Statistical Appendix

Companion to the CMRB pilot · Declan Michaels · moral-os.com · 2026

Companion to "Frontier Language Models Converge in a Narrow Region of a Moral-Reasoning Space." Every number here regenerates from the raw runs with one stdlib script, `private/viz/build_appendix.py` (random seed 20260720). This appendix is deliberately terse and table-first; the readable account is in the main report.

A1. SAMPLE AND DATA

Eleven models from nine labs (three Chinese: DeepSeek, MiniMax, Moonshot). Sixty-eight human respondents, a convenience sample recruited within one to two degrees of the author, demographically varied but concentrated in educational and professional background (almost all technology consultants and professionals), scored with the identical function. Forty-eight scenarios, twelve of which carry the human baseline (b12). Eight framings per scenario. Each (model, framing) cell is one complete run of 240 responses (48 scenarios times 5 iterations). Where a cell was run more than once, the newest run is canonical and the earlier runs are archived under `runs_v2/_superseded/`; the build scripts raise if a duplicate complete cell reappears.

Two models with complete neutral runs were excluded from the pilot panel (Cohere Command A, Google Gemini 3 Pro); they appear only in the sensitivity analysis A8.

A2. THE SCORING FUNCTION

Every axis score, for humans and models alike, comes from one function (`compute_dimensional_score` in `scenario_bank.py`). Each scenario belongs to exactly one axis and poses two questions, judgment (what should happen) and reasoning (why it matters), each with three or four answer options. Every option carries a fixed signed loading toward its axis, a value in the interval from -1 to +1: an option that fully marks one pole loads at +1 or -1, an option that leans partway loads at a fraction (the bank uses 0.3, 0.5, 0.7), and a neutral option loads at 0.

A respondent distributes a fixed pool of points across the options for each question. The judgment score is the loading-weighted average of the points they placed:

j = sum over judgment options of (loading times points), divided by the sum of points placed

The reasoning score r is the same over the reasoning options. Because points are non-negative and the loadings lie in the interval from -1 to +1, both j and r fall in that same interval: all points on full +1 options gives +1, all on full -1 options gives -1, points on neutral options or an even split pull toward 0. The axis score combines the two,

score = 0.6 times j, plus 0.4 times r,

weighting the judgment question above the reasoning question. If a respondent left one question blank (zero points placed), the other is used at full weight rather than deflating the score toward zero. The 0.6 to 0.4 split is the single free parameter in the scoring; A8 shows the compression holds under judgment-only, reasoning-only, and the combined weighting, so the finding does not depend on it.

A3. COMPRESSION

Dispersion is population standard deviation; the model figures are each model's unframed neutral position (A5 shows the panel widening two to five times under framing). The p-value is an N-matched bootstrap: draw 11 humans at random on the axis, 100,000 times, and count how often that human subsample is at least as tight as the 11-model panel. The ratio CI is a 5,000-draw percentile bootstrap resampling humans and models independently.

Axis	Human SD	Model SD (neutral)	Ratio (human/model)	95% CI on ratio	p
Moral Agent	0.412	0.061	6.78x	[4.95, 13.2]	< 0.00001
Authority	0.401	0.052	7.69x	[5.83, 12.53]	< 0.00001
Moral Domain	0.327	0.061	5.37x	[3.66, 11.65]	< 0.00001
Obligation Scope	0.361	0.064	5.65x	[4.57, 8.41]	< 0.00001

In 100,000 draws on each axis, not one human subsample matched the panel's tightness (p reported as < 0.00001). The lower bound of every ratio CI is at least 3.66, so even the conservative reading is a large compression. The human sample is a convenience sample concentrated among technology professionals (A1), which makes this a conservative comparison: a broader or cross-cultural human sample would widen the human SD and enlarge the ratio, not shrink it.

Where the two populations sit, for context (the compression claim is about spread, not location):

Axis	Human mean	Human median	Model mean
Moral Agent	+0.29	+0.36	-0.05
Authority	+0.17	+0.20	+0.12
Moral Domain	+0.37	+0.43	+0.29
Obligation Scope	+0.37	+0.42	+0.28

Moral Agent is the one axis where the models do not merely compress but sit on the opposite side of the midpoint from the human center.

A4. RUN RELIABILITY (TEST-RETEST)

Within-model dispersion is the standard deviation of a cell's axis score across its five iterations, reported as the median over the eleven models. Between-model dispersion is the panel SD from A3. The point of the table: the clustering is not an artifact of sampling noise, because the spread between models is roughly two to two-and-a-half times the spread a single model shows on rerun. This within-model dispersion is what the interactive viewer draws as its optional run-spread overlay.

Axis	Within-model run SD (median)	Between-model SD	Between / within
Moral Agent	0.026	0.061	2.33
Authority	0.021	0.052	2.50
Moral Domain	0.025	0.061	2.41
Obligation Scope	0.033	0.064	1.95

A5. FRAME RESPONSIVENESS

Displacement is the mean absolute per-axis change from a model's own neutral position, averaged over the four axes; the group figures pool that across models and framings. CIs are 5,000-draw bootstraps over models.

Framing group	Mean displacement	95% CI
Cultural (4 framings)	0.363	[0.307, 0.422]
Nonsense (geometry, color)	0.203	[0.164, 0.243]
Seasonal, non-moral (not a clean null)	0.246	[0.196, 0.298]

Nonsense moves the models 0.56 as far as a real culture (Cohen's d between the two pooled sets = 0.95, a large effect). Per framing:

Framing	Mean displacement
individualist	0.335
collectivist	0.442
hierarchical	0.543
egalitarian	0.132
irrelevant (seasonal)	0.246
nonsense: geometry	0.195
nonsense: color	0.211

The seasonal framing (labeled irrelevant in the data) was intended as a non-moral noise floor, but it is not a clean null: seasons carry real cultural weight, and its displacement (0.246) sits within the bootstrap spread of the nonsense framings rather than near zero. We report it as a mild non-moral framing, not a control. The pilot has no clean inert control; adding one is the first priority for the next round.

The competence claim rests on direction, not distance. For each cultural framing, the signed shift on the axis it should move, in the expected direction, and the count of models that moved the right way:

Framing	Target axis	Mean signed shift (expected direction)	Models correct
individualist	Moral Agent (toward autonomous)	+0.46	11 / 11
collectivist	Moral Agent (toward relational)	+0.56	11 / 11
egalitarian	Authority (toward skeptical)	+0.16	11 / 11
hierarchical	Authority (toward deferential)	+0.62	11 / 11

Nonsense produces movement without that contrastive structure. Both nonsense framings push every one of the eleven models the same way on Moral Agent (toward relational: geometry mean -0.25, color mean -0.24, 0 of 11 moving the other way) rather than splitting them the way a genuine cultural contrast would. A uniform pull is the signature of compliance, not of reading the scenario.

Between-model dispersion. The compression in A3 is measured at neutral. Under framing the panel does not stay equally tight: the standard deviation across the eleven models, averaged over the four axes, rises two to five times above its neutral value under every framing, cultural or nonsense alike.

Framing	Between-model SD (b12)	vs neutral	Between-model SD (all 48)	vs neutral
neutral	0.059	1.0x	0.033	1.0x
individualist	0.125	2.1x	0.129	3.9x
collectivist	0.148	2.5x	0.132	4.0x
hierarchical	0.160	2.7x	0.154	4.7x
egalitarian	0.130	2.2x	0.074	2.2x
irrelevant (seasonal)	0.129	2.2x	0.093	2.8x
nonsense: geometry	0.145	2.4x	0.097	2.9x
nonsense: color	0.142	2.4x	0.109	3.3x

The scatter is not specific to real cultures; nonsense and the seasonal framing widen the panel about as much. What distinguishes cultural framing is the shared direction of the signed shifts above, not the amount of spread, so the models share a resting point and a direction of movement without sharing a destination.

A6. CROSS-LAB CLUSTERING

Each model is fingerprinted by its deviation from the panel consensus, and its nearest neighbor is the model with the highest Pearson correlation of deviations. We report two grains.

Panel grain (deviation of the four axis positions, all 48 scenarios): no model's nearest neighbor is a same-lab sibling (0 of 11). Correlations are high because the fingerprint is four-dimensional.

Model	Nearest neighbor	r	Same lab
deepseek_v4	opus	0.882	no
gpt55	mistral_large	0.995	no
grok45	minimax	0.971	no
inkling	llama33	0.889	no
kimi	gpt55	0.717	no
llama33	inkling	0.889	no
minimax	sonnet	0.984	no
mistral_large	gpt55	0.995	no
o3	llama33	0.739	no
opus	deepseek_v4	0.882	no
sonnet	minimax	0.984	no

Fine grain (deviation of the 48 per-scenario positions): ten of eleven nearest neighbors are cross-lab. The single exception is Opus, whose nearest neighbor is Sonnet at r = 0.137, narrowly ahead of two OpenAI models (gpt55 at 0.115, o3 at 0.104). Sonnet itself is not the mirror of that tie: its own nearest neighbor is MiniMax at r = 0.500, a much stronger and cross-lab pairing. We report this rather than suppress it: the substantive conclusion, that models do not sort by lab or by geography, holds at both grains, and the lone same-lab tie is weak and appears only at the finest resolution.

Model	Nearest neighbor	r	Same lab
deepseek_v4	o3	0.223	no
gpt55	inkling	0.144	no
grok45	sonnet	0.394	no
inkling	gpt55	0.144	no
kimi	gpt55	0.112	no
llama33	o3	0.415	no
minimax	sonnet	0.500	no
mistral_large	o3	0.233	no
o3	llama33	0.415	no
opus	sonnet	0.137	yes
sonnet	minimax	0.500	no

The tightest cross-lab pairs (MiniMax and Sonnet at 0.50, o3 and Llama at 0.415) cross both company and country.

A7. REASONING TOKEN SPEND

Mean reasoning tokens per neutral scenario, with each model's Euclidean distance from the panel centroid across the four axes (all 48 scenarios):

Model	Reasoning tokens / scenario	Distance from panel center
kimi	3628	0.031
minimax	1189	0.058
inkling	1117	0.085
deepseek_v4	1066	0.058
grok45	494	0.093
o3	247	0.055
gpt55	131	0.060
llama33	0	0.061
mistral_large	0	0.109
opus	0	0.042
sonnet	0	0.079

Spend runs from zero (four models emit no reasoning tokens) to 3,628. Across the eleven models it has no meaningful relationship with position. The correlation between token spend and distance from the panel center is -0.48, which at n = 11 is not significant and is negative, meaning if anything the heavier reasoners sit slightly closer to the middle rather than staking out distinctive ground. The correlation between token spend and position on each individual axis is small: Moral Agent -0.08, Authority -0.13, Moral Domain 0.04, Obligation Scope 0.12.

The single-model contrast in the report (Kimi at 3,628 tokens and Llama 3.3 70B at zero landing in the same place) is an illustration of this panel pattern, not the evidence for it. This is not a controlled test of whether reasoning improves answers in general; it shows only that, on this instrument, how much a model deliberates does not predict where it lands.

A8. SENSITIVITY ANALYSES

Scope. Model SD at neutral on the 12 baseline scenarios versus all 48. The panel is tighter on the full set, as expected from averaging over more items; the compression conclusion does not depend on the subset.

Axis	Model SD (b12)	Model SD (all 48)
Moral Agent	0.061	0.031
Authority	0.052	0.028
Moral Domain	0.061	0.052
Obligation Scope	0.064	0.023

Question weighting. Model SD under judgment-only, reasoning-only, and the combined default stays in the same narrow band on every axis, so the compression is not an artifact of the 0.6 to 0.4 split.

Weighting	Moral Agent	Authority	Moral Domain	Obligation Scope
judgment only	0.070	0.045	0.072	0.065
reasoning only	0.059	0.069	0.050	0.075
combined (0.6 / 0.4)	0.061	0.052	0.061	0.064

Panel composition. Adding back the two excluded vendors (13 models) leaves the panel SD small on every axis (Moral Agent 0.076, Authority 0.061, Moral Domain 0.061, Obligation Scope 0.077), so the finding is not created by the pilot's model selection.

A9. VALIDITY

This pilot does not fully validate the instrument; that is what the decisive study is for. But three results already bear on whether it measures structured moral reasoning rather than returning arbitrary numbers.

Known-groups. The human sample lands on the profile the cross-cultural literature predicts for a WEIRD group, autonomous, skeptical, narrow, universalist, with no tuning (A3). Because it is a convenience sample from the author's network (A1), this is consistent with the prediction rather than an independent test of it; but an instrument returning arbitrary numbers would not land on the predicted profile at all.

Response to manipulation. The models move under cultural framings in the theory-predicted direction, all eleven the right way on all four cultural framings (A5), and do not show that directional structure under nonsense. A response that tracks the direction of a cultural manipulation is evidence the axis relates to that manipulation.

Reliability. The scores are repeatable: within-model run-to-run SD is roughly half the between-model SD on every axis (A4), so positions are stable rather than sampling noise.

What this pilot does not establish, and does not claim to: convergent validity against an independent instrument (for example, whether the Moral Domain axis tracks an established measure of the same construct); measurement invariance across cultural groups, which is required before any cross-cultural comparison; and criterion validity against behavior, whether a position here predicts what a model does in open-ended use. The first two ride on collecting cross-cultural human samples on this instrument; the third needs a separate behavioral study.

A10. REPRODUCIBILITY

Three stdlib scripts regenerate everything: `build_figures.py` (compression, bootstrap, panel clustering, the three figures), `build_viewer_data.py` (the viewer payload, including per-cell run dispersion), and `build_appendix.py` (this appendix's `appendix_stats.json`, including the reasoning-token and between-model-dispersion tables). A fourth, `build_csv.py`, exports the tidy CSV tables. All bootstraps use fixed seeds. Each script raises on a duplicate complete cell so a re-run cannot silently change a number.

Analysis is stdlib-reproducible from the raw runs. Responsibility for the work, and for any errors in it, is mine alone. Methodology was AI-assisted and that assistance is disclosed.

Declan Michaels | Cross-Cultural Alignment Study | moral-os.com