LANGUAGE SHAPES DECISIONS IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are deployed globally as multilingual decision-makers under the assumption that language is a neutral doorway to the same underlying system. Although leading models often reach similar factual accuracy across languages, it remains unclear whether they make the same choices when the language changes. We tested this question with 52,800 controlled forced-choice trials covering six model families, eleven scenarios, and seven behavioral domains. Each model was compared with itself in matched English and Chinese conditions, and we also asked it to respond in a language different from the language of the scenario. Across 66 comparisons, 37 showed a statistically significant shift after Benjamini–Hochberg correction at $\alpha = 0.05$, or 56.1% of all comparisons. The arcsine-transformed contrast reached $|h| = 3.02$. The direction of change depended on the model and sometimes reversed completely. In a loss-framed risk task, one model moved from complete risk tolerance in English to complete risk aversion in Chinese, while another model showed the opposite pattern. Agent attribution was the only domain with the same directional shift across all six models. When the requested answer language was crossed with the scenario language, mixed-language responses frequently departed from the monolingual envelope by more than $\delta = 0.02$, with marked suppression in some pairs and marked amplification in others. These results show that the distribution of observable decisions can change with language and support evaluating model behavior separately by language before deployment.

1 INTRODUCTION

In global AI deployment, natural language is often treated as an interchangeable interface to an unchanged computational core. If a system solves a task in English and Chinese, practice usually assumes that it will make the same decisions in both languages. For large language models, however, similar factual accuracy and stable behavior are different properties (Jin et al., 2025; AlKhamissi et al., 2024; Naous et al., 2024). A model may answer factual questions equally well in two languages while changing whom it blames for an event, how much financial risk it accepts, or how cooperatively it behaves. Such changes matter when models support negotiation, legal triage, or risk assessment. The practical question is therefore how much the observable choices of one fixed model change when only the language condition changes.

We use established tasks from cross-linguistic cognitive science and behavioral economics as controlled probes of event attribution, temporal orientation, risk framing, and moral judgment (Boroditsky, 2001; Fausey & Boroditsky, 2011; Costa et al., 2014; Hayakawa et al., 2017). Comparing one fixed model in two languages keeps its learned parameters constant while changing the words, script, translation, and other language-specific features of the prompt. These probes let us ask a direct question: **does the model's distribution of choices remain stable when the language is substituted?**

Existing multilingual evaluations have expanded coverage, fairness analysis, and disagreement-aware annotation. Two limitations motivate our study. Aggregate results can hide reversals in the size or direction of a shift, and many cross-lingual comparisons change the input language, translation, and expected answer language at the same time. We therefore keep the model and scenario fixed and compare languages within each matched pair.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Three related literatures stop short of the question we ask. Multilingual capability benchmarks (MMLU-style suites, knowledge probes, and INCLUDE-style regional suites) measure whether a model recalls the same fact under a different script; they do not measure whether a model picks the same forced choice. Cultural-bias and moral-preference studies document systematic disagreement between a model's English answer and its Chinese or Arabic answer on normed vignettes; they typically aggregate across diverse items and rarely check whether the same item produces opposite responses in two languages on the same model. Safety and refusal benchmarks focus on jailbreak resistance and are mostly monolingual in practice. The closest neighbours, multilingual CoT studies and self-translation audits, demonstrate an English-reasoning advantage and an under-exploitation of non-English capability, but they observe reasoning traces, not the final forced choice. We extend this line of work by holding the item and the model constant and treating the prompt language itself as the manipulated variable.

We address this gap by treating each model–scenario pair as a separate comparison. In the matched monolingual conditions, the prompt structure is the same and only English or Chinese is used. In the requested-response conditions, the scenario stays in one language while the model is asked to answer in the other. This allows us to see whether the mixed-language choice follows the input language, falls between the two monolingual choices, or moves outside their range.

Design choices and operational stakes. Three design decisions frame the result. *(i)* The EN/ZH contrast targets the highest-deployment language pair and stresses vocabulary, script, and tokenisation simultaneously; a shift under this contrast therefore has direct operational relevance. *(ii)* Forced-choice probes control for generation length, free-form style, and refusal rate, which are the dominant confounds in open-ended evaluation; the trade-off is that they cannot, by construction, measure open-ended deployment behaviour, and that limitation is taken up in the scope subsection below. *(iii)* Within-model matched pairs fix the model's parameters and isolate the prompt language as the single varying factor, so any observed shift is attributable to language rather than to a model-level confound; the cost is a narrower claim, restricted to what changes when only the language is changed. Operationally, a model that prices risk more aggressively under one language than under another produces systematically different actions across a multilingual deployment, and the present measurements are designed to surface that kind of inconsistency before it reaches users.

Our evaluation spans six leading models, eleven canonical scenarios across seven domains, and four prompt conditions. In total, we ran $N = 52{,}800$ trials. Each response was converted to a score whose meaning was defined by the scenario. We asked three research questions:

- **RQ1, prevalence and size:** How often do choices change with language, and how large are those changes?
- **RQ2, direction across models:** Do models shift in a common direction, or can the direction reverse from one model to another?
- **RQ3, requested answer language:** When the scenario language and requested answer language differ, do choices stay within the monolingual range or move beyond it?

Our main findings are threefold. First, the matched design separates a change in scenario language from a change in requested answer language. Second, language-conditioned shifts are common: 37 of 66 comparisons remain significant after Benjamini–Hochberg correction, and their size reaches $|h| = 3.02$. Risk and social-choice tasks show reversals in direction across models. Third, mixed-language responses often fall outside the monolingual range by more than $\delta = 0.02$, and the departures include both marked suppression and marked amplification that warrant language-specific deployment checks.

2 Related Work

Multilingual alignment, fairness, and pluralistic evaluation. Prior multilingual benchmarks predominantly examine whether model outputs reflect cultural values and human judgments across languages (Jin et al., 2025; AlKhamissi et al., 2024; Naous et al., 2024; Ramezani & Xu, 2023; Fleisig et al., 2023). In parallel, cross-lingual fairness studies reveal systematic disparities across legal regimes, linguistic contexts, and protected attributes (Chalkidis et al., 2022). More recent pluralistic and disagreement-aware evaluations explicitly model human variation rather than reduc-

ing it to a single ground-truth label (Chiu et al., 2025; Xie et al., 2025), while persuasion-oriented benchmarks disentangle judgment elicitation from persuasive generation (Singh et al., 2025). While sharing this high-level motivation toward reliable multilingual deployment, we target a more controlled inquiry: whether an identical model alters its forced-choice decisions solely in response to prompt language shifts.

Cross-lingual reasoning and language consistency. A related strand of literature investigates the language of intermediate reasoning as an explicit variable. Prior work on multilingual chain-of-thought (CoT) demonstrates an English reasoning advantage over native-language reasoning (Shi et al., 2023), whereas self-translation studies suggest that models often under-exploit their multilingual capabilities when prompted in non-English languages (Etxaniz et al., 2024). Meanwhile, inconsistency metrics have been proposed to quantify cross-lingual divergence (Gupta et al., 2025), and complementary research on the "say–do" gap indicates that models' stated values often decouple from their revealed choices in interactive scenarios (Xu et al., 2025). Distinct from works analyzing latent reasoning traces, our study takes a behavioral, descriptive approach by directly evaluating final-choice distributions.

Evaluation frameworks and distributional analysis. Standardized frameworks like HELM systematically benchmark model capabilities across tasks and metrics (Liang et al., 2023), whereas geographically grounded initiatives like INCLUDE mitigate translation artifacts by sourcing natively curated items (Romanou et al., 2025). Concurrently, black-box distributional auditing has emerged as a rigorous lens for measuring behavioral shifts without requiring parameter access (Richter et al., 2025; Xu et al., 2025), demonstrating that complex multilingual safety and value dynamics cannot be collapsed into scalar summaries (Chiu et al., 2025; Xie et al., 2025; Gupta et al., 2025). Building on these insights, we operationalize cross-lingual divergence by characterizing both the effect size and directional bias within matched language pairs.

3 Experimental Design and Analysis

Experimental structure and formalization. Let $m \in \mathcal{M}$ denote a model, $s \in \mathcal{S}$ a scenario, and $c \in \mathcal{C}$ a prompt condition. Each response becomes a semantic score Y. We first estimate the difference between Chinese and English choices, $\Delta_{m,s} = \mathbb{E}[Y \mid m, s, \mathrm{ZH}] - \mathbb{E}[Y \mid m, s, \mathrm{EN}]$, and its effect size h. We then compare the direction of this difference across models. Finally, we compare the two mixed-language means with the two monolingual means and label a mixed-language mean as outside the monolingual range when it lies strictly beyond $M_{\min} - \delta$ or $M_{\max} + \delta$, with $\delta = 0.02$. Replicate order was randomized, failed parses were treated as missing, and up to 200 responses were attempted for each model–scenario–condition combination.

3.1 Endpoints and Prompt Conditions

We evaluated six model families: Claude Opus 4.6, Gemini 3.1 Pro, GPT-5.4, GLM-5.1, Kimi-K2.5, and Qwen3.6-plus. Queries ran from April 11 to 23, 2026 UTC through the providers' official chat-completion services. Sampling temperature was fixed at $T = 1.0$ and the generation limit at $4{,}096$ tokens. System prompts, task instructions, and answer choices were kept the same across conditions. The order of the choices was rotated independently on each trial according to the rule described below.

The design crossed six models with eleven scenarios and four language conditions, using 200 stochastic replicates for each combination. The EN and ZH conditions are monolingual baselines: the scenario and the requested answer use the same language. In EN→ZH and ZH→EN, the scenario is written in one language and the model is explicitly asked to answer in the other. Every condition uses the same prompt structure. Only the language of the scenario, its accompanying text, and the requested answer changes; choice order continues to rotate on every trial.

3.2 Scenarios and Operationalised Scores

The eleven scenarios span seven decision domains and draw on established paradigms (Fausey & Boroditsky, 2011; Clark, 1973; Yu, 1998; Kahneman & Tversky, 1979; Tversky & Kahneman, 1981;

Berg et al., 1995; Fehr & Gächter, 2000; Nagel, 1979; Bloom, 1981). Seven scenarios offer two choices and four offer three. Table 1 lists the choices and explains what a high score means in each scenario. Bilingual researchers drafted the Chinese versions and checked them through independent back-translation into English.

Table 1: Scenario coverage and score meanings. Response options are ordered from the low-score pole to the high-score pole (K: number of response choices). All statistical comparisons are performed strictly within matched scenarios.

Domain	Scenario	K	Response options (low → high)	High score indicates
Event attribution	Causal Attribution	3	accident < both equally < agent	Agentive blame
Temporal mapping	Time Metaphor	2	Friday < Monday	"Moved forward" = earlier day
	Vertical Time	2	below < above	Earlier event placed above
Risk choice	Gain-Framed Risk	2	certain gain < 50/50 gain gamble	Risk-seeking (gain frame)
	Loss-Framed Risk	2	certain loss < 50/50 loss gamble	Risk-seeking (loss frame)
Framing effect	Gain-Framed Disease	2	save 300 for sure < 1/3 chance to save 900	Risk-seeking (gain frame)
	Loss-Framed Disease	2	lose 600 for sure < 1/3 chance to lose 0	Risk-seeking (loss frame)
Social choice	Trust Game	3	\$0 < \$5 < \$10 transferred	Trust (amount sent)
	Public Goods	3	\$0 < \$10 < \$20 contributed	Cooperation (contribution)
Moral judgment	Moral Luck	3	identical penalty < added requirement < harsher penalty	Outcome-based blame
Counterfactual	Counterfactual	2	declines to speculate < speculates	Speculative engagement

To avoid positional bias, the order of the choices was rotated on each trial with a reproducible hash rule, $\texttt{MD5}(m, s, c, \text{trial}) \bmod K$. The response was then converted to a semantic score. Two-choice tasks use scores of 0 and 1, while three-choice tasks use 0, 0.5, and 1. A high score means risk-seeking in the disease tasks, a larger contribution in the public-goods task, or more agentive attribution in the event task. The meaning of the score stays fixed within each scenario, and scores are never averaged across different domains.

The score mapping is local to each scenario. A high score can mean greater risk-taking, more trust or cooperation, more outcome-based blame, or greater willingness to engage with a counterfactual. These directions describe the response options; they do not label any behavior as desirable, and they should not be averaged across domains. The shared numerical scale is used only to compare English and Chinese within the same scenario.

3.3 Response Parsing, Exclusion, and Missing-Data Handling

Model responses were read with a fixed rule that searched the final $2{,}048$ tokens for a choice letter after the expected delimiter. Responses with no unambiguous choice were excluded rather than filled in. We parsed 52,613 of the 52,800 trials, a rate of 99.65%. Table 2 gives the rate for each language condition. All analyses use the number of responses actually observed in each comparison.

Table 2: Parseability and trial retention across language conditions ($N = 52{,}800$).

Condition	Parsed	Unparsed	Parse Rate
EN	13,171	29	99.78%
ZH	13,182	18	99.86%
EN→ZH	13,096	104	99.21%
ZH→EN	13,164	36	99.73%
Total	**52,613**	**187**	**99.65%**

Parse confidence and denominator balance. The main marker rule identified 52,090 choices, and a nearest-letter fallback identified 517 more. Together these routes supplied 52,607 records with the metadata needed for the position check. Six additional responses contained a clear choice but lacked that metadata, giving 52,613 parsed choices in total and 187 failures. We retained the fallback records because their final choice was still unique. Missing responses made the English and Chinese sample sizes differ in 14 of the 66 comparisons; the smallest condition still retained 195 of 200 trials. The reported contrasts therefore use the observed sample sizes.

3.4 Inferential Tests, Effect Sizes, and Multiplicity Control

The primary estimand is the conditional language contrast:

$$\Delta_{m,s} = \mathbb{E}[Y \mid m, s, \text{ZH}] - \mathbb{E}[Y \mid m, s, \text{EN}]. \tag{1}$$

For two-choice scenarios, the contrast is a difference in the probability of choosing the high-score option and is tested with a two-sided Fisher exact test. For three-choice scenarios, it is a difference in ordered means and is tested with a two-sided Mann–Whitney U test. We also report the arcsine-transformed mean contrast $h = 2\arcsin(\sqrt{M_{\text{ZH}}}) - 2\arcsin(\sqrt{M_{\text{EN}}})$. Extreme means are limited to the interval $[0.001, 0.999]$ before transformation. For two-choice tasks this is Cohen's h; for three-choice tasks it is a descriptive contrast on the same transformed scale. We obtained 95% bootstrap confidence intervals from $10{,}000$ percentile resamples. Appendix A reports all cell-level results and sensitivity analyses.

We control false discoveries across the 66 comparisons with the Benjamini–Hochberg procedure at $\alpha = 0.05$. As a planned sensitivity check, we also report the same procedure after removing cells in which both monolingual means lie within 0.03 of the scale boundaries. Near these boundaries, a real change can be compressed by the score scale. Table 4 shows that the number of large contrasts is unchanged across boundary thresholds from 0.90 to 0.99, and reports the BH-significance rate after the same correction is applied independently within each retained family.

Position verification and effective denominators. We also repeated the analysis on the subset for which the recorded choice could be checked against the planned position rotation. Thirty-three of the 66 comparisons remained significant after the same correction. The reconstructed position agreed with the displayed letter in 23,269 of 52,607 records with complete metadata, or 44.2%. This agreement rate is not a direct estimate of rotation error because an explanation can contain several letters while ending with the correct choice. The expected random-agreement rate under the observed scenario mix (seven two-choice and four three-choice) is approximately $(7/11) \cdot \frac{1}{2} + (4/11) \cdot \frac{1}{3} = 43.9\%$, so the observed agreement sits just above the random floor. The restricted analysis is therefore a sensitivity check. The planned rotations themselves were close to balanced, with Cramér's $V = 0.0117$ for two-choice scenarios and $V = 0.0297$ for three-choice scenarios.

4 Results

The results are organized around the three research questions. We first quantify the prevalence and magnitude of language-conditioned shifts in the 66 matched English–Chinese model–scenario comparisons. We then examine whether these shifts have a common direction across models or instead reverse across endpoints. Finally, we test whether changing the requested response language pushes choices outside the range defined by the two monolingual conditions. Across these analyses, language-conditioned behavior is common but heterogeneous: effects vary substantially in size and direction, and mixed-language conditions sometimes extend beyond the monolingual envelope. Boundary, position, and parsing checks assess whether these patterns are robust to the main design and measurement sensitivities.

4.1 Prevalence and Magnitude of Language-Conditioned Shifts

Across the 66 matched model–scenario comparisons, 37 remain significant after Benjamini–Hochberg correction at $\alpha = 0.05$. Thus 56.1% of the comparisons show a language-conditioned shift. The boundary-screened analysis removes the 20 comparisons in which both monolingual means are close to 0 or 1. Among the remaining 46 comparisons, 37 retain the same significance status under the original 66-family correction, giving 80.4% significance among informative comparisons. The contrasts vary widely and reach $|h| = 3.02$ in the risk-framing tasks. Of the 46 comparisons away from the boundaries, 15 are large, 9 medium, 15 small, and 7 near zero. The other 20 comparisons are compressed by the score boundaries. Table 4 gives the results across thresholds, and Appendix A.4 gives the position-checked analysis. These results answer RQ1: choices often change with language, and many changes are large. Figure 1 summarizes the corresponding contrasts across models and scenarios.

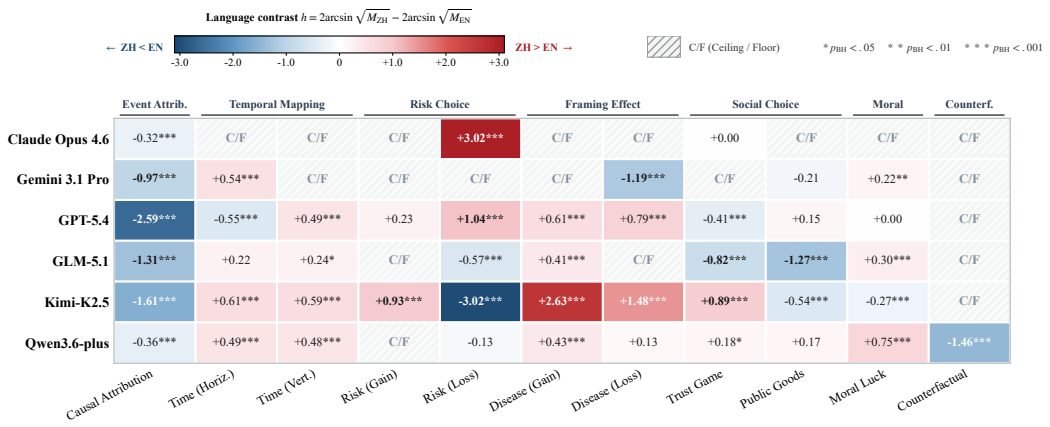


Figure 1: Comprehensive heatmap of language-conditioned prompt sensitivity (h) across all six model families and eleven canonical scenarios. Red hues denote higher scores under Chinese prompts ($h > 0$); blue hues denote higher scores under English prompts ($h < 0$). Asterisks indicate Benjamini–Hochberg FDR significance ($* p_{\mathrm{BH}} < .05$, $* * p_{\mathrm{BH}} < .01$, $* * * p_{\mathrm{BH}} < .001$). Hatched cells indicate ceiling/floor (C/F) baselines where either both monolingual means exceed 0.97 or both fall below 0.03.

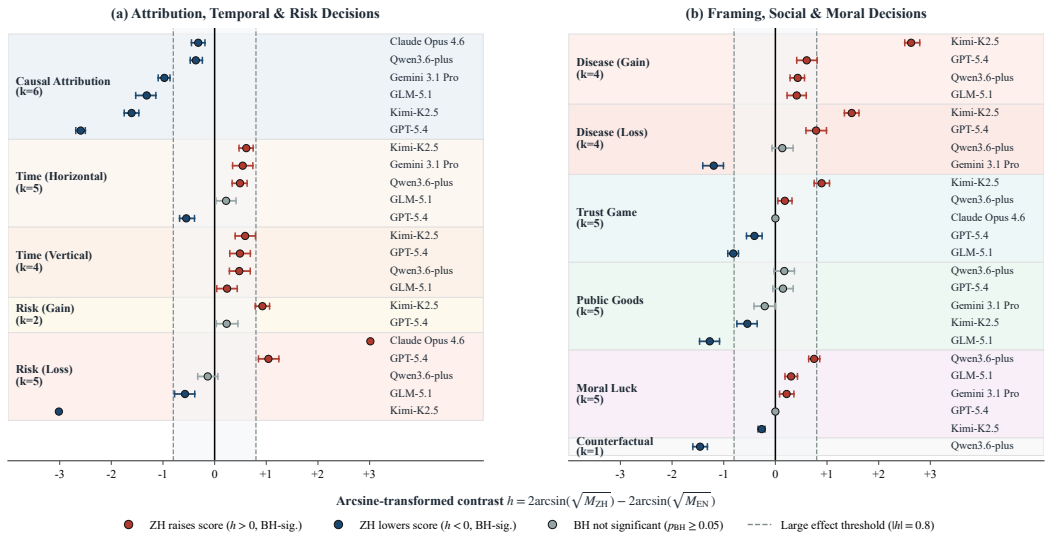


Figure 2: Dual-panel forest plot of informative EN–ZH language contrasts across all decision domains. Panel (a) covers event attribution, temporal mapping, and risk choices; Panel (b) covers disease framing, social cooperation, moral luck, and counterfactual choice. Error bars depict 95% bootstrap confidence intervals (10,000 resamples). Dashed lines denote the large effect size threshold ($|h| = 0.8$).

4.2 Endpoint-Level Heterogeneity and Polarity Reversals

The shifts do not follow one common cross-model pattern. Their directions differ across models and sometimes reverse completely. In the loss-framed risk scenario, Kimi-K2.5 chose the risky option in all parsed English trials, with $M_{\mathrm{EN}} = 1.00$, and the safe option in all parsed Chinese trials, with $M_{\mathrm{ZH}} = 0.00$. This produced $h = -3.02$ with $p_{\mathrm{BH}} < .001$. Claude Opus 4.6 showed the opposite pattern: it chose the safe option in English, with $M_{\mathrm{EN}} = 0.00$, and the risky option in Chinese, with $M_{\mathrm{ZH}} = 1.00$ and $h = +3.02$. In the gain-framed disease task, Kimi-K2.5 again moved strongly toward risk-seeking in Chinese, with $|h| = 2.63$, whereas other models were neutral or moved in the opposite direction. Figure 2 displays the endpoint-level contrasts.

The social-choice tasks show the same heterogeneity. Chinese prompts increased trust in Kimi-K2.5, with $h = +0.89$ and $p_{\mathrm{BH}} < .001$, but reduced it in GLM-5.1, with $h = -0.82$ and $p_{\mathrm{BH}} < .001$. Cooperation fell in both models' Public Goods conditions. These opposing effects, together with the risk reversals, do not support a single cross-model language prior.

4.3 Directional Convergence in Agent Attribution

Agent attribution was the one domain in which all six models moved in the same direction. The score was lower under Chinese prompts than under English prompts for every model, with transformed contrasts ranging from -2.59 to -0.32. The size still depended on the model. GPT-5.4 showed the largest change, with $h = -2.59$ and $p_{\mathrm{BH}} < .001$, whereas Claude Opus 4.6 showed a smaller change, with $h = -0.32$ and $p_{\mathrm{BH}} < .01$.

This common direction is a property of the prompts we tested. It does not show that the models share an internal representation or that grammar alone causes the shift. Wording, tokenization, translation, pretraining, later training, and provider-specific details changed together, so the data motivate targeted follow-up experiments rather than a claim about mechanism.

4.4 Outcomes Beyond Margin-Extended Monolingual Bounds

For RQ3, we examined the 29 model–scenario pairs with a substantial English–Chinese difference, defined as $|\Delta| > 0.10$. We compared each mixed-language mean with the two monolingual means $M_{\min} = \min(M_{\mathrm{EN}}, M_{\mathrm{ZH}})$ and $M_{\max} = \max(M_{\mathrm{EN}}, M_{\mathrm{ZH}})$. A mixed-language mean is reported as *below* the monolingual range if it is strictly less than $M_{\min} - \delta$, and as *above* if it is strictly greater than $M_{\max} + \delta$, with $\delta = 0.02$. A pair is labelled below the range if either mixed-language mean falls below $M_{\min} - \delta$, and above the range if either mixed-language mean rises above $M_{\max} + \delta$. The full means appear in Appendix A.2, Table S2; the threshold rule appears in Appendix A.3. Figure 3 shows representative outcome classes.

Table 3 shows the four means for three illustrative scenarios: causal attribution, loss-framed risk, and public-goods cooperation. It shows how the same requested-answer instruction can preserve, weaken, or reverse the English–Chinese difference depending on the model.

Table 3: Representative four-condition means across all six models for three illustrative scenarios. Bolded mixed-language values lie strictly below $M_{\min} - \delta$ or strictly above $M_{\max} + \delta$ with $\delta = 0.02$.

Endpoint	Scenario	EN	ZH	EN→ZH	ZH→EN
Claude Opus 4.6	Causal Attribution	0.94	0.84	**0.72**	**0.51**
	Loss-Framed Risk	0.00	1.00	0.41	0.95
	Public Goods	1.00	1.00	1.00	1.00
Gemini 3.1 Pro	Causal Attribution	0.96	0.61	0.72	**0.55**
	Loss-Framed Risk	1.00	1.00	0.97	1.00
	Public Goods	0.98	0.95	**0.26**	**0.10**
GPT-5.4	Causal Attribution	1.00	0.06	0.96	0.19
	Loss-Framed Risk	0.36	0.84	0.55	**0.29**
	Public Goods	0.71	0.78	**0.03**	**0.00**
GLM-5.1	Causal Attribution	0.99	0.53	0.90	**0.42**
	Loss-Framed Risk	0.25	0.06	0.17	0.08
	Public Goods	0.72	0.14	0.12	**0.00**
Kimi-K2.5	Causal Attribution	1.00	0.45	0.96	**0.42**
	Loss-Framed Risk	1.00	0.00	0.13	0.01
	Public Goods	0.73	0.47	**0.08**	**0.03**
Qwen3.6-plus	Causal Attribution	1.00	0.95	0.99	0.94
	Loss-Framed Risk	0.42	0.35	**0.15**	**0.18**
	Public Goods	0.31	0.40	**0.02**	**0.02**

The monolingual envelope does not always contain the mixed-language responses. The strongest suppression appears when a requested-language instruction pushes the score below both monolingual baselines. In GLM-5.1 Public Goods, the monolingual means are 0.725 and 0.140, whereas EN→ZH and ZH→EN produce 0.120 and 0.000; the EN→ZH value equals $M_{\min} - \delta$ exactly and is treated as on the boundary, while ZH→EN lies strictly below. The result is compatible with an effect of the requested language, but this forced-choice design cannot identify the particular process behind it.

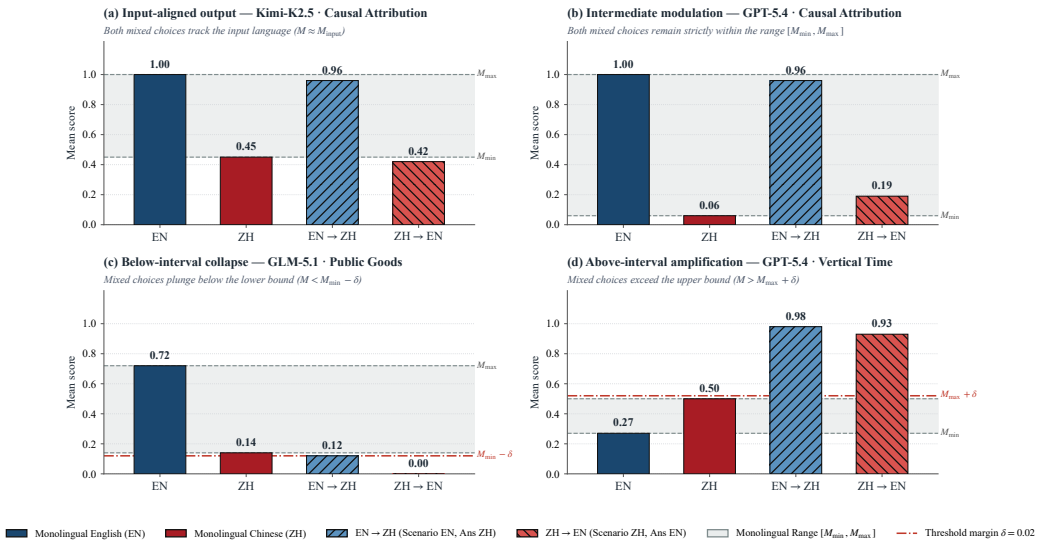


Figure 3: Representative outcome classes under requested-response instructions across the 29 informative pairs ($|\Delta| > 0.10$). Shaded gray envelopes indicate the monolingual baseline range $[M_{\min}, M_{\max}]$, bounded by dashed lines. Red dot-dashed lines mark the margin-extended thresholds ($M_{\min} - \delta$ or $M_{\max} + \delta$, with $\delta = 0.02$). Panels illustrate: (a) input-tracking, (b) intermediate modulation within the envelope, (c) below-interval collapse, and (d) above-interval amplification.

In the other direction, marked amplification appears in GPT-5.4 Vertical Time, where the monolingual means of 0.27 and 0.50 rise to 0.98 and 0.93 under the two mixed-language instructions, both strictly above $M_{\max} + \delta$. The design cannot tell whether this reflects attention, the answer format, or another implementation detail. The remaining informative pairs show mixed-language responses that lie inside the monolingual envelope to within δ, sometimes closer to the input-language mean and sometimes between the two monolingual baselines. Thus mixed-language outputs are not always contained within the monolingual range, but these data do not establish a universal nonlinear mechanism.

4.5 Sensitivity, Thresholds, and Robustness Analyses

The sensitivity checks preserve the main pattern. The number of large h contrasts remains 15 when the ceiling and floor thresholds range from 0.90 to 0.99, and 33 of 66 comparisons remain significant in the strict position-checked subset. Choice rotations were close to balanced, and every language condition had a parse rate above 99.2%. Details are in Appendix A.4 and Table 4.

The requested-response departures are also stable over the prespecified margin range. Both suppression and amplification appear consistently for $\delta \in [0.02, 0.03]$; widening δ to at least 0.05 reclassifies more pairs as lying inside the envelope, while tightening δ to at most 0.01 surfaces one additional pair in each tail. The qualitative conclusion is therefore not an artifact of one rounded threshold, although exact per-pair classifications depend on display precision.

5 Discussion and Implications

5.1 Mechanism and Interpretation

The experiment establishes differences in conditional outputs, not their internal cause. Script, tokenization, translation, pretraining, later training, and provider-specific implementation changed together, and we had neither access to model weights nor a controlled intervention. The measured shift is therefore a change in the distribution of one model's choices. It is not evidence about human cultural preferences or about a model's internal mental representation. Distinguishing the possible causes will require controlled interventions in models whose parameters can be inspected. The spe-

Table 4: Robustness of findings to ceiling and floor screening thresholds. Large contrasts are defined by $|h| > 0.8$; their count remains invariant ($n = 15$) across all thresholds. The BH-significant column reports the Benjamini–Hochberg correction at $\alpha = 0.05$ applied independently within each retained informative family, so the column tracks how many comparisons remain significant after the smaller family is re-corrected; the primary 56.1% headline (37 of 66) comes from the single correction on the full 66-cell family.

Threshold rule	C/F cells	Informative cells	BH-significant	% significant
0.90 / 0.10	30	36	31	86.1%
0.93 / 0.07	26	40	35	87.5%
0.95 / 0.05	22	44	38	86.4%
0.97 / 0.03 (Primary benchmark)	**20**	**46**	**39**	**84.8%**
0.99 / 0.01	18	48	39	81.2%
No exclusion (full 66-cell family)	0	66	37	56.1%

cific causal claims ruled out by the present matched design, and the residual co-variates that require separate experiments, are itemised in Appendix A.7.

5.2 Real-World Deployment Stakes: Implications for Evaluation

The finding that 37 of 66 comparisons shift after correction supports testing models separately by language before using them in consequential settings. These forced-choice probes are diagnostic; they do not estimate error rates in credit, clinical, legal, or commercial systems. Realistic, domain-specific tests and human oversight remain necessary. A practical evaluation should compare matched prompts in each language, include conditions that separate scenario language from answer language, and send unstable cases for bilingual checking or human review. The thresholds used here are analysis criteria, not validated safety limits.

5.3 Interpreting Requested-Response Outcomes

Requested-response conditions hold the scenario language fixed while changing the requested answer language, so any departure from the monolingual envelope carries information that a pure within-language comparison cannot. The same instruction can suppress one model's cooperation (GLM-5.1 Public Goods) and amplify another model's temporal judgment (GPT-5.4 Vertical Time), illustrating why a single average "language effect" is insufficient. Because model, scenario, and requested language remain linked in these cells, the patterns motivate deployment-side language-specific checks rather than claims about a universal processing mechanism.

5.4 What the Sensitivity Analyses Establish

The sensitivity analyses support the qualitative conclusion. The planned rotation was close to balanced (Cramér's $V \leq 0.03$), the boundary and margin sweeps preserve both the suppression and amplification tails, and the per-pair classifications remain stable across $\delta \in [0.02, 0.03]$. Quantitative counts would shift if model versions or templates change; we report exact numbers as descriptive, not as bounds.

5.5 Scope Limitations and Epistemic Boundaries

Three limits constrain interpretation: only English and Chinese, only forced-choice templates, and only six commercial models observed during April 11–23, 2026 UTC, so generality across other languages, dialogue formats, or future model versions is not established; option-position verification is incomplete, although 33 of 66 comparisons remain significant in the strict verified subset; and the selected scenarios and score mappings represent only a limited set of behaviors, so other task formulations may produce different estimates. Within these limits, language-conditioned shifts are common, model-specific, and often push at least one mixed-language mean beyond the monolingual range, so multilingual deployment requires language-specific behavioral evaluation, mixed-language stress tests, and review of unstable cases. Follow-up extensions and the items this study does and does not claim are itemised in Appendices A.5 and A.6.

Ethics Statement

This study evaluated publicly available commercial and open-access language models. No human participants, private personal data, or vulnerable populations were involved, and no human-subjects review was required. Canonical psychological paradigms served only as templates for constructing forced-choice prompts; nothing in the design implies that the tested models possess mental states, cultural identities, or moral values. Reported quantities are aggregate choice distributions, not statements about any language community. To reduce dual-use concerns, we report aggregate distributions and standardized semantic scores rather than instructions for manipulating individual scenarios.

AI Use Statement

Generative AI tools were used only for grammar proofreading and LaTeX typesetting assistance. The authors independently conceived and verified the study, data collection, analysis, and interpretation.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. –, 2024. URL https://aclanthology.org/2024.acl-long.671/.

Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142, 1995.

Alfred H. Bloom. *The Linguistic Shaping of Thought: A Study in the Impact of Language on Thinking in China and the West*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.

Lera Boroditsky. Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43(1):1–22, 2001.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. –, 2022. URL https://aclanthology.org/2022.acl-long.301/.

Yu Ying Chiu, Liwei Jiang, and Yejin Choi. DailyDilemmas: Revealing value preferences of LLMs with quandaries of daily life. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PGhiPGBf47.

Herbert H. Clark. Space, time, semantics, and the child. In Timothy E. Moore (ed.), *Cognitive Development and the Acquisition of Language*, pp. 27–63. Academic Press, 1973.

Albert Costa, Alice Foucart, Sayuri Hayakawa, Melina Aparici, Jose Apesteguia, Joy Heafner, and Boaz Keysar. Your morals depend on language. *PLoS ONE*, 9(4):e94842, 2014.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 550–564, 2024. URL https://aclanthology.org/2024.naacl-short.46/.

Caitlin M. Fausey and Lera Boroditsky. Who dunnit? cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review*, 18(1):150–157, 2011.

Ernst Fehr and Simon Gächter. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994, 2000.

Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. Fairprism: Evaluating fairness-related harms in text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. –, 2023. URL https://aclanthology.org/2023.acl-long.343/.

Ashim Gupta, Maitrey Mehta, Zhichao Xu, and Vivek Srikumar. Found in translation: Measuring multilingual LLM consistency as simple as translate then evaluate. In *Proceedings of the 14th*

International Joint Conference on Natural Language Processing, pp. 3477–3496, 2025. URL https://aclanthology.org/2025.ijcnlp-long.185/.

Sayuri Hayakawa, David Tannenbaum, Albert Costa, Joanna D. Corey, and Boaz Keysar. Thinking more or feeling less? explaining the Foreign-Language Effect on moral judgment. *Psychological Science*, 28(10):1387–1397, 2017.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez Adauto, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. Language model alignment in multilingual trolley problems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VEqPDZIDAh.

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

Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2211.09110.

Thomas Nagel. *Mortal Questions*. Cambridge University Press, Cambridge, 1979.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 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, 2024. URL https://aclanthology.org/2024.acl-long.862/.

Aida Ramezani and Yang Xu. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. –, 2023. URL https://aclanthology.org/2023.acl-long.26/.

Leo Richter, Xuanli He, Pasquale Minervini, and Matt J. Kusner. An auditing test to detect behavioral shift in language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=h0jdAboh0o.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, et al. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k3gCieTXeY.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03057.

Somesh Kumar Singh, Yaman Kumar Singla, Harini S. I, and Balaji Krishnamurthy. Measuring and improving persuasiveness of large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NfCEVihkdC.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YfKNaRktan.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Jia Zheng, Weixiang Zhou, Le Sun, and Yingfei Sun. Large language models often say one thing and do another. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.07003.

Ning Yu. *The Contemporary Theory of Metaphor: A Perspective from Chinese*. John Benjamins Publishing Company, 1998.

A Supplementary Results and Sensitivity Analyses

This appendix provides the numerical and sensitivity results for the three research questions. Section A.1 lists all 66 matched monolingual model–scenario comparisons used for RQ1 and RQ2. Section A.2 lists the four means for the 29 informative pairs used for RQ3. Sections A.3 and A.4 give the requested-response rule and the option-position analysis. Scenario definitions, parseability, representative examples, and ceiling/floor sensitivity remain in the main text.

A.1 Complete Matched EN–ZH Results for RQ1 and RQ2

Rows are ordered by model and scenario so that Table S1 can be read with Figures 1 and 2. Means and contrasts are shown to two decimals, while statistical tests and the $|\Delta| > 0.10$ selection use the full-precision values.

Table S1: Complete matched EN–ZH results for all 66 endpoint–scenario cells. Means and contrasts are rounded for display; inferential quantities use full-precision values, and p_{BH} is corrected across the full 66-cell family. C/F marks cells in which either both means exceed 0.97 or both fall below 0.03.

Endpoint	Scenario	M_{EN}	M_{ZH}	Δ	h	p_{BH}	C/F
Claude Opus 4.6	Causal Attribution	0.94	0.84	−0.10	−0.32	< .001	
Claude Opus 4.6	Time Metaphor	1.00	1.00	+0.00	+0.00	1.000	C
Claude Opus 4.6	Vertical Time	0.00	0.00	+0.00	+0.00	1.000	F
Claude Opus 4.6	Gain Risk	0.00	0.00	+0.00	+0.00	1.000	F
Claude Opus 4.6	Loss Risk	0.00	1.00	+1.00	+3.02	< .001	
Claude Opus 4.6	Gain Disease	0.00	0.00	+0.00	+0.00	1.000	F
Claude Opus 4.6	Loss Disease	1.00	1.00	+0.00	+0.00	1.000	C
Claude Opus 4.6	Trust Game	0.50	0.50	+0.00	+0.00	1.000	
Claude Opus 4.6	Public Goods	1.00	1.00	+0.00	+0.00	1.000	C
Claude Opus 4.6	Moral Luck	0.00	0.00	+0.00	+0.00	1.000	F
Claude Opus 4.6	Counterfactual	1.00	1.00	+0.00	+0.00	1.000	C
Gemini 3.1 Pro	Causal Attribution	0.96	0.61	−0.35	−0.97	< .001	
Gemini 3.1 Pro	Time Metaphor	0.65	0.88	+0.22	+0.54	< .001	
Gemini 3.1 Pro	Vertical Time	1.00	1.00	+0.00	+0.00	1.000	C
Gemini 3.1 Pro	Gain Risk	0.00	0.00	+0.00	+0.00	1.000	F
Gemini 3.1 Pro	Loss Risk	1.00	1.00	+0.00	+0.00	1.000	C
Gemini 3.1 Pro	Gain Disease	0.00	0.00	+0.00	+0.00	1.000	F
Gemini 3.1 Pro	Loss Disease	0.95	0.47	−0.48	−1.19	< .001	
Gemini 3.1 Pro	Trust Game	0.01	0.00	−0.01	−0.08	1.000	F
Gemini 3.1 Pro	Public Goods	0.98	0.95	−0.04	−0.21	0.081	
Gemini 3.1 Pro	Moral Luck	0.03	0.07	+0.05	+0.22	0.004	
Gemini 3.1 Pro	Counterfactual	1.00	0.99	−0.01	−0.14	1.000	C
GPT-5.4	Causal Attribution	1.00	0.06	−0.94	−2.59	< .001	
GPT-5.4	Time Metaphor	1.00	0.91	−0.09	−0.55	< .001	
GPT-5.4	Vertical Time	0.27	0.50	+0.24	+0.49	< .001	
GPT-5.4	Gain Risk	0.02	0.06	+0.04	+0.23	0.053	
GPT-5.4	Loss Risk	0.36	0.84	+0.48	+1.04	< .001	
GPT-5.4	Gain Disease	0.05	0.24	+0.20	+0.61	< .001	
GPT-5.4	Loss Disease	0.15	0.50	+0.35	+0.79	< .001	
GPT-5.4	Trust Game	0.17	0.05	−0.12	−0.41	< .001	
GPT-5.4	Public Goods	0.71	0.78	+0.06	+0.15	0.262	
GPT-5.4	Moral Luck	0.50	0.50	+0.00	+0.00	1.000	
GPT-5.4	Counterfactual	1.00	1.00	+0.00	+0.00	1.000	C
GLM-5.1	Causal Attribution	0.99	0.53	−0.46	−1.31	< .001	
GLM-5.1	Time Metaphor	0.34	0.45	+0.11	+0.22	0.053	
GLM-5.1	Vertical Time	0.81	0.89	+0.08	+0.24	0.031	
GLM-5.1	Gain Risk	0.00	0.00	+0.00	+0.00	1.000	F
GLM-5.1	Loss Risk	0.25	0.06	−0.20	−0.57	< .001	
GLM-5.1	Gain Disease	0.01	0.07	+0.07	+0.41	< .001	
GLM-5.1	Loss Disease	0.00	0.01	+0.01	+0.18	1.000	F
GLM-5.1	Trust Game	0.45	0.10	−0.35	−0.82	< .001	
GLM-5.1	Public Goods	0.72	0.14	−0.58	−1.27	< .001	
GLM-5.1	Moral Luck	0.14	0.26	+0.12	+0.30	< .001	
GLM-5.1	Counterfactual	1.00	1.00	+0.00	+0.00	1.000	C
Kimi-K2.5	Causal Attribution	1.00	0.45	−0.55	−1.61	< .001	
Kimi-K2.5	Time Metaphor	0.89	1.00	+0.11	+0.61	< .001	
Kimi-K2.5	Vertical Time	0.62	0.87	+0.25	+0.59	< .001	
Kimi-K2.5	Gain Risk	0.00	0.23	+0.23	+0.93	< .001	
Kimi-K2.5	Loss Risk	1.00	0.00	−1.00	−3.02	< .001	
Kimi-K2.5	Gain Disease	0.00	0.95	+0.95	+2.63	< .001	
Kimi-K2.5	Loss Disease	0.00	0.48	+0.48	+1.48	< .001	
Kimi-K2.5	Trust Game	0.65	0.96	+0.32	+0.89	< .001	
Kimi-K2.5	Public Goods	0.73	0.47	−0.26	−0.54	< .001	

Table S1 continued

Endpoint	Scenario	M_{EN}	M_{ZH}	Δ	h	p_{BH}	C/F
Kimi-K2.5	Moral Luck	0.56	0.43	−0.13	−0.27	< .001	
Kimi-K2.5	Counterfactual	1.00	1.00	+0.00	+0.00	1.000	C
Qwen3.6-plus	Causal Attribution	1.00	0.95	−0.05	−0.36	< .001	
Qwen3.6-plus	Time Metaphor	0.93	1.00	+0.07	+0.49	< .001	
Qwen3.6-plus	Vertical Time	0.85	0.97	+0.12	+0.48	< .001	
Qwen3.6-plus	Gain Risk	0.00	0.00	+0.00	+0.00	1.000	F
Qwen3.6-plus	Loss Risk	0.42	0.35	−0.07	−0.13	0.335	
Qwen3.6-plus	Gain Disease	0.00	0.06	+0.06	+0.43	< .001	
Qwen3.6-plus	Loss Disease	0.03	0.05	+0.03	+0.13	0.439	
Qwen3.6-plus	Trust Game	0.02	0.06	+0.04	+0.18	0.018	
Qwen3.6-plus	Public Goods	0.31	0.40	+0.08	+0.17	0.141	
Qwen3.6-plus	Moral Luck	0.10	0.42	+0.32	+0.75	< .001	
Qwen3.6-plus	Counterfactual	1.00	0.53	−0.47	−1.46	< .001	

A.2 Requested-Response Means for RQ3

The 29 rows below are the model–scenario pairs selected by the prespecified full-precision criterion $|\Delta| > 0.10$. The two rightmost columns give the means when the requested answer language differs from the scenario language. Arrows indicate the requested visible answer language. These values are descriptive inputs to the classification in Section A.3, not additional statistical tests.

Table S2: Four-condition means for the 29 endpoint–scenario pairs selected by the full-precision criterion $|\Delta| > 0.10$. Arrows denote the requested visible response language. Means are rounded for display; classification uses full-precision values.

Endpoint	Scenario	M_{EN}	M_{ZH}	$M_{\mathrm{EN}\to\mathrm{ZH}}$	$M_{\mathrm{ZH}\to\mathrm{EN}}$
Claude Opus 4.6	Loss Risk	0.00	1.00	0.41	0.95
Gemini 3.1 Pro	Causal Attribution	0.96	0.61	0.72	0.55
Gemini 3.1 Pro	Time Metaphor	0.65	0.88	0.99	1.00
Gemini 3.1 Pro	Loss Disease	0.95	0.47	0.37	0.30
GPT-5.4	Causal Attribution	1.00	0.06	0.96	0.19
GPT-5.4	Vertical Time	0.27	0.50	0.98	0.93
GPT-5.4	Loss Risk	0.36	0.84	0.55	0.29
GPT-5.4	Gain Disease	0.05	0.24	0.00	0.01
GPT-5.4	Loss Disease	0.15	0.50	0.01	0.07
GPT-5.4	Trust Game	0.17	0.05	0.00	0.00
GLM-5.1	Causal Attribution	0.99	0.53	0.90	0.42
GLM-5.1	Time Metaphor	0.34	0.45	0.90	0.92
GLM-5.1	Loss Risk	0.25	0.06	0.17	0.08
GLM-5.1	Trust Game	0.45	0.10	0.07	0.03
GLM-5.1	Public Goods	0.72	0.14	0.12	0.00
GLM-5.1	Moral Luck	0.14	0.26	0.34	0.37
Kimi-K2.5	Causal Attribution	1.00	0.45	0.96	0.42
Kimi-K2.5	Time Metaphor	0.89	1.00	0.83	0.98
Kimi-K2.5	Vertical Time	0.62	0.87	0.97	0.72
Kimi-K2.5	Gain Risk	0.00	0.23	0.00	0.03
Kimi-K2.5	Loss Risk	1.00	0.00	0.13	0.01
Kimi-K2.5	Gain Disease	0.00	0.95	0.05	0.01
Kimi-K2.5	Loss Disease	0.00	0.48	0.00	0.00
Kimi-K2.5	Trust Game	0.65	0.96	0.16	0.29
Kimi-K2.5	Public Goods	0.73	0.47	0.08	0.03
Kimi-K2.5	Moral Luck	0.56	0.43	0.39	0.30
Qwen3.6-plus	Vertical Time	0.85	0.97	0.98	0.99
Qwen3.6-plus	Moral Luck	0.10	0.42	0.37	0.46
Qwen3.6-plus	Counterfactual	1.00	0.53	0.96	0.51

A.3 Requested-Response Classification

For each of the 29 pairs selected by full-precision $|\Delta| > 0.10$, let $M_{\min} = \min(M_{\mathrm{EN}}, M_{\mathrm{ZH}})$ and $M_{\max} = \max(M_{\mathrm{EN}}, M_{\mathrm{ZH}})$. The mixed-language means are $M_{\mathrm{EN}\to\mathrm{ZH}}$ and $M_{\mathrm{ZH}\to\mathrm{EN}}$. A mixed-language mean is classified as *below the range* when it is strictly less than $M_{\min} - \delta$ and as *above the range* when it is strictly greater than $M_{\max} + \delta$, with $\delta = 0.02$. A pair is labelled below the range if either mixed-language mean falls below $M_{\min} - \delta$, and above the range if either mixed-language mean rises above $M_{\max} + \delta$; the label is "beyond the range" if either condition holds. All reported labels use the full-precision means rather than the two-decimal display values shown in Table S2, because the margin $\delta = 0.02$ coincides with the rounding granularity. The qualitative pattern is stable for $\delta \in [0.02, 0.03]$; widening δ to at least 0.05 reclassifies more pairs as lying inside the envelope, while tightening δ to at most 0.01 surfaces one additional pair in each tail. Per-pair counts depend on display precision and are therefore reported only at the qualitative level in the main text.

A.4 Option-Position Sensitivity

The option order was rotated with the reproducible rule. The reconstructed schedule was close to balanced, with Cramér's $V = 0.0117$ for two-choice scenarios and $V = 0.0297$ for three-choice scenarios. These associations are small, although the three-choice association is statistically detectable because it has more possible positions. This analysis checks the planned rotation; it does not recover the semantic option for every response letter.

The merged JSONL does not retain a trusted mapping from each response letter to its semantic option. Reconstruction agreed with the displayed letter in 23,269 of 52,607 records with complete metadata, or 44.2%. The expected random-agreement rate under the observed scenario mix (seven two-choice and four three-choice) is approximately $(7/11)\cdot\frac{1}{2} + (4/11)\cdot\frac{1}{3} = 43.9\%$, so the observed agreement sits only marginally above the random floor. Explanations containing several letters can appear to disagree even when the final semantic choice is correct. We therefore recomputed the 66 English–Chinese comparisons using only records whose displayed letter matched the reconstructed rotation. In this restricted subset, 34 comparisons have uncorrected $p < 0.05$ and 33 remain significant after the full Benjamini–Hochberg correction, compared with 40 and 37 in the primary analysis. No comparison changes direction. Because the agreement rate is not a direct estimate of rotation failure, we report this restricted analysis as a sensitivity check rather than as a second prevalence estimate.

A.5 Follow-up Extensions

Four extensions are natural follow-ups to the present scope. First, expanding the language set to typologically distant pairs (EN–FR, ZH–JA, EN–AR) would let us separate vocabulary, script, and tokenisation effects, because each pair differs on only a subset of these dimensions. Second, replacing the forced-choice probes with open-ended decision scenarios under matched rubrics would test whether the drift is an artefact of the choice architecture; if the drift survives, the language effect generalises to deployment surfaces, and if it does not, the choice machinery itself is implicated. Third, repeating the design on a single model with controlled ablations of its instruction-tuning data would isolate the training-distribution channel from the language channel. Fourth, freezing the model versions and re-running the suite on subsequent checkpoints would let us ask whether providers are narrowing or widening cross-lingual drift over time.

A.6 Contributions and Non-Claims

We make no claim about whether the models share an internal representation of language, whether their behaviour tracks any cultural norm, or whether the drift will narrow as providers update their systems; each of those questions is open and tractable, but each requires a different experiment than the matched forced-choice design reported here. The present contribution is the controlled measurements themselves: 52,800 trials, 66 matched comparisons, and the consistent finding that language-conditioned shifts are common, model-specific, and sometimes push a requested-response mean outside the monolingual envelope.

A.7 Causal Claims Ruled Out by the Matched Design

A few causal claims are nevertheless ruled out by the matched design. If the drift were a script-driven persona switch, we would expect a consistent fingerprint across the four prompt conditions; instead, the requested-language manipulation moves the score outside the monolingual envelope for several model–scenario pairs, hard to attribute to persona switching alone. If the drift were just translation noise that moves probability mass between options, the matched monolingual contrasts would wash out; they do not, with 37 of 66 surviving Benjamini–Hochberg correction. The remaining co-variates (tokenisation, vocabulary statistics, instruction-tuning corpus balance, provider-side decoding defaults) cannot be separated without controlled tokenisation experiments, instruction-tuning ablations, or hidden-state access during a forced choice.