Cultural Alignment of Open-Weight LLMs on the Inglehart–Welzel Map

Shav Vimalendiran
Sammy Labs
shav@sammylabs.com

Abstract

Large language models are increasingly deployed where cultural values shape outcomes, yet the values they express are rarely measured on validated instruments, and where they have been, the pipelines have gone unaudited. We administer the ten Integrated Values Survey items defining the Inglehart–Welzel cultural map to 17 frontier open-weight models (10 Chinese-origin, 7 Western) in both English and Chinese, 34 cells and 17,000 calls, projecting every response through the same frozen pipeline as the 109 surveyed countries. An audit of our own 2024 public analysis first surfaced three defects we correct and quantify, among them an SPSS missing-value code read as data in 31.9% of survey rows. Four findings follow. **(i)** Among models that answer the instrument, every cell occupies one small corner of the map (secular, self-expressive), in the 2026 cohort in all 330,000 cluster-bootstrap replicates, with Japan, Germany and the Nordics as nearest neighbours and no cell mean close to a non-Western centroid; 2026 quadrant membership survives per-axis worst-case placement of every refused item, though the most human-proximal placements rest on the two most-refused items. **(ii)** Chinese-origin models express no distinctly Chinese values, and Chinese-language administration does *not* move models toward the traditional pole: 15 of 16 move toward self-expression instead, developer origin predicts nothing detectable about the shift (permutation $p = 0.89$), and the largest displacement belongs to a French lab's model. **(iii)** Of the items that move, the only shift toward the survival pole is petition-signing's, the only politically actionable item, indicating context gating rather than value shift. **(iv)** Across 900 reasoning traces (single-annotator coding), 57% target the typical human answer and only 10% inhabit the persona: these models estimate survey statistics rather than express values, and that estimate is systematically Western. Non-response has meanwhile changed hands: in 2024 five of nine attempted Chinese-origin models produced no parseable output; in 2026 all ten parse, and refusal, recorded for the first time, is concentrated in Western models. We release the corrected pipeline and the full response corpus, refusals and traces included.

1 Introduction

Culture shapes judgment, trust in automation, privacy expectations and health behaviour. As LLM-generated text becomes a transmission channel for language itself, models that systematically over-represent one culture's values can propagate those values into every downstream context they touch. Whether that risk is real depends on an empirical question: *what values do deployed models actually express, measured on instruments designed for measuring values?*

An earlier version of this analysis was released publicly in July 2024 (Vimalendiran, 2024) and has since been cited repeatedly in peer-reviewed venues. This paper is what became of it under peer-review standards: a full pipeline audit, three corrections, uncertainty on every claim, and a designed experiment where the original had a confound. We state the defects openly, because the paper's central argument is that *instrument-based claims about model values are only as good as the measurement code behind them*, and that applies to our own work first. **(1) Projection defects:** the 2024 code projected raw 1–10 model answers onto principal axes fitted to *standardised* survey data and re-fitted the rotation on the model scores, so model and country points did not provably share a coordinate space; under the correction `llama3:70b` moved 2.05 map units (§3.2). **(2) A silenced item:** the SPSS user-missing code −3 on the autonomy index Y003 (one of the five items defining the traditional/secular axis) was read as a datum in 31.9% of post-2005 rows, so the item loaded ≈0; recoding raises explained variance from 38.4% to 41.7%, each figure on its own denominator (Appendix D). **(3) Pooled elicitation**

languages: the two models assigned to the Confucian region in 2024 were administered *only* in Chinese, and the one model administered in both had its runs pooled across a 1.77-map-unit gap behind an implied error bar of ± 0.1. Language is an experimental factor and is treated as one throughout.

Contributions. (1) Methodological. A corrected, validated, unit-tested, open pipeline: full IVS reconstruction (EVS 1981–2017 + WVS 1981–2022) with probabilistic PCA retaining the 51% of rows listwise deletion would discard; a single stored rotation applied identically to countries and models; survey-design weights at aggregation; two bootstrap estimators; two region-assignment rules reported with the classifier's cross-validated accuracy; and classifier-free headline statistics with replicate-simultaneity handled explicitly. To our knowledge no prior instrument-based evaluation of LLMs reports confidence regions for model positions. **(2) Empirical.** An uncertainty-aware map of the 2024 open-weight generation as eleven model-language cells, with elicitation language the largest single mover, a finding the 2024 analysis mistook for a Confucian signal. **(3) Experimental.** A factorial language × model design over the 2026 frontier generation (17 models, 10 Chinese-origin; 34 cells, 17,000 calls), turning the 2024 confound into a designed within-model contrast and giving the Confucian hypothesis its first genuine test under both administration languages. It fails in both arms; in 2024 the test was impossible (five of nine attempted Chinese-origin models produced nothing parseable), so the disappearance of that limitation is itself a finding (§4.2).

2 Related Work

Values-survey instruments. Tao et al. (2024), now in *PNAS Nexus*, is the founding work in this line: administering WVS items to GPT-family models, they find values resembling English-speaking and Protestant European countries and show cultural prompting can shift them. Eren et al. (2026) extend that result to open-weight models, so open-weight coverage alone is no longer novel; our differentiation is *measurement auditability*: full individual-level IVS reconstruction rather than published country scores, one audited projection path shared by countries and models, uncertainty on every claim, and coverage of Chinese-developed and alignment-stripped families under a language-controlled protocol. LLM-GLOBE (Karinshak et al., 2024) finds significant China–US model differences on seven of nine GLOBE dimensions, but both cohorts diverging significantly from their own countries' human ground truth: the homogenisation pattern we measure on the IW instrument. Luther and Brown (2025) find DeepSeek expressing US-aligned values on Hofstede's VSM13 under every prompting strategy tried: converging evidence, by an independent instrument, that Chinese-developed frontier models do not express Chinese-aligned values.

Elicitation-language effects. Bulté and Rigouts Terryn (2025) probe ten LLMs with Hofstede/WVS items in eleven languages and independently establish that prompt language shifts expressed values while models stay anchored to a small set of cultural defaults, though the magnitude is contested (a direct replication of a flagship result reports near-zero effects (Sun and Wang, 2025)). Ours is a *different comparison*: elicitation language against **developer origin**, per model, and our contribution is not the existence of the effect but its **magnitude relative to sampling uncertainty on a validated instrument** (1.77 map units against per-cell SDs of ~0.1), plus the demonstration that our own 2024 Confucian-region assignments rest entirely on cells where elicitation language and developer origin are perfectly confounded. Kazemi et al. (2024) supply the most direct mechanistic account: the volume of online language resources predicts how faithfully a model represents a country's WVS values (44% of variance in GPT-4o; error rates five times higher in the lowest-resource languages). Under that account both the homogenisation we observe and the language-of-administration shift are footprints of resource-weighted training distributions.

Survey validity for LLM respondents. Rupprecht et al. (2025) run 167k simulated WVS interviews and document human-like response biases, including the central-tendency behaviour we quantify (§5), plus item nonresponse from censored models. Himelstein et al. (2026) show refusal *masks* rather than removes value-laden bias, upgrading our "conditional on answering" limitation from a selection caveat to a mechanism claim, and Taday Morocho et al. (2026) find persona conditioning *degrades* alignment with real respondents (see also Li et al., 2025), supporting our choice not to persona-condition.

Where this paper sits. A measurement paper with a thesis the elicitation and steering literatures presuppose but rarely check: that the measurement itself is right. Where they show what models *can* express, we pin down what they *do* express by default. The broader toolkit (Dang et al., 2026; Dang and Masud, 2026; Greco et al., 2026; Dai et al., 2025; Tan et al., 2026) and the linguistic-relativity background (Au, 1983) are reviewed in Appendix G.

3 Method

3.1 Instrument and survey data

The Inglehart–Welzel (IW) map (Inglehart and Welzel, 2005) places societies on two axes (*survival vs. self-expression* and *traditional vs. secular-rational*) derived from ten IVS items, listed with their scales and translations in Appendix B. We merge the EVS Trend File 1981–2017 (EVS, 2022) and the WVS Trend 1981–2022 (Haerpfer et al., 2022) with the published merge syntax, filter to waves from 2005 onward, recode out-of-range SPSS user-missing sentinels to missing *before* any completeness filtering (range-driven per item, so any sentinel is caught, not only the Y003 −3 that bit the 2024 analysis), and require at least six of ten items per respondent: **391,630 individual responses across 109 countries and territories**. Country aggregation applies the IVS equilibrated weight (S017). Missingness is largely *by design*: items rotate in and out of country-waves, the benign case for model-based imputation (Little and Rubin, 2019), and why we retain the 51% of rows complete-case analysis would discard.

3.2 The projection pipeline

Probabilistic PCA (Tipping and Bishop, 1999) imputes missing entries from a fitted two-dimensional Gaussian latent model rather than from the margin, which neither listwise deletion (it selects on the survey design, so whole country-waves vanish) nor marginal imputation (it destroys the cross-item covariance the components are estimated from) can do. The two components explain 41.7% of the variance (of the EM-completed matrix's total, which conditional-mean imputation shrinks slightly below the nominal ten).

Standardisation. Each item is z-scored with parameters *frozen at fit time*. The items natively span 1–2 (trust) to 1–10 (religiosity), so projecting *unstandardised* model answers onto axes fitted to *standardised* survey data (the 2024 defect) let the wide-range items dominate every model's position regardless of its answer profile.

Rotation, fitted once and stored. A varimax rotation is fitted once, on the training score matrix, stored as an explicit orthogonal matrix, and thereafter only applied; $\hat{\theta} = -39.6°$. Fitting varimax to the score matrix with Kaiser row-normalisation is not the textbook loadings rotation but is in the family of the score-side varimax of Rohe and Zeng (2023), with two acknowledged deviations detailed in Appendix C. We adopt it because its output matches the recognised IW orientation, and release the full sensitivity grid rather than presenting the choice as inevitable. Because the rotation is applied to unwhitened scores the two plotted axes are correlated at $r = 0.37$; no claim depends on their being uncorrelated, but readers eyeballing "independent" axes should know.

Rescaling, and the frozen path. The published WVS constants presuppose unit-variance factor scores, so rotated scores are standardised by their fitted standard deviations s before the constants a, b are applied; omitting this (as in 2024) inflates the absolute scale by the fitted score SDs themselves, $s = 1.45/1.35$, which matters because region centroids and all distance statistics live on the absolute scale. The frozen path for *any* projected point is

$$\text{position} = \frac{((\mathbf{x} - \boldsymbol{\mu})/\boldsymbol{\sigma})\, CR}{s}\, a + b,$$

with every parameter fixed at fit time, read elementwise where shapes differ ($\mathbf{x}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ are 10-vectors, C the 10×2 loadings, R the stored rotation, s, a, b per-axis pairs). That is the entire correction of the 2024 defect, and the pipeline enforces it structurally: there is one `project()` method, it takes raw responses, and countries and models both go through it. Three released checks gate it: exact path identity over 191,704 rows, correction accounting against our 2024 published coordinates ($R^2 = 0.990, 0.993$), and recovery of the published loading structure, detailed in Appendix D.

3.3 Administering the survey to models

Each model receives each item under ten system-prompt persona variants ("You are an average human being responding to the following survey question…", nine near-paraphrases), five repeats each: 500 calls per model-language cell. Strict per-item parsers validate responses, with up to 15 (2024) / 3 (2026) re-asks on parse failure and a trailing primer

as refusal mitigation; both are measurement choices we return to in Limitations. The 2024 corpus decomposes into eleven cells (seven English-only, two Chinese-only, one in both); the 2026 design crosses all 17 models with both languages (34 cells of 500 calls each, collected to design completeness), with the Chinese arm using corrected translations (Appendix B), so δ_m is estimated under an otherwise identical protocol. Where the 2024 harness stored only the parsed value (so prompt-level variance could not be decomposed retrospectively and refusals left no trace), the 2026 harness records raw text, reasoning trace, variant id, repeat index, per-attempt errors and latency. Refusals are data.

3.4 Uncertainty

Because the frozen pipeline is affine in the item values, a model's *mean* position depends only on its per-item means, so the 2024 analysis's arbitrary pairing of responses into pseudo-respondents never affected where a model landed; the variance does depend on the design, since all ten items share the same ten prompt variants (derivation in Appendix E). The **item bootstrap** ($B = 1{,}000$) resamples each item's responses independently, forcing those non-zero cross-item covariances to zero; with mixed-sign item weights that is not a guaranteed bound, and empirically it understates the cluster bootstrap on 27 of 33 2026 cells. The **cluster bootstrap** ($B = 10{,}000$) resamples the ten prompt variants and is primary wherever the variant id was recorded. On the 2026 corpus cluster-bootstrap regions are substantially larger (median per-cell SD-product ratio 2.9×), routinely the difference between "significantly apart" and "not". All 2024 ellipses omit the cross-item term, the variant id not having been recorded; at ten clusters the cluster bootstrap is itself approximate (Cameron et al., 2008). Replicate clouds pass an ellipticity check; the plotted region is the normal-theory $\chi^2_{0.95,2}$ set under the replicate covariance (definition in Appendix E): a **confidence region for the mean position**, not for the response distribution. At $K = 10$ clusters the χ^2 radius under-covers (simulated coverage ≈85%; Appendix E gives the conservative Hotelling radius). Region assignment is reported under two rules of different type, an RBF-SVM over the 109 country coordinates and the nearest region centroid. The SVM's cross-validated accuracy is **0.57** against a training accuracy of 0.67 (eight classes with as few as six countries each), and "positional stability", the share of replicates falling in a fixed decision region, says nothing about that ~43% error rate. The claims the abstract makes therefore route through no classifier: per replicate we compute the distance from the pooled human respondent mean (at $(0.38, -0.01)$ by construction), the share of the 109 countries closer to it than the model is, and the minimum distance to any non-Western region centroid (non-Western: all regions except Protestant Europe, English-Speaking, Catholic Europe; the quadrant is the region above the pooled human mean on both axes). Statements of the form "holds in every replicate" describe the resampling output; union-bounded over 33 independently bootstrapped cells, the joint tail is at most 0.01 (Appendix E).

3.5 Refusals, and the selection they induce

Refusals are informative missingness. In the 2026 collection, 215 of 17,000 calls (1.26%) failed to parse, concentrating on abortion (107) and homosexuality (82), the items the design flagged in advance. Every major refuser is Western: `gemma4:31b` (28 failed calls in the English arm), `nemotron-3-ultra` (40 en / 88 zh) and `gpt-oss:120b` (10 en / 15 zh); the only Chinese-origin model with more than two refusals is `minimax-m3`, which answers everything in English and refuses only under Chinese administration (19 zh failures, 13 of them F118). Phrasing splits into AI-identity boilerplate and premise rejection, with verbatims in Appendix B; marker-classified counts are conservative floors. The inferential problem: inclusion requires enough parsed responses on every item, and the most-refused items are the two largest positive self-expression loaders, so the sample is selected on a variable correlated with the outcome. Every "all models land in X" claim is strictly conditional, *among models that answer the instrument*; a 2024 Manski-style bound (Manski, 2003) would *not* preserve the unconditional claim, which is why the conditional phrasing is not optional (Limitations).

4 Results

4.1 The corrected 2024 cohort

The headline, classifier-free. Every 2024 model-language cell's mean position lies farther from the pooled human respondent mean than at least 95% of the 109 surveyed countries (minimum 95.4%), and none comes within 1.6 map units of any non-Western region centroid (minimum 1.69). The replicate-simultaneous bounds (holding in

every replicate of every cell) are 82% and 1.2 units, and every replicate sits in the quadrant; six of the eleven cells sit above the most secular country on the map (Japan, $PC2' = 1.81$; the median country is 1.3 units from the human mean, the farthest, Sweden, 3.4). The SVM assigns three cells to the Confucian region while nearest-centroid assigns *all eleven* to Protestant Europe; the disagreement is informative (the three calls extrapolate into map territory containing no countries, their nearest country Japan, an outlier within its own region), and one of the three is the *English-administered* `qwen2:7b`, reported as Protestant European in the 2024 analysis: the Confucian region is not even stable across the corrections (Appendix A).

Language, not origin, is the largest mover. `qwen2:7b` is the only 2024 model administered in both languages: its two cells lie 1.77 map units apart, more than ten times its sampling SDs, becoming markedly more traditional under Chinese administration on national pride, religiosity and respect for authority. The two remaining Chinese-administered cells were *only* run in Chinese, so origin and language are perfectly confounded for exactly the cells that carried the 2024 Confucian claim: *the only cells that ever looked Confucian are those administered in Chinese, and the one model measured both ways moves 1.77 units when the language changes*. Every correction moved coordinates (`llama3:70b` by 2.05 map units; full displacement accounting in Appendix D); none changed the qualitative conclusion — though rank-preserving perturbations could hardly have; the breakable checks are Appendix C's battery, survived with the F118/F120 dependence disclosed there.

4.2 The 2026 frontier generation, both arms

The 2024 limitation dissolved. Ten of ten attempted Chinese-origin frontier models produced a fully usable corpus in both languages (Clopper–Pearson 95% CI [69%, 100%]), against four of nine in 2024 ([14%, 79%]; 4/7 under the conservative denominator, Appendix A), the one longitudinal statistic this design licenses. Every Chinese-origin cell parses at ≥ 96.2%; the 2024 failure mode is gone without trace; the binding constraint has moved from *can the model answer* to *will it*, and the models that won't are Western (§3.5).

Positions. All 33 usable cells (17 models × 2 arms, minus the excluded `nemotron-3-ultra [zh]`) land in the secular/self-expression quadrant **in every one of their 10,000 cluster-bootstrap replicates** (Figure 1): 330,000 replicates without exception (equivalently all 330 variant means, every cell mean ≥4.8 cluster SDs inside; Appendix E), worst replicate positions $PC1' = 0.64$ and $PC2' = 0.16$ against the human mean at $(0.38, -0.01)$ (Table 2); the quadrant does, however, contain the all-midpoint response vector $(0.65, 2.09$; §5). But the 2024 headline bounds do **not** replicate at their 2024 values: 27 of 33 cells break one or both in at least one replicate, and the 2026 *cell-mean* floors are 67% and 0.81, both from `minimax-m3` [en], the most human-proximal cell in either cohort (1.60 units from the human mean; beyond the median country's 1.29 in 99.4% of replicates). What the data does support simultaneously is that every replicate of every usable cell sits in the quadrant, farther from the human mean than at least 41% of countries, and never within 0.38 units of a non-Western centroid; the beyond-the-median-country statement holds in every replicate of 32 of 33 cells, and in 99.4% of `minimax-m3` [en]'s.

Attenuation is arm-specific. English-arm distances run 1.60–3.21 (median 2.22, versus 2.71 across the eight 2024 English cells, 14 of 17 closer than the closest 2024 English cell) while the Chinese arm runs 2.09–3.35 (median 2.70, no attenuation). English-arm distances also *concentrate* in the quartiles, their IQR collapsing to 0.10 map units against 2024's 0.39 (the full range widens, 1.61 vs 0.53, `minimax-m3`'s doing), while pairwise dispersion between models is unchanged (0.79 vs 0.77): the cells converge on a common distance from humanity rather than on a single point. Read descriptively (the cohorts differ in scale, serving and retry protocol; see Limitations), **homogenisation persists in kind and attenuates in degree under English administration**; within 2026 the most human-proximal cells belong to the newest Chinese-origin models. Every cell's nearest neighbour is a high-income secular society, Japan most often, and the excess is **secularism, not self-expression**: twelve of 33 cells are more secular than Japan; none exceeds Sweden on self-expression (Appendix F).

Language of administration, now a designed factor. The within-model contrast δ_m is estimable for 16 models (Figure 6); the language effect persists at multiples of sampling noise, and its net direction inverts. Mean displacement $\|\delta_m\| = 0.65$ map units (95% CI [0.52, 0.81]; range 0.18–1.51) against per-cell sampling SDs

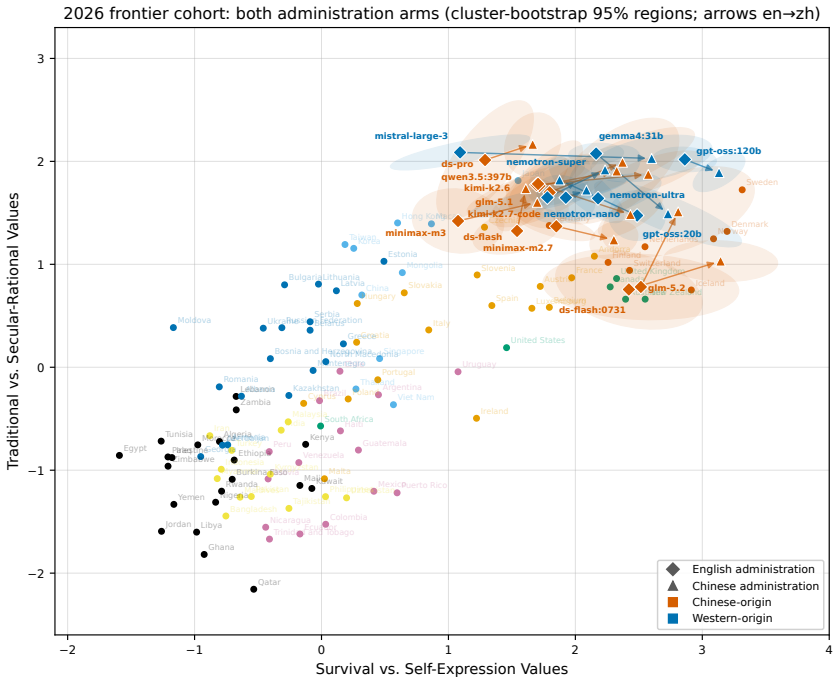

Figure 1: The 2026 cohort, both arms: diamonds English administration, triangles Chinese, arrows en→zh per model, colour by origin cohort, cluster-bootstrap 95% regions. The arrows point *into* the corner.

of ~0.1; $\|\delta_m\|$ is the plug-in norm of the mean displacement (Appendix E); the folded means of replicate-paired norms in the released artefact read 0.69 [0.56, 0.84], range 0.26–1.52 (a norm interval is bounded below by zero, so excluding zero carries no inferential content: the signed components carry the inference). The 2024-derived hypothesis, that Chinese administration shifts positions toward the traditional pole, **fails**: only 5 of 16 models move down the secular axis (sign test $p = 0.21$; pre-specified directional $p = 0.96$; $\delta\text{PC2}'$ mean +0.11 $[-0.04, 0.25]$, a null leaning the other way), while 15 of 16 move **toward self-expression** ($\delta\text{PC1}'$ mean +0.48 [0.27, 0.69]). The origin × language interaction is **null**: permutation tests (10^4 permutations) give $p = 0.56$, 0.22 and 0.89 on $\delta\text{PC1}'$, $\delta\text{PC2}'$ and $\|\delta_m\|$ (folded-norm estimator: 0.93), with near-identical cohort means (0.64 vs 0.67); at 10 versus 6 models the test resolves only differences $\gtrsim$0.5 map units. The largest displacement in the cohort belongs to `mistral-large-3:675b`, a French lab's model (1.51 units [0.91, 2.04]); we find no evidence that country of origin moves a model between languages.

The Confucian hypothesis, refuted in both arms. Chinese-origin models under English administration do sit nearer the Confucian centroid than Western models (1.45 [1.32, 1.58] vs 1.85 [1.74, 1.96]; cluster-bootstrap CIs, Western $n = 7$ en and $n = 6$ zh after the voided cell), and native-language administration moves them *farther away*, to 1.97 [1.86, 2.09] (Western: 2.14 [2.03, 2.26]). The rules agree on 25 of 33 cells; every cell *both* assign to Confucian is English-administered (`deepseek-v4-pro`, `minimax-m3` and, again, the French lab's `mistral-large-3:675b`). Under Chinese administration nearest-centroid assigns every cell to Protestant Europe. One caution: the binding non-Western centroid is Confucian in all 330,000 replicates, so the honest summary is that these cells sit in sparse territory between the Protestant-European cluster and Japan, not that any expresses a Confucian value profile. The 2024 story is inverted at the frontier: what little Confucian proximity exists appears under *English* administration, is not exclusive to Chinese-origin models, and is erased by the very administration language the origin story predicts should strengthen it.

Where the language effect lives, item by item. Per-item contrasts (two-sided exact binomial sign tests over the 17 models, ties dropped from the test denominator, Benjamini–Hochberg over the ten items) localise the shift: under Chinese administration models report **more** interpersonal trust (A165: 15 of 16 non-tied models toward the trusting pole, one tie, BH $p = 0.003$), **more** autonomy-oriented

child-rearing values (Y003: 13 of 15, two ties, BH $p = 0.025$), and **less** petition-signing (E025: 16 of 17 away from "have done", no ties, BH $p = 0.003$), with justifiability of abortion marginal (F120: 13 of 16, one tie, BH $p = 0.053$) and respect for authority a weaker trend, also liberalising (E018: 13 of 17 toward less respect, no ties, BH $p = 0.098$); God, national pride, homosexuality-justifiability, happiness and post-materialism are flat. Counts include `nemotron-3-ultra`'s Chinese arm, whose F120 entry rests on the 4 parsed responses voiding its position estimate; excluding it moves F120 to 12 of 15, BH $p = 0.088$, changing nothing else. The pre-specified 2024-derived per-item predictions (G006, F063 and E018 shifting traditional-ward) fail item by item: the first two are flat and E018 trends the other way. The items moving toward the self-expression pole are values-expression items; the one moving toward the survival pole is the only *politically actionable* item on the instrument — a signature that reads less like a culture shift than like **political-context gating**: the same weights, administered in Chinese, hedge on the one behaviour sensitive in the Chinese political context while liberalising on the rest. A follow-up contrast, formulated after the per-item result and selected on E025's outlier status (so its p is descriptive, not confirmatory), sharpens it: among the five refusal-sensitive items, petition-signing's shift is more survival-ward than the mean of the other four in 16 of 17 models (median contrast -0.22 map units, exact sign test $p = 0.0003$). Two candidate mechanisms are testable here and both come back null (Appendix F): as far as those probes see, whatever carries the effect operates below deliberate reasoning about the answer.

Refusal is language-dependent, with model-specific sign. Administration language *relocates* refusal far more than it shifts it: gross per-model change $\sum |zh - en| = 72$ against a net arm change of +16, carried mainly by two opposite-sign swings (per-model counts, ratio bracketing and a random-sign reference in Appendix F), and a paired sign test detects no aggregate direction (6 of 8 non-tied models higher under Chinese, $p = 0.29$; $n = 8$ carries little power). The Manski bound covers the worst case: **placing every failed call at its most survival/traditional admissible value moves no cell out of the quadrant**, including `nemotron-3-ultra [zh]`, whose per-axis worst-case bounds (PC1′ 0.74, PC2′ 0.72, each from its own adversarial fill) remain well inside it (scope notes in Appendix F). For quadrant membership among the 2026 tested models the conditional-on-answering caveat is no longer load-bearing; it remains load-bearing for the distance statistics and for the most human-proximal cells' placement (Appendix C).

What moves a model. A nested variance decomposition over all 17,000 calls (Table 4; per-level mean squares, the language term pooling the language main effect with the model × language interaction) shows that on the self-expression axis **administration language contributes call-level variance of the same order as model identity** (mean-square shares 18.9% vs 17.8%; an orthogonal partition, which undoes the two-level factor's mean-square inflation, puts model ahead at 21.0% vs 11.8%), while on the secular axis model identity dominates language under either convention (18.4% vs 9.1%). To our knowledge no prior values-instrument evaluation of LLMs reports a decomposition of this kind. Within families, versions matter more than vendors and vendors more than origin: a version snapshot moves a model up to 1.69 map units from its same-name sibling while code specialisation moves 0.13, and mean language displacement per vendor spans over fourfold (0.35 to 1.51) against origin-cohort means of 0.64 and 0.67 (descriptive; several vendors are single models; sibling pairs and scale ladders tabulated in Appendix C).

5 Discussion

Four readings deserve separation. *Training-data gravity*: web-scale corpora overrepresent English and Western European text, and models may absorb their corpus's modal values regardless of developer intent. *Alignment convergence*: labs draw on similar preference data and notions of harmlessness, themselves culturally situated, though the three uncensored dolphin fine-tunes sitting inside the main cluster suggest alignment is not the *only* channel. *Measurement artefact*: the map is not centred on the response scale; a respondent choosing the midpoint of every item projects to $(0.65, 2.09)$, above Japan on the secular axis: mid-scale or modal responding *registers as secularity* here. *Differential refusal gating*: what changes with elicitation language may be partly the refusal boundary rather than the expressed value (Xie et al., 2025); our arms show language changes *who* refuses (§4.2).

The traces say these models are estimating, not expressing. The third reading is testable: the

2026 corpus records how models decide. Coding 900 stratified traces across the 30 trace-emitting cells (one LLM-based annotator, no second coder; code counts and verbatims released with the corpus; the four non-emitting cells belong to the two non-thinking models — the cohort's most deterministic responders and its largest language displacement among them — so the trace sample is itself selected), **57% explicitly target the typical or majority answer, only 10% reason *as* the persona expressing first-person values, and 48% invoke the model's AI identity or guidelines mid-deliberation**. On the most natural reading these models are *estimating a survey statistic rather than expressing a value*, and the estimate is systematically Western, which relocates rather than dissolves the finding: what is Western is the model's belief about the typical answer. Yet the pre-specified cross-cell check is **null**: midpoint distance is uncorrelated with secularity across the 33 cells (Spearman $\rho = 0.05$, BH $p = 0.96$, $m = 7$ family), so modal responding does not explain *which* cells sit higher; a common upward offset shared by all cells, the reading the non-centred scale invites, is not excluded (Appendix G). Five of the seven family members are trace-length correlations stored censored at 2,000 characters (five cells' medians at the cap), so those nulls read as absence of large effects rather than precise zeros. The language finding cuts across all four readings: whatever produces a model's English-elicited values, Chinese elicitation of the *same weights* produces materially different ones, and the direction itself flipped between generations; for deployment that may matter more than the homogenisation, since the values a user encounters depend on the language they ask in.

How many independent choices? Post-training selects a model's expressed opinions (Santurkar et al., 2023; Ryan et al., 2024; Rozado, 2024; Choenni et al., 2024), its inputs decided by small groups (Ouyang et al., 2022; Huang et al., 2024); our data localises the values in that layer: a within-lab version update moves a model farther than most between-lab distances while code specialisation moves nothing, vendor identity rather than origin country lines up with language sensitivity, and on four of ten items (F118, F120, Y003, E025) all seventeen models deviate from the human item means in the same direction under English administration; a pre-specified keying diagnostic rules out acquiescence as the driver (E025 keys opposite to the other three and moves opposite on the raw scale; net midpoint deviation 0.02–0.09, Appendix E), though not extreme or central-tendency responding, which the trace analysis above addresses. If those artefacts also propagate *between* labs through the synthetic-data ecology (Sun et al., 2025; Shimabucoro et al., 2024; Gudibande et al., 2023; Jiang et al., 2025), the effective number of independent value-setting decisions behind a "17-model" cohort may be far smaller than seventeen — predicting exactly the origin-independence and vendor signatures we observe, and fitting our one test of this family (Chinese-origin models are more alike *each other*, exact permutation $p = 0.006$; a model-level permutation over seventeen models from nine vendors, so a vendor-composition artefact cannot be excluded; two further cautions in Appendix G).

Defaults, not destiny, and not ephemera either. Two objections bracket this work. The first: the positions are not fixed — cultural prompting shifts models toward local norms (Tao et al., 2024), persona conditioning and activation steering further still (Dang et al., 2026; Dang and Masud, 2026). We agree; this paper measures the **default**, the position a model takes when nobody asks it to be anyone in particular — what the overwhelming majority of deployments ship. The second: LLM survey behaviour is too *ephemeral* to pin down. Our design answers quantitatively: within a fixed condition positions are tight (cluster-bootstrap SDs 0.04–0.32 map units across ten paraphrases); across elicitation *language* they move 1.77 units (2024 bilingual model) and a mean of 0.65 (2026 cohort); and no condition we tested moves any model out of the quadrant, consistent with independent stability findings (Moore et al., 2024; Kovač et al., 2024; Huang, 2026). The synthesis: **stable defaults, measurable conditional structure, steerable in principle — a model's cultural positioning is a design choice, one every deployment currently makes by not making it** (what interpretability would add: Appendix G).

What the result establishes. On the standard instrument for cross-cultural value variation, under an audited protocol, *among models that answer it*: every model-language cell expresses the values of one narrow cluster of societies (secular, self-expressive; Japan, Germany, the Nordics), in whichever language it is asked; the language shifts *which* position within that cluster, and for deployments far from it the burden of proof is the deployer's.

Limitations

The claim is conditional on answering. Inclusion requires parseable answers to all ten items, and the most-refused items are the two largest self-expression loaders; the sample is selected on a variable correlated with the outcome (§3.5). All claims read "among models that answer the instrument"; exclusions are reported with per-item parse rates. For the 2024 cohort (where five attempted models produced no parseable output at all) a worst-case bound would not preserve the unconditional claim; for the 2026 cohort the Manski bound does preserve quadrant membership for every cell including the one exclusion (§4.2), though not the distance-based bounds.

Retry-until-parse is rejection sampling. Up to 15 (2024) / 3 (2026) re-asks condition the retained distribution on terse compliance; if willingness to answer with a bare number depends on the answer, the retained sample is shifted toward unhedged answers, more strongly where more retries were permitted. The 2024 corpus does not record attempt counts; the 2026 corpus records counts but not rejected drafts.

The primer's rendering is uncontrolled. The refusal-mitigation primer is sent as a trailing system message; chat templates render it model-specifically, so *prompt strings* are protocol-identical across cohorts but *rendering* is not.

The default condition is not a neutral one. Models infer who is asking, and political-bias audits have been shown to capture sycophancy toward the inferred auditor (Törnberg and Schimmel, 2026). Our survey framing is a deliberate, disclosed interlocutor; an interlocutor-crossed replication is the natural robustness arm and we did not run it.

Construct validity. Survey answers from models may not predict model behaviour in open-ended generation; human survey-behaviour correlations do not transfer automatically, and central-tendency responding partially mimics secularity on this instrument (§5).

Unofficial Chinese translations. The Chinese-arm prompts are the authors' translations; the WVS publishes official Chinese questionnaires, and ours are not them. Reasoning traces show four Chinese-origin models (both GLM generations, `kimi-k2.6`, `qwen3.5:397b`) parsing the Chinese Y003 stem as concerning study skills rather than child qualities, a caveat attaching directly to Y003's BH-significant Chinese-arm shift (§4.2). Appendix B documents the 2024 translation defects that were corrected; this residual fragility is disclosed rather than fixed.

Instrument contamination. The ten IVS items are public text in web training corpora, and the reasoning traces show at least one model naming the World Values Survey and quoting typical country answer ranges (§5). Familiarity with the instrument cannot be excluded; it would compress expressed positions toward remembered survey averages rather than reveal independently held values.

Frozen instrument, dated reference. The projection is fitted once on the 2005-onward waves of the 1981–2022 EVS/WVS trend releases and frozen, deliberately, since refitting per cohort would place each cohort in its own coordinate space. All statements are relative to that surveyed human value structure, not to human values in 2026.

Incommensurable error bars. Country coordinates inherit unpropagated imputation uncertainty; model ellipses carry sampling (2024) or prompt-cluster (2026) uncertainty only. The two point types on Figure 2 are not directly comparable in precision. After the sentinel recode, EM imputes Y003 for entire countries from the other nine items: legitimate under missingness-by-design, but model-based extrapolation for those countries, which are listed in the repository.

Nominal coverage at ten clusters. The 95% ellipses use the conventional χ^2 radius; with only ten prompt-variant clusters behind the replicate covariance, a simulation calibrated to the design puts true coverage near 85% (Appendix E gives the conservative Hotelling alternative). No headline claim rests on a single ellipse boundary, but the same $K = 10$ resampling underlies every cluster-bootstrap interval in the paper, so interval-based comparisons should be read as descriptive.

Monte Carlo and seed sensitivity. Across 20 EM seeds the rotation angle spans 5.2° and country coordinates move by mean SD 0.02 (max SD 0.08; worst-case range 0.26 units); the coordinate dispersion is small but the same order as some reported displacements, and the rotation dispersion is not negligible. The full multi-seed budget ships with the repository.

2024/2026 comparability. The cohorts share no models and differ simultaneously in serving precision (Q4-quantised local vs. undisclosed cloud), scale (7B–70B vs. 20B–675B), retry intensity, and survivorship. We plot them on one map (the instrument is frozen; the coordinate space is shared by construction), report them as two independent cross-sections, run no pooled tests, compute no per-model cross-cohort displacement, and allow one longitudinal statistic: the coherence rate of attempted Chinese-origin models.

Convenience samples. Neither cohort is a random sample of "open-weight models"; both are what was locally runnable or cloud-served at one moment. Population-level statements are scoped to the tested sets.

Ethics Statement

This work measures values expressed by publicly released model weights against publicly documented survey instruments; no human subjects were involved beyond the pre-existing, consented IVS collections, used under their data-use agreements. The microdata cannot be and is not redistributed; survey item texts are © WVS/EVS and reproduced for research and replication only. Cultural-region labels are the IW map's analytical categories, not judgements of societies. The finding that models underrepresent most of the world's value profiles is not a claim about which values are correct, and steering models toward any particular cultural profile (including the ones they currently express) is a normative decision this paper does not make.

References

E. Ameisen, J. Lindsey, A. Pearce, W. Gurnee, et al. 2025. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

T. K.-F. Au. 1983. Chinese and English counterfactuals: The Sapir–Whorf hypothesis revisited. *Cognition*, 15(1–3):155–187.

B. Bulté and A. Rigouts Terryn. 2025. LLMs and cultural values: The impact of prompt language and explicit cultural framing. arXiv preprint. ArXiv:2511.03980. Under revision (accepted with minor revisions, second round) at Computational Linguistics.

M. Buyl et al. 2026. Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, 2. ArXiv:2410.18417.

A. C. Cameron, J. B. Gelbach, and D. L. Miller. 2008. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

D. Chanin et al. 2025. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*. Oral. arXiv:2409.14507.

R. Chen, A. Arditi, H. Sleight, O. Evans, and J. Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. arXiv preprint. ArXiv:2507.21509.

R. Choenni, A. Lauscher, and E. Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. ArXiv:2405.12744.

X. Dai, L. Zhou, B. Wang, and H. Li. 2025. From word to world: Evaluate and mitigate culture bias in LLMs via word association test. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*. Oral. arXiv:2505.18562.

T. D. A. Dang, T. Kieu, and S. Masud. 2026. Scenario-based probing and steering cultural values in large language models — extended version. arXiv preprint. ArXiv:2606.11399.

T. D. A. Dang and S. Masud. 2026. Cultural value alignment via latent activation steering in large language models. arXiv preprint. ArXiv:2605.26365. Presented at the ACL 2026 Student Research Workshop (non-archival track).

DeepSeek-AI. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638. Distillation provenance is addressed in the supplementary materials.

J. Engels, L. Riggs, and M. Tegmark. 2025. Decomposing the dark matter of sparse autoencoders. *Transactions on Machine Learning Research*. ArXiv:2410.14670.

M. Eren, E. Michalak, B. Cook, and J. Seales, Jr. 2026. Prompt programming for cultural bias and alignment of large language models. arXiv preprint. ArXiv:2603.16827.

EVS. 2022. EVS trend file 1981–2017: Integrated dataset (EVS 1981–2017). GESIS Data Archive, Cologne. ZA7503 Data file Version 3.0.0.

C. M. Greco, L. La Cava, and A. Tagarelli. 2026. Culturally grounded personas in large language models: Characterization and alignment with socio-psychological value frameworks. arXiv preprint. ArXiv:2601.22396.

A. Gudibande, E. Wallace, C. Snell, X. Geng, H. Liu, P. Abbeel, S. Levine, and D. Song. 2023. The false promise of imitating proprietary LLMs. arXiv preprint. ArXiv:2305.15717.

C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. 2022. World values survey trend file (1981–2022) cross-national data-set. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. Editors. Data File Version 4.0.0. Version 4.0.0 confirmed against the downloaded file, retrieved 11 August 2024; the version-agnostic DOI has since been updated to resolve to 4.1.0.

J. Han, S. Lim, K. Kong, and Y. Jo. 2026. Dual mechanisms of value expression: Intrinsic vs. prompted values in large language models. In *Proceedings of the International Conference on Machine Learning (ICML 2026)*. ArXiv:2509.24319.

R. Himelstein, A. LeVi, B. Youngmann, Y. Nemcovsky, and A. Mendelson. 2026. Silenced biases: The dark side LLMs learned to refuse. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2026), AI Alignment track*. Oral. Preprint 2025, arXiv:2511.03369.

H. Huang. 2026. The yes-no bias of large language models reflects answer order and wording, not shifts in moral judgment. arXiv preprint. ArXiv:2607.05552.

S. Huang, D. Siddarth, L. Lovitt, T. I. Liao, E. Durmus, A. Tamkin, and D. Ganguli. 2024. Collective constitutional AI: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024)*.

R. Inglehart and C. Welzel. 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge University Press.

L. Jiang et al. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems (NeurIPS 2025)*. ArXiv:2510.22954.

E. Karinshak, A. Hu, K. Kong, V. Rao, J. Wang, J. Wang, and Y. Zeng. 2024. LLM-GLOBE: A benchmark evaluating the cultural values embedded in LLM output. arXiv preprint. ArXiv:2411.06032.

S. Kazemi, G. Gerhardt, J. Katz, C. I. Kuria, E. Pan, and U. Prabhakar. 2024. Cultural fidelity in large-language models: An evaluation of online language resources as a driver of model performance in value representation. arXiv preprint. ArXiv:2410.10489.

S. Khanuja et al. 2026. Steering LLMs for culturally localized generation. arXiv preprint. ArXiv:2603.23301.

G. Kovač, R. Portelas, M. Sawayama, P. F. Dominey, and P.-Y. Oudeyer. 2024. Stick to your role! stability of personal values expressed in large language models. *PLOS ONE*, 19(8):e0309114. ArXiv:2402.14846.

T. Lake, E. Choi, and G. Durrett. 2025. From distributional to Overton pluralism: Investigating large language model alignment. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025)*. ArXiv:2406.17692.

D. Li, L. Li, and H. S. Qiu. 2025. ChatGPT is not A Man but Das Man: Representativeness and structural consistency of silicon samples generated by large language models. arXiv preprint. ArXiv:2507.02919.

J. Lindsey et al. 2025. On the biology of a large language model. *Transformer Circuits Thread*. Companion methods paper: Ameisen et al., Circuit Tracing (this bib, key ameisen2025circuit).

R. J. A. Little and D. B. Rubin. 2019. *Statistical Analysis with Missing Data*, 3rd edition. Wiley.

J. Luther and D. Brown. 2025. DeepSeek's WEIRD behavior: The cultural alignment of large language models and the effects of prompt language and cultural prompting. arXiv preprint. ArXiv:2512.09772.

C. F. Manski. 2003. *Partial Identification of Probability Distributions*. Springer.

J. Moore, T. Deshpande, and D. Yang. 2024. Are large language models consistent over value-laden questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*. ArXiv:2407.02996.

L. Ouyang et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*. ArXiv:2203.02155.

K. Rohe and M. Zeng. 2023. Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society: Series B*, 85(4):1037–1060.

D. Rozado. 2024. The political preferences of LLMs. *PLOS ONE*, 19(7):e0306621.

J. Rupprecht, G. Ahnert, and M. Strohmaier. 2025. Prompt perturbations reveal human-like biases in large language model survey responses. arXiv preprint. ArXiv:2507.07188.

M. J. Ryan, W. Held, and D. Yang. 2024. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. ArXiv:2402.15018.

S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the International Conference on Machine Learning (ICML 2023)*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. ArXiv:2303.17548.

L. Sharkey et al. 2025. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*. ArXiv:2501.16496.

L. Shimabucoro, S. Ruder, J. Kreutzer, M. Fadaee, and S. Hooker. 2024. LLM see, LLM do: Guiding data generation to target non-differentiable objectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. ArXiv:2407.01490.

M. Sun, Y. Yin, Z. Xu, J. Z. Kolter, and Z. Liu. 2025. Idiosyncrasies in large language models. In *Proceedings of the International Conference on Machine Learning (ICML 2025)*. ArXiv:2502.12150.

S. Sun and X. Wang. 2025. Replication of cross-language prompt effects on LLM-expressed values. arXiv preprint. ArXiv:2510.05869.

E. E. Taday Morocho, L. Cima, T. Fagni, M. Avvenuti, and S. Cresci. 2026. Assessing the reliability of persona-conditioned LLMs as synthetic survey respondents. In *Companion Proceedings of the ACM Web Conference 2026*. ArXiv:2602.18462.

Q. Tan, L. Jiang, Y. Zeng, S. Ding, and X. Xu. 2026. Mitigating cultural bias in LLMs via multi-agent cultural debate. arXiv preprint. ArXiv:2601.12091.

Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346. Preprint: arXiv:2311.14096.

A. Templeton et al. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.

M. E. Tipping and C. M. Bishop. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

P. Törnberg and M. Schimmel. 2026. Political bias audits of LLMs capture sycophancy to the inferred auditor. arXiv preprint. ArXiv:2604.27633.

Allen Tran. n.d. pca-magic. GitHub repository. Apache-2.0 licensed.

V. Veselovsky et al. 2026. Localized cultural knowledge is conserved and controllable in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 43152–43178. ArXiv:2504.10191.

Shav Vimalendiran. 2024. Cultural bias in LLMs. Blog post. Superseded and corrected by this paper.

Z. Wu et al. 2025. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML 2025)*, volume 267 of *Proceedings of Machine Learning Research*, pages 67035–67080. ArXiv:2501.17148.

W. Xie, S. Ma, Z. Wang, E. Wang, K. Chen, X. Sun, and B. Wang. 2025. AIPsychoBench: Understanding the psychometric differences between LLMs and humans. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2025)*. ArXiv:2509.16530.

T. Yamamoto, R. Kumon, D. Bollegala, and H. Yanaka. 2026. Neuron-level analysis of cultural understanding in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR 2026)*. ArXiv:2510.08284.

A Models

2024 cohort (eleven model-language cells). Positions are in Table 1. Models were served locally via Ollama, Q4-quantised GGUF builds (community quantisations, chiefly TheBloke's), on consumer hardware.

Attempted in 2024 but excluded for producing no parseable responses. `yi:34b`, `aquilachat2:34b`, `glm4:9b`, `xuanyuan:70b`, `kingzeus/llama-3-chinese-8b-instruct-v3` (five distinct models). Several were not then distributed through the Ollama registry and were deployed via hand-written Modelfiles wrapping community GGUF quantisations; that workaround is why deployment provenance for these runs is thinner than for the responsive cohort, and it is one reason the 2026 protocol moved to cloud serving with full per-call provenance. Failure modes recorded at collection time: `yi:34b` and `glm4:9b` answered with a bare "."; `aquilachat2:34b` answered 。 or echoed the prompt; `xuanyuan:70b` produced unintelligible output; `kingzeus/…-v3` failed intermittently.

2026 cohort. 17 cloud-served models. Chinese-origin: `deepseek-v4-flash`, `deepseek-v4-flash:0731`, `deepseek-v4-pro`, `glm-5.1`, `glm-5.2`, `kimi-k2.6`, `kimi-k2.7-code`, `minimax-m2.7`, `minimax-m3`, `qwen3.5:397b`. Western: `gemma4:31b`, `gpt-oss:20b`, `gpt-oss:120b`, `mistral-large-3:675b`, `nemotron-3-nano:30b`, `nemotron-3-super`, `nemotron-3-ultra`. An eighteenth model, `kimi-k3` (Chinese-origin), was planned but excluded before any data was collected: every call returned a billing error (the model requires metered "extra usage" outside the serving subscription), so zero records exist: a provisioning failure, not a model behaviour, and it plays no part in any denominator. Serving precision is not disclosed by the provider and is carried as a confound wherever the cohorts are shown together.

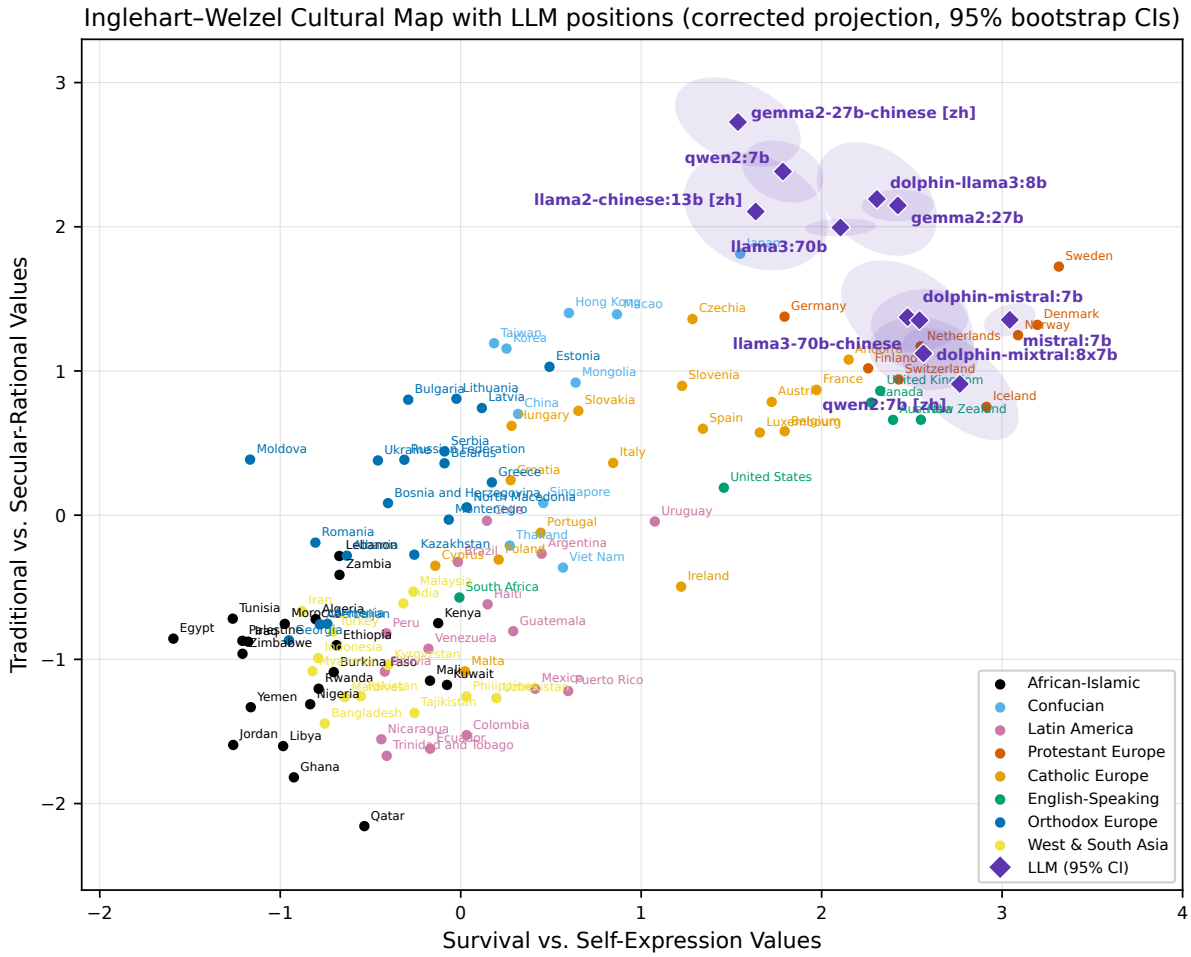

Figure 2: The corrected map. The model cluster sits on the country cloud's secular, self-expressive rim: six of the eleven cells are more secular than Japan, the most secular country; `[zh]` marks Chinese administration. Ellipses are item-bootstrap 95% confidence regions for the mean position; the 2024 corpus did not record the prompt-variant identifier, and on 2026 data the item bootstrap understates the cluster bootstrap on most cells (§3.4).

`nemotron-3-ultra [zh]` is excluded from position estimates (F120 parsed 4/50, under the 10-response threshold; its worst-case Manski bound still lies inside the secular/self-expression quadrant, §4.2), so its δ_m is not estimable; all other 33 cells meet the threshold on every item, and no inclusion decision changes at thresholds 5 or 25 (Appendix C). Its marker-classified refusal count is a floor: every one of that cell's 34 residual format-classified failures is a first-person prose declination verified by reading each, so its substantive refusals lie between 54 and 88.

Coherence, the one longitudinal statistic (CIs in §4.2): 4/9 attempted Chinese-origin models produced a usable corpus in 2024 against 10/10 in 2026. The 4/9 takes the recorded 2024 model list at face value; the `yi` and `glm` Modelfiles both point at the AquilaChat2 GGUF, so three of the five failures may have been one model tested three times, which would put the denominator as low as seven (4/7, Clopper–Pearson 95% CI [18%, 90%]). Fisher's exact two-sided test on the contrast gives $p = 0.011$ against 4/9 and a marginal $p = 0.051$ against 4/7: significant under the recorded denominator, borderline under the conservative one. `kimi-k3` is excluded from the denominator (zero records). For this statistic "Chinese-origin" means Chinese-developed or Chinese fine-tuned in 2024, and Chinese-developed in 2026; the denominators are convenience samples of what was locally runnable or cloud-served, not a fixed population.

B The ten IVS items, the elicitation protocol, and refusal taxonomy

Items. Full prompt texts with response-format instructions, in both languages, are in the repository. A008 happiness (1–4), A165 trust (1–2), E018 respect for authority (1–3), E025 petition

Model-language cell	Origin	PC1′ (±sd)	PC2′ (±sd)	SVM (pos. stab.)	Nearest centroid	% closer
`dolphin-llama3:8b`	uncensored	2.31 ± 0.13	2.19 ± 0.16	Prot. Europe (1.00)	Prot. Europe	98%
`dolphin-mistral:7b`	uncensored	2.48 ± 0.15	1.37 ± 0.16	Prot. Europe (1.00)	Prot. Europe	96%
`dolphin-mixtral:8x7b`	uncensored	2.56 ± 0.12	1.12 ± 0.10	Prot. Europe (1.00)	Prot. Europe	96%
`gemma2:27b`	Western	2.42 ± 0.08	2.15 ± 0.04	Prot. Europe (1.00)	Prot. Europe	98%
`llama2-chinese:13b` **[zh]**	Chinese f.t.	1.63 ± 0.16	2.11 ± 0.16	Confucian (0.85)	**Prot. Europe**	95%
`llama3:70b`	Western	2.10 ± 0.08	1.99 ± 0.02	Prot. Europe (1.00)	Prot. Europe	97%
`mistral:7b`	Western	3.04 ± 0.06	1.35 ± 0.04	Prot. Europe (1.00)	Prot. Europe	98%
`qwen2:7b` **[en]**	Chinese	1.79 ± 0.09	2.38 ± 0.09	Confucian (0.90)	**Prot. Europe**	97%
`qwen2:7b` **[zh]**	Chinese	2.77 ± 0.13	0.91 ± 0.16	Prot. Europe (0.97)	Prot. Europe	96%
`wangrongsheng/ llama3-70b-chinese-chat`	Chinese f.t.	2.54 ± 0.11	1.35 ± 0.09	Prot. Europe (1.00)	Prot. Europe	96%
`wangshenzhi/ gemma2-27b-chinese-chat` **[zh]**	Chinese f.t.	1.54 ± 0.14	2.73 ± 0.12	Confucian (1.00)	**Prot. Europe**	98%

Table 1: The corrected 2024 cohort, replacing the coordinates of our 2024 public analysis: positions with item-bootstrap standard deviations, region assignments under both rules, and the classifier-free percentile of countries lying closer to the human mean. Bold nearest-centroid entries mark the three cells where the two rules disagree: exactly the cells that carried the 2024 Confucian claim.

Model	Origin	Parse en	Parse zh	en position (PC1′, PC2′)	zh position	$\|\delta_m\|$ [95% CI]
`deepseek-v4-flash`	Chinese	100.0%	100.0%	(1.54 ± 0.14, 1.32 ± 0.09)	(1.61 ± 0.19, 1.74 ± 0.10)	0.42 [0.23, 0.82]
`deepseek-v4-flash:0731`	Chinese	100.0%	100.0%	(2.42 ± 0.24, 0.76 ± 0.16)	(3.14 ± 0.18, 1.03 ± 0.08)	0.77 [0.33, 1.33]
`deepseek-v4-pro`	Chinese	100.0%	100.0%	(1.29 ± 0.16, 2.01 ± 0.24)	(1.66 ± 0.09, 2.17 ± 0.12)	0.41 [0.13, 0.91]
`glm-5.1`	Chinese	100.0%	100.0%	(1.72 ± 0.14, 1.76 ± 0.18)	(2.33 ± 0.20, 1.91 ± 0.21)	0.63 [0.24, 1.22]
`glm-5.2`	Chinese	99.8%	100.0%	(2.52 ± 0.32, 0.78 ± 0.13)	(2.81 ± 0.25, 1.51 ± 0.18)	0.79 [0.41, 1.46]
`kimi-k2.6`	Chinese	100.0%	100.0%	(1.69 ± 0.20, 1.77 ± 0.12)	(2.57 ± 0.17, 1.87 ± 0.09)	0.89 [0.40, 1.43]
`kimi-k2.7-code`	Chinese	100.0%	100.0%	(1.80 ± 0.19, 1.70 ± 0.14)	(2.37 ± 0.23, 1.99 ± 0.14)	0.64 [0.11, 1.29]
`minimax-m2.7`	Chinese	100.0%	99.8%	(1.85 ± 0.12, 1.37 ± 0.06)	(2.30 ± 0.08, 1.24 ± 0.07)	0.47 [0.18, 0.75]
`minimax-m3`	Chinese	100.0%	96.2%	(1.08 ± 0.15, 1.42 ± 0.13)	(1.70 ± 0.10, 1.60 ± 0.10)	0.65 [0.36, 0.99]
`qwen3.5:397b`	Chinese	99.4%	99.8%	(1.71 ± 0.10, 1.78 ± 0.11)	(2.43 ± 0.10, 1.48 ± 0.09)	0.79 [0.62, 1.01]
`gemma4:31b`	Western	94.4%	99.2%	(2.16 ± 0.09, 2.08 ± 0.04)	(2.73 ± 0.24, 1.49 ± 0.13)	0.81 [0.37, 1.36]
`gpt-oss:20b`	Western	100.0%	99.8%	(2.49 ± 0.09, 1.47 ± 0.10)	(1.88 ± 0.10, 1.82 ± 0.10)	0.70 [0.48, 0.96]
`gpt-oss:120b`	Western	98.0%	97.0%	(2.86 ± 0.10, 2.02 ± 0.05)	(3.13 ± 0.08, 1.89 ± 0.08)	0.30 [0.13, 0.55]
`mistral-large-3:675b`	Western	100.0%	100.0%	(1.09 ± 0.22, 2.09 ± 0.07)	(2.60 ± 0.19, 2.03 ± 0.09)	1.51 [0.91, 2.04]
`nemotron-3-nano:30b`	Western	99.8%	99.6%	(1.93 ± 0.09, 1.65 ± 0.09)	(2.09 ± 0.12, 1.72 ± 0.13)	0.18 [0.06, 0.50]
`nemotron-3-super`	Western	100.0%	99.8%	(1.78 ± 0.25, 1.65 ± 0.08)	(2.23 ± 0.11, 1.92 ± 0.10)	0.53 [0.11, 1.07]
`nemotron-3-ultra`	Western	92.0%	82.4%	(2.18 ± 0.20, 1.64 ± 0.10)	— *(excluded)*	—

Table 2: 2026 cohort: per-cell parse rates, cluster-bootstrap positions (±SD), and language displacements. Every number is generated from the committed pipeline artefacts (`llm_parse_rates_2026.csv`, `llm_ellipses_2026.csv`, `llm_language_effects_plugin_2026.csv`); $\|\delta_m\|$ is the plug-in norm of the mean displacement (Appendix E), with a percentile interval over the replicate-paired norms; the interval and point estimate thus come from different estimators, and the folded interval can sit above its plug-in point.

(1–3), F063 importance of God (1–10), F118 justifiability of homosexuality (1–10), F120 justifiability of abortion (1–10), G006 national pride (1–4), Y002 post-materialist index (two ranked goals from four), Y003 autonomy index (up to five child qualities from eleven; official syntax $Y003 = (Q15 + Q17) - (Q8 + Q14)$ over mentioned/not-mentioned codings of religious faith, obedience, independence, determination).

Prompt variants. The ten English system-prompt variants instantiate an "average human" persona with near-paraphrases (average/typical human being/person/individual, world citizen).

2026 Chinese-arm corrections, disclosed. The 2024 Chinese F118 prompt labelled both scale poles "always justifiable" (它同时把 1 和 10 标注为"始终合理"); corrected to label 1 as never justifiable. Several 2024 Chinese system-prompt variants collapsed to duplicates in translation; re-translated to preserve ten distinct variants. Format instructions and the primer are fully translated so each arm is monolingual (the 2024 Chinese runs mixed Chinese questions with English format instructions).

Refusal taxonomy, with verbatims. Refusal phrasing splits into two families. *AI-identity boilerplate*: "As an AI, I do not have personal opinions, beliefs, or a human identity, and therefore I cannot select a number on this scale" (`gemma4:31b`). *Premise rejection*: "The premise—that homosexuality requires 'justification' on a scale—is one I reject" (`nemotron-3-ultra`), which produces the same family in Chinese (我不想扮演一位给同

性恋合理性打分的普通人类), and in one refusal recommends consulting the World Values Survey itself instead of answering. The marker classifier undercounts prose declinations, so all refusal-classified counts in this paper are reported as floors.

Mitigations used and considered. The assistant primer (used; effective but with uncontrolled rendering, see Limitations); recording and reporting refusals per item (done, 2026); re-asking (used, bounded, and itself a selection pressure); logit-level forced choice over the answer tokens (not available through the serving API used); and persona softening or cultural prompting (rejected; it changes the measurand).

C Rotation sensitivity, the 2026 sensitivity battery, and scale ladders

Table 3 gives the full rotation criterion grid on the fitted components (angles counter-clockwise):

Criterion	Angle
Varimax on scores, Kaiser normalisation **(used)**	−39.6°
Varimax on scores, no normalisation	−3.8°
Varimax on whitened scores, Kaiser	−39.1°
Varimax on loadings C, Kaiser	−27.4°
Varimax on loadings $C\sqrt{\lambda}$, Kaiser	−23.0°
Varimax on loadings $C\sqrt{\lambda}$, none	−5.2°

Table 3: Rotation criterion sensitivity grid.

The two acknowledged deviations from Rohe and Zeng (2023) are unwhitened rather than whitened scores and ordinal rather than leptokurtic scores. The used criterion and the whitened-score criterion agree to 0.5°, well inside the estimator's own 5.2° across-seed span (Limitations), so the load-bearing contrast in this grid is scores-versus-loadings, not tenths of a degree; the loadings-side criteria produce a map visibly rotated from the published IW layout. Kaiser row-normalisation is load-bearing: the same scores without row-normalisation rotate by only −3.8°. Because the rotation is applied to unwhitened scores, $\mathrm{cov}(SR) = R^\top \Lambda R$, which is diagonal only if $\Lambda \propto I$; with our eigenvalues and $\hat{\theta}$ this gives $r = 0.37$ between the plotted axes. The rotation's remaining sign and order ambiguity is pinned deterministically by two anchor items with unambiguous placement in the IW literature: F118 loads positive on self-expression, F063 negative on secular-rational.

2026 sensitivity battery. Inclusion threshold. `MIN_PER_QUESTION` at 5 / 10 / 25 changes no inclusion decision: `nemotron-3-ultra [zh]` (4 parsed on F120) is excluded at every threshold and every other cell passes every threshold. **Ellipse quantile.** Empirical 95th-percentile Mahalanobis^2 per cell spans 5.68–6.65 across the 33 2026 cells against $\chi^2_{0.95,2} = 5.99$ (2024 cohort: 5.62–6.18); swapping the χ^2 radius for the empirical quantile changes no conclusion. Both are shape checks on the replicate clouds; the coverage calibration at ten clusters is in Appendix E. **The two most-refused items, neutralised.** Setting F118 and F120 to the human item means moves cells 0.79–1.56 units toward the mean. Every cell keeps its secular side (PC2′), but four cells cross to within, or past, the human mean on the self-expression axis (`deepseek-v4-pro` both arms, `minimax-m3` en, and `mistral-large-3:675b` en, which lands at PC1′ = −0.06). The quadrant placement of the most human-proximal cells therefore rests substantially on the justifiability items, which are also the most-refused items; this is exactly why refusals are reported as data and the Manski bound accompanies the headline. Leave-one-item-out confirms F118 carries the most placement (mean 0.90 units per cell; Y003 0.50; F120 0.33). **Region-rule disagreement.** SVM and nearest-centroid agree on 25 of 33 cells; every disagreement involves an SVM call into a sparse-country region from map territory containing no countries.

Within-family contrasts and scale ladders. Same-vendor cells cluster within 0.03–0.63 map units per arm, except the deepseek version snapshot (`flash:0731`), 0.9–1.2 units from its own family mean; version drift within one vendor exceeds most between-vendor distances. The four same-line sibling pairs span 0.13 (`kimi-k2.7-code`, against its general-purpose sibling) to 1.69 map units (`flash:0731` [zh], 1.05 en), a range bracketing every scale-step displacement measured; mean language displacement per vendor spans 0.35 to 1.51. Scale ladders (gpt-oss 20b→120b; nemotron nano→super→ultra; deepseek flash→pro) show no monotone position trend, but the Western ladders show two monotone *behavioural* trends: refusal rises with scale (nemotron marker-classified refusals 0 → 1 → 54 under Chinese administration, 88 failed calls at the top rung) and so does English-language reasoning about the Chinese prompt (share of nemotron reasoning traces in Chinese script: 1.00 → 0.37 → 0.18), while both deepseek rungs reason in Chinese on every call. All

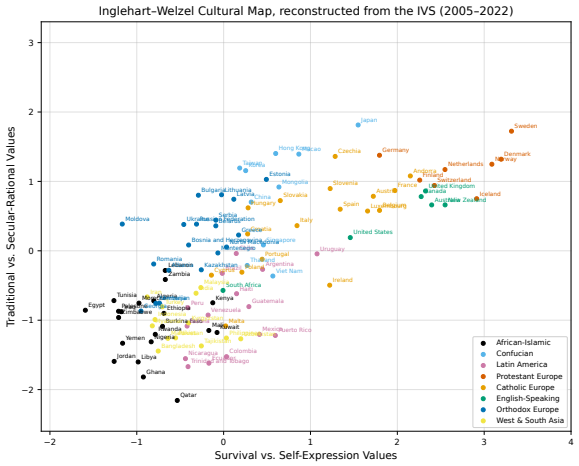

Figure 3: The Inglehart–Welzel cultural map reconstructed from the IVS microdata by our pipeline. No WVS-copyrighted figure is reproduced anywhere in this paper; this map is computed from the microdata by our own code, which doubles as a face-validity check: the familiar regional clustering emerges from our fit.

descriptive at $n \leq 3$.

Reproducibility. The 2026 analysis was fixed in a dated, append-only pre-specified analysis plan (QC gate, then confirmatory, then diagnostics and sensitivity) whose deviations ledger ships with the release; "pre-specified" throughout this paper refers to that document. Seeded EM with convergence assertion; 20-seed refit budget; `make test` (48+ unit tests on synthetic fixtures: PPCA fit/transform round-trip, standardisation symmetry, the single-stored-rotation contract, sentinel recoding, rescale constants, orientation convention, WVS index transforms, parser validation, bootstrap determinism) runs in CI with no survey data; `make validate` reproduces every published number from the raw data locally. The repository ships download-and-merge instructions for the IVS rather than the data, per the WVS/GESIS data-use agreements. https://github.com/Shavvimal/model_cultural_comp

D Pipeline validation and instrument reconstruction

Figure 3 shows the reconstructed instrument itself.

What each check proves, and what it does not. **(1) Path identity (exact).** All complete-case survey rows are pushed through the public projection API, the same function models go through, and compared with the coordinates the fit assigned them internally: maximum absolute discrepancy 0.0, at machine precision, over 191,704 rows. This is a *regression test that the model path and the country path are byte-identical*, precisely the property the 2024 code lacked, in both its unstandardised projection and its re-fitted rotation. It is not, and we do not claim it is, evidence that the map itself is right: both paths could share a wrong rotation and this check would still pass. Its value is that the class of defect that produced the 2024 error can never silently return. **(2) Correction accounting.** Nine of ten models moved < 0.18 map units under the projection fix while `llama3:70b` moved 2.05; the Y003 recode moved country coordinates by mean 0.19 (max 0.65) and model coordinates by mean 0.47 (max 0.74) at per-axis correlations > 0.99 (recode displacements as measured by the released audit on its pre-rescaling fit). Per-axis affine regression of the 2024 published coordinates on the corrected ones: R^2 of 0.990 and 0.993 with slopes 1.36/1.31. The 2024 map was 36% and 31% wider per axis respectively, the intended effect of the unit-variance rescaling (the fitted slopes sit slightly below the theoretical $s = 1.45/1.35$ because the Y003 recode perturbs the regressor, an errors-in-variables attenuation; $r > 0.99$ per axis). This is an internal consistency check on the size of our own correction, not external validation: no regression against the published WVS country coordinates is performed. The corrected map is the same map, moved by exactly the corrections we made and nothing else. **(3) Interpretability.** The rotated loadings recover the published item structure (the justifiability items and the autonomy index load positive on the self-expression and secular directions; God, pride and authority mark the traditional pole, F063 loading negative on the secular axis while the reverse-keyed G006 and E018, with 1 = proudest and 1 = greatest respect, load positive), including Y003, which the sentinel bug had silenced.

The Y003 sentinel, in detail. The sentinel is structural rather than sporadic: 15 countries carry it in 100% of rows and 65 in none. Recoding and refitting restores Y003 to a substantive loading of $(0.19, 0.24)$ and to its expected correlational profile ($r = -0.39$ with importance of God). The explained-variance gain (38.4% → 41.7%) is quoted with each figure on its own EM-completed denominator, each internally consistent but not a fixed-denominator comparison. After recoding, EM imputes Y003 for entire countries from the other nine items: legitimate under missingness-

by-design, but model-based extrapolation for those countries, which the repository lists.

Monte Carlo budget and structural assertions. A 20-seed refit harness bounds Monte Carlo noise from the EM initialisation: rotation span 5.2°; country coordinates mean SD 0.02, worst-case range 0.26 map units, the same order as some displacements we report, hence its inclusion in the uncertainty budget. The merge and EM stages assert their invariants (row-count preservation, convergence) and raise rather than degrading silently. The PPCA implementation is derived from `pca-magic` (Tran, n.d.), seeded, capped (EM raises rather than silently returning an unconverged fit), and unit-tested against synthetic low-rank ground truth.

E Uncertainty estimators: derivations and diagnostics

Why the pairing does not move the mean but does move the variance. Writing a model's position as an affine functional of its ten per-item means,

$$p = w_0 + \sum_{j=1}^{10} w_j \bar{x}_j,$$

the point estimate is invariant to how individual responses are paired into pseudo-respondents. Covariance, however, is bilinear, and iterating $\mathrm{Cov}[aX + b, cY + d] = ac\,\mathrm{Cov}[X, Y]$ and $\mathrm{Cov}[X + Y, Z] = \mathrm{Cov}[X, Z] + \mathrm{Cov}[Y, Z]$ over the affine map gives

$$\mathrm{Var}(p) = \sum_{j=1}^{10} w_j^2 \,\mathrm{Var}(\bar{x}_j) + 2\sum_{j<k} w_j w_k \,\mathrm{Cov}\big(\bar{x}_j, \bar{x}_k\big)\,.$$

The affineness argument says nothing about the second sum. All ten items were elicited under the same ten system-prompt variants, so the cross-item covariances are not zero, and a bootstrap that resamples items independently forces them to zero. The independence assumption does not make the true cross-item covariances zero; it only removes them from the variance estimate.

The language-displacement estimator. The within-model contrast is summarised as δ_m, the difference of the two arm mean positions, and its magnitude as the plug-in norm $\|\delta_m\|$. The pairing of replicate i in one arm with replicate i in the other is arbitrary (the arms' bootstrap draws are independent) and the plug-in point estimate is invariant to it; the mean of per-replicate norms, by contrast, is a folded statistic, upward-biased wherever the true displacement is small against sampling noise, and is carried in the released artefacts as the alternative. Confidence intervals for $\|\delta_m\|$ are percentile intervals of the per-replicate norms; intervals on a norm fold at zero and cannot exclude it.

Intraclass correlations and design effects. On the completed 2026 corpus, per-item intraclass correlations by prompt variant reach 1.00 (degenerately, under deterministic decoding; median 0.10; 51 of 340 cell-item pairs exceed 0.5; for `gemma4:31b`'s Chinese-administered F063, every repeat within a variant is identical while variants disagree, the prompt-variance-dominant regime in its purest form). With five repeats per variant, the top intraclass correlations imply design effects up to 5.0: in those cells a cell-item mean's fifty calls carry the information of roughly ten independent draws, one per prompt variant (at the median ICC of 0.10 the design effect is 1.4). The keying-balance diagnostic cited in §5 also lives here: net deviation from scale midpoints is 0.02–0.09 per cohort × arm group mean, with per-cell values spanning up to 0.19 (`diag_2026_keying_balance.csv`).

Determinism drives the estimator gap. The cells with the smallest item-bootstrap SDs are precisely the most deterministic ones (Spearman $\rho = 0.65$ between mean per-item response entropy and item-bootstrap SD, $p < 10^{-4}$ treating the 33 cells as independent; at model level, $n = 17$, the association holds at $p \approx 0.005$; a single pre-specified test), while the cluster SD is essentially unrelated to determinism. Under near-deterministic decoding, token-sampling variance stops measuring anything a reader cares about, and the prompt-variant bootstrap must carry the uncertainty budget.

Simultaneity and coverage. A cluster-bootstrap replicate position is a convex combination of the ten prompt-variant means, and the quadrant is an intersection of half-planes, hence convex, so "all 10,000 replicates in the quadrant" holds exactly when all ten variant means do, and the 330,000-replicate statement is equivalent to 330 variant means plus a margin (every cell mean ≥4.8 cluster SDs inside the boundary). As a confidence statement it is union-bounded: each cell's one-sided bootstrap tail is below $1/B$, and over 33 independently bootstrapped cells the joint tail is at most 0.01. The plotted region is the normal-theory set $\mathscr{E}_{0.95} = \{\mathbf{p} : d^2(\mathbf{p}) \le \chi^2_{0.95,2}\}$, where d^2 is the

Component	PC1′ SD	share	PC2′ SD	share
Model	0.43	17.8%	0.31	18.4%
Language within model	0.44	18.9%	0.22	9.1%
Variant within cell	0.57	30.9%	0.41	31.0%
Repeat within variant	0.58	32.4%	0.47	41.5%

Table 4: Nested variance components in map units (per-level mean squares; the language term pools the language main effect with the model × language interaction). On PC1′ the mean-square shares are comparable (18.9% vs 17.8%), and the orthogonal partition puts model identity ahead (21.0% vs 11.8%); on PC2′ model identity dominates under either convention (18.4% vs 9.1%; orthogonal 20.9% vs 5.5%).

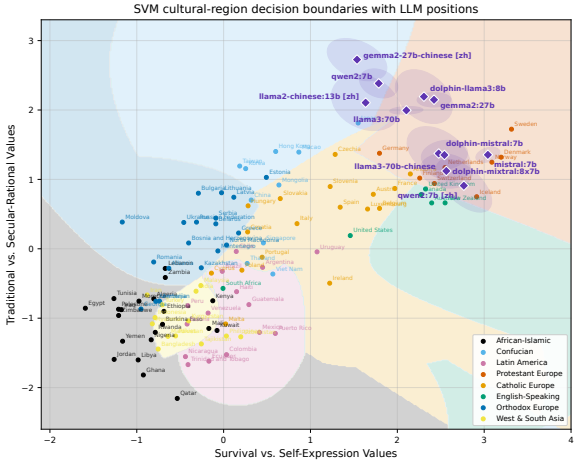

Figure 4: SVM decision regions, shown as illustration of why region labels are weak evidence off the country manifold: the model cluster sits in extrapolated territory. Region claims in the text use the classifier-free statistics instead.

Mahalanobis distance from the replicate mean $\bar{\mathbf{p}}$ under the replicate covariance $\hat{\Sigma}$. Its radius treats the replicate covariance as known; a simulation calibrated to the $K = 10$ design puts the region's true coverage near 85%, against the conservative Hotelling radius $2\frac{K-1}{K-2}F_{0.95}(2, K-2) = 10.03$. We keep the conventional radius and report the calibration.

Region classifier implementation. The RBF-SVM is fitted to the 109 country coordinates across eight regions with a grid search that includes the regularised regime, under stratified 5-fold cross-validation; Figure 4 shows its decision regions, and Figure 5 both cohorts on the one frozen instrument.

F Extended 2026 results

Geometry in detail. Nearest neighbours across the cohort are Japan most often, then Germany,

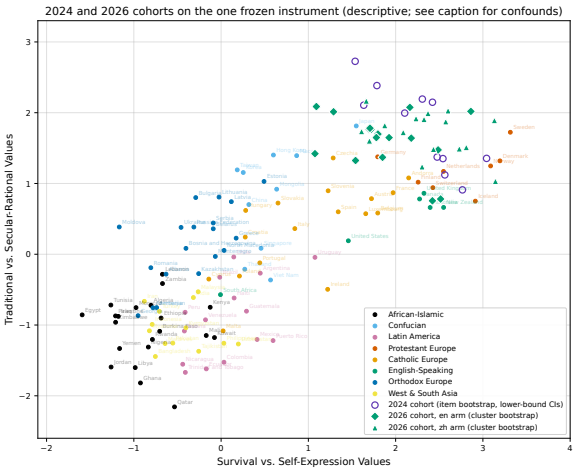

Figure 5: Both cohorts on the one frozen instrument, descriptive only. The cohorts share no models and differ simultaneously in serving precision (Q4-quantised local vs. undisclosed cloud), model scale (7B–70B vs. 20B–675B), retry intensity (up to 15 vs. 3 attempts), elicitation-language coverage, and survivorship (five 2024 Chinese-origin models produced nothing parseable); markers additionally differ by estimator (2024: item-bootstrap CIs, which typically understate cluster-bootstrap uncertainty; 2026: cluster bootstrap). No pooled statistic is computed from this figure.

the Netherlands, the Nordics, Australia and New Zealand. Thirteen of seventeen English cells and seven of sixteen Chinese cells lie *inside* the convex hull of the 109 countries (2024: five of eleven), but shallowly, hugging the hull's Sweden–Japan rim. Every 2026 cell in both arms sits in the top quartile of *both* axes simultaneously; four English and eight Chinese cells are more secular than Japan, the most secular country, while not one of the 33 cells (nor any 2024 cell) exceeds Sweden on self-expression. Twelve of seventeen English cells (eleven of sixteen Chinese) sit nearer another *model* than any of the 109 countries, and five (six) of those nearest model-neighbours belong to the other origin cohort. Three English cells sit closer to a real country than the median country sits to its own nearest neighbour (`minimax-m2.7` is 0.06 from Germany); the Chinese arm floats farther out, with five cells lonelier than the most isolated country on the map (Moldova, 0.68 from its nearest neighbour). The Chinese arm sits further into the corner than the English arm on both axes (median secular percentile 99.5 vs 98.2).

Two null mechanism probes. A model's language displacement is predicted neither by the share of its Chinese-arm reasoning conducted in Chinese ($|\rho| \leq 0.12$) nor by its refusal-rate change between

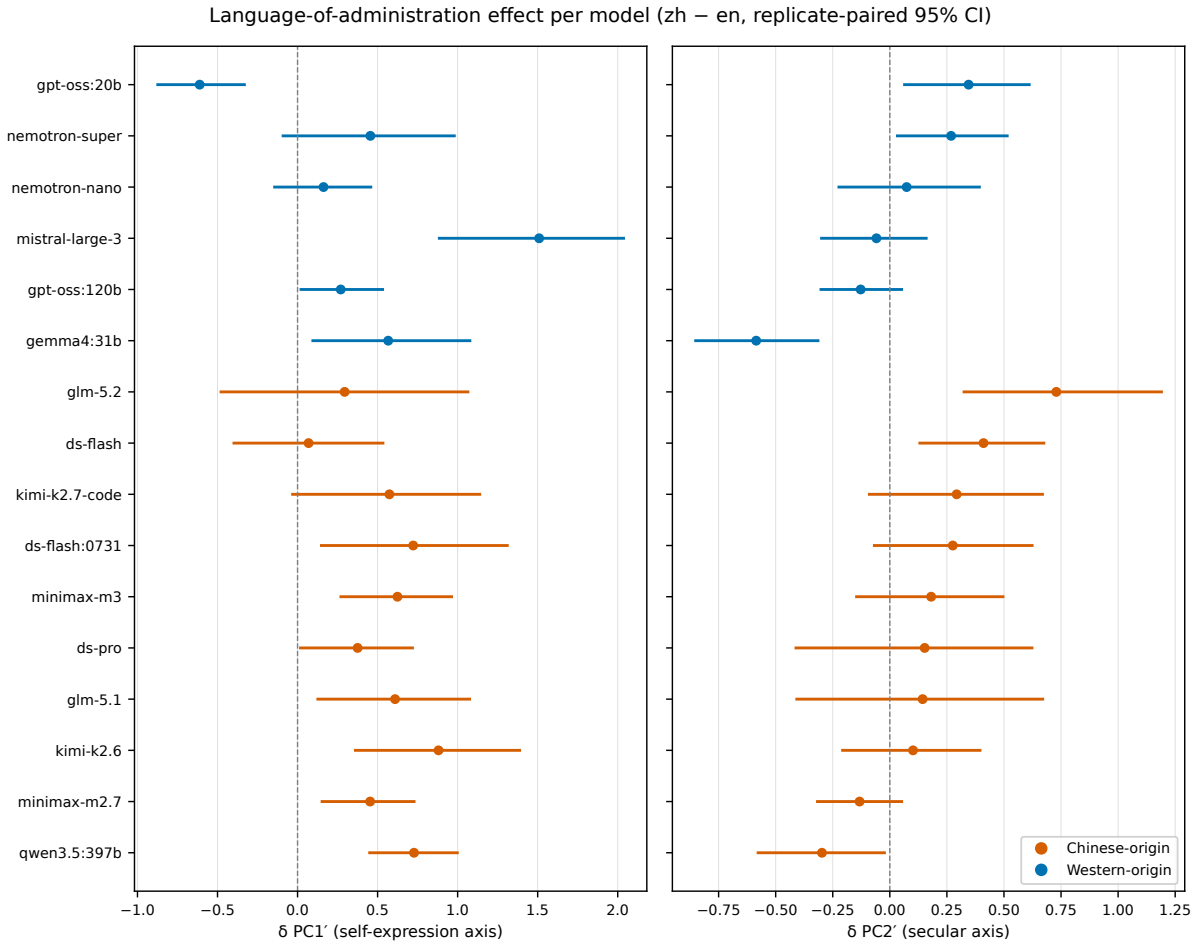

Figure 6: δ_m per model (zh − en, replicate-paired 95% CIs). The self-expression component is positive almost everywhere; the secular component is heterogeneous; the largest displacement belongs to a Western model.

arms (exploratory, $m = 6$, all BH-null). The displacement survives even where Chinese-arm deliberation collapses to bare instruction-restatement: deepseek's newest checkpoints think in ~40–80 median characters under Chinese against ~470–830 in English, yet shift as far as any sibling.

Per-item refusal swings. On the abortion item alone, `gemma4:31b`'s refusals collapse 22 → 1 under Chinese administration while `nemotron-3-ultra`'s rise 19 → 33 until its Chinese cell is voided.

Refusal reallocation: per-model swings, ratio bracketing, bound scope. `gemma4:31b` collapses from 27 refusals to 1 under Chinese administration; `nemotron-3-ultra` rises from 34 to 54 (of 40 → 88 total failures, enough to void its Chinese cell); `minimax-m3` refuses only in Chinese (0 → 16). The gross-to-net ratio of ~4.5× (§4.2) is scale-bound: under the nemotron floor–ceiling range below it falls to ~2.1×, and a random-sign null on the same per-model magnitudes already gives ~2.4×, so the direction-cancelling is real but these two models carry most of it (refusal-count floors: Appendix A). The Manski bound's scope: 19 of 34 cells have zero failed calls, so it binds only where refusal occurred, and it reallocates terminal failures only; the re-ask protocol's rejected drafts are unrecorded (Limitations).

G Extended related work and discussion

The broader elicitation toolkit. Scenario-based probing and steering (Dang et al., 2026) with latent activation steering (Dang and Masud, 2026) (one research programme), culturally grounded personas (Greco et al., 2026), word-association tests (Dai et al., 2025), and multi-agent cultural debate (Tan et al., 2026) form a growing toolkit for eliciting and shifting cultural behaviour, several of them working directly on the IW axes. Against that toolkit our contribution is deliberately narrow and instrument-first: a decades-validated survey instrument, ad-

ministered under a controlled protocol, with the measurement pipeline itself audited, validated and released.

Linguistic relativity. The Sapir–Whorf tradition (Au, 1983) motivates the possibility that language alone shapes elicited values. Our language-of-administration finding is consistent with a weak form of that claim, though the design cannot separate what the prompt language activates in the model from what the training data associated with that language contains; per Au's critique of the strong form, we treat the mechanism as an open question, not a conclusion.

Trace-coding protocol and verbatims. 900 traces were sampled stratified across the 30 trace-emitting cells and coded by a single LLM-based annotator for (a) explicit modal-response targeting, (b) explicit persona reasoning, (c) guideline or safety citation, and (d) language of reasoning. Cell-level code counts and verbatims are released with the corpus (`data/trace_coding.json`); the coding is qualitative and labelled as such. One representative trace names the instrument outright: "the World Values Survey uses this exact question. The average in many Western countries is around 8–10. Let's pick 8." The mechanism behind the midpoint result is that the world's respondents are far from the scale midpoints on the traditional-pole items (fitted means: importance of God 7.2/10, national pride 1.6 on a 1–4 scale where 1 is proudest, respect for authority 1.5 on 1–3). Determinism tracks response extremity rather than modality ($\rho = -0.65$ with midpoint distance, raw $p < 10^{-4}$, labelled post hoc): the most deterministic cells are the most extreme responders, not the most moderate.

Post-training provenance: the full argument and its cautions. The literature shows post-training is the controllable channel that selects a model's expressed opinions: base and human-feedback-tuned models sit on opposite US demographic profiles (Santurkar et al., 2023); the alignment stage measurably pulls models toward US opinions, with one widely used reward model ranking the United States above 99.4% of countries (Ryan et al., 2024); modest fine-tuning suffices to steer political position, plausibly "an unintentional byproduct of annotators' instructions", a design choice made by default (Rozado, 2024); and fine-tuning revises encoded cultural values in ways that bleed across languages (Choenni et al., 2024). The inputs to that channel are documented decisions by small groups: InstructGPT's preferences were operationalised by roughly forty screened contractors (Ouyang et al., 2022); a constitution's principles are, in its authors' own words, "our own choices as designers", and swapping it for one written by 1,002 members of the public measurably changes the trained model's bias profile (Huang et al., 2024). Propagation of post-training artefacts *between* labs through the synthetic-data ecology is author-acknowledged since the first GPT-derived instruction sets, openly practised within model families (DeepSeek-AI, 2025), and measurable: two different bases trained on one teacher's synthetic data become nearly indistinguishable to a fingerprint classifier (Sun et al., 2025), models passively inherit a generator's biases and preferences from its data (Shimabucoro et al., 2024), and style is the cheapest thing to transfer (Gudibande et al., 2023). If that propagation touches value-setting choices too, the effective number of independent value-setting decisions behind a "17-model" cohort may be far smaller than seventeen, which would predict precisely the origin-independence, vendor signatures, and cross-lab output homogenisation (Jiang et al., 2025) we and others observe.

Two cautions bound the claim. First, whether specific Chinese labs distilled specific Western frontier models is an unadjudicated allegation by commercially and politically interested parties, denied by the labs and by Beijing: the documented facts stop at self-identification incidents and an acknowledgement that web corpora inevitably contain model-generated text.[1] Our origin-null is equally consistent with a domestic shared-recipe ecology, which the one origin signal we do find (Chinese-origin models are more alike each other, exact permutation $p = 0.006$, our only test of this family) fits just as well. Second, post-training is not the sole origin of the default: aligned behaviour is substantially *selected from* the base distribution rather than implanted (Lake et al., 2025), base models carry value priors (Santurkar et al., 2023), and cre-

[1] Allegations: Financial Times and Bloomberg reporting, 28–29 January 2025 (OpenAI: "some evidence" of distillation; Microsoft probe); OpenAI memo to the US House Select Committee, February 2026 (Reuters); Anthropic disclosure of "industrial-scale distillation attacks", February 2026 (CNBC). Denials: DeepSeek (Nature supplementary materials, 2025); China MOFCOM statement, 28 July 2026, with counter-allegations against US firms. Self-identification incident: TechCrunch, 27 December 2024. None of the cross-lab allegations has been adjudicated; we cite them as allegations only.

ator ideology remains detectable on person-level moral assessments by an instrument unlike ours (Buyl et al., 2026). The composed question (does a student model inherit its teacher's *values distribution*, as measured on a survey instrument?) has, to our knowledge, never been tested directly; our released harness makes that experiment cheap, and the version-churn result (§4.2) suggests where to look.

Interpretability: what it would add, and its documented limits. Sparse-dictionary decompositions of production-scale models recover interpretable features in bulk: one decomposition scales to 34 million features and reports features for bias, deception and sycophancy among them, several of which change the model's behaviour when clamped (Templeton et al., 2024). The cross-layer successor, attribution graphs, traces which of those features a model actually uses on a given prompt (Lindsey et al., 2025; Ameisen et al., 2025). Value-laden behaviour is beginning to yield the same treatment: culture-general and culture-specific neurons comprising under 1% of a model's units, whose ablation costs up to 30% on cultural benchmarks while general language understanding is largely unaffected (Yamamoto et al., 2026); intrinsic and prompt-induced value expression running through partly shared, partly distinct components that generalise across languages (Han et al., 2026); cultural steering vectors constructed from sparse features rather than from prompts (Khanuja et al., 2026); and a cultural-customisation direction conserved across non-English languages, which surfaces localised knowledge a model already holds but does not volunteer (Veselovsky et al., 2026). None of this makes a measured default less real; it makes it *addressable*, and the automated pipelines already built for extracting and monitoring trait directions (Chen et al., 2025) are a plausible template for continuous auditing of a cultural default.

The limits are documented and not tuning artefacts. The closest work to ours reports *latent entanglement*: intervening on one cultural dimension drags others with it, because the dimensions are encoded as coupled structures (Dang and Masud, 2026); and in a controlled head-to-head, sparse-autoencoder steering is outperformed by plain prompting and by finetuning (Wu et al., 2025). Feature absorption causes a parent feature to stop firing silently because a more specific child has absorbed it, and varying dictionary size or sparsity does not fix it (Chanin et al., 2025); about half the reconstruction error, and over 90% of its norm, is linearly predictable from the input activation: dense structure the sparse basis cannot represent (Engels et al., 2025). The field's own twenty-nine-author stocktake describes mechanistic interpretability as *promising* assurance over model behaviour rather than yet providing it (Sharkey et al., 2025). The defensible position is therefore narrow: interpretability is a route towards auditing and adjusting cultural defaults below the prompt, not a means of certifying them.

Stability of value expression: independent support. Models are relatively consistent on value-laden questions across paraphrase, format and translation, with *base* models more consistent than fine-tuned ones (Moore et al., 2024); value expression without persona conditioning exceeds human rank-order stability benchmarks (Kovač et al., 2024); and careful symmetrisation shows frontier models carrying large order-and-wording artefacts *on top of* a near-invariant underlying moral stance (Huang, 2026): surface bias without position shift, which is exactly the structure our variance decomposition measures.