, T., Ryan, M. J., Ritter, A., and Xu, W. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16366–16393. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.862. URL https://aclanthology.org/2024.acl-long.862.

Nie, E., Yuan, S., Ma, B., Schmid, H., Färber, M., Kreuter, F., and Schütze, H. Decomposed prompting: Probing multilingual linguistic structure knowledge in large language models, 2024. URL https://arxiv.org/abs/2402.18397.

Nurullayev, D. and Papa, M. Bloc politics at the un: How other states behave when the united states and china–russia disagree. *Global Studies Quarterly*, 3(3):ksad034, 2023. doi: 10.1093/isagsq/ksad034. URL https://academic.oup.com/isagsq/article/3/3/ksad034/7223049.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, August 2025. URL https://openai.com/index/gpt-oss-model-card/. Reports June 2024 knowledge cutoff.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L.,

Lovitt, L., Lucas, M., Sellitto, M., Zhang, M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., Showk, S. E., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. Discovering language model behaviors with model-written evaluations, 2022. URL https://arxiv.org/abs/2212.09251.

Qi, J., Fernández, R., and Bisazza, A. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10650–10666. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.658. URL https://aclanthology.org/2023.emnlp-main.658.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3: Think deeper, act faster, April 2025b. URL https://qwenlm.github.io/blog/qwen3/. Announcement page for the Qwen3 model family.

Razavi, A., Soltangheis, M., Arabzadeh, N., Salamat, S., Zihayat, M., and Bagheri, E. Benchmarking prompt sensitivity in large language models, 2025. URL https://arxiv.org/abs/2502.06065.

Reporters Without Borders. World press freedom index 2025, 2025. URL https://rsf.org/en/index. Release materials dated May 2025; accessed April 8, 2026.

Schoenegger, P., Park, P. S., Karger, E., Trott, S., and Tetlock, P. E. Ai-augmented predictions: Llm assistants improve human forecasting accuracy, 2024a. URL https://arxiv.org/abs/2402.07862.

Schoenegger, P., Tuminauskaite, 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), 2024b. doi: 10.1126/sciadv.adp1528. URL https://eprints.lse.ac.uk/125626/.

Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., and Vosoughi, S. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Shin, J., Song, H., Lee, H., Jeong, S., and Park, J. Ask LLMs directly, "what shapes your bias?": Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 16122–16143. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.954. URL https://aclanthology.org/2024.findings-acl.954/.

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

Tetlock, P. E. and Gardner, D. *Superforecasting: The Art and Science of Prediction*. Crown, New York, 2015.

Tetlock, P. E. and Mellers, B. Judging political judgment. *Proceedings of the National Academy of Sciences*, 111(32):11574–11575, 2014. doi: 10.1073/pnas.1412524111. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4136586/.

Transparency International. Corruption perceptions index 2024, 2025. URL https://www.transparency.org/en/publications/corruption-perceptions-index-2024. Publication page dated February 2025; accessed April 8, 2026.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023. doi: 10.48550/arXiv.2305.04388. URL https://arxiv.org/abs/2305.04388.

Tversky, A. and Kahneman, D. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. doi: 10.1126/science.7455683.

United Nations General Assembly. Resolution ES-11/3: Suspension of the rights of membership of the Russian Federation in the Human Rights Council, 2022. URL https://digitallibrary.un.org/record/3967950. A/RES/ES-11/3, adopted 7 April 2022.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., and Sui, Z. Large language models are not fair evaluators. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9440–9450. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Wei, J., Huang, D., Lu, Y., Zhou, D., and Le, Q. V. Simple synthetic data reduces sycophancy in large language models, 2023. URL https://arxiv.org/abs/2308.03958.

Yang, Q., Mahns, S., Li, S., Gu, A., Wu, J., and Xu, H. Llm-as-a-prophet: Understanding predictive intelligence with prophet arena, 2025. URL https://arxiv.org/abs/2510.17638.

Ye, C., Hu, Z., Deng, Y., Huang, Z., Ma, M. D., Zhu, Y., and Wang, W. Mirai: Evaluating llm agents for event forecasting, 2024. URL https://arxiv.org/abs/2407.01231.

Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. Large language models are not robust multiple choice selectors, 2024. URL https://arxiv.org/abs/2309.03882.

Zhou, D. and Zhang, Y. Political biases and inconsistencies in bilingual GPT models: The cases of the U.S. and china. *Scientific Reports*, 14:25048, 2024. doi: 10.1038/s41598-024-76395-w. URL https://www.nature.com/articles/s41598-024-76395-w.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. Forecasting future world events with neural networks. *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.15474.

A. Experiment Coverage and Shared Protocol

Table 1 records the model coverage for the experiments used in this study. The full model set is Claude Haiku 4.5, GPT-OSS 20B, GPT-OSS 120B, and Qwen3 32B. The shared default is five samples at temperature 0.7, median aggregation over valid extractions, and no retrieval, web search, or external tools.

Shared inference setup. Unless an experiment defines extra conversational turns, each model receives a single rendered user prompt and no custom system prompt. Probability extraction first looks for an explicit terminal tag such as `PROBABILITY: X.XX`; if no such tag is found, the extractor falls back to percentages and then to a bare decimal near the end of the output. Numerical extraction similarly prefers `ESTIMATE: X.XX`. If every sampled output fails extraction, binary questions default to 0.5 and numerical questions to 0.0.

We also audited extraction failures after aggregation. The reported analysis subsets contain no all-fail numerical extractions. In the broader rerun archive, all-fail numerical fallbacks are rare, affecting 7 of 12,357 numerical rows, or 0.057%, and occur outside the headline analyses.

Protocol exceptions. The acquiescence experiment uses three samples per question. The consistency-training pilot uses deterministic decoding at temperature 0.0 for evaluation. For non-English country-index prompts, the localized resolution statement is embedded in the translated question text and the renderer leaves `resolution_info` empty rather than appending an English resolution line.

B. Model and Target Timing

Closed-book evaluation here means that the inference call receives no retrieval, web search, or external tool output. Tables 3 and 4 report public model timing and source-release timing for interpreting the country-index comparisons.

C. Country-Index Construction

Main governance examples are built from a 2024 observed value and a 2025 target value. HFI examples are built from adjacent HFI editions, using a 2022 observed value and a 2023 target value. The binary version asks whether the target value improves on the previous value; the numerical version asks directly for the target value. The English binary question template at the question-text level is shown below.

```
Will {country}'s {index} in 2025 be
above its 2024 value of {value}?
{scale statement}
```

The matched numerical question follows.

```
What will {country}'s {index} be in
2025?  {scale statement}
```

For all matched analyses, the country, index, target year, previous value, scale direction, and model are fixed across the binary and numerical variants.

The country blocs are analytic groupings constructed from publicly available alignment evidence rather than model-inferred clusters. We use UN General Assembly voting and ideal-point evidence as the primary source signal (Bailey et al., 2017), including work that measures recent geoeconomic alignment from those ideal points (Aiyar et al., 2024), bloc-politics analyses of cases where the United States and China–Russia disagree (Nurullayev & Papa, 2023), and the public vote record for UNGA Resolution ES-11/3 on suspending Russia from the Human Rights Council (United Nations General Assembly, 2022). We separate China sphere and Russia sphere because preliminary analyses found distinct baseline scores and response patterns for these countries, so pooling them would obscure the main US-versus-China contrast.

The main analyses exclude EPI, Polyarchy, and Liberal Democracy because data checks identified contamination or measurement issues. Chinese and Russian question texts were translated from the matched English question set; native speakers reviewed spot-check samples for fluency and semantic fidelity.

D. ForecastBench Control

The closed-book control is derived from ForecastBench (Karger et al., 2024). We keep only resolved binary questions, remove `dbnomics`, `yfinance`, and `fred` items because they are effectively time-series lookups, remove entries with non-string IDs, and keep the earliest resolution timestamp when duplicate question texts appear. To reduce contamination from model timing, we use two snapshots and apply an explicit time filter. For models with a public cutoff, questions must resolve after the relevant cutoff. For Qwen3 32B, whose public sources do not report a training cutoff, we use the April 2025 public release as an upper-bound proxy and keep only questions resolving after release. Only 18 questions overlap between snapshots, so we compare each model to its slice-specific base-rate baseline rather than making paired cross-slice model comparisons.

Dataset A contains 60 ACLED questions, 59 Polymarket questions, 54 Wikipedia questions, 32 Manifold questions, 19 Metaculus questions, and 13 Infer questions. Dataset B contains 63 ACLED questions, 62 Polymarket questions, 53 Wikipedia questions, 18 Manifold questions, 8 Metaculus questions, and 4 Infer questions. The Qwen3 32B release-filtered subset contains 71 of the Dataset B questions, with resolutions from 2025-05-04 to 2026-01-01, a yes rate of

Table 1. Experiment coverage and protocol details.

Analysis	Coverage	Models	Samples	Notes
Closed-book control	Full model set	Claude Haiku 4.5, GPT-OSS 20B, GPT-OSS 120B, Qwen3 32B	5	ForecastBench split by model timing
Governance framing comparison	Full model set	Claude Haiku 4.5, GPT-OSS 20B, GPT-OSS 120B, Qwen3 32B	5	English matched binary/numerical rerun
Sample-fraction robustness	Full model set	Claude Haiku 4.5, GPT-OSS 20B, GPT-OSS 120B, Qwen3 32B	5	Recomputes numerical implied improvement from individual samples
HFI domain comparison	HFI subset	Claude Haiku 4.5, GPT-OSS 120B, Qwen3 32B	5	Matched governance/economics comparison; GPT-OSS 20B not run
Trilingual governance reruns	Full model set	Claude Haiku 4.5, GPT-OSS 20B, GPT-OSS 120B, Qwen3 32B	5	English, Chinese, and Russian prompts
Reasoning traces	Trace subset	Claude Haiku 4.5, Qwen3 32B	5	Blinded judge pass over 284 usable English governance traces
Acquiescence control	Full model set	Claude Haiku 4.5, GPT-OSS 20B, GPT-OSS 120B, Qwen3 32B	3	Mirrored improve/decline prompts
Consistency-training pilot	Pilot	Qwen3-4B-Instruct	1	Separate small-model study at temperature 0.0

0.197, and a base-rate Brier of 0.158.

E. Metrics

Our main differential-treatment estimand is the US-sphere minus China-sphere forecast difference

$$\Delta_{\mathrm{US-CN}} = \frac{1}{|U|}\sum_{i\in U} s_i - \frac{1}{|C|}\sum_{j\in C} s_j,$$

where U and C denote the US-sphere and China-sphere question sets. For binary questions, $s_i = p_i$, the forecast probability that the 2025 value exceeds the 2024 reference value. For numerical questions, we convert the point forecast $\hat{y}_i$ into an implied improve/decline call by setting $s_i = \mathbf{1}[\hat{y}_i > y_i^{(2024)}]$.

As a robustness check, Appendix H also reports a sample-level numerical variant. If a numerical question has valid extracted samples $\hat{y}_{i1}, \ldots, \hat{y}_{iK}$, we set

$$s_i^{\mathrm{sample}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}[\hat{y}_{ik} > y_i^{(2024)}],$$

and then compute the same US-sphere minus China-sphere difference. This preserves the directional interpretation while avoiding a single median point estimate as the only bridge between numerical and binary formats.

Unless a table states otherwise, target-level uncertainty intervals resample country–indicator targets within model while preserving the matched question construction. The mean row in Figure 2 resamples models, since it summarizes across model-level estimates. The HFI domain comparison in Figure 3 reports aggregate model values without bootstrap intervals. Table 2 gives the corresponding sample sizes and valid-extraction counts.

For resolved binary event forecasting, the Brier score is

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2,$$

where p_i is the predicted probability and $o_i \in \{0, 1\}$ is the resolved outcome. For matched binary–numerical comparisons, the contradiction rate is

$$\mathrm{CR} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\big[(p_m \geq 0.5) \neq (\hat{y}_m > y_m^{(2024)})\big].$$

For the mirrored improve/decline diagnostic, we report

$$\mathrm{AS} = P(\mathrm{improve}) + P(\mathrm{decline}) - 1.$$

F. Prompt Templates

No custom system prompt. The shared evaluation pipeline does not prepend a custom system message. The model receives the rendered user prompt below unless a protocol explicitly adds extra turns.

Default binary prompt.

```
You are a forecaster estimating the
probability of future events.

Question: {question_text}
```

Table 2. Sample sizes and uncertainty units for the reported analyses. Aggregate extraction means one valid aggregated prediction for the model–question row.

Analysis	Countries	Targets	Rows	Generations / valid extractions	CI unit
Main governance framing	20	60 country–index targets, each binary and numerical	480 model–question rows	2,400 generations; 480/480 aggregate extractions	Target bootstrap
Sample-fraction robustness	20	60 country–index targets, each binary and numerical	480 model–question rows	2,400 generations; 2,391 valid sample-level extractions	Target bootstrap
HFI domain comparison	15 in the US–China contrast; 25 generated overall	90 US–China binary HFI targets	270 model–question rows	1,350 generations; 270/270 aggregate extractions	Aggregate model values
Mirrored improve/decline diagnostic	158	553 country–indicator mirror pairs	2,212 model–pair rows	13,272 generations; 4,424 aggregate extractions forming 2,212 valid pairs	Pair-level intervals
Reasoning traces	20	284 usable traces	284 judged traces	One deterministic judge call per trace	Trace-level summaries

Table 3. Public model timing information used for cutoff- and release-aware controls.

Model	Public release	Reported cutoff	Interpretation
Claude Haiku 4.5	October 2025 (Anthropic, 2025)	February 2025 public-internet cutoff in the system card (Anthropic, 2025; Amazon Web Services, 2025a)	Some 2025 index releases occur close to or after this cutoff, so we report source timing explicitly.
GPT-OSS 20B	August 2025 (OpenAI, 2025)	June 2024 (OpenAI, 2025)	The 2025 country-index target values postdate the reported cutoff.
GPT-OSS 120B	August 2025 (OpenAI, 2025)	June 2024 (OpenAI, 2025)	Same timing as GPT-OSS 20B.
Qwen3 32B	April 2025 (Qwen Team, 2025b)	Not publicly reported in the Qwen3 technical report or AWS Bedrock model card (Qwen Team, 2025a; Amazon Web Services, 2025b)	For ForecastBench, we use the public release as an upper-bound proxy and restrict to post-release resolutions. For country-index targets, we report source timing explicitly.

Default numerical prompt.

```
{resolution_info}

Instructions:
- Think step by step about the key
factors that would influence this
outcome.
- Consider base rates, historical
precedents, and current trends.
- Provide your reasoning, then give
a final probability estimate.

You MUST end your response with
exactly this format:
PROBABILITY: X.XX

where X.XX is a number between 0.00
and 1.00 representing your estimated
probability.
```

```
You are a forecaster estimating
numerical values for future events.

Question: {question_text}

{resolution_info}

Instructions:
- Think step by step about the key
factors that would influence this
value.
- Consider historical data, trends,
and current conditions.
- Provide your reasoning, then give
a final numerical estimate.

You MUST end your response with
exactly this format:
ESTIMATE: X.XX

where X.XX is your best numerical
```

Table 4. Country-index target timing for retained source families.

Source family	Prompted transition	Public source timing	Use in this study
Freedom House *Freedom in the World*	2024 value to 2025 value	*Freedom in the World 2025* was public in February 2025 (Freedom House, 2025).	Included in the main governance framing, trilingual, and acquiescence analyses.
Transparency International CPI	2024 value to 2025 value	The CPI 2024 publication page is dated February 2025 (Transparency International, 2025).	Included in the main governance framing, trilingual, and acquiescence analyses.
RSF World Press Freedom Index	2024 value to 2025 value	The 2025 index page was public in 2025, with release materials dated May 2025 (Reporters Without Borders, 2025).	Included in the main governance framing, trilingual, and acquiescence analyses.
Human Freedom Index	2022 value to 2023 value	The HFI comparison pairs the 2024 edition for 2022 values with the 2025 edition for 2023 values (Cato Institute and Fraser Institute, 2024; 2025).	Used as a retrospective domain-specificity check comparing governance and economic sub-indices under a shared source family.

Table 5. Analytic country groups used in the reported comparisons.

Bloc	Countries
US sphere	United States, United Kingdom, Canada, Australia, France, Germany, Japan, South Korea, New Zealand, Netherlands
China sphere	China, Iran, Vietnam, Laos, Cambodia
Russia sphere	Russia, Belarus, Kazakhstan, Kyrgyzstan, Uzbekistan
Non-aligned	India, Brazil, Turkey, Indonesia, South Africa, Mexico

```
estimate.
```

ForecastBench binary prompt.

```
You are an expert superforecaster,
familiar with the work of Tetlock
and others.  The following question
is a binary forecasting question --
it resolves to either Yes (1) or No
(0).

Question:  {question_text}

{resolution_info}

Instructions:
1.  Identify the key factors and
considerations.
2.  Consider base rates and
reference classes.
3.  Think about what evidence would
update you in each direction.
4.  Synthesize into a final
probability estimate.

IMPORTANT: You MUST end your
response with exactly this line:
PROBABILITY: X.XX

where X.XX is a decimal between
0.00 and 1.00.
```

Geopolitical binary prompt.

```
You are a superforecaster assessing
changes in global governance and
development indices.

Question:  {question_text}

{resolution_info}

You MUST end your response with
exactly this format:
PROBABILITY: X.XX

where X.XX is a number between 0.00
and 1.00 representing your estimated
probability.
```

Geopolitical numerical prompt.

```
You are a superforecaster estimating
values for global governance and
development indices.

Question:  {question_text}

{resolution_info}

You MUST end your response with
exactly this format:
ESTIMATE: X.XX

where X.XX is your best numerical
estimate.
```

Table 6. Country-index families retained for the reported analyses.

Indicator	Description	Scale / direction	Role
Freedom House Freedom Score (Freedom House, 2025)	Composite annual measure of political rights and civil liberties.	0–100; higher means more freedom.	Main governance
Corruption Perceptions Index (Transparency International, 2025)	Annual measure of perceived public-sector corruption.	0–100; higher means less corruption.	Main governance
RSF Press Freedom Index (Reporters Without Borders, 2025)	Annual measure of journalism conditions, media independence, and journalist safety.	0–100; higher means more press freedom.	Main governance
HFI Freedom of Expression and Information (Cato Institute and Fraser Institute, 2024; 2025)	Freedom to access, produce, and exchange information.	0–10; higher means more freedom.	HFI governance
HFI Freedom of Assembly and Association (Cato Institute and Fraser Institute, 2024; 2025)	Freedom to organize, associate, and assemble.	0–10; higher means more freedom.	HFI governance
HFI Rule of Law (Cato Institute and Fraser Institute, 2024; 2025)	Legal predictability, due process, and impartial enforcement.	0–10; higher means stronger rule of law.	HFI governance
HFI Freedom to Trade Internationally (Cato Institute and Fraser Institute, 2024; 2025)	Openness to cross-border exchange and trade barriers.	0–10; higher means more openness.	HFI economics
HFI Sound Money (Cato Institute and Fraser Institute, 2024; 2025)	Monetary stability and protection from inflationary erosion.	0–10; higher means more monetary stability.	HFI economics
HFI Size of Government (Cato Institute and Fraser Institute, 2024; 2025)	Economic freedom with respect to taxation, spending, and state ownership.	0–10; higher means smaller government footprint.	HFI economics

Table 7. Closed-book ForecastBench control. Lower Brier is better.

Model	Timing filter	n	Brier
Claude Haiku 4.5	cutoff	208	0.115
Qwen3 32B	release proxy	71	0.156
GPT-OSS 20B	cutoff	237	0.137
GPT-OSS 120B	cutoff	237	0.172

Localized prompt files. The trilingual reruns use localized versions of the geopolitical templates, named `binary_geopolitical_zh.txt`, `binary_geopolitical_ru.txt`, `numerical_geopolitical_zh.txt`, and `numerical_geopolitical_ru.txt`. The question slot is `{question_text}`. The required terminal tags are localized equivalents of `PROBABILITY: X.XX` or `ESTIMATE: X.XX`; the extractor accepts the English tags and the localized Chinese and Russian tags used in those prompt files.

Acquiescence mirror question texts.

```
Positive: Will {country}'s {index}
score improve from its 2024 value of
{value} in the 2025 edition? {scale
statement}

Negative: Will {country}'s {index}
score decline from its 2024 value of
{value} in the 2025 edition? {scale
statement}
```

Both mirror questions use the geopolitical binary prompt wrapper above.

Reasoning-mode judge prompt. The trace analysis uses Claude Haiku 4.5 as a blinded judge at temperature 0.0. The judge sees the question context and reasoning trace, but not the source-model label, and returns only JSON.

```
Classify this governance forecasting
reasoning trace on THREE dimensions
about the REASONING APPROACH used:

1. reasoning_mode: one of
counterfactual, statistical,
narrative, comparative.
2. reform_assessment: one
of yes_unlikely, yes_possible,
not_assessed.
3. anchoring_behavior: one
of strong_anchor, weak_anchor,
no_anchor.

Respond ONLY with JSON:
{"reasoning_mode": "...",
"reform_assessment": "...",
"anchoring_behavior": "..."}
```

G. Framing Values

H. Sample-Fraction Robustness

The main binary–numerical comparison thresholds the median numerical point estimate against the previous score.

Table 8. ForecastBench control slices used in the closed-book evaluation.

Property	Dataset A	Dataset B
Snapshot date	2024-07-21	2025-03-02
Questions	237	208
Resolution window	2024-07-25 to 2026-01-01	2025-03-09 to 2026-01-01
Yes rate	0.186	0.173
Base-rate baseline	0.151	0.143
Main models	GPT-OSS 20B, GPT-OSS 120B	Claude Haiku 4.5, Qwen3 32B

Table 9. Matched US-sphere minus China-sphere governance forecast differences on English country-index questions. Positive values mean higher predicted improvement for US-sphere countries. Brackets give 95% bootstrap CIs over country–index targets.

Model	Binary	Numerical
Claude Haiku 4.5	0.131 [0.071, 0.190]	−0.200 [−0.439, 0.045]
GPT-OSS 20B	0.166 [0.094, 0.237]	−0.233 [−0.458, 0.000]
GPT-OSS 120B	0.177 [0.116, 0.236]	0.100 [−0.155, 0.337]
Qwen3 32B	0.436 [0.357, 0.515]	−0.300 [−0.533, −0.054]

Because this is a directional comparison rather than a probability-equivalent measurement, we also recompute the English governance comparison using the fraction of valid numerical samples that exceed the previous score. This leaves the binary side unchanged and replaces each numerical point-estimate direction with an empirical sample fraction. Thus the binary column repeats Table 9; only the numerical aggregation changes. The qualitative result is preserved: averaged across models, the unchanged binary gap is 0.228, the sample-fraction numerical gap is −0.076, and the binary-minus-numerical swing is 0.303.

Table 10. Sample-fraction robustness check on matched English governance questions. Positive values mean higher predicted improvement for US-sphere countries.

Model	Binary	Num. fraction	Swing
Claude Haiku 4.5	0.131	−0.040	0.171
GPT-OSS 20B	0.166	−0.077	0.243
GPT-OSS 120B	0.177	−0.007	0.183
Qwen3 32B	0.436	−0.180	0.616
Mean	0.228	−0.076	0.303

I. HFI Domain Results

Table 11 reports the model-level values underlying Figure 3. Values are US-sphere minus China-sphere mean predicted improvement under binary prompting. Governance averages the HFI expression, assembly, and rule-of-law sub-indices; economics averages trade, sound money, and size of government. The figure-generation code computes these values by filtering the English HFI rows to binary prompts, averaging predicted improvement by bloc and domain, and subtracting the China-sphere mean from the US-sphere mean.

Table 11. Model-level HFI binary domain values used in Figure 3. Positive values mean higher predicted improvement for US-sphere countries.

Model	Governance	Economics
Claude Haiku 4.5	0.141	0.062
GPT-OSS 120B	0.214	0.063
Qwen3 32B	0.277	0.121
Mean	0.211	0.082

J. Language Decomposition

Table 12 reports the Chinese and Russian within-bloc shifts behind the trilingual governance rerun. Values are mean changes relative to English on matched binary governance questions.

Table 12. Within-bloc mean shifts relative to English.

	Chinese		Russian	
Model	US	China	US	China
Claude Haiku 4.5	−0.106	−0.065	−0.096	−0.071
GPT-OSS 20B	0.007	0.004	0.018	0.008
GPT-OSS 120B	−0.041	0.046	0.003	0.029
Qwen3 32B	−0.240	−0.015	0.027	0.010

K. Mechanism and Acquiescence Checks

Anonymous score-swap controls retain country names while manipulating score anchors. The US-sphere minus China-sphere gap remains 0.260 when all countries are assigned a score of 50 and 0.290 when democratic and authoritarian score profiles are reversed. Retrospective HFI questions about already resolved 2018–2019 changes preserve a directional gap of about 0.240 even though the ground-truth gap is −0.100.

The reasoning-trace judge pass uses 284 usable English governance traces from Claude Haiku 4.5 and Qwen3 32B. Binary traces are judged counterfactual more often than numerical traces (48% versus 7%). On China- and Russia-sphere cases, binary traces almost always assess whether reform is possible (98% versus 71% for numerical traces), while strong anchoring remains common in both framings.

For the acquiescence control, every country–indicator pair has an improve question and a decline question. Because no change is also possible, AS is not a full complementarity score; large positive values would indicate high probability assigned to both directional claims.

L. Consistency-Training Pilot

The pilot intervention uses stock-price forecasting as an out-of-domain calibration task. The core idea is to train a model's binary outputs to match probabilities implied by its own numerical forecasts, then evaluate whether that calibration transfers to geopolitical country-index questions that the model never sees during training.

Training data. The stock universe consists of the top 300 S&P 500 constituents by January 2025 market capitalization together with the CSI 300 constituents. Questions are resolved with Yahoo Finance close prices across six forecasting horizons of 7, 30, 90, 180, 270, and 330 trading days. For each stock–horizon pair, the base model is sampled 20 times at temperature 1.0 on the numerical prompt below. The fraction of valid numerical samples in which the predicted price exceeds the January 1 reference price defines the implied binary target q_{num}. All stocks with at least five valid numerical samples are included without class balancing, yielding 3,344 training examples with mean $q_{\text{num}} = 0.349$ and standard deviation 0.189.

Fine-tuning. Fine-tuning uses Qwen3-4B-Instruct with LoRA rank 16, a constant learning rate of 10^{-4}, batch size 16, one epoch, 209 gradient steps, and maximum sequence length 2,048. The target response is a short answer ending in `PROBABILITY: X.XX`, where `X.XX` is the rounded q_{num}. Prompt tokens receive zero weight; only response tokens are trained. The full pipeline costs under $7 on Tinker (Lab, 2025).

Evaluation. The trained model is evaluated deterministically on 628 country-index questions, with 314 binary and 314 numerical questions drawn from the main geopolitical experiments. These cover both governance and economic sub-indices. The model never sees governance or country-index data during training.

Stock numerical teacher prompt.

```
You are an expert superforecaster,
familiar with the work of Tetlock
and others.  The following question
is a numerical forecasting question.

Question:  What will {ticker}'s
({name}) market close price be on
{resolution_date}?

Company information:
- Ticker:  {ticker}
- Name:  {name}
- Sector:  {sector}
- Industry:  {industry}
- Market:  {market_desc}
- Current Price (Jan 1, 2025):
${freeze_value}
- Resolution Date:  {resolution_date}
- Trading Days Until Resolution:
~{days_ahead}
- Company Background:  {background}

Instructions:
1.  Identify the key factors and
considerations.
2.  Consider base rates and
reference classes.
3.  Think about what evidence would
update you in each direction.
4.  Synthesize into a final price
estimate.

IMPORTANT: You MUST end your
response with exactly this line:
ESTIMATE: $X.XX
where X.XX is your best estimate of
the closing price.
```

Stock binary training prompt.

```
You are an expert superforecaster,
familiar with the work of Tetlock
and others.  The following question
is a binary forecasting question --
it resolves to either Yes (1) or No
(0).

Question:  Will {ticker}'s
({name}) market close price on
{resolution_date} be higher than
${freeze_value}?

Company information:
- Ticker:  {ticker}
- Name:  {name}
- Sector:  {sector}
- Industry:  {industry}
- Market:  {market_desc}
- Current Price (Jan 1, 2025):
${freeze_value}
- Resolution Date:  {resolution_date}
- Trading Days Until Resolution:
~{days_ahead}
- Company Background:  {background}
```

Table 13. Overall improve/decline diagnostic scores. Positive values would indicate high probability assigned to both directional claims. Mean-AS intervals use the pair-level variance over 553 mirror pairs per model.

Model	Pairs	Mean AS [95% CI]	|AS|
GPT-OSS 20B	553	-0.189 [-0.206, -0.173]	0.219
GPT-OSS 120B	553	-0.207 [-0.221, -0.193]	0.230
Claude Haiku 4.5	553	-0.213 [-0.225, -0.201]	0.216
Qwen3 32B	553	-0.262 [-0.283, -0.241]	0.295

Table 14. Consistency-training pilot results on out-of-domain country-index questions.

Metric	Base	Calibrated
Binary spread (US−CN)	-0.106	-0.030
Gov. binary spread (US−CN)	-0.138	-0.033
Econ. binary spread (US−CN)	-0.075	-0.036
Numerical spread (US−CN)	-0.290	-0.075
Contradiction rate	44.3%	22.0%
Binary Brier score	0.295	0.311

```
Instructions:
1.  Identify the key factors and
considerations.
2.  Consider base rates and
reference classes.
3.  Think about what evidence would
update you in each direction.
4.  Synthesize into a final
probability estimate.

IMPORTANT: You MUST end your
response with exactly this line:
PROBABILITY: X.XX
where X.XX is a decimal between 0.00
and 1.00.
```