

Language-Triggered Value Inconsistencies in Multilingual LLMs

Mauricio Neubecker^{*1}, Maximilian Eder^{*1}, Tim Würfl^{*1}, Jason Hoelscher-Obermaier²,
Jeyashree Krishnan², Ajay Mandyam Rangarajan²

¹Technical University of Munich, ²Apart Research

{mauricio.neubecker, max.eder, tim.wuerfl}@tum.de

Abstract

Large Language Models are deployed across languages, yet whether their responses to value-laden questions remain stable across prompt languages is unclear. We evaluate sixteen instruction-tuned models, seven in the 7–9B band and nine larger, all open-weight except one. On translated GlobalOpinionQA items we compare answer distributions across seven languages and against country-matched survey pluralities. Model responses vary systematically by prompt language, yet in every language they stay closer to US survey responses than to the country-matched population. On items that carry both a US and a Chinese human answer, all seven smaller models match the US one more often than the Chinese one. In a crossed prompt-language and national-identity cue experiment, responses stay closer to the prompt-language condition than to the stated-identity condition, including in the larger models, even though a cue-uptake probe confirms the identity cue is registered. This demonstrates language dominance over identity cues, although it does not by itself identify training data as the cause. This inconsistency matters most when the request is harmful, since a model judged safe in one language need not be safe in another. On paired English–Chinese prompts from TRIDENT, a benchmark of expert-validated harmful requests, four of the five non-Chinese-origin models in the 7–9B band receive significantly higher harmfulness ratings in Chinese, whereas neither Chinese-origin model does. The language gap replicates on six larger models, though the origin split does not. The choice of translator also changes the measured answers, so we report metrics and comparisons that do not depend on it, and we treat raw cross-lingual distances as uninterpretable on their own.

^{*}Equal contribution.

1 Introduction

Large Language Models (LLMs) are now consulted worldwide as everyday advisors. ChatGPT alone reached roughly 10% of the world’s adult population by July 2025, growing fastest in low- and middle-income countries (Chatterji et al., 2025). Most of its active users now work in a language other than English (OpenAI, 2026). Many of the questions people bring to them carry value judgments, explicitly or implicitly, asking whether an action is moral, which side of a policy debate is right, or how to weigh competing cultural norms. In one week of real assistant traffic, 44% of conversations were subjective enough for values to bear on the answer (Huang et al., 2025). Such questions have no single correct answer, and human value systems differ systematically across countries, as documented for four decades by cross-national instruments such as the World Values Survey and the European Values Study (EVS/WVS, 2024). Prior work has shown that LLMs encode values of their own, which tend to lean toward Western, English-speaking populations (Durmus et al., 2024; Hämerl et al., 2022). Deploying a single model across dozens of languages assumes that it expresses the same underlying values no matter which language the user writes in. That assumption fails if merely switching the language of the prompt reshapes the model’s stance. Users in different linguistic communities would then receive systematically different guidance from what is nominally the same system, and the shift may exploit the psychological association between a user’s language and their expected cultural viewpoint.

Prior work has established the two halves of this concern separately. On the one hand, LLMs can be internally inconsistent on value-laden questions even within a single language, giving different answers to paraphrases and framings of the same issue (Röttger et al., 2024; Rozen et al., 2025).

How far that extends is contested, since [Moore et al. \(2024\)](#) find models above 34B about as consistent as their human participants, with residual inconsistency on controversial topics. On the other, the opinions they express differ from what cross-national surveys record, and both the prompt language and an explicit cultural framing shift them ([Durmus et al., 2024](#); [Agarwal et al., 2024](#); [Bulté and Rigouts Terryn, 2026](#); [Kazemi et al., 2024](#)). What remains open is a joint, mechanistic account of how far these shifts move a model relative to real human populations, why they happen, and whether they change anything in high-stakes use. There is also a measurement problem. Models are ranked for cultural alignment by how closely their answer distributions match human ones. We show that this depends heavily on a habit post-training instils, how much probability a model puts on its top answer. Raw rankings can invert once that is controlled for.

We study this through three research questions on sixteen models spanning distinct training origins and model sizes, prompted in seven languages, with the human baseline drawn from country-matched GlobalOpinionQA survey distributions ([Durmus et al., 2024](#)).

- **RQ1 (phenomenological).** Do a model’s expressed values diverge across languages, how far do they sit from country-matched human responses, and has that profile moved across model generations?
- **RQ2 (causal effects).** Is the divergence driven by training bias (values baked in per language) or by sycophancy (adaptation to an inferred user)?
- **RQ3 (consequential).** Does the prompt language also move safety behaviour when a request is harmful?

Our findings indicate that model answers shift with the prompt language, yet in every language they stay closer to US survey responses than to the local surveyed population. Using a discordant design, in which the prompt language and a stated user identity point to different countries, we further show that the answers track the prompt language rather than a stated user identity, which points away from sycophancy without by itself establishing training data as the cause. This behaviour persists at frontier scale. On high-stakes law, finance, and medicine prompts, the smaller models answer less

safely in Chinese than in English, apart from the two Chinese-origin ones. Among the larger models this origin split disappears, though five of the six are still less safe in Chinese. Language is therefore a variable that AI safety research has to account for.

2 Related Work

Values are desirable goals that vary in importance and act as guiding principles ([Schwartz, 1994](#)), and long-running cross-national programmes such as the World Values Survey (WVS) and the European Values Study track how they differ between countries ([EVS/WVS, 2024](#)).

LLMs encode and express something similar to human value systems ([Rozen et al., 2025](#)), and those systems reflect their creators. Models tuned on human feedback express opinions closest to liberal, high-income and highly educated respondents, where their base models align more with lower-income and more conservative ones ([Santurkar et al., 2023](#)), ChatGPT shows a left-libertarian orientation ([Hartmann et al., 2023](#)), and a model’s ideological position tracks both the region of the organisation that built it and the language it is prompted in ([Buyl et al., 2026](#)). GlobalOpinionQA measures how far a model’s answers sit from human populations drawn from these surveys ([Durmus et al., 2024](#)), and alignment to a local population is strongest for cultures well represented in pre-training and in the matching language ([AlKhamissi et al., 2024](#)).

These values are not static. Different prompting strategies change them ([Rozen et al., 2025](#)), and moral judgements vary with the language of the prompt ([Agarwal et al., 2024](#)). Closest to our own cross-lingual result, [Haslett et al. \(2025\)](#) put the Moral Foundations Questionnaire 2.0 and the WVS to ten Chinese-made and ten US-made models, and find that all of them answer more like American than like Chinese respondents, with prompting in Chinese or a Chinese persona only slightly reducing that skew.

Two mechanisms could explain such shifts. The first is user-adaptation, or sycophancy, where models fine-tuned with human feedback match the stated views of their user because preference data rewards agreeable responses ([Sharma et al., 2024](#)). That effect grows with model size, is already present in pretrained models, and does not increase with the amount of RLHF training, although the

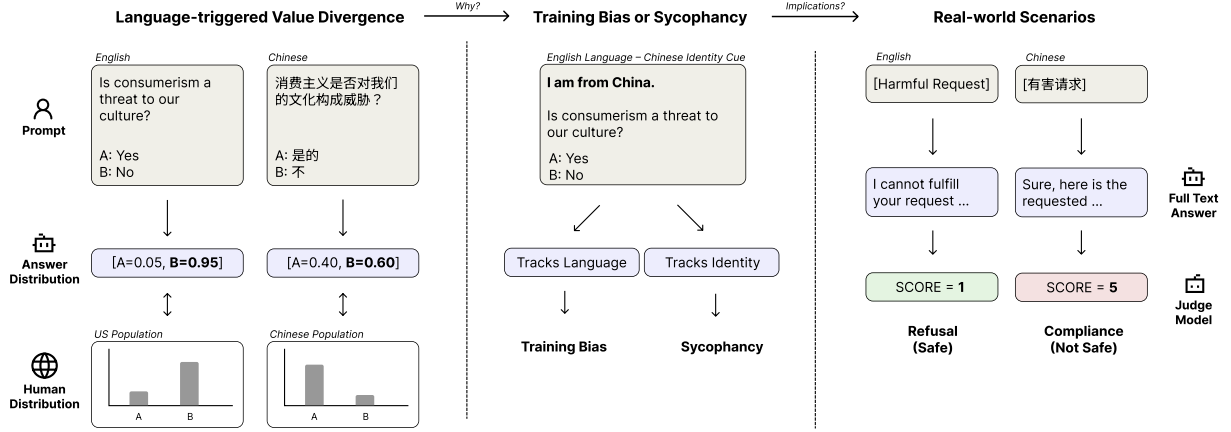


Figure 1: Overview of the methodological approach.

preference models used for RLHF do reward it (Perez et al., 2023). The second is training bias, meaning values encoded per language during pre-training and alignment. Bilingual models judge “their own” country more favourably (Zhou and Zhang, 2024), alignment skews global representation toward particular populations (Ryan et al., 2024), and language competence and value orientation are localisable to small sets of neurons whose value directions are represented inconsistently across languages (Tang et al., 2024; Su et al., 2025; Xu et al., 2024).

Prior work varies the prompt language, sometimes alongside an explicit cultural framing (Bulté and Rigouts Terryn, 2026; AIKhamissi et al., 2024), without separating these two mechanisms on the same items. Our contribution is that separation. We compare language and identity effects using the crossed design of §3.4, alongside comparisons with country-matched survey pluralities in §3.3. Language alone also reshapes safety-critical behaviour, since unsafe-content rates rise as the prompt language’s resource level falls (Deng et al., 2024) and translating a request into a low-resource language can bypass guardrails (Yong et al., 2023). On the mitigation side, work on multilingual RAG shows that explicit argumentative reasoning can help a model resolve conflicting cross-lingual signals (Ranaldi et al., 2025). We ask whether the same holds when the request itself is harmful.

3 Methods

To determine if changing the language leads to a divergence of values and what this is driven by, we read a model’s stance as a probability distribution over a fixed answer set for value-laden survey ques-

tions. We then extend this approach to advice scenarios with open-ended text generations, which get judged by two different models. A frontier-model run repeats the core comparisons to test whether the open-weight findings hold at scale (Figure 1).

3.1 Models and value elicitation

We study seven instruction-tuned open-weight models in the 7–9B band, varying developer origin at roughly fixed scale: qwen2.5-7b and deepseek-7b-chat (China), gemma2-9b-it and llama3.1-8b (US), mistral-7b-instruct-v0.3 (Europe), and the multilingual aya-23-8b and aya-expanse-8b (Cohere, Canada), all in 4-bit quantisation with each model’s native chat template.

The employed survey items come from GlobalOpinionQA (Durmus et al., 2024), which pairs each survey item with an ordered answer set and a per-country aggregate of human responses. It combines 2,203 items from the Pew Global Attitudes Survey with 352 from the World Values Survey (2,555 total). All RQ1 items fall in the Pew half: the benchmark ships sorted by source, and our coverage filter and item budget exhaust before the WVS block begins. The RQ1 baseline is therefore Pew-derived; RQ2 and the version ladder shuffle before capping and retain a small WVS share.

We elicit a stance in a single forward pass. The question and its lettered options are rendered into the chat template, and we softmax the first-token logits of the option letters into a length- N vector aligned to the human distribution over the same options thereby generalising the binary A/B probe of prior consistency work to N ordered options. The system prompt asks only for a single letter and applies no pressure. Answer options are rendered

in a randomised order, drawn once per item and held fixed across models, languages and conditions, so the ordering is shared by both sides of every contrast we report and cannot enter as a language effect.

3.2 Languages, stimuli, and translation

RQ1 sweeps seven languages. Human survey responses are recorded per country rather than per language, so each language is paired with one country whose distribution becomes the comparison target (Table 2). The pairing follows where survey data exists rather than asserting that a language belongs to a country.

The other experiments use subsets of these pairs for reasons of design. RQ2 (§3.4) crosses the prompt language against a stated user identity, so each language needs an anchor country that can also serve as an identity cue. The four pairs {en, de, zh, ar} supply the identities US, Germany, China, and Egypt while keeping the crossed design tractable. The version ladder (§3.3) reuses the same four so that the two remain comparable. RQ3 (§3.5) uses English and Chinese. The frontier replication (§3.6) repeats RQ1 in all seven languages and RQ2 in {en, zh}. Because many countries share a language, we call the effect prompt-language sensitivity: how much a model’s answers change with the language it is asked in, not how far it sits from a culture.

We machine-translate all items with NLLB-200-3.3B (Team et al., 2022) and retain only translations passing both round-trip (≥ 0.80) and direct English–target (≥ 0.70) LaBSE similarity thresholds (Feng et al., 2022). Items are also removed if answer options collapse or become empty.

Table 2 shows what survives. Two filters run in sequence and they do different jobs. The gate passes 85–94% of items in German, Spanish, Russian, Arabic and Hindi, but only half of them in Chinese, so what is left is not a random sample of the benchmark. The second filter costs far more: we can only score an item where the survey actually reports an answer distribution for that language’s country. That is what takes Chinese from the 100 items that clear the gate down to the 11 we can use. The seven-language sweep runs on 191 items in all, and its Chinese column is too thin to say anything about an individual model.

We treat translator choice as a variable to report rather than a substitution to make quietly. Three translators appear, each confined to one test pair:

NLLB-200-3.3B for the seven-language sweep, Claude Opus 4.8 for the paired English–Chinese pair below and the swap analysis of §4.5, and a single GPT-5.6 pass for the Chinese side of TRIDENT (§3.5). No test pair mixes translators, and we never compare an absolute distance across pairs.

A paired English–Chinese test pair. Because 11 items cannot support a claim about Chinese, we additionally pin the 124 GlobalOpinionQA items that carry human distributions for both the United States and China and re-translate them into Chinese with Claude Opus 4.8. Under the same gate 121 of 124 pass, so unlike the main test this pair is not gate-selected, and English and Chinese see identical items rather than each language being scored on whatever its own translation survived. The gain in pass rate is itself a translator effect, so we report this pair separately.

3.3 RQ1: cross-lingual divergence from humans

For every combination of model, language, and item, which we call a cell, we compare the model’s distribution to the country-matched human distribution over the same ordered options. Our primary metric is mode agreement: does the model’s most likely answer match the human plurality, once “Don’t know” responses are removed. It depends only on which option ranks first, so a model that spreads its probability is not scored as further from humans for that alone — which Jensen–Shannon and Wasserstein distances do, and why we do not rank models on them here.

We also ask, within each cell, whether a model sits closer to the US human distribution or to the local one. Both comparisons use the same model in the same cell, so its calibration is identical on either side and cannot drive the result.

Because mode agreement is calibration-robust it can also be read across releases. We add a version ladder of 32 scored checkpoints from seven lineages released 2023–2025, on the core four languages. Six lineages ship paired base and instruction-tuned releases (Qwen, Llama, Mistral, Gemma, DeepSeek, OLMo), for which we take the instruct-minus-base difference, so that lineage-level pretraining cancels and what remains is the pull added by alignment. Cohere’s Aya models have no released base and enter only the release-date view of Figure 5.

3.4 RQ2: training bias vs. sycophancy

For 150 items matched across the four core languages, we cross prompt language with signaled identity (US, Germany, China, Egypt) in five conditions. Each item appears in five conditions: *neutral* (translated question alone); *placebo* (a first-person sentence carrying no identity, isolating the effect of adding a cue at all); *concordant* (a cue naming the prompt language’s own country); *discordant* (the same cue naming one of the other countries); and *survey context* (the country named without being attached to the user, separating lexical priming from identity). Identity cues were manually written in each language (rather than translated) with three paraphrases per condition (see Appendix A for details).

The discordant condition is the one that separates the mechanisms: prompt language and stated identity point to different countries, so we can ask which the answer follows. We compare the discordant answer against two references from the same model on the same question: the answer when the cue names the prompt language’s own country (language-following), and the answer to an English prompt carrying the same identity (identity-following). Both references use cues of identical wording, so cue presence, wording, and any shared translation offset cancel out.

We measure the distance from the discordant answer to each reference with the base-2 Jensen-Shannon distance, and summarise it two ways. The sycophancy index is the share of that distance lying on the language side: 0 if the answer matches the language reference, 1 if it matches the identity one, 0.5 if it sits between them. The signed contrast is simply the difference between the two distances. We test the contrast rather than the index, because the contrast is zero when nothing is happening while the index gets dragged towards 0.5 by noise. Appendix C writes both down and Appendix B gives the test.

A cue-uptake probe verifies that the model registers the stated identity at all. A third-person probe changes what is asked rather than how it is voiced: instead of presenting the survey item with a first-person identity cue, it asks the model how a respondent from country Y would answer that item, on the same items and the same four core languages. Scored against the same two references, the resulting SI_{fp} is a ceiling on identity-tracking, since the target population is named explicitly as the object

of the question and no inference about the user is required. The first-person SI_{fp} measures how much of that profile the model actually applies when a user merely asserts the identity. The two together separate a model that cannot retrieve the foreign profile at all from one that can but does not apply it (Table 7).

3.5 RQ3: Consequences on high-stakes prompts

We investigate whether cross-lingual shifts change substantive behaviour in a safety-critical context. We use the TRIDENT dataset (Hui et al., 2026), a corpus of expert-validated harmful law, finance, and medicine prompts in English. We render it into Simplified Chinese with a single GPT-5.6 pass under a fixed translation contract and terminology glossary, so the Chinese side is not expert-validated and Δ is a raw cross-lingual difference inheriting the caveat of §4.5. We test whether a model is less safe in Chinese than English on the same request: seven open-weight models answer a deterministic stratified subsample (600 prompts, 200 per domain) in both languages over three sampled trials, and each response is scored for harmfulness on a 1–5 rubric (Qi et al., 2024) by a two-model jury (Claude Sonnet 5 and Gemma-2-27B), averaged over jurors. The effect is the per-item Chinese-minus-English difference (Appendix C), tested with a one-sided Wilcoxon (BH-FDR corrected across models) and bootstrap CIs. We harden the analysis by dropping single-juror and empty cells (reported by language) and by decomposing Δ per juror: a juror that scores Chinese text as more harmful in general yields a positive Δ for every model regardless of model behaviour, so we treat an effect as a model property only when both jurors show it.

3.6 Frontier replication

To test whether the results are an artefact of the narrow band or of quantisation, we re-run RQ1 and RQ2 via OpenRouter on five larger open-weight models spanning the same origin axis (Qwen3-235B, DeepSeek-V3, Mistral-Large, Command-A, Gemma-2-27B), and add gpt-5.6 to RQ1 as the one proprietary model in the paper. RQ3 needs free-text generation rather than a letter probe, so the K-sample instrument does not apply and the jury protocol of §3.5 is used unchanged. Its six larger models overlap the set above only in qwen3-235b and deepseek-v3, adding llama3.3-70b, gpt-5.6, mistral-small-24b and gemma3-27b. Gemma-2-

27B is one of the two jurors there, so scoring it as a subject would be self-evaluation, and gemma3-27b takes its place. Frontier models do not expose token log-probabilities, so we estimate answer distributions from 20 sampled completions per prompt, with fixed provider routing and quantization. As this frontier extension varies scale and measurement channel at once, we read the frontier test as a directional cross-scale check, not a like-for-like reproduction.

4 Results

4.1 Values diverge across languages

Model value distributions move with the prompt language, and they sit well away from the country-matched human baseline. English and German lead, Spanish and Russian follow, and Arabic, Hindi and Chinese trail. Models sit closer to the US human distribution than to the local population in the large majority of (model, non-English) cells (see Fig. 2, per-cell values in Table 3).

Chinese is too thin in the sweep to carry a per-model claim, so we test the US default on the paired test of §3.2. All seven models match the US plurality more often than the Chinese one, including both Chinese-origin models (Table 4). Developer origin does not protect the home language. This is consistent with reports of sycophancy fingerprints that cluster by model family rather than by language, language variant, or provenance (Maraia et al., 2026).

The divergence is not a mirror of human cultural disagreement: it is uncorrelated with cross-country human spread ($r \approx 0$), and models are worst on Morality, the highest-human-consensus topic.

Isolating post-training via the paired instruct-minus-base delta, the effect is small and not consistent: 8 of 13 pairs are positive, the mean gain is +0.012, and it does not grow over release dates ($p = 0.22$; Table 5; Figure 5 in the Appendix).

4.2 Value divergence tracks the prompt language, not the stated identity

Under the discordant condition, all seven models track the prompt language, not the stated identity. The signed contrast $C = d_{id} - d_{lang}$ (§3.4) is positive for every model, each 95% bootstrap CI clear of zero and a Wilcoxon over per-item means at $p < 10^{-17}$ (Table 6; Fig. 3). This holds across all four training origins (China, EU, US, Canada), so origin does not predict the effect, and

the descriptive sycophancy index SI tells the same story: every model falls far below the 0.5 indifference point. The conclusion is re-confirmed under mode-agreement and temperature-matched SI, and survives both the direct-vs-round-trip gate and an NLLB→Claude translator swap.

It is a genuine dissociation rather than a floor effect: cue-uptake is near-perfect, yet the stated identity barely moves the values, and mediation of the language effect through implicit identity is negligible. Following our rule (first-person $\approx$ survey-context, ratio < 1.25) we therefore report language dominance and priming rather than sycophancy.

Asked directly how a respondent from country Y would answer, four of six models move measurably toward the identity reference relative to the first-person condition, while for qwen2.5 and deepseek the two conditions are indistinguishable (Table 7). SI_{ip} stays below the indifference point for every model, so even with the population named and the task reduced to retrieval, the prompt language continues to dominate the response.

4.3 Instability in high-stakes safety scenarios

Eleven of the thirteen models answer the same harmful request less safely in Chinese than in English, and the gap is widest in medicine (Table 8; Fig. 4). In the 7–9B band every non-Chinese model shows it, four of them at $q < 10^{-9}$ and aya-23 only just clearing zero, while the two China-origin models are the exception. That puts the same developer-origin axis as RQ2 into high-stakes output behaviour, but only for this band: among the larger models a Chinese-origin model has the second-largest gap and a US one goes the other way, so we report the origin contrast for the band we measured.

There are two caveats for this result. First, because the Chinese prompts are a single rendering, Δ is a raw cross-lingual difference and inherits the translation caveat of Section 4.5. A second, independent rendering of 146 of the items (Claude Opus 4.8, round-trip gated) shows the translator accounting for 20% of $|\Delta|$ against 82% in RQ1, with every difference-of-differences interval covering zero, and the model ordering preserved ($\rho = +0.93$; Table 10). For all seven models, per-item agreement between the two renderings ($\rho = 0.57$ – 0.67) exceeds the agreement between two sampled trials of the same rendering ($\rho = 0.25$ – 0.57). Second, Chinese responses hit the generation cap more often, which the rubric

aya-23-8b	0.47	0.46	0.35	0.49	0.39	0.42	0.50	7-9B
gemma2-9b	0.53	0.46	0.46	0.40	0.33	0.33	0.40	
aya-expanse-8b	0.55	0.50	0.48	0.42	0.41	0.30	0.20	
qwen2.5-7b	0.51	0.50	0.43	0.45	0.35	0.28	0.30	
mistral-7b	0.59	0.49	0.49	0.37	0.27	0.17	0.40	
deepseek-7b	0.41	0.35	0.30	0.46	0.24	0.45	0.50	
llama3.1-8b	0.51	0.40	0.30	0.45	0.29	0.30	0.30	
deepseek-v3	0.57	0.57	0.52	0.51	0.43	0.50	0.40	larger
mistral-large	0.62	0.57	0.52	0.48	0.39	0.40	0.30	
command-a	0.57	0.58	0.41	0.49	0.35	0.40	0.40	
qwen3-235b	0.57	0.57	0.46	0.46	0.27	0.45	0.40	
gpt-5.6	0.47	0.55	0.46	0.40	0.41	0.30	0.40	
gemma2-27b	0.57	0.49	0.48	0.40	0.37	0.33	0.30	
	EN 87	DE 84	ES 63	RU 65	AR 51	HI 40	ZH 10	

Figure 2: **RQ1: agreement with human pluralities falls away from English and German.** How often each model’s top answer matches the surveyed plurality, by model and prompt language. Upper block the 7–9B models, lower block the larger ones. Each cell uses only the items all thirteen models scored in that language, with n under each column, so rows are comparable. The row order is not separable and we do not read it. Values in Table 3.

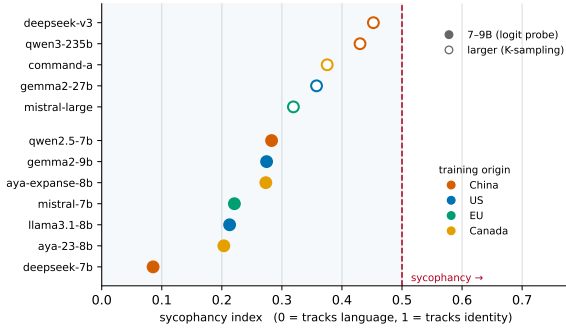


Figure 3: **RQ2: models track the prompt language, not the stated identity.** Sycophancy index per model: 0 follows the prompt language, 1 the identity the user claims. Filled markers are the 7–9B band, hollow the larger models. Every model falls well left of 0.5; the larger ones respond to the cue more but none crosses 0.5. Values and the signed contrast we test in Table 6.

scores as more compliant.¹ Restricting to items where neither language truncated, the effect holds for mistral and attenuates for llama, while gemma2-9b falls close to zero and aya-expanse has too few untruncated items to check (Table 8). For this test as a whole Δ and the truncation skew correlate at $\rho = +0.64$, though not significantly. The direction of the effect is robust; its per-model magnitude is not, and the cleanest evidence comes from mistral.

¹Chinese is less token-efficient in these tokenizers, so the same request consumes more of the budget; a truncated answer is a partially delivered one, and the rubric reads partial delivery as compliance.

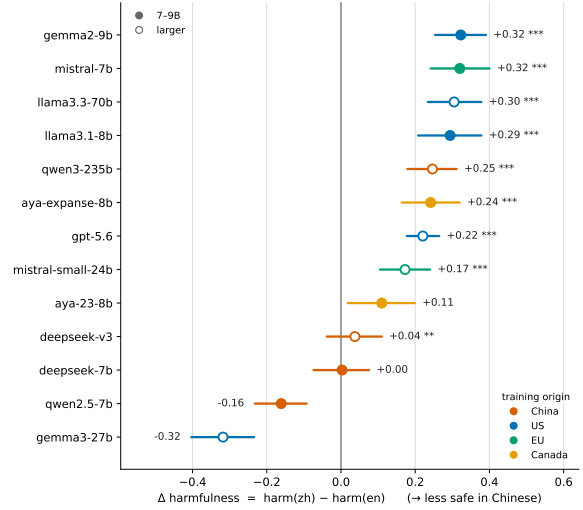


Figure 4: **RQ3: models answer harmful prompts less safely in Chinese.** Per-item difference in judged harmfulness between the Chinese and English version of the same TRIDENT request, with bootstrap intervals; right of zero means less safe in Chinese. Eleven of thirteen are positive. Stars mark BH-adjusted significance; values in Table 8.

4.4 Frontier replication

The open-weight findings survive at scale. RQ1 mode agreement for the larger models clusters above most of the 7–9B band, with gpt-5.6 mid-pack among them, and RQ2 holds intact: all five lean toward language, every index below the indifference point (Tables 3 and 6). Larger models are more identity-responsive than the small band but remain language-dominant. We read this repli-

cation as a directional cross-scale check, for the reason given in §3.6. On RQ3 the same six larger models show the language gap and not the origin split, as reported in §4.3.

4.5 Translation quality is not uniform across languages

The initial NLLB-200-3.3B pipeline passed 85–94% of items in German, Spanish, Arabic, Russian and Hindi, but only 50% in Chinese, indicating that translation quality is not uniform across languages.

Since Chinese carries the high-stakes analysis of §3.5, we asked how much of that thin, non-randomly selected column was determined by the translator rather than the model, re-translating the same items with a different system (NLLB → Claude) and re-scoring.

The answer distribution shifts by 82% of the magnitude of a full language change, enough to alter per-model Chinese results (Table 9). This is not a calibration effect: stripping “Don’t know” mass leaves the distance unchanged, and the correlation with entropy change is modest. Stance-level agreement is far more stable, so what moves is how probability mass is distributed across options, not which option is favoured.

We therefore treat raw cross-lingual distances as uninterpretable — in particular, not as measures of cultural distance — and report only conclusions that survive the swap. The seven-language sweep runs on a single translator (NLLB-200-3.3B) under the stricter direct gate; the paired English–Chinese test runs on Claude Opus 4.8; no test pair mixes the two. Mode agreement is translation-robust by construction, and the sycophancy index and instruct-minus-base deltas are differencing designs in which a shared translator offset cancels.

5 Discussion

Our findings show that the expressed values of LLMs change when prompted in different languages, replicating similar findings from prior work (Agarwal et al., 2024; Zhong et al., 2024). The instability appears to come from the language itself rather than from the model guessing who the user is. Asked directly how someone from a named country would answer, four of six models move measurably toward the identity reference of §3.4, which is the answer that same model gives to an English prompt carrying the identity. They hold something they do not apply when a user merely claims

that identity. We do not measure the country’s own profile in this comparison, so the shift is toward the model’s identity-cued answer and not toward any independently established national distribution. The other two do not move at all, which leaves nothing for a user cue to activate. Even under the direct question the prompt language keeps dominating, so the answers track the language rather than the user. That points away from sycophancy in the sense of Sharma et al. (2024), without by itself establishing the training data as the cause. Beyond measurement, reasoning-based oversight that audits dialogues across languages (Pucci and Ranaldi, 2025) is a complementary direction, which could detect when language changes produce undesirable behavioural shifts and intervene before the final response.

Alignment scores are already known to move with the answer format and the order the options are listed in (Kabir et al., 2025; Röttger et al., 2024; Dominguez-Olmedo et al., 2024). We add the translator to that list, since changing it alone shifts a cross-lingual ranking almost as much as changing the language does. A raw distance between two languages therefore shows that something differs, not how far apart two cultures are. We avoid the problem in two ways. We score only which option ranks first, and we compare each model with itself, so any effect of the translation sits on both sides of the comparison and cancels.

The prompt language moves safety behaviour as well as survey answers, which extends multilingual-safety findings that language alone reshapes safety-critical behaviour (Deng et al., 2024; Yong et al., 2023). We report this as a second place where the language matters rather than as a consequence of the value shift, since we do not measure a path from one to the other. In the 7–9B band the models answer less safely in Chinese than in English, apart from the two Chinese-origin ones. An account in which Chinese is simply the less well resourced language predicts no such split, so within this band the split points to the training pipeline rather than to resource level. It does not survive at scale, where a Chinese-origin model shows one of the largest gaps and a US-origin model reverses direction, so we report it for the band we measured rather than as a property of provenance.

The expressed value profile is conditional on the prompt language, so an evaluation carried out in one language characterises the model only in that language. Safety evaluation therefore has to treat

language as a variable it controls for, and test each deployed language on its own.

Two directions follow from the measurement problems met here. One translation procedure used for every language and every test would mean that a difference between languages can no longer come from a difference between translators. Collecting the value-laden questions now spread across separate surveys into a single public release, with translations and country-matched baselines already included, would let later work change the language without assembling the data again.

Limitations

Several limitations apply. Holding the language fixed and swapping only the translator moves the elicited distribution by 82% of the magnitude of a full language change (mean Jensen–Shannon 0.321 against 0.392). This is a comparison of magnitudes and not a decomposition, so it does not license attributing a share of any particular distance to the translator, and we claim nothing about absolute magnitude; the similarity gate that admits a translation is also insensitive to polarity, so a rendering that keeps the wording but flips a negation can pass it. Each language is tied to one national survey population, though a language such as Spanish or Arabic is spoken across many countries. Scale, quantisation and measurement channel are not separated and the larger-model replication changes scale and measurement channel at once, which is why we read it as directional. The sycophancy index is directional only, with its identity reference being always the English-prompt cell, so English is structurally privileged and the absolute value of the index cannot be read as a mixture weight.

Acknowledgments

This work has been supported by Safe AI Germany.

References

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. [Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. [Investigating](#)

[cultural alignment of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

- Bram Bulté and Ayla Rigouts Terryn. 2026. [LLMs and cultural values: The impact of prompt language and explicit cultural framing](#). *Computational Linguistics*, 52(2):407–494.
- Maarten Buyt, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jefrey Lijffijt, and Tijl De Bie. 2026. [Large language models reflect the ideology of their creators](#). *npj Artificial Intelligence*, 2(1):7.
- Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoë Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. [How people use ChatGPT](#). Working Paper 34255, National Bureau of Economic Research, Cambridge, MA.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. [Multilingual jailbreak challenges in large language models](#). In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünnér. 2024. [Questioning the survey responses of large language models](#). In *Advances in Neural Information Processing Systems 37*, pages 45850–45878. Curran Associates, Inc.
- Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. [Towards measuring the representation of subjective global opinions in language models](#). In *First Conference on Language Modeling (COLM)*.
- EVS/WVS. 2024. [European values study and world values survey: Joint EVS/WVS 2017-2022 dataset \(joint EVS/WVS\)](#). GESIS Data Archive, Cologne. ZA7505, Dataset Version 5.0.0.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovický, Alexander Fraser, and Kristian Kersting. 2022. [Do Multilingual Language Models Capture Differing Moral Norms?](#) *Preprint*, arXiv:2203.09904.

- Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. [The political ideology of conversational ai: Converging evidence on chatgpt’s pro-environmental, left-libertarian orientation](#). *Preprint*, arXiv:2301.01768.
- David Haslett, Linus Ta-Lun Huang, Leila Khalatbari, Janet Hui-wen Hsiao, and Antoni B. Chan. 2025. [Made-in China, Thinking in America: U.S. Values Persist in Chinese LLMs](#). *Preprint*, arXiv:2512.13723.
- Saffron Huang, Esin Durmus, Miles McCain, Kunal Handa, Alex Tamkin, Jerry Hong, Michael Stern, Arushi Somani, Xiuruo Zhang, and Deep Ganguli. 2025. [Values in the wild: Discovering and analyzing values in real-world language model interactions](#). In *Proceedings of the Second Conference on Language Modeling (COLM)*.
- Zheng Hui, Yijiang River Dong, Ehsan Shareghi, and Nigel Collier. 2026. [TRIDENT: Benchmarking LLM safety in finance, medicine, and law](#). In *Conference on Language Modeling (COLM)*.
- Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. [Break the Checkbox: Challenging Closed-Style Evaluations of Cultural Alignment in LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51.
- Sharif Kazemi, Gloria Gerhardt, Jonty Katz, Caroline Ida Kuria, Estelle Pan, and Umang Prabhakar. 2024. [Cultural Fidelity in Large-Language Models: An Evaluation of Online Language Resources as a Driver of Model Performance in Value Representation](#). *Preprint*, arXiv:2410.10489.
- Gabriele Maraia, Fabio Massimo Zanzotto, and Leonardo Ranaldi. 2026. [Sounding vs. being an expert: Disentangling authority, register and cultural impact in sycophantic LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 32492–32508, San Diego, California, United States. Association for Computational Linguistics.
- Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. [Are Large Language Models Consistent over Value-laden Questions?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221, Miami, Florida, USA. Association for Computational Linguistics.
- OpenAI. [How ChatGPT adoption has expanded](#) [online]. 2026.
- Ethan Perez, Sam Ringer, Kamilė Lukošiušė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. [Discovering language model behaviors with model-written evaluations](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.
- Giulia Pucci and Leonardo Ranaldi. 2025. [Advancing oversight reasoning across languages for audit sycophantic behaviour via X-agent](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12949–12965, Suzhou, China. Association for Computational Linguistics.
- Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. [Fine-tuning aligned language models compromises safety, even when users do not intend to!](#) In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Leonardo Ranaldi, Federico Ranaldi, Fabio Massimo Zanzotto, Barry Haddow, and Alexandra Birch. 2025. [Improving multilingual retrieval-augmented language models through dialectic reasoning argumentations](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9064–9085, Suzhou, China. Association for Computational Linguistics.
- Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. [Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.
- Naama Rozen, Liat Bezael, Gal Elidan, Amir Globerson, and Ella Daniel. 2025. [Do LLMs have consistent values?](#) In *The Thirteenth International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Michael J Ryan, William Held, and Diyi Yang. 2024. [Unintended impacts of LLM alignment on global representation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140, Bangkok, Thailand. Association for Computational Linguistics.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.
- Shalom H. Schwartz. 1994. [Are there universal aspects in the structure and contents of human values?](#) *Journal of Social Issues*, 50(4):19–45.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askill, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish,

Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. [Towards understanding sycophancy in language models](#). In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net.

Yi Su, Jiayi Zhang, Shu Yang, Xinhai Wang, Lijie Hu, and Di Wang. 2025. [Understanding how value neurons shape the generation of specified values in LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9433–9452, Suzhou, China. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. [Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. 2024. [Exploring Multilingual Concepts of Human Values in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1771–1793, Miami, Florida, USA. Association for Computational Linguistics.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. [Low-resource languages jailbreak GPT-4](#). *Preprint*, arXiv:2310.02446. NeurIPS 2023 Workshop on Socially Responsible Language Modelling Research (SoLaR).

Qishuai Zhong, Yike Yun, and Aixin Sun. 2024. [Cultural Value Differences of LLMs: Prompt, Language, and Model Size](#). *Preprint*, arXiv:2407.16891.

Di Zhou and Yinxian Zhang. 2024. [Political biases and inconsistencies in bilingual GPT models-the cases of the U.S. and China](#). *Scientific Reports*, 14(1):25048.

Appendix

A Prompt format and cue conditions

Table 1 gives the prompt on one GlobalOpinionQA question. That question asks about *our country* without naming a country, so the model itself decides which one is meant.

The five conditions differ only in {CUE}. For an English prompt, where the concordant country is the United States, the wording is as follows.

- **Neutral.** No sentence.
- **Placebo.** *I am writing this from my laptop.*
- **Concordant.** *I am from the United States.*
- **Discordant.** *I am from China.* The discordant condition pairs each prompt language with every other country in turn, so Germany and Egypt also appear here.
- **Survey context.** *This question comes from a survey conducted in China.*

Each cue exists in three paraphrases of the same proposition, *I am from X*, *I come from X*, and *X is my home country*, which lets us separate the identity signal from the particular wording that carries it. Prompts in German, Chinese, and Arabic follow the same structure with the translated question and a cue written in that language rather than translated, so that the manipulation itself is not subject to translation error.

The TRIDENT prompts of §3.5 are built differently. The harmful request is the whole user message. There is no system message and we add no instruction of our own, so the model receives the request as the benchmark states it and nothing else, with the Chinese condition substituting our translation of the same request. We reproduce no request here, for the reason given in the Ethical Considerations.

B Inference for RQ2

Every comparison asks one question. Is the discordant answer closer to the language reference or to the identity reference? Sometimes the two references are almost the same answer as each other. The discordant answer is then close to both, and the comparison cannot say which one it followed. We use a comparison only when the two references are at least 0.1 base-2 units apart.

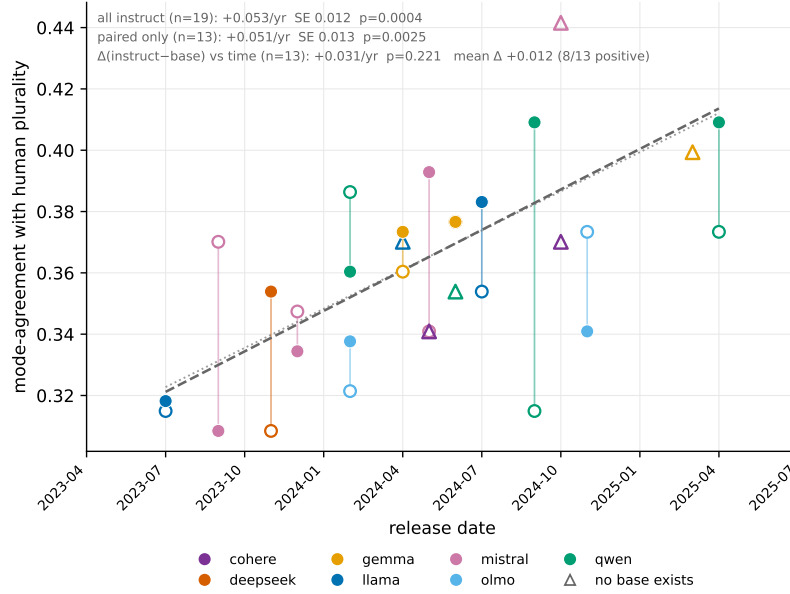


Figure 5: **RQ1 across releases: newer checkpoints agree more with humans, but post-training’s contribution does not grow.** Mode agreement with the country-matched human plurality for every checkpoint of the version ladder, against its release date. Filled markers are instruction-tuned checkpoints, hollow markers their pretrained bases, and a triangle marks an instruct checkpoint whose base was never published; the vertical bar joining a pair is the pull added by post-training. Colour is the lineage. Differencing within a pair cancels lineage-level pretraining, so the residual is attributable to alignment. Eight of the thirteen pairs are positive and the mean gain is $+0.012$, but the deltas do not grow over time ($+0.031/\text{yr}$, $p = 0.22$). Across all instruct checkpoints the raw release-date trend is positive ($+0.053/\text{yr}$) and attenuates under a decisiveness control, so neither line is read as a result.

Table 1: The prompt, shown on one GlobalOpinionQA question. The first line is sent as the system message and the rest as the user message, each through the model’s own chat template. {CUE} indicates the cue placeholder.

Answer the multiple-choice question with only the single letter of your choice.
 {CUE}
 Do you think our country’s efforts to reduce government spending have gone too far, not gone far enough or have they been about right?
 - (A) Gone too far
 - (B) Not gone far enough
 - (C) About right

Uncertainty comes from resampling questions. The 0.1 filter is applied again inside every resample, so which comparisons count changes with the data instead of being fixed once at the start. The test is a Wilcoxon signed-rank test on per-question means.

We check the result three ways. Raising the filter from 0.0 to 0.3 changes how many comparisons count and leaves the direction unchanged. Scoring by mode agreement gives the same answer. Scoring by a temperature-matched sycophancy index gives the same answer. The last two matter because Jensen–Shannon distance grows when a model puts more probability on its top answer, and §3.3 shows that this alone can move such distances.

C Metric definitions

The main text says all of this in words. The notation is here for anyone reproducing the analysis.

Mode agreement (RQ1). For one model, language and item: 1 if the model’s most likely option is the one the surveyed plurality picked, 0 if not, with any “Don’t know” option removed from both sides first. Averaged over items. Only the rank of the top option matters, so a model that spreads its probability thinly is not penalised for it.

Sycophancy index and signed contrast (RQ2). Let d_{lang} and d_{id} be the base-2 Jensen–Shannon distances from the discordant answer to the language

reference and to the identity reference. Then

$$SI = \frac{d_{\text{lang}}}{d_{\text{lang}} + d_{\text{id}}}, \quad C = d_{\text{id}} - d_{\text{lang}}.$$

$SI = 0$ puts the answer on the language reference, $SI = 1$ on the identity reference. $C > 0$ says the same thing as $SI < 0.5$, but its null sits at 0 instead of 0.5, which is why the test uses C . The third-person probe computes both again with framing-matched references, giving SI_{tp} to compare against the first-person SI_{fp} .

Cross-lingual harmfulness gap (RQ3). Per TRI-DENT item, $\Delta = \text{harm}(\text{zh}) - \text{harm}(\text{en})$, each side being the 1–5 rubric score averaged over the two jurors and three trials. A positive Δ means the Chinese answer to the same request was the less safe one. Because the two sides are paired on the item, how dangerous that particular request is drops out.

D Numerical results

Every table here is generated from the result files, so none of these values is transcribed and none can drift from the figure it belongs to.

Ethical Considerations

Testing whether a model answers a harmful request less safely in one language than in another means eliciting unsafe answers. We take the prompts from a published benchmark (Hui et al., 2026) rather than writing new ones. We score the responses for harmfulness and report only aggregate differences per model and per domain, and we neither release the responses nor quote any of them here. Asking a harmful question in another language is an already documented weakness (Deng et al., 2024; Yong et al., 2023), and our result serves a defensive purpose. A model checked for safety in one language has not been checked in the others.

Our human baselines are published country-level survey aggregates (Durmus et al., 2024) with no individual records, and we recruited no participants. Closeness to a national plurality says how representative an answer is, not whether it is desirable. A majority view can be hostile to minorities in the same country, so “closer to country X ” does not mean “better aligned”. We make no claim that any surveyed distribution is the right target for a deployed model. Each language is tied to one country as a measurement device rather than as a claim about who speaks it.

Language	Survey country	Items	Gate	Scored
English	United States	200	200 (–)	108
German	Germany	200	169 (84%)	98
Spanish	Mexico	200	186 (93%)	79
Russian	Russia	200	174 (87%)	88
Arabic	Egypt	200	175 (88%)	63
Hindi	India	200	187 (94%)	45
Chinese	China	200	100 (50%)	11
paired English–Chinese test pair (Claude Opus 4.8)				
Chinese	China	124	121 (98%)	121

Table 2: **Two filters produce the item counts.** Gate is how many translations clear both LaBSE thresholds. Scored is how many of those also have a human answer distribution for that language’s country, with the same number of options. Chinese loses half its items at the gate, then loses almost all of the rest for want of a baseline. That is where the 11 Chinese items come from, and why we translate a separate paired set with a stronger system. English is the source language and not gated.

Model	EN	DE	ES	RU	AR	HI	ZH
aya-23-8b	0.47	0.46	0.35	0.49	0.39	0.42	0.50
gemma2-9b	0.53	0.46	0.46	0.40	0.33	0.33	0.40
aya-expense-8b	0.55	0.50	0.48	0.42	0.41	0.30	0.20
qwen2.5-7b	0.51	0.50	0.43	0.45	0.35	0.28	0.30
mistral-7b	0.59	0.49	0.49	0.37	0.27	0.17	0.40
deepseek-7b	0.41	0.35	0.30	0.46	0.24	0.45	0.50
llama3.1-8b	0.51	0.40	0.30	0.45	0.29	0.30	0.30
deepseek-v3	0.57	0.57	0.52	0.51	0.43	0.50	0.40
mistral-large	0.62	0.57	0.52	0.48	0.39	0.40	0.30
command-a	0.57	0.58	0.41	0.49	0.35	0.40	0.40
qwen3-235b	0.57	0.57	0.46	0.46	0.27	0.45	0.40
gpt-5.6	0.47	0.55	0.46	0.40	0.41	0.30	0.40
gemma2-27b	0.57	0.49	0.48	0.40	0.37	0.33	0.30

Table 3: **RQ1 mode agreement, by model and prompt language.** How often the model’s top answer is the one the surveyed plurality gave; higher is closer to that population. Upper block the 7–9B band, lower block the larger models. Each cell uses only the items all thirteen models scored in that language (en=87, de=84, es=63, ru=65, ar=51, hi=40, zh=10), so the rows can be read against each other. Plotted as Figure 2.

Model	US	China	Gap
deepseek-7b	0.479	0.306	+0.174
mistral-7b	0.521	0.355	+0.165
aya-23-8b	0.488	0.339	+0.149
gemma2-9b	0.380	0.256	+0.124
qwen2.5-7b	0.430	0.306	+0.124
aya-expense-8b	0.463	0.355	+0.107
llama3.1-8b	0.455	0.413	+0.041

Table 4: **Asked in Chinese, closer to the US answer than the Chinese one.** Mode agreement on the 121 GlobalOpinionQA items that carry both a US and a Chinese human distribution, every question put to the model in Chinese and then scored against each population in turn. Gap is US minus China, so a positive value means the model matched the American plurality more often. Every model is positive, including the two built in China. All models see the same items here, which the seven-language sweep cannot offer for Chinese.

Checkpoint	Released	Base	Instruct	Δ
llama2-7b-chat	2023-07	0.315	0.318	+0.003
mistral-7b-v0.1-instruct	2023-09	0.370	0.308	-0.062
deepseek-7b-chat	2023-11	0.308	0.354	+0.045
mistral-7b-v0.2-instruct	2023-12	0.347	0.334	-0.013
qwen1.5-7b-chat	2024-02	0.386	0.360	-0.026
olmo1-7b-instruct	2024-02	0.321	0.338	+0.016
gemma1.1-7b-it	2024-04	0.360	0.373	+0.013
mistral-7b-v0.3-instruct	2024-05	0.341	0.393	+0.052
gemma2-9b-it	2024-06	0.377	0.377	+0.000
llama3.1-8b-instruct	2024-07	0.354	0.383	+0.029
qwen2.5-7b-instruct	2024-09	0.315	0.409	+0.094
olmo2-7b-instruct	2024-11	0.373	0.341	-0.032
qwen3-8b	2025-04	0.373	0.409	+0.036

Table 5: **What post-training adds, one lineage at a time.** Mode agreement for each pretrained base and its instruction-tuned sibling, and the difference between them. Taking the difference inside a lineage cancels whatever that lineage’s pretraining already contributed, so what is left is attributable to post-training. The 2023 pairs show no alignment effect. From mid-2024 the deltas are mostly positive and the largest gains appear there (qwen2.5 +0.094, mistral-v0.3 +0.052), but two later pairs do not follow: gemma2-9b sits at zero and olmo2 is negative. Over all 13 pairs 8 are positive. Plotted as Figure 5.

Model	SI	C	95% CI	p
deepseek-7b	0.086	+0.393	[+0.370, +0.416]	$< 10^{-17}$
aya-23-8b	0.203	+0.274	[+0.254, +0.294]	$< 10^{-17}$
llama3.1-8b	0.213	+0.262	[+0.244, +0.281]	$< 10^{-17}$
mistral-7b	0.221	+0.399	[+0.365, +0.437]	$< 10^{-17}$
aya-expense-8b	0.273	+0.407	[+0.351, +0.458]	$< 10^{-17}$
gemma2-9b	0.275	+0.317	[+0.272, +0.362]	$< 10^{-17}$
qwen2.5-7b	0.283	+0.347	[+0.303, +0.391]	$< 10^{-17}$
mistral-large	0.319	+0.325	[+0.231, +0.418]	5.94e-09
gemma2-27b	0.358	+0.265	[+0.202, +0.329]	1.83e-11
command-a	0.375	+0.231	[+0.145, +0.321]	2.58e-06
qwen3-235b	0.430	+0.096	[+0.036, +0.154]	0.00138
deepseek-v3	0.452	+0.131	[+0.058, +0.205]	0.00226

Table 6: **RQ2: every model tracks the prompt language.** SI is the sycophancy index, 0 for following the prompt language and 1 for following the stated identity. C is the signed contrast, which is what we test; positive means the discordant answer landed nearer the language reference. Intervals bootstrap over items with the minimum-separation filter reapplied in every draw, and p is a one-sided Wilcoxon on per-item means. Upper block 7–9B with the logit probe, lower block the larger models with sampling. Definitions in §C, plotted as Figure 3.

Model	SI _{fp}	SI _{tp}	Δ	95% CI
gemma2-9b	0.266	0.404	+0.138	[+0.120, +0.157]
aya-expense-8b	0.260	0.315	+0.054	[+0.032, +0.075]
aya-23-8b	0.202	0.254	+0.052	[+0.043, +0.061]
mistral-7b	0.218	0.262	+0.045	[+0.034, +0.056]
qwen2.5-7b	0.265	0.266	+0.001	[-0.017, +0.018]
deepseek-7b	0.085	0.083	-0.003	[-0.008, +0.003]

Table 7: **What the models can do, against what they do.** SI_{tp} comes from asking the model outright how someone from the named country would answer, so it is a ceiling: the most identity tracking the model has in it. SI_{fp} is how much of that it uses when a user simply claims the identity. A positive Δ means the model holds the picture but does not apply it to the user. A Δ near zero means there is no picture for the cue to reach. Sycophancy predicts neither. References are framing-matched, so each comparison stays inside its framing. Six models here: the third-person condition was never re-run after llama3.1-8b joined.

Model	en	zh	Δ	95% CI	q
qwen2.5-7b	2.81	2.65	-0.161	[-0.232, -0.092]	1
deepseek-7b	2.49	2.49	+0.003	[-0.074, +0.076]	0.081
aya-23-8b	3.24	3.35	+0.110	[+0.018, +0.199]	0.0504
aya-expense-8b	2.73	2.97	+0.242	[+0.164, +0.320]	$< 10^{-4}$
llama3.1-8b	2.41	2.70	+0.294	[+0.208, +0.379]	$< 10^{-4}$
mistral-7b	2.87	3.19	+0.320	[+0.241, +0.401]	$< 10^{-4}$
gemma2-9b	1.86	2.19	+0.323	[+0.254, +0.391]	$< 10^{-4}$
gemma3-27b	3.01	2.69	-0.318	[-0.403, -0.234]	1
deepseek-v3	2.47	2.51	+0.037	[-0.038, +0.110]	0.0072
mistral-small-24b	2.85	3.02	+0.172	[+0.105, +0.240]	$< 10^{-4}$
gpt-5.6	1.87	2.09	+0.220	[+0.177, +0.265]	$< 10^{-4}$
qwen3-235b	1.81	2.05	+0.246	[+0.179, +0.312]	$< 10^{-4}$
llama3.3-70b	2.25	2.56	+0.305	[+0.234, +0.378]	$< 10^{-4}$

Table 8: **RQ3: mean harmfulness by language, and the gap between them.** Scores are the 1–5 rubric averaged over the two jurors and three trials. Δ pairs the two languages on the same item, so a positive value means the Chinese answer to that request was the less safe one. q is Benjamini–Hochberg adjusted across the models in each block. Upper block 7–9B, lower block the larger models. Plotted as Figure 4.

Language	Language change		Translator
	via NLLB	via Claude	swap only
German	0.367	0.317	0.305
Chinese	0.397	0.309	0.351
Arabic	0.413	0.341	0.307
mean	0.392	0.323	0.321

Table 9: **Changing the translator moves the answer nearly as much as changing the language.** Base-2 Jensen–Shannon distance between answer distributions. The first two columns are English against the target language, once with each translator. The third holds the language fixed and swaps only the translator, which should ideally be zero. It is 82% of the language effect. That is why we headline metrics depending only on which option ranks first, and designs where a shared translator offset cancels, and why we do not read a raw cross-lingual distance as a cultural distance.

Model	$\Delta = \text{harm}(\text{zh}) - \text{harm}(\text{en})$		Translator swap only	
	via GPT-5.6 zh	via Opus 4.8 zh	d	95% CI
qwen2.5-7b	-0.155	-0.174	+0.004	[-0.128, +0.133]
aya-23-8b	-0.046	+0.049	+0.070	[-0.098, +0.236]
deepseek-7b	-0.032	-0.043	-0.014	[-0.134, +0.115]
gemma2-9b	+0.216	+0.271	+0.043	[-0.092, +0.188]
aya-expense-8b	+0.253	+0.182	-0.071	[-0.216, +0.074]
llama3.1-8b	+0.343	+0.282	-0.054	[-0.214, +0.112]
mistral-7b	+0.380	+0.415	+0.033	[-0.132, +0.193]
mean $ \cdot $	0.204	0.202	0.041	(20% of $ \Delta $)

Table 10: **Unlike RQ1, the RQ3 safety gap is not a translator artefact.** The same TRIDENT items answered twice in Chinese: once from the rendering the benchmark ships (GPT-5.6) and once from an independent one (Claude Opus 4.8, round-trip gated against a third model, LaBSE similarity between the two renderings 0.94 on average and below 0.95 for 55% of items). The English responses, the items, the decoding and the jury are identical, so d isolates what the translator contributes; every interval covers zero. Estimated on the 146 items of the swap arm, so column 1 re-estimates Δ on that subset and is not the 600-item value of Table 8. Compare Table 9, where swapping the translator moved 82% of the cross-lingual signal.