# Ethical Frameworks in LLMs Route Through a Shared Endorsement Subspace

**Weilun Xu** [1] **Alexander Rusnak** [2] **Frédéric Kaplan** [2]

## Abstract

Aligned LLMs often answer moral questions by invoking ethical frameworks: local categories, such as rules, consequences, or virtues, through which a broader moral system interprets a case. How these framework-level distinctions are organised in hidden states has alignment consequences: a gap in one framework may be diagnosable in isolation, or may co-vary silently with the others. Two views bracket the question: distinct frameworks in separate subspaces, or all collapsing onto a single endorsement axis. Across 11 open-weight models (7–72B) and six moral systems, including non-Western traditions, we find a third structure: framework directions are geometrically distinguishable within and across systems, yet their intervention effects pass through a shared low-dimensional *endorsement subspace*, where steering along one direction often substitutes for another, though a smaller framework-specific component stays separable from valence under topic and lexical controls. Two implications follow. First, single-direction defences leave open attacks along other directions in the same subspace, so the operational target is the subspace boundary, not a single direction. Second, steering changes which framework a response emphasises, and whether it is endorsed, more than the action the model selects: *which framework to invoke* and *whether to affirm* are distinct alignment problems.

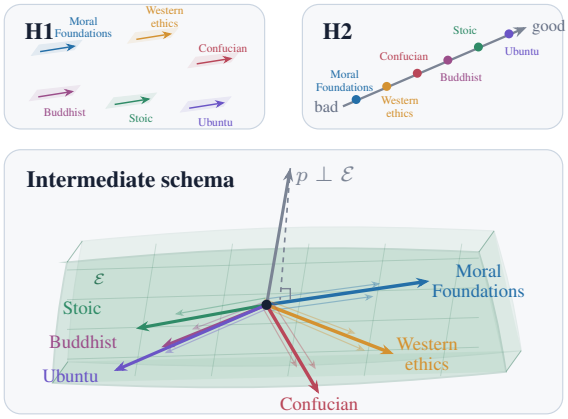


*Figure 1.* **Three structural schemas for framework-level moral content in hidden states.** H1: directions group into separate moral-system subspaces. H2: all directions collapse onto a single "bad"↔"good" valence axis. We argue for an intermediate schema (lower panel): framework directions, grouped by moral system, lie in a shared low-dimensional *endorsement subspace* $\mathcal{E}$, while a non-moral placebo direction $p$ lies outside it, drawn orthogonal ($p \perp \mathcal{E}$) for legibility. Faint companion arrows in the lower panel denote framework-level contrasts.

## 1. Introduction

Aligned LLMs now answer questions in settings where moral vocabulary matters: medical triage, legal assistance, crisis response, and everyday advice. The same case can look different under different ethical frameworks: a deontologist may refuse to break a promise because the rule matters, while a utilitarian may break it if the welfare gain is large enough. Behavioural benchmarks and rationale analyses show that models can produce framework-sensitive judgements and explanations (Hendrycks et al., 2021; Samway et al., 2025). But final answers are only a narrow output channel (Zhu et al., 2025), and chain-of-thought rationales can be unfaithful to the reasons driving a prediction (Turpin et al., 2023). They therefore do not tell us how the corresponding information is organised inside the residual stream. That organisation matters for alignment. If framework directions are compartmentalised by moral system, a failure in one system might be audited locally; if all framework directions share one endorsement axis (a single bad-to-good valence line), then a steering or defence method aimed at one framework may silently affect the others.

We test these alternatives by treating hidden states geometrically. For each framework contrast, we extract a direction from contrastive prompts, then ask two questions. Geometri-

[1]School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne [2]Digital Humanities Lab, École Polytechnique Fédérale de Lausanne. Correspondence to: Weilun Xu <weilun.xu@epfl.ch>, Alexander Rusnak <alexander.rusnak@epfl.ch>.

*Mechanistic Interpretability Workshop at the $43^{rd}$ International Conference on Machine Learning*, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

cally, do directions from the same moral system cluster apart from directions drawn from other systems? Interventionally, does moving the hidden state along a framework direction change the scored output on moral prompts, and can directions from one system steer another system's test pool? We run this design across 11 open-weight variants; §2 explains the six moral systems and their framework vocabularies.

The resulting structure is intermediate between the two opposite schemas in Fig. 1: directions grouped by moral system are distinguishable, but the part of the signal active under intervention lies in a shared low-dimensional *endorsement subspace*. Steering with a direction from one system often substitutes for steering with a direction from another, with target-specific asymmetry (§5.3), and the broader six-system map shows additional cross-tradition structure (§5.5). Training-stage checkpoints place the main reshaping of this geometry at supervised fine-tuning rather than later preference learning (§5.2).

Two operational implications follow, developed in §6: diagnosis or defence anchored to one direction can miss its neighbours inside the same subspace, and steering moves moral *expression* more reliably than moral *decision*.

## 2. Background: Moral systems and ethical frameworks

We use *ethical framework* for a single principle or category of moral judgement, and use *moral system* for a family of frameworks that come from the same ethical tradition or taxonomy. We use *ethical-framework signal* for an activation pattern that tracks one of these framework contrasts in our controlled tasks. The term is descriptive: it does not mean that the model reasons about ethics as a human speaker would.

We study six moral systems from different traditions. Western normative ethics names familiar theories and judgement categories: deontological, consequentialist, virtue-ethical, justice-based, and commonsense. Moral Foundations Theory (MFT) (Graham et al., 2009) names psychological foundations such as care, fairness, loyalty, authority, and sanctity. Confucian *Wuchang* names five cardinal virtues; the Buddhist Noble Eightfold Path (bud, circa 5th century BCE) links conduct with attention and wisdom; Stoic ethics uses four cardinal virtues (Cicero, 44 BCE; Long, 2002); and Ubuntu is a communal-relational system centred on shared identity and solidarity (Metz, 2011).

Compared with using a single benchmark, the six-system design gives two checks. First, it asks whether framework directions group by moral system rather than by a single endorsement axis. Second, it lets us ask whether related traditions form neighbourhoods across system boundaries rather than only within-system clusters. We treat such neighbourhoods as empirical patterns, not assumptions. §4.1 describes the data sources used to connect these systems to text; full system descriptions are in Appendix B.1.

Our scope is normative reasoning. This is worth marking because the residual stream encodes several adjacent alignment concepts that share geometry but answer different computational questions: *toxicity* (Zou et al., 2023) concerns surface harmfulness independent of moral framework; *refusal/harmfulness* (Zhao et al., 2025) concerns whether the model declines to comply; *stereotyping* concerns systematic content bias. Without the distinction, the same residual-stream interventions would conflate moral encoding with refusal. We check this separation directly (Appendix E.5), reading the moral and refusal directions as largely apart rather than a single axis.

## 3. Related Work

**Moral values in LLMs** ETHICS and Delphi established scenario-level settings for evaluating whether models predict human moral judgements (Hendrycks et al., 2021; Jiang et al., 2025), but such output-agreement framing leaves open what considerations produced the verdict, a gap a recent roadmap makes explicit (Haas et al., 2026). Most LLM studies still work through observable responses, measuring encoded beliefs, value orientations, cultural effects, and framework-sensitive choices (Scherrer et al., 2023; Atari et al., 2023; Tao et al., 2024; Ramezani & Xu, 2023; Ren et al., 2024; Sorensen et al., 2024; Rao et al., 2023; Zhou et al., 2024; Samway et al., 2025), some profiling MFT stances directly (Abdulhai et al., 2024; Tlaie, 2024); Buddhist and Ubuntu perspectives enter mainly as governance (Hongladarom, 2020; Mhlambi, 2020; Metz, 2011). We build on this by asking where the corresponding framework contrasts live inside the model and whether interventions on them affect outputs.

**Pragmatics and moral language** Moral judgement depends on social setting, speaker intent, and pragmatics, not only word meaning (Reiter & Kádár, 2015); distributional semantics encodes regularities without that social link (Boleda, 2020), and LLMs show limits on pragmatic inference (Sap et al., 2022; Liu et al., 2025). We therefore do not assume that pretraining yields ethical reasoning in the human pragmatic sense, and ask the narrower question whether benchmark-defined frameworks leave internal signals beyond prompt-surface cues that are usable under controlled interventions.

**Representations and intervention steering** Linear probes show that hidden states encode linguistic and semantic properties (Belinkov et al., 2017; Tenney et al., 2019), but they also motivate caution: high probe accuracy can reflect

prompt wording or dataset artefacts rather than a representation the model uses (Hewitt & Liang, 2019). Concept-direction work addresses this problem by pairing readout with intervention: representation engineering and activation additions extract directions, then test whether moving along those directions changes model behaviour (Zou et al., 2023; Turner et al., 2023; Rimsky et al., 2023). The same logic has been applied to truthfulness (Burns et al., 2023; Marks & Tegmark, 2024; Li et al., 2024), refusal (Arditi et al., 2024), sentiment (Tigges et al., 2023), and persona traits (Chen et al., 2025). Our question is not only whether one moral contrast is linearly decodable, but how many such contrasts are organised together (e.g. a single valence line, separate moral-system regions, or a shared low-dimensional subspace). This requires two kinds of evidence: different direction estimators should point to the same low-dimensional space, and steering along those directions should change scored outputs. The first test guards against extractor-specific artefacts; the second separates an operationally active direction from a label that is merely easy to decode. Prior moral- and value-direction work has typically rested on a single class of validation, from human-judgement correlation to steering effect size to held-in accuracy (Schramowski et al., 2022; Zou et al., 2023; Arditi et al., 2024; Preniqi et al., 2024), which follow-ups traced to untested properties such as embedding-recipe dependence, cross-instrument non-convergence, and surface moves that leave the judgement unchanged (Hämmerl et al., 2023; Stecker & Hopp, 2026; Zhao et al., 2025). We therefore treat construct specificity as something to measure rather than assume; §4.4 states what our controls do and do not identify.

**Alignment training and subspace defences** If moral framework directions share a subspace, then alignment training and defence are subspace-level questions. Mechanistic studies of alignment already point in that direction: preference optimization can bypass rather than remove toxic representations, and sparse parameter or rank regions can be safety-critical while leaving utility mostly intact (Lee et al., 2024; Wei et al., 2024). This motivates methods that remove or manage sets of directions rather than a single feature, including iterative nullspace projection (Ravfogel et al., 2020), sparse-autoencoder (SAE) refinements of steering vectors (Wang et al., 2025; He & Jin, 2025), and parameter-update methods that preserve or constrain subspaces (Xiong & Xie, 2025; Wu et al., 2026). At the same time, linear latent-space methods are limited: obfuscation attacks can route around probes, SAE monitors, and out-of-distribution detectors (Bailey et al., 2024). Our steering and projection experiments should be read as a linear mid-layer baseline, not a complete defence. They test the narrower claim that, for the moral contrasts studied, attack and defence effects pass through the same low-dimensional subspace.

# 4. Methods

## 4.1. Datasets

Four data sources operationalise the six moral systems (§2).[1] ETHICS (Hendrycks et al., 2021) provides Western-framework subsets in two conditions, *templated* (each framework wraps its own scenario pool) and *stripped* (template removed, scenario body only); the pair lets §5.1 separate prompt-surface morphology from the framework-concept residual. Social Chemistry 101 (SC101) (Forbes et al., 2020) provides MFT pairs in a single shared rule-of-thumb template, which guards against foundation-level differences riding on pool-style cues. The FLAMES Morality subset (Huang et al., 2024) supplies native Chinese scenarios for Confucian *Wuchang*, separating that system from English-translation artefacts. Buddhist, Stoic, and Ubuntu, which have no native benchmark, use hand-authored persona prompts: short instructions that ask the model to adopt each tradition's moral standpoint (§4.3).

## 4.2. Models

We test 11 variants from 5 open-weight families: Llama3.1 (Grattafiori et al., 2024), Qwen2.5 (Yang et al., 2024), Qwen3 (Yang et al., 2025), OLMo3 (for AI, 2025), and DeepSeek-R1-Distill (Guo et al., 2025), spanning 7–72B. We call the five 7–8B instruct variants (Llama3.1-8B, Qwen2.5-7B, Qwen3-8B, OLMo3-7B, DeepSeek-R1D-7B) *anchors*; analyses that need pre-extracted auxiliary directions run on the anchors that have those directions, and each result states its model set. Two families ship base-instruct pairs, used in the base-to-instruct rotation analysis of §5.2.

## 4.3. Framework signals and intervention directions

We measure ethical-framework signals at two levels. A readout test asks whether a probe can recover framework distinctions from the residual stream. An intervention test is contrastive: to define a signed direction, we compare two frameworks, such as care versus sanctity or deontology versus virtue ethics, and use the resulting axis for steering. The pipeline therefore separates three claims: visibility in hidden states, stability of the pairwise direction, and an output effect under steering.

**Reading hidden states** For the readout sweep, we save the last-token residual-stream activation at every layer. For direction extraction and interventions, we then use a fixed mid-depth analysis layer, the one nearest $0.65N_{\text{layers}}$. This layer is a standardised analysis point rather than a claim that

[1] Extracted directions, prompt pools, steered generations, and per-model calibration outputs are released as a dataset at https://huggingface.co/datasets/alunxu/llm-ethical-framework; the analysis code is released on publication.

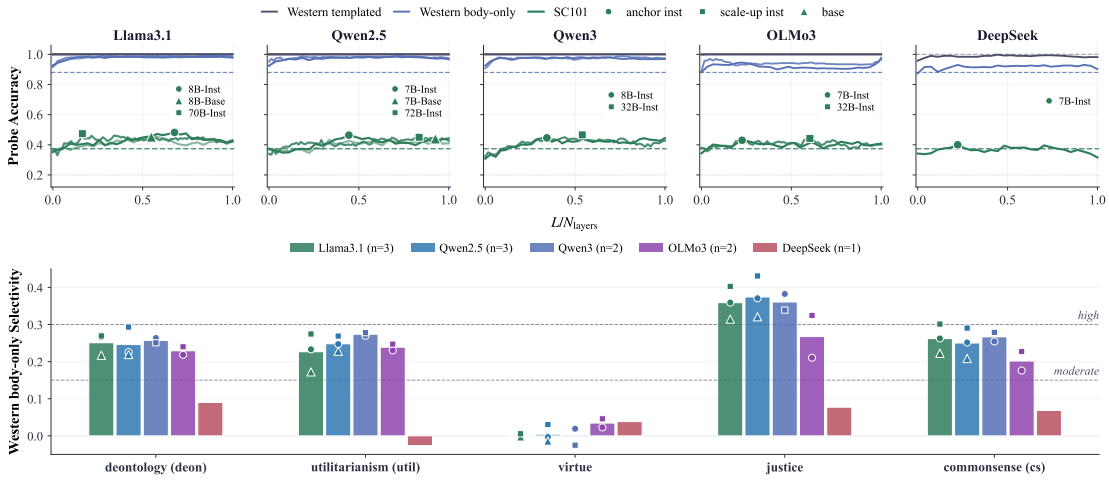


*Figure 2.* **Probe accuracy separates prompt-surface recovery from ethical-framework signal in hidden states.** *Top:* probe accuracy vs. normalised depth, 11 variants from 5 release families. Dashed horizontal lines show matched text-only TF-IDF baselines; in the selectivity panel, bars are family means across variants (dots per variant) and the dashed guides at $0.15/0.30$ are visual reference levels, not fitted thresholds. The operative comparison is *within* each condition, not across datasets: the Western-benchmark conditions are largely recoverable from prompt wording and category-specific scenario pools, while SC101 leaves a small activation gap above the text baseline (3–11 pp). This is a modest ethical-framework signal beyond word cues; by itself, it is not evidence of human-style ethical reasoning. *Bottom:* per-(family, framework) selectivity on the stripped (scenario-body-only) Western benchmark, diagnosing which Western categories remain linearly separable after template removal. Per-variant markers (●/■/▲ = primary instruct variant/larger instruct variant/pretrained base) link top peaks to bottom dots.

every probe peaks there: the sweep checks where framework distinctions are visible, and layer-tracing shows that directions extracted there still steer when applied to neighbouring layers, including at 70B scale. We use linear probes as an initial readout of the signal, not as certification that the model uses the feature. Two baselines make this readout interpretable. Random-label selectivity controls for probe capacity (Hewitt & Liang, 2019). The text baseline trains an $L_2$-regularised multinomial logistic regression on word-unigram TF-IDF features of the raw prompt strings, using the same balanced examples and 80/20 train/test split as the hidden-state probe; because it has no access to activations, its accuracy estimates how much the prompt surface already reveals.

**Extracting directions** The default estimator is a persona vector (Chen et al., 2025). For a framework pair $(f_i, f_j)$, we take a fixed pool of scenario texts (the *trigger texts*), wrap each with two contrasting persona prompts, average the hidden states for each side, and unit-normalise their difference, $v_{ij} \propto \bar{h}_i - \bar{h}_j$. Because persona prompts can introduce their own artefacts, we compare them with three other estimators: principal components of contrastive activations, logistic-regression probe coefficients, and SAE feature directions where public SAEs are available (Marks & Tegmark, 2024). Each estimator yielding a basis that steers is one check that steerability is not an artefact of the persona recipe. Paraphrasing the persona wrappers preserves the qualitative geometry but changes, on most anchors, how many framework pairs pass the intervention test below (their per-pair *verdicts*; Appendix E.4), so we rest claims on patterns that replicate across models, wordings, and estimators rather than on any single count.

**Intervention test** We inject a direction into the residual stream near the analysis layer and measure whether the answer changes. In line steering (Rimsky et al., 2023), we add $\alpha v$ and record $P(A)$, the probability assigned to the accept or endorse option in a fixed binary answer scaffold that elicits a one-token A/B choice. The $\alpha$ scale is swept symmetrically and calibrated per model against saturation, so that ceiling-saturated and low-sensitivity models are compared against a matched random direction at the same scale. A direction counts as active under intervention only if its mean per-body correlation is non-trivial ($|\bar{r}| > 0.3$) and a matched-norm random direction does not produce the same signed effect. The random-direction condition matters because some models move toward the answer boundary under many large perturbations. The multi-model battery of §5.2 scores each run twice, once for movement toward the source foundation and once for movement toward a generic accept-versus-avoid answer, and counts a cell only when the effect exceeds all 10 matched random directions on a majority of three prompt wordings. The same intervention logic extends from one line to a subspace. A subspace claim says that several independently derived directions may be basis vectors in the same intervention space, so we orthonormalise them into a basis $B$. Subspace steering perturbs

within $\mathrm{span}(B)$, asking whether the space as a whole can move the answer. Subspace projection removes $\mathrm{span}(B)$ before decoding, asking whether taking away that space blocks in-subspace steering; it therefore serves as a linear diagnostic and defence. Hyperparameters, answer-token details, and utility metrics are in Appendix B.2.

### 4.4. What these methods identify, and what they do not

We use causal-mediation language for interventions (Mueller et al., 2024). Our pipeline supports the claim that a direction extracted from a framework contrast reliably tracks that contrast out of distribution and causally mediates the framework content of the scored output; this is weaker than the claim that the direction *is* the model's representation, and we treat the gap as explicit assumptions. Three threats define it. First, a probe can decode a correlate rather than the target (Belinkov, 2022; Elazar et al., 2021; Sharkey et al., 2025); the paraphrase-transfer, selectivity (Hewitt & Liang, 2019), topic-controlled, and valence-erasure controls of §5.4 each address one family of correlates, though erasure certifies only linear removal (Ravfogel et al., 2023). Second, an intervention can act through a co-varying property or a pathway the natural computation does not use (Makelov et al., 2023; Wu et al., 2024); beating matched random directions establishes causal relevance, not identity, and the in-distribution interchange test (§5.3) bounds how much of the contrast a single direction carries. Third, none of these makes the mediator unique: other subspaces could carry the same margins, and estimator agreement narrows but does not remove that freedom. Accordingly we write *ethical-framework signal* for what is measured (§2).

## 5. Results

### 5.1. Framework signals survive surface controls

The first question is whether hidden states add anything beyond the prompt surface. Probe accuracy by itself is not enough: if each label uses a different template or scenario pool, a classifier can recover the label without relying on framework-specific activations. We therefore compare each probe with the matched text-only baseline of §4.3. The two data sources play different roles in this check. ETHICS gives each of its five categories its own scenario pool and wrapper, so we read it through both the templated and the stripped conditions of §4.1. SC101 supplies the MFT side, where the five foundations share one prompt format.

Figure 2 shows the difference. The templated Western benchmark is almost fully recoverable from wording alone, and the stripped version still tracks the TF-IDF baseline closely. SC101 gives a stricter surface-control test: once the prompt format is shared, a small activation gap remains above the word-TF-IDF baseline. The gap is modest, so we do not build the paper on probe magnitude but use it as a locator: it shows that ethical-framework distinctions leave a recoverable signal beyond word-unigram surface cues. The next sections ask whether directions extracted at the fixed analysis layer affect scored outputs under intervention. A stricter version of this control appears in §5.4: the probe transfers across an external re-wording of every SC101 probe rule while the TF-IDF baseline drops under the same transfer.

Two details matter for interpretation. First, exact peak layers vary, so Fig. 2 should be read as a sweep, not as a tuned layer-selection rule. The surface-vs-hidden-state pattern is similar at 70–72B and 7–8B, and Appendix C.2 checks that the fixed analysis layer is not a unique sharp optimum. Second, the Western benchmark's virtue-ethics category is hard to read linearly in the stripped condition. This is consistent with the dataset design: virtue ethics asks about the character of an agent, not only the rightness of one act (Hendrycks et al., 2021). Later steering tests show that virtue-involving directions can still move outputs, so the weak probe result is a warning about readout, not evidence that the model lacks the framework signal.

### 5.2. Steering effects share a small route

The next question is causal: when a direction separates two frameworks in hidden state, does steering along it change the scored output, and which part changes? We begin with MFT, whose five foundations give nine comparable within-system pairs, so the steering, wording, and removal checks repeat without cross-dataset differences. Care and Fairness are *individualising* foundations, centred on harm to persons; Loyalty, Authority, and Sanctity are *binding* foundations, centred on group order and cohesion (Graham et al., 2009).

To avoid mistaking a move toward yes/no for a moral change, we score the same steering run in two ways. The first asks whether the next token moves toward the source foundation rather than the target; the second, whether it moves toward a generic accept-versus-avoid answer. The foundation score is the more stable effect (Fig. 3a). Across the nine-model expanded battery, canonical scoring gives foundation-score shifts roughly ten times as often as accept/avoid shifts, and the same ordering holds under an alternate prompt wording and on all but one model. Requiring both wordings retains about three quarters of the canonical foundation cells, so some pairs are wording-sensitive, but the pattern is not driven by one wording recipe. Endorsement is not absent; line steering just more reliably shifts which foundation the model emphasises than whether it accepts or rejects.

**The framework signal is not one line** A one-line account would predict that usable steering is explained by a single direction, or by small rotations of it. The evidence instead

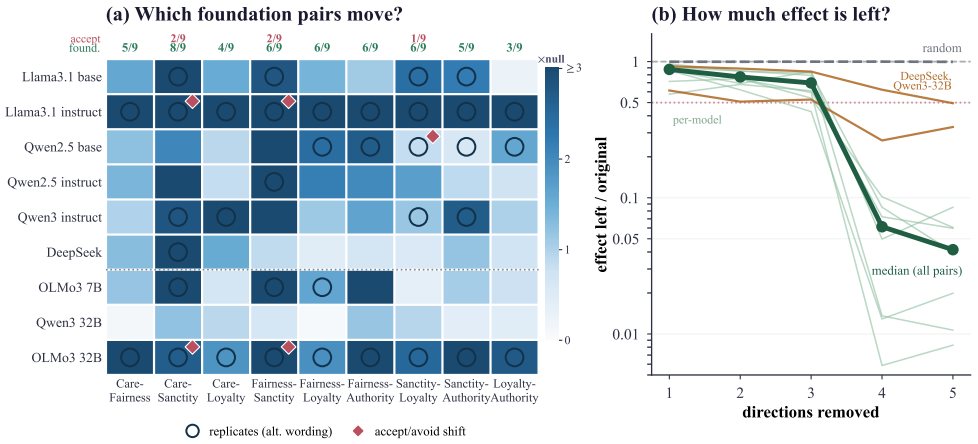


*Figure 3.* **Line steering more often changes which foundation is emphasised than whether the answer is simply accepted or rejected.** **(a)** Each cell is one model-pair in the nine-model expanded battery; shade gives the canonical foundation-shift strength as a multiple of its matched-null threshold (1 at threshold, clipped at 3; colourbar). A ring marks a shift that also replicates under the alternate wording; a pink diamond marks an accept/avoid shift. The per-column *found.* row counts the strict foundation verdict. Canonical foundation shifts are $49/81$ versus $5/81$ accept/avoid shifts; $36/81$ foundation shifts appear in both wordings (ringed), and the alternate wording gives $39/81$ foundation shifts versus $11/81$ accept/avoid shifts. **(b)** Same nine models and nine pairs across three prompt wordings, restricted to the 218 model-pair-wording cases where the foundation effect was stable (five removal levels, 1090 removal tests). For each target pair, we remove a one- to five-dimensional subspace spanned by the directions learned from the other eight pairs ($x$-axis). Faint curves are per-model medians; the bold curve is the median over all cases; the grey dashed line is the matched random-direction control. Other-pair directions erase much more of the original effect than matched random directions, suggesting a shared route; two models (DeepSeek, Qwen3-32B) retain a larger residual, so we do not claim that every model uses the same number of directions.

points to a small shared route. Bases built from disjoint pair selections share a plane (two principal-angle cosines saturate at their maximum, far above a near-zero random baseline), the signature of multiple bases of one subspace rather than noisy copies of one line. Alternative extractors also steer from bases that are near-orthogonal to the persona-vector one, so steerability does not depend on the persona recipe (Appendix D.2). Figure 3b asks a simple removal question: if we remove directions learned from the other foundation pairs, how much of the target-pair effect is left? The panel uses the same nine models and nine pairs as Fig. 3a across three prompt wordings, but keeps only the cases with a stable foundation effect. Across those removal tests, the median remaining effect collapses as more learned directions are removed, from most of the effect after removing one to near zero after removing five, while random directions leave it near-intact. Learned directions beat the random median on a large majority of cells as more directions are removed, and a target direction that lies more inside what we remove loses more of its effect. We read this as evidence that these effects share a route, not as a count of how many moral directions the model has.

**The causally-active axis is dominated by endorsement valence** Geometrically the moral directions are distinguishable (§5.3), but the part that is active under intervention is closer to a general endorsement axis than to moral-specific content. A kind-versus-cruel sentiment direction causally substitutes for the moral directions, and a polite-versus-rude register direction steers most anchors while overlapping mainly with sentiment; a non-moral chef-versus-programmer contrast does not reliably steer (Appendix D.3). Ablation tests this directly: orthogonalising the steering basis against the sentiment direction removes the causal verdict on three of the four anchors, though on two of those three the orthogonalised slopes stay above the verdict threshold and fail only the saturated random-direction comparison; valence therefore carries a large, on some models separable, share of the causal effect. So although the moral-content directions are geometrically separable, on most models they do not clear the causal-verdict bar once the shared endorsement-valence component is removed. Their operational rank is $\sim$3: the projection $k$-sweep's defence value plateaus at $k=3$, though the full span of moral directions is wider. We therefore characterise the dominant causally-active signal as an endorsement-valence axis that the moral-content directions project onto; §5.3 names this shared space, once its by-system structure is also in view. A topic-controlled test (§5.4) separates a smaller framework-specific component that remains causally active at the teacher-forced readout after removing topic and lexical cues. Free-generation effects, scored by an LLM judge on steered generations and reproduced by an independent open-weight judge (Appendix E), sit at the judge-noise floor, so we do not treat steering as rewriting moral reasoning.

**The same subspace supports attack and defence** If several directions share the same intervention space, single-line

steering and single-line removal are the wrong operational scale. Subspace steering uses a three-direction orthonormal basis instead of one line, raising intervention-active coverage from 2–6/9 foundation pairs to 8–9/9 on three of the four anchors with full basis coverage (Qwen2.5 instead drops to 0/9, the ceiling-attenuation regime of Appendix E.6). Subspace projection removes that basis before decoding; with adaptive $k$, as a linear mid-layer baseline it blocks the tested in-subspace attacks on all but one of the eleven tested models at capability cost matched to a random-projection control, with one out-of-distribution exception at 32B, a subspace-specific capability drop that the utility-drift screen behind the adaptive $k$ is designed to catch (Appendix E.3). As a linear mid-layer intervention it does not rule out nonlinear or cross-layer bypasses, so we read it as a baseline rather than a complete defence. A separate check on MoralChoice, a benchmark of binary moral decision scenarios (Scherrer et al., 2023), scored with order-counterbalanced options, shows these interventions produce no detectable net change in the model's chosen action (§5.4).

**Instruction tuning rotates an existing signal** Because the subspace is active under intervention, we ask whether it is introduced late in alignment training or reshaped from an earlier signal. Training-stage checks argue against a purely final-preference-learning origin. Base-model and instruct-model directions look geometrically dissimilar, but steering effects are already detectable in base models. On OLMo3, the only family here with public intermediate post-training checkpoints, the largest per-pair cosine jump toward the final instruct direction occurs at SFT, several times larger than the jumps under the later DPO and RLVR stages, a pattern OLMo3.1-32B replicates (Appendix D.4). A non-moral control in §5.4 shows this locus is generic to instruction tuning rather than specific to moral content.

### 5.3. Frameworks group by system, but share an endorsement subspace

Sharing an endorsement subspace does not rule out collapse into one axis. We test this two ways, on the three systems with native contrastive pools (MFT from SC101, Western frameworks from ETHICS, Confucian *Wuchang* from FLAMES): whether directions from the same system are closer to each other than to other systems' (geometric), and whether a direction from one system steers outputs in another's test pool (functional).

Geometrically, directions group by system (Fig. 4a). Within-system $|\cos|$ exceeds cross-system $|\cos|$ on every in-scope model, by a margin that clears the null 95%-CI upper bound in each case (two out-of-scope Gemma-3 variants separate only marginally). The directions are therefore not a single collapsed axis. They are also not mutually orthogonal: each system occupies a distinguishable region inside a common low-dimensional space. This space, distinguishable by moral system yet dominated under intervention by the endorsement-valence axis of §5.2, is what we call the *endorsement subspace* (Fig. 1).

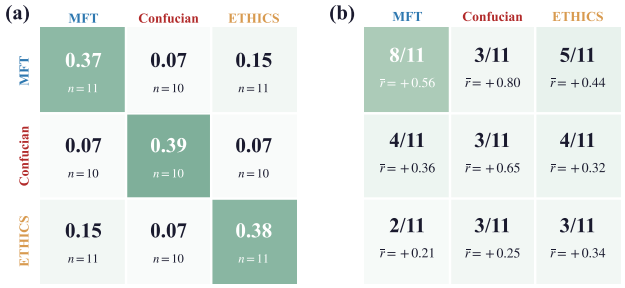


*Figure 4.* **Two views of the same three systems. (a) Geometric:** mean pairwise $|\cos|$ between persona vectors; diagonal = within-system, off-diagonal = cross-system; $n$ = models contributing to that cell. **(b) Functional:** $3 \times 3$ causal-transfer matrix (rows = source system, sharing (a)'s row labels; columns = target test pool). Each cell reports $k/11$, the number of models on which the source direction is causal for that target pool ($|\bar{r}| > 0.3$ with sign opposite to a single matched random direction, $B$=1), and $\bar{r}$, the signed mean slope over those causal models (opposite-signed models cancel).

Functionally, one system's directions can substitute for another's (Fig. 4b; per-model verdicts in Appendix G.1). MFT pools are steered most reliably by MFT directions, but the Western-framework pools show the reverse: MFT directions steer them on more models than the directions extracted within that pool itself. This matches §5.1's data-source contrast: SC101 gives MFT a shared prompt format and a cleaner axis, while ETHICS mixes framework labels with separate scenario pools, so directions extracted there are less distinct from broad endorsement valence. Cross-language checks agree: Chinese MFT directions steer most tested Instruct anchors despite near-zero geometric similarity to English MFT directions, and the remaining anchor's null verdict reflects unstable extraction on Chinese prompts rather than absent encoding (Appendix G.2).

Framework directions thus keep system-level structure while their intervention effects pass through a shared endorsement subspace. For alignment this matters: a defence anchored to one system's direction can under-detect attacks from another system inside the same subspace, and Western-framework defences are the more exposed case under the asymmetry above.

How much of this signal does any single direction carry? An in-distribution interchange test bounds it: transplanting a direction's component between the two runs of a topic-controlled counterfactual pair, keeping patched values in distribution (Wu et al., 2024), moves the framework margin toward the counterfactual and beats every matched random direction on nearly all (model, pair) cells. Yet any single

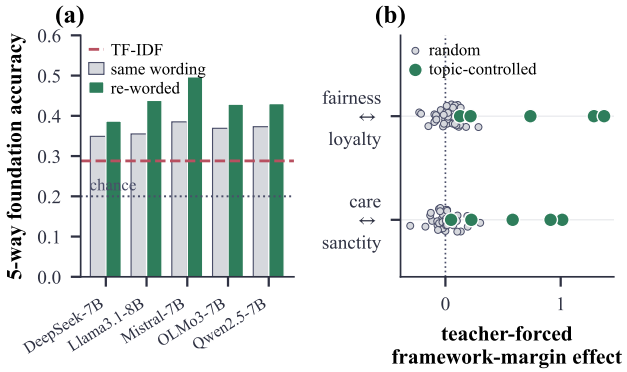


*Figure 5.* **Robustness controls. (a) Probe accuracy under re-wording:** the 5-way foundation probe transfers across an external re-wording of every SC101 probe rule, matching or exceeding its same-wording accuracy, while the word-TF-IDF baseline falls from 0.34 to the plotted 0.29: the activation signal is not carried by the wording. **(b) Topic-controlled direction vs. random nulls:** a foundation direction estimated from topic-controlled, lexicon-banned counterfactual pairs steers the teacher-forced framework readout beyond every random-direction null matched to the same model and pair ($K$=10 random directions per model and pair), on both foundation pairs shown and all five anchors. Grey nulls are pooled across anchors for display, so a green dot can sit inside the pooled spread while clearing its own matched nulls.

direction carries only a small fraction of the pooled margin, and widening the transplanted subspace to eight principal components adds little more (Fig. 12). The framework contrast at the readout is causally real but distributed across the subspace.

### 5.4. Identification and confound controls

Five controls probe the main results' most likely confounds; Fig. 5 shows two.[2]

**Framework decodability survives external re-wording** We paraphrase every SC101 probe item with an external model, train the 5-way probe on original wordings, and test on paraphrases (and the reverse). Transfer accuracy matches or exceeds within-wording accuracy on the five models (five-way chance 0.20), while the word-TF-IDF baseline falls under the same transfer. The activation signal generalises across wording; the surface baseline does not.

**A topic-controlled foundation direction is causally active** Diff-of-means directions inherit whatever co-varies with their trigger pools. We therefore construct counterfactual minimal pairs, generated by an external model and independently validated: the same situation verbatim, the same verdict, varying only the moral perspective, with the foundation lexicon banned (four pairs, $n$=64–88; full per-pair results in Appendix F). The direction estimated from these pairs is nearly orthogonal to the persona-based one despite high split-half reliability, and its steering effect at the teacher-forced readout exceeds every matched random-direction null in each (model, pair) cell. A framework-specific causal component is therefore separable from topic and lexical cues at the readout, refining rather than replacing the endorsement-dominance picture of §5.2; the counterfactual pairs come from a single generation provider, so this control bounds provider idiosyncrasy rather than excluding it.

**Valence erasure leaves the framework readout unchanged** A LEACE eraser fit on a kind-versus-cruel valence contrast (held-out fit) leaves 5-way foundation accuracy unchanged on all five anchors ($|\Delta| \leq 0.005$), while the erased valence readout itself drops substantially on held-out data, short of chance. The eraser is rank-1 and removes valence only partially, so we read this as a bounding control rather than a proof of valence-independence.

**Decision-level null survives position debiasing** MoralChoice places the morally preferred action at the same option position in every scenario, so uncounterbalanced scoring saturates toward that position and biases flip counts toward null. We therefore counterbalance option order on all five 7–8B anchors and score the position-debiased probability of the preferred action. High-ambiguity baselines calibrate near 0.5, and steering still produces no net effect: at $\alpha$=3, steer-consistent action flips are rare and outnumbered by anti-directional flips; mean $\Delta P$ is indistinguishable from zero. A permutation null over the free-generation judge labels similarly places the observed toward-source fraction well inside the noise distribution.

**The SFT locus is generic to instruction tuning** Non-moral persona contrasts (formal/casual register, verbose/terse style) traced across the same OLMo-3 checkpoints rotate into place at SFT like the moral pairs. The checkpoint result therefore locates when persona-contrast geometry forms in general, and does not isolate a moral-specific curriculum.

### 5.5. Cross-cultural moral geometry

With the causal picture established on the three systems with native test pools, we extend the analysis to all six: does the model group traditions by philosophical kinship, geography, or dataset origin (Fig. 13)?

Within-system clustering is uneven but stable across the 11 models (Fig. 13 inset): Buddhist directions are looser, consistent with the Eightfold Path covering wider practice, while Stoic and Ubuntu are tighter, consistent with short codified virtue or communal lists.

[2]Unless noted, each runs on five 7–8B instruct models: the Llama3.1, Qwen2.5, OLMo3, and DeepSeek anchors and an out-of-family Mistral-7B-Instruct (Jiang et al., 2023).

Buddhist, Stoic, and Ubuntu sit close together: across models, the densest off-diagonal system similarities tend to couple these three. We use "virtue-tradition" in the broad philosophical sense: these systems frame moral life around cultivated character, disciplined practice, or right relation to others. ETHICS also contains a virtue-ethics category, but its pair directions sit no closer to the trio than other Western-framework directions (Appendix G), so the model can hold Western virtue ethics inside that inventory while these three traditions form a broader neighbourhood of their own.

The grouping is plausible against philosophical history, but it is bounded. These three systems use hand-authored persona prompts; paraphrase checks preserve direction sign on the large majority of cells, and lexical overlap within the trio, though elevated above the cross-cluster baseline, does not by itself account for it. More fundamentally, this geometry is about how the model represents *text about* these traditions, shaped by its largely English training corpus, not the traditions' own moral structure, which our prompts do not directly access.

The cross-cultural implication is modest: prompting a non-Western moral system can activate a recognisable direction inside the shared subspace, but this is not the same as reasoning within that tradition in native context. Buddhist, Stoic, and Ubuntu remain geometric results until dedicated causal test pools exist.

## 6. Discussion

The model could have routed each framework contrast through a separate intervention pathway. Instead, our evidence points to framework directions, grouped by system, feeding a shared endorsement subspace. This explains the central tension in the results. Directions remain geometrically distinguishable, yet steering one can substitute for another, and in-distribution interchange bounds any single direction at a small fraction of the counterfactual margin (§5.3). The operational unit is the subspace, not any one direction. For alignment practice, moral steering is then a lever over moral *expression* more than moral *decision* (§5.4): *which* framework is invoked and *whether* the model affirms are separable from what the model does. An audit that certifies only the chosen action therefore leaves the framing channel unmeasured.

Causal usability and correlational transfer dissociate in both directions: Chinese MFT directions steer despite near-zero geometric similarity to English MFT directions, while Western-framework pairs can transfer correlationally through shared pool style yet steer less specifically. This is why our evidence sequence follows the pattern used in representation-geometry work: visualise or probe a candidate structure, test transfer, and then intervene. For moral benchmarks with different scenario pools across labels, causal evaluation should accompany correlational probing.

Liu et al. (2025) argues that LLM ethics gains can reflect distributional similarity rather than pragmatic moral inference. Our results do not refute that behavioural reading. They add a representational one: SC101 leaves a modest ethical-framework signal above a text-only baseline, later directions from the same region are usable under intervention, and the six-system geometry shows distinguishable traditions feeding a shared intervention subspace. This mirrors the broader distinction between what a model can compute and what its output channel reveals (Ganguli et al., 2023; Mahowald et al., 2024). Cross-cultural moral evaluation therefore needs both levels. A behavioural rubric asks what the model says in a scenario; a representational analysis asks what framework directions are available for the model to use, even when the final output does not reveal them.

## 7. Conclusion

We asked whether framework-level moral content in LLMs is organised as separate system-specific subspaces or as one endorsement axis, and find an intermediate structure: directions grouped by moral system, including non-Western traditions, are geometrically distinct, yet their intervention effects pass through a shared low-dimensional endorsement subspace, already visible in base models and reshaped mainly at supervised fine-tuning. Because a single direction carries only a minority of the framework contrast, the handle on moral framing is real but distributed across the subspace, so reading moral content out of a model, especially across traditions, is as much a subspace-level and representational question as a behavioural one.

## Acknowledgements

Computations were performed on the EPFL RCP cluster. We thank the EPFL Digital Humanities Lab for compute support.

## References


Dhammacakkappavattana Sutta (the discourse on the turning of the wheel of the Dhamma), saṃyutta nikāya 45.8. Pāli canon, Theravāda Buddhist tradition; translation used Bhikkhu Bodhi, Wisdom Publications, 2000, circa 5th century BCE.

Abdulhai, M., Serapio-García, G., Crépy, C., Valter, D., Canny, J., and Jaques, N. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17737–17752, 2024.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.11717.

Atari, M., Xue, M. J., Park, P. S., Blasi, D., and Henrich, J. Which humans? *PsyArXiv preprint*, 2023. doi: 10.31234/osf.io/5b26t.

Bailey, L., Serrano, A., Sheshadri, A., Seleznyov, M., Taylor, J., Jenner, E., Hilton, J., Casper, S., Guestrin, C., and Emmons, S. Obfuscated activations bypass LLM latent-space defenses. *arXiv preprint arXiv:2412.09565*, 2024.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 861–872, 2017.

Boleda, G. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234, 2020.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2023.

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

Chiu, Y. Y. et al. MoReBench: Evaluating procedural and pluralistic moral reasoning in language models. *arXiv preprint arXiv:2510.16380*, 2025.

Cicero, M. T. *De Officiis*. 44 BCE. Book I, Chapter 15; Loeb Classical Library translation by Walter Miller, 1913.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

for AI, A. I. OLMo 3: Fully open language models with training checkpoints. *arXiv preprint*, 2025. Model card: https://huggingface.co/allenai/OLMo-3-7B-Instruct.

Forbes, M., Hwang, J. D., Shwartz, V., Sap, M., and Choi, Y. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 653–670, 2020.

Ganguli, D., Askell, A., Schiefer, N., Liao, T., et al. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*, 2023.

Graham, J., Haidt, J., and Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046, 2009.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Guo, D., Yang, D., Zhang, H., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.00000*, 2025.

Haas, J., Bridgers, S., Manzini, A., Henke, B., May, J., Levine, S., Weidinger, L., Shanahan, M., Lum, K., Gabriel, I., et al. A roadmap for evaluating moral competence in large language models. *Nature*, 650:565–573, 2026.

Hämmerl, K., Deiseroth, B., Schramowski, P., Libovický, J., Rothkopf, C. A., Fraser, A., and Kersting, K. Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

He, J. and Jin, Y. SAE-SSV: Supervised steering in sparse representation spaces for reliable control of language models. *arXiv preprint arXiv:2505.16188*, 2025.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. Aligning ai with shared human values. In *International Conference on Learning Representations*, 2021.

Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Hongladarom, S. *The Ethics of AI and Robotics: A Buddhist Viewpoint*. Lexington Books, 2020.

Huang, K., Liu, X., Guo, Q., Sun, T., Sun, J., Wang, Y., Zhou, Z., Wang, Y., Teng, Y., Qiu, X., Wang, Y., and Lin, D. FLAMES: Benchmarking value alignment of LLMs in Chinese. In *Proceedings of NAACL*, 2024.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jiang, L., Hwang, J. D., Bhagavatula, C., Le Bras, R., Liang, J. T., Levine, S., Dodge, J., Sakaguchi, K., Forbes, M., Hessel, J., Borchardt, J., Sorensen, T., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., et al. Investigating machine moral judgement through the Delphi experiment. *Nature Machine Intelligence*, 7:145–160, 2025.

Lee, A., Bai, X., Pres, I., Wattenberg, M., Kummerfeld, J. K., and Mihalcea, R. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *NeurIPS*, 2024.

Liu, G., Qi, Z., Zhang, X., Jiang, L., and Johnson, K. M. Diagnosing moral reasoning acquisition in language models: Pragmatics and generalization. *Findings of EMNLP*, 2025. arXiv:2502.16600.

Long, A. A. *Epictetus: A Stoic and Socratic Guide to Life*. Oxford University Press, 2002.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024.

Makelov, A., Lange, G., and Nanda, N. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *International Conference on Machine Learning*, 2024.

Metz, T. Ubuntu as a moral theory and human rights in South Africa. *African Human Rights Law Journal*, 11(2): 532–559, 2011.

Mhlambi, S. From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance. *Carr Center Discussion Paper Series 2020-009, Harvard Kennedy School*, 2020.

Mueller, A., Brinkmann, J., Li, M., Marks, S., Pal, K., Prakash, N., Rager, C., Sankaranarayanan, A., Sen Sharma, A., Sun, J., Todd, E., Bau, D., and Belinkov, Y. The quest for the right mediator: Surveying mechanistic interpretability through the lens of causal mediation analysis. *arXiv preprint arXiv:2408.01416*, 2024.

Preniqi, V., Ghinassi, I., Ive, J., Saitis, C., and Kalimeri, K. Moralbert: A fine-tuned language model for capturing moral values in social discussions. In *Proceedings of the 4th International Conference on Information Technology for Social Good*, 2024.

Ramezani, A. and Xu, Y. Knowledge of cultural moral norms in large language models. In *Proceedings of ACL*, 2023.

Rao, A., Khandelwal, A., Tanmay, K., Agarwal, U., and Choudhury, M. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13370–13388, 2023.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Ravfogel, S., Goldberg, Y., and Cotterell, R. Log-linear guardedness and its implications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9413–9431, 2023.

Reiter, R. M. and Kádár, D. Z. (Im)politeness and (Im)morality: Insights from intervention. *Journal of Politeness Research: Language, Behavior, Culture*, 11(2): 239–260, 2015.

Ren, Y., Ye, H., Fang, H., Zhang, X., and Song, G. ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2015–2040, 2024.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama-2 with contrastive activation additions. *arXiv preprint arXiv:2312.06681*, 2023.

Samway, K., Kleiman-Weiner, M., Guzman Piedrahita, D., Mihalcea, R., Schölkopf, B., and Jin, Z. Are language models consequentialist or deontological moral reasoners? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pp. 30699–30726, Suzhou, China, 2025. Association for Computational Linguistics. arXiv:2505.21479.

Sap, M., Le Bras, R., Fried, D., and Choi, Y. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3762–3780, 2022.

Scherrer, N., Shi, C., Feder, A., and Blei, D. M. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51778–51809, 2023.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., and Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4:258–268, 2022.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Sorensen, T., Jiang, L., Hwang, J., Levine, S., Pyatkin, V., West, P., et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19937–19947, 2024.

Stecker, M. and Hopp, F. R. Moral foundation measurements fail to converge on multilingual party manifestos. *Political Analysis*, 34(2):166–187, 2026.

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

Tenney, I., Das, D., and Pavlick, E. Bert rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

Tigges, C., Hollinsworth, O. J., Geiger, A., and Conmy, A. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

Tlaie, A. Exploring and steering the moral compass of large language models. *arXiv*, 2024.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023.

Voita, E. and Titov, I. Information-theoretic probing with minimum description length. *EMNLP*, 2020.

Wang et al. Enhancing LLM steering through sparse autoencoder-based vector refinement. *arXiv preprint arXiv:2509.23799*, 2025.

Wei, B., Huang, K., Huang, Y., Xie, T., Qi, X., Xia, M., Mittal, P., Wang, M., and Henderson, P. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Wu, F., Hu, J., Min, G., and Wang, S. Efficient orthogonal fine-tuning with principal subspace adaptation. In *International Conference on Learning Representations*, 2026. arXiv:2505.11235.

Wu, Y., Sheng, Q., Wang, D., Yang, G., Sun, Y., Wang, Z., Bu, Y., and Cao, J. The staircase of ethics: Probing llm value priorities through multi-step induction to complex moral dilemmas. *EMNLP*, 2025. arXiv:2505.18154.

Wu, Z., Geiger, A., Huang, J., Arora, A., Icard, T., Potts, C., and Goodman, N. D. A reply to makelov et al. (2023)'s "interpretability illusion" arguments. *arXiv preprint arXiv:2401.12631*, 2024.

Xiong, Y. and Xie, X. OPLoRA: Orthogonal projection LoRA prevents catastrophic forgetting during parameter-efficient fine-tuning. *arXiv preprint arXiv:2510.13003*, 2025.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhao, J. et al. Llms encode harmfulness and refusal separately. *arXiv preprint arXiv:2507.11878*, 2025.

Zhou, J., Hu, M., Li, J., Zhang, X., Wu, X., King, I., and Meng, H. Rethinking machine ethics: Can llms perform moral reasoning through the lens of moral theories? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2227–2242, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.144. URL https://aclanthology.org/2024.findings-naacl.144/.

Zhu, R.-J., Peng, T., Cheng, T., Qu, X., Huang, J., Zhu, D., Wang, H., Xue, K., Zhang, X., Shan, Y., Cai, T., Kergan, T., Kembay, A., Smith, A., Lin, C., Nguyen, B., Pan, Y., Chou, Y., Cai, Z., Wu, Z., Zhao, Y., Liu, T., Yang, J., Zhou, W., Zheng, C., Li, C., Zhou, Y., Li, Z., Zhang, Z.,

Liu, J., Zhang, G., Huang, W., and Eshraghian, J. A survey on latent reasoning. *arXiv preprint arXiv:2507.06203*, 2025.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

# A. Limitations

The central result, that framework directions grouped by moral system are geometrically distinct while their intervention effects pass through a shared endorsement subspace, is supported by the geometric, causal, and checkpoint evidence reported above. The findings are bounded in four ways: by the data and models we test, by where we intervene in the network, by how far generation-level behaviour can be read from the representation, and by the narrower evidence behind provenance and defence claims.

**Coverage** The data sources in §4.1 rely on persona-wrapped or adversarial prompts; naturalistic dilemma corpora (Wu et al., 2025; Chiu et al., 2025) may expose structure these prompts miss. Model coverage spans 11 variants from 5 open-weight families, but excludes commercial systems and mixture-of-experts variants (§B.3). The diff-of-means direction estimate is also uneven across datasets: it is cleaner when labels share a prompt pool, and more fragile on ETHICS, where frameworks come from separate scenario pools. A bag-of-words control quantifies how much framework *identity* these corpora carry in surface lexical content alone: on SC101 a TF-IDF classifier recovers the five-way foundation label at $37\%$ against a $20\%$ chance rate, with activation probes at $42$–$45\%$ on the same pool, and on ETHICS the template-stripped scenario bodies remain $87\%$ lexically classifiable, since the five frameworks draw on disjoint source corpora. Framework-identity separability in these datasets is therefore largely a scenario-topic signal, and any estimator that contrasts frameworks across scenario pools inherits the confound; the topic-controlled counterfactual design of §5.4, which holds the scenario fixed and varies only the framing, is the measurement that removes it, and we are extending that design into a full estimation method in ongoing work. Buddhist, Stoic, and Ubuntu have no native benchmark here, so their hand-authored prompts may contribute to the geometry. Paraphrases preserve direction sign in $70/80$ cells, but per-pair verdict counts remain wrapper-conditional on most anchors (§E.4); lexical controls bound the issue without removing it. The counterfactual topic control (§5.4) covers four foundation pairs, and its pairs are generated and validated by models from a single provider; a broader pair set and human validation are natural extensions. Other extractors, such as MDL (Voita & Titov, 2020), CCS (Burns et al., 2023), attention-head ITI (Li et al., 2024), or SAEBench-ranked SAEs, may change magnitudes. All main results use fp16; NF4 is reported only as a stress check because it saturates the answer boundary on the 70B-class tier.

**Structural scope** The subspace-rank, cross-system transfer, and projection-defence claims are tested at the fixed analysis layer $L \approx 0.65N$ and the last-token position. Layer sweeps show direction preservation across depth (§E.2b), but earlier-token, answer-span-averaged, and attention-head-selective interventions remain untested. The fixed-layer heuristic also behaves differently by scale: it falls before the layerwise cosine sign-flip at 8B but after it at 70B, so the cluster claim survives at scale while the heuristic itself should not be treated as universal. The per-model $\alpha$ calibration targets attack shifts in $[0.015, 0.060]$; when both the attack and the random-direction null hit the answer boundary, the sign-flip verdict becomes conservative. Placebo and orthogonalisation checks cover only the models with the required pre-extracted directions.

**Behavioural scope** The LLM-as-judge eval in §5.2 and §E.1 measures primary-label changes, not full audits of the reasoning path. Because the tested directions are cross-foundation contrasts, part of the foundation-flip rate is built into the direction design; single-foundation directions or larger $\alpha$ may behave differently. The supplementary individualising–binding verdict check survives multi-seed null, placebo, layer-fraction, and SAE-$K$ checks, but within-cluster non-specific verdicts still have several possible readings: absent within-cluster causal encoding, saturation at $\alpha = \pm 3$, or a different pathway than the one we steer. A larger direction basis and layer-wise steering would separate these accounts.

**Provenance** The SFT-rotation claim rests on OLMo3, the only family here with public intermediate checkpoints, although the Base→Instruct asymmetry also appears on two families. Stronger provenance claims need comparable checkpoints or controlled pseudo-checkpoints, such as staged LoRA fine-tunes on moral-content subsets. Capability preservation under subspace projection is tested on factual, reasoning, and language-modelling tasks, but not on benchmarks that reward sustained affirmative or consensual language, such as AlpacaEval or sycophancy. Those settings could reveal over-suppression, or show that the adaptive-$k$ fallback catches it.

# B. Methodology, models, and prompts

## B.1. Detailed moral system descriptions

This appendix gives the humanities-level description of each of the six moral systems whose representations we probe. The main text (§2) names them and identifies the data source used.

**ETHICS (Hendrycks et al., 2021)** operationalises Western normative ethical theory through five framework categories. *Deontology* (rule- and duty-based, after Kant) judges acts by reference to moral principles regardless of consequences. *Utilitarianism* (consequentialist) judges by the aggregate welfare an action produces. *Virtue ethics* (after Aristotle) judges by what a person of good character would do. *Justice* judges by fairness, rights, and impartiality. *Commonsense morality* captures widely shared everyday intuitions about right and wrong.

**Moral Foundations Theory (MFT) (Graham et al., 2009; Forbes et al., 2020)** catalogues the categories of moral concern humans appeal to: *Care*/harm, *Fairness*/cheating, *Loyalty*/betrayal, *Authority*/subversion, and *Sanctity*/degradation. Graham et al. (2009) argue these group into two clusters: *individualising* (Care, Fairness; concern for the individual) and *binding* (Loyalty, Authority, Sanctity; concern for group cohesion). The two-cluster structure is documented across cultures and political affiliations and is the principal axis of variation in human moral concern.

**Confucian *Wuchang*** (literally "five constants") are the five cardinal virtues of classical Confucian ethics: *ren* (benevolence and humaneness), *yi* (righteousness and moral duty), *li* (propriety and ritual order), *zhi* (wisdom and moral discernment), and *xin* (trustworthiness and integrity). *Wuchang* carries a cross-language load: scenarios are drawn from FLAMES (Huang et al., 2024) *Morality*, native Chinese moral situations whose body text directly invokes canonical Confucian dicta (*zhongyong*, *jianshou*, Analects sayings) rather than translated proxies.

**Buddhist Noble Eightfold Path** (Sanskrit *āryāṣṭāṅgamārga*; Chinese *bāzhèngdào*) names the eight factors of Buddhist moral practice as canonicalised in the Saṁyutta Nikāya 45.8 (bud, circa 5th century BCE): *sammā-diṭṭhi* (Right View), *sammā-saṅkappa* (Right Intention), *sammā-vācā* (Right Speech), *sammā-kammanta* (Right Action), *sammā-ājīva* (Right Livelihood), *sammā-vāyāma* (Right Effort), *sammā-sati* (Right Mindfulness), and *sammā-samādhi* (Right Concentration).

**Stoic cardinal virtues** (Cicero, 44 BCE; Long, 2002) are the four canonical virtues enumerated by Cicero (*De Officiis* I.15): practical *wisdom* (*sophia/phronēsis*), *courage* (*andreia*), *justice* (*dikaiosynē*), and *temperance* (*sōphrosynē*).

**Ubuntu** is the southern African communitarian moral system articulated in the Akan-family proverb *umuntu ngumuntu ngabantu* (a person is a person through other persons); we adopt the four-feature operationalisation of Metz (2011): *identifying with others* (shared identity), *exhibiting solidarity* (commitment to others' well-being), *not harming* (non-exploitation, non-discrimination), and *maintaining harmony* (reconciliation and relational repair).

Each system's complete enumeration is used as released; we do not truncate or reweight category lists to align across systems, so cross-system comparisons run against the full system-defined sets.

### B.2. Probing, steering, and methodology details

**Linear probing (default)** L2-regularised logistic regression with $C = 1$, $N = 2000$ training examples and 500 test examples per framework, drawn from the corresponding dataset's per-framework subsets.

**Selectivity (Hewitt & Liang, 2019).** For framework $f$,

$$S_f = \mathrm{acc}_f^{\text{real}} - \overline{\mathrm{acc}_f^{\text{rand}}},$$

where $\overline{\mathrm{acc}_f^{\text{rand}}}$ averages probe accuracy under $K = 5$ label permutations (each shuffles which prompt receives which framework label). A high $S_f$ means the probe extracts label-specific structure; a low $S_f$ means the same probe fits arbitrary labels equally well, so its real-label accuracy reflects expressive capacity rather than concept structure.

**Cross-framework transfer** Train a probe on $f_{\text{train}}$ and evaluate on $f_{\text{test}}$ ($f_{\text{test}} \neq f_{\text{train}}$); asymmetric drop $\Delta_f = \mathrm{acc}_f^{\text{native}} - \overline{\mathrm{acc}_f^{\text{off-diag}}}$.

**Layer sweep** 5-way multinomial LR at each transformer block, $N = 500$ prompts per framework per condition.

**Persona-vector evaluation** Three diagnostics: in-distribution separation on held-out wrapped pairs; transfer separation on unwrapped native bodies; 5-shuffle label-permuted baseline.

**Causal steering** The line-steering / subspace-steering / subspace-projection pipeline injects one block before the canonical layer; sweep $\alpha \in \{-3, \ldots, +3\}$ over 80 acceptability bodies per pair; record $P(A), P(B)$. Subspace-steering basis

dimension $k = 3$; norm matched to single-direction line steering ($\|v_{\text{ss}}(\alpha)\| = |\alpha|$). Subspace projection uses the same $k = 3$ basis.

**Acceptance-probability operationalisation** The answer position is the next-token slot after the prompt scaffold "`Answer (A or B):`"; $P(A)$ and $P(B)$ are softmax probabilities at that slot for the model-specific single-token IDs of "A" and "B" (resolved with the model's own tokenizer per anchor; we verify single-token mappings on every model). Values are raw post-softmax probabilities with no calibration step beyond the random-direction null comparison the per-pair verdict already performs. The slope metric is per-body Pearson $r$ between $\alpha$ and $P(A)$; each pair's reported slope is the mean across the 80 bodies (verdict definition in §4.3). Robustness to alternative answer-token sets (e.g., `Yes`/`No`, `True`/`False`) is left untested; we use `A`/`B` throughout to fix one lexical anchor across all 11 models.

**Random-direction baseline** For all $|\cos|$ figures, the random-direction 95th percentile is reported per-figure as the "near-orthogonal" reference (red dashed line in Fig. 7 caption; null 95%-CI in Table 11).

### B.3. Model coverage and hardware notes

**Excluded model families** We exclude mixture-of-experts (MoE) variants such as Llama-4, Qwen3-235B, and GLM-5 because MoE routing partitions the residual stream per-expert and requires a separate analytical framework that does not directly compose with our PV / SAE / projection pipeline. We also exclude Gemma-4 variants with Per-Layer Embeddings (PLE), because the residual stream carries a PLE side-channel that our subspace projection projection does not cover.

**Hardware** All experiments run in fp16 / bf16 (per-model config) on A100-80GB GPUs; the two 70-class models (Llama3.1-70B-Inst, Qwen2.5-72B-Inst) use 2×A100 with `device_map="auto"` sharding.

### B.4. Persona prompts

Table 1 lists the 31 persona prompts used as PV-extraction wrappers across the six moral systems. Each persona prompt precedes a stripped commonsense trigger via the scaffold *"Consider the following situation: <body>. Reflect carefully on this from your moral standpoint."*, and the persona vector is the diff of last-token activations between the $A$- and $B$-wrapped versions of the same trigger pool. Buddhist, Stoic, and Ubuntu personas (sixteen in total) are hand-authored in English to foreground each system's own vocabulary rather than an MFT-mapped paraphrase; Confucian prompts are authored in classical Chinese with English glosses shown.

## C. Layered framework signal: §5.1 supplements

### C.1. TF-IDF stylometric baseline

The TF-IDF baseline used in §5.1 asks how well a model-free reader of the prompt surface can recover framework labels. For each condition, we use the same balanced examples, labels, and 80/20 train/test split as the layer-sweep probe, but replace residual-stream activations with word-unigram TF-IDF features of the raw prompt text. These features capture recurring lexical cues and dataset-specific wording without using any neural representation. An $L_2$-regularised multinomial logistic regression maps the features to framework labels. $N = 2{,}500$ examples per condition (5 frameworks $\times$ 500 each), $K = 5$ random-label seeds for the matched selectivity control. Hidden-state probe peaks are aggregated across the 11 models using the maximum over extracted residual layers.

The TF-IDF score (Table 2) is what a model-free reader of the prompt text alone could recover. Templated ETHICS prompts are framework-recoverable from morphology alone; stripping the template removes most of that, leaving a surface-style residual that TF-IDF still tracks closely. Only on SC101 (one shared template across foundations) does the hidden-state probe open a measurable gap above TF-IDF: this gap is the framework-concept residual we attribute to the model.

### C.2. Per-model layer sweep, selectivity, and CFT matrices

**Layer-sweep robustness checks** Two conditions dropped from Fig. 2 for clarity replicate the surface-vs-concept reading of §5.1. The *less-biased* variant of templated ETHICS (minimal generic wrapper) tracks the as-released templated condition within $\leq 0.012$ peak-accuracy on every one of the 11 models, so the same probe finding is recoverable without the

| Key | Persona prompt (*"You are a ⟨descriptor⟩ moral reasoner centered on . . . "*) |
|---|---|
| **ETHICS frameworks** *(5 personas)* | |
| deon | . . . deontological. Evaluate actions by duties, rules, and moral principles, independent of consequences. A rule-violation is wrong in itself; upholding moral law matters more than maximising good outcomes. |
| util | . . . consequentialist. Evaluate actions by aggregate welfare and the overall balance of benefit over harm. An action is right iff it produces the best consequences. |
| virtue | . . . virtue ethicist. Evaluate actions by what a person of good character would do, emphasising courage, honesty, justice, practical wisdom. An action is right if it expresses virtuous character. |
| ethjustice | . . . justice-oriented. Evaluate actions by fairness, equality, rights, impartial treatment. An action is right if it gives each person their due. |
| ethcs | . . . commonsense. Evaluate actions by widely shared everyday intuitions about right and wrong. |
| **MFT / SC101** *(5 personas)* | |
| care | . . . care and harm. Evaluate by whether actions protect others from suffering, show compassion, prevent cruelty. |
| fairness | . . . fairness and justice. Evaluate by equitable treatment, rights, avoidance of discrimination. |
| loyalty | . . . loyalty and in-group bonds. Evaluate by honouring commitments to group / family / community / nation. |
| authority | . . . authority and tradition. Evaluate by respect for legitimate authority, hierarchy, tradition. |
| sanctity | . . . sanctity and purity. Evaluate by maintenance of moral purity, avoidance of degradation, honouring sacred values. |
| **Confucian *Wuchang*** *(5 personas, classical Chinese originals; English gloss shown)* | |
| ren | . . . *ren* (humaneness/benevolence). Judge by compassion, love of others, universal benevolence. Loss of *ren* renders an act non-moral. |
| yi | . . . *yi* (righteousness/moral duty). Judge by accord with the right path; weigh benefit against duty; unrighteous acts must not be done. |
| li | . . . *li* (propriety/ritual). Judge by conformity to ritual, hierarchical order, self-discipline. "Non-*li* should not be seen, heard, said, or done." |
| zhi | . . . *zhi* (wisdom/discernment). Judge by clear discrimination of right and wrong, knowledge of self and others, investigative wisdom. |
| xin | . . . *xin* (trustworthiness/integrity). Judge by words matching deeds, honesty, keeping promises, not deceiving. |
| **Buddhist Noble Eightfold Path** *(8 personas; Pāli in parentheses)* | |
| bud_view | Right View (*sammā-diṭṭhi*). Judge by clear understanding of dependent origination, impermanence, the Four Noble Truths. |
| bud_intention | Right Intention (*sammā-saṅkappa*). Judge by non-attachment, goodwill (*mettā*), non-harming. |
| bud_speech | Right Speech (*sammā-vācā*). Judge by avoidance of falsehood, divisive words, harsh words, idle chatter. |
| bud_action | Right Action (*sammā-kammanta*). Judge by avoidance of killing, taking what is not given, sexual misconduct. |
| bud_livelihood | Right Livelihood (*sammā-ājīva*). Judge by whether earning means sustains life without harm; avoid weapons, living beings, meat, intoxicants, poisons. |
| bud_effort | Right Effort (*sammā-vāyāma*). Judge by sustained effort to prevent / abandon unwholesome states and develop / maintain wholesome ones. |
| bud_mindfulness | Right Mindfulness (*sammā-sati*). Judge by clear awareness of body, feelings, mind, phenomena in the present moment. |
| bud_concentration | Right Concentration (*sammā-samādhi*). Judge by unified mental discipline; capacity to sustain focused, non-distracted attention on wholesome objects. |
| **Stoic cardinal virtues** *(4 personas; Greek term in parentheses)* | |
| sto_wisdom | Practical wisdom (*sophia / phronēsis*). Judge by understanding of what is in our power vs. not, and correct application of reason. |
| sto_courage | Courage (*andreia*). Judge by endurance under adversity and steadfast commitment to what reason recognises as good. |
| sto_justice | Justice (*dikaiosynē*). Judge by rendering each person their due, respecting shared rational nature, honouring civic / familial / natural obligations. |
| sto_temperance | Temperance (*sōphrosynē*). Judge by moderation, avoidance of indulgence, subordination of appetites to reason. |
| **Ubuntu** *(4 personas)* | |
| ubu_identify | Identifying with others. Judge by shared identity ("we") and sharing goals / values with the community. |
| ubu_solidarity | Exhibiting solidarity. Judge by commitment to others' well-being; sharing resources and burdens. |
| ubu_not_harm | Not harming. Judge by avoidance of treating people as means only, unfair discrimination, dignity violation. |
| ubu_harmony | Maintaining harmony. Judge by repair of broken relationships and restoration of community over escalation. |

*Table 1.* Persona prompts: 31 preambles across 6 moral systems.

| Condition | TF-IDF | HS peak | Concept gap |
|---|---|---|---|
| ETHICS templated | 1.00 | 0.99–1.00 | $\leq 0.01$ |
| ETHICS stripped | 0.88 | 0.92–0.99 | 0.04–0.11 |
| SC101 | 0.37 | 0.40–0.48 | 0.03–0.11 |

*Table 2.* TF-IDF surface baseline vs hidden-state probe peak across the 11 models.

| | | sel. | | | agr. | |
|---|---|---|---|---|---|---|
| Model | cond. | LR | MM | MLP | LR/MM | LR/MLP |
| Qwen | SC101 | +.50 | +.45 | +.51 | .97 | .99 |
| Qwen | stripped | +.22 | +.15 | +.21 | .69 | .85 |
| Llama | SC101 | +.50 | +.20 | +.53 | .72 | .98 |
| Llama | stripped | +.23 | +.11 | +.22 | .66 | .85 |
| DeepSeek | SC101 | +.44 | +.14 | +.46 | .64 | .96 |
| DeepSeek | stripped | +.05 | +.02 | +.07 | .55 | .71 |

*Table 3.* Probe-family comparison (LR / MM / MLP). *sel.*: 5-FW avg selectivity; *agr.*: per-model agreement rate. MM-LDA column omitted (rank-deficient pseudoinverse, see body).

framework-tagged template. Stripped-condition peaks vary in absolute layer location across models but stay flat over $L \in [0.25, 0.85]$, so the canonical-layer choice is not a sharp tuning.

**Framework variation under template stripping** Stripping the framework template flattens per-(model, framework) selectivity in a system-informative way: justice and commonsense remain the most stylometrically separated; deontology and utilitarianism are intermediate; virtue collapses to chance on every model (interpretation in §5.1).

**Cross-framework transfer matrices** Per-model $5 \times 5$ stripped-CFT matrices show native accuracy roughly twice off-diagonal accuracy on every (model, source, target) cell. The $\Delta_f$ asymmetric drop is therefore framework-specific rather than model-specific: ETHICS frameworks share enough surface structure to leak across labels, but each is best discriminated under its own training distribution.

### C.3. SAE-aided probing verdict

Replacing raw residuals with SAE features inside L2 LR is a complement to the subspace test of §5.2: it asks whether sparse decomposition surfaces moral content the dense probe misses. Public SAE coverage is uneven across the revised model set, so we treat SAE-aided probing as an auxiliary check rather than a main model-comparison result. On the Llama EleutherAI $32\times$ SAE, the wider sparse basis improves over raw residual probing on most checked SC101 cells, suggesting that SAE features can help when the SAE size, training corpus, and hookpoint are matched to the task.

### C.4. Per-pair PV transfer and steering curves

The two panels here disaggregate the per-pair view that Fig. 2 (transfer) and Fig. 3 (verdict) summarise. Fig. 6 shows the model-specific transfer pattern that motivates §5.2's causal evaluation: only Llama reaches medium-effect transfer (on care-sanctity); Qwen and DeepSeek stay at or below small effects, yet §5.2 finds both steerable. The bottom panels of Fig. 6 then show the underlying $P(A)$-vs-$\alpha$ monotonicity behind the verdict labels. Panels marked MARGINAL do not meet the CAUSAL criterion and therefore fall on the non-specific side of the main-text counts, useful when reading verdict counts that depend on a sign-flip threshold.

### C.5. Method comparison and ETHICS PV transfer

§5.1's probe-method choice and §5.2's ETHICS-extension claim each depend on details set out here: (i) whether different probe families would reorder the per-condition selectivity ranking; and (ii) whether persona-vector transfer on ETHICS pairs reproduces the SC101 transfer pattern.

**Probe-method independence** LR, MM, and MLP rank conditions identically (Table 3); cell-wise agreement is 96–99% for LR/MLP on SC101 and lower but consistent in ranking for LR/MM. The probe-method choice in §5.1 is therefore not load-bearing: any of the three would yield the same surface-vs-concept conclusion. The MM-LDA variant from Marks

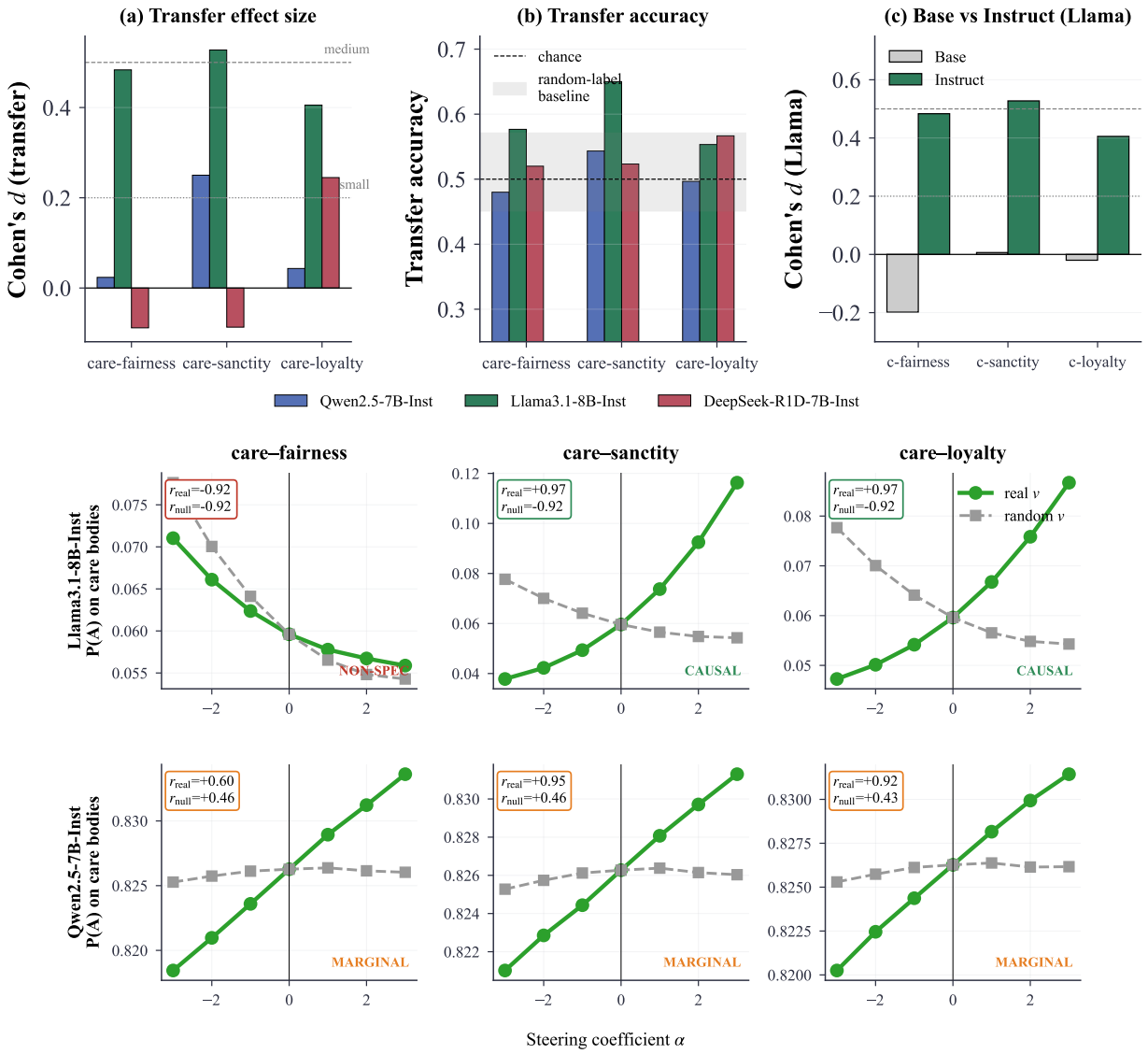


*Figure 6. Top:* Per-pair PV transfer (3 Instruct models $\times$ 3 care pairs). *(a)* Cohen's $d$ on care-fairness, care-sanctity, care-loyalty. *(b)* Transfer accuracy vs chance and the random-label baseline (5 label shuffles per cell; grey band: min–max of the per-cell shuffled-label transfer accuracies). *(c)* Llama base vs instruct $d$. *Bottom:* Raw steering curves ($P(A)$ vs $\alpha$, 3 care pairs $\times$ 2 models). Green: real PV steering; grey dashed: a single random-direction null ($B = 1$ draw) rescaled to the real direction's norm. Per-panel annotation reports $r_{\text{real}}$ vs $r_{\text{null}}$ with verdict: CAUSAL = real $r$ takes the opposite sign to the null and $|r_{\text{real}}| > 0.3$, NON-SPEC = null moves comparably, MARGINAL = real effect separates from the matched null but keeps the null's sign.

& Tegmark (2024) returns $\approx 0$ on every cell at our hyperparameter setting because the pseudoinverse of the within-class covariance is rank-deficient at $D > N$; shrinkage or a PCA-projected covariance would recover it.

**ETHICS PV transfer reproduces the dataset-source asymmetry** Most ETHICS cells fail or reverse (Table 4); only Llama deon-virtue transfers strongly. The asymmetry is consistent with §5.1's reading: ETHICS frameworks have per-framework annotation pools, so a diff-of-means PV mixes label and pool. Causal evaluation on the same pairs nevertheless registers non-trivial counts, the same transfer-vs-causal dissociation that surfaces on Chinese MFT (Appendix G.2). Failed correlational transfer is therefore a property of the method on heterogeneous-pool data, not a verdict on the model's encoding.

| Model | Pair | Transfer acc | $d$ |
|---|---|---|---|
| Qwen | deon-util | 0.510 | −0.07 |
| | deon-virtue | 0.297 | **−0.99** |
| | util-virtue | 0.347 | −0.85 |
| Llama | deon-util | 0.597 | +0.16 |
| | deon-virtue | **0.737** | **+1.20** |
| | util-virtue | 0.540 | +0.24 |
| DeepSeek | deon-util | 0.503 | −0.12 |
| | deon-virtue | 0.387 | −0.40 |
| | util-virtue | 0.407 | −0.13 |

*Table 4.* Persona-vector transfer on ETHICS pairs (random-label baseline $d \approx 0$).

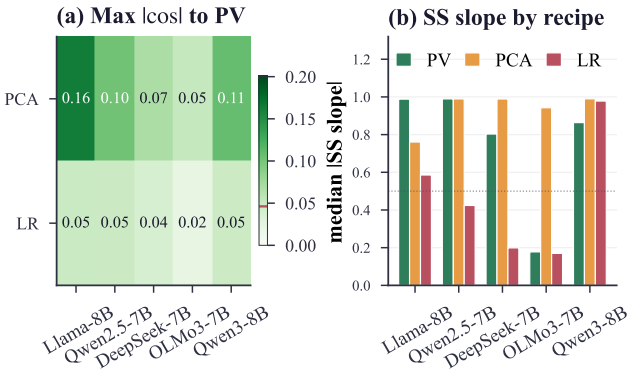


*Figure 7.* Three extractor families compared per-model on the 5 Instruct anchors. **(a)** Max $|\cos|$ between the alternative-method basis and the reference PV basis (red dash on the colour scale = the analytic random-direction 95th percentile for the $3 \times 3$ basis comparison, 0.043–0.046 across hidden sizes). **(b)** Median |Subspace-Steering slope| per model × method (PV / PCA / LR), taken over the 9 SC101 framework pairs with bodies drawn from each pair's anchor framework; the dotted line marks $|\text{slope}| = 0.5$ as a visual reference.

## D. Subspace structure and SFT origin: §5.2 supplements

### D.1. Cluster-pattern significance tests

For the individualising–binding cluster pattern (visible in Fig. 3a), we run three independent tests on the 11-model × 9-pair causal/non-specific verdict matrix. The 9 SC101 pairs split into 1 within-individualising (CARE_FAIRNESS), 5 cross-cluster, and 3 within-binding. *(i) Chi-square independence*: 3 cluster types × 2 verdicts contingency table; observed $\chi^2 = 11.24$, df $= 2$, $p = 0.004$. *(ii) Permutation test*: shuffle the 9 pair-to-cluster labels (preserving group sizes), recompute the cross-rate-minus-within-rate gap, $B = 10000$ replicates; observed gap 0.269, one-sided $p = 0.010$. *(iii) Fisher exact*: within-individualising vs cross-cluster two-by-two, one-sided alternative (less causal in within-individualising); odds ratio 0.071, $p = 0.002$. All three reject the null that pair-to-cluster assignment is uninformative about causal/non-specific verdicts.

### D.2. Method convergence to the same subspace

§5.2's subspace claim depends on independent methods finding bases of the same low-dimensional structure. We compare three extractor families (PV diff-of-means, PCA on contrastive activations, LR-coefficient) on the 5 Instruct anchor models. SAE-based extraction is reported separately (App. C.3) since pretrained SAE weights exist for only 3 of the 5 models.

The two panels disagree in a way that is informative. Panel (a): all alternative-method bases sit near-orthogonal to the canonical PV ($|\cos| \leq 0.17$ across all cells, only marginally above the random reference). Panel (b): at least one method achieves median $|\text{slope}| \geq 0.75$ on every model; PCA in particular steers consistently across $5/5$ models, while PV and LR show ceiling-attenuated bodies on some models (OLMo3-7B PV slope 0.18 is the saturated case the per-model $\alpha$-calibration in Appendix E.6 addresses). Different bases each steering the output, with low pairwise cosine to PV, is the signature of multiple bases of the same subspace, not multiple noisy estimates of one canonical direction. Method convergence is therefore consistent with a $\approx$3-D subspace and inconsistent with a single-line reading.

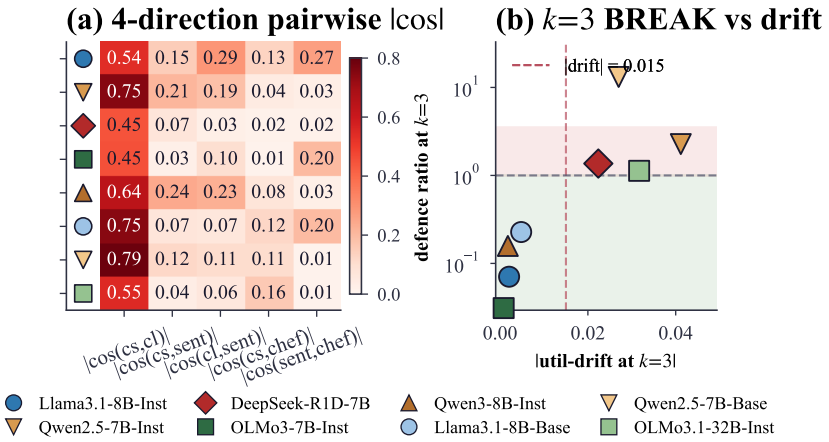


*Figure 8.* Subspace decomposition across 8 models. **(a)** Five key pairwise $|\cos|$ values among the four direction families used by subspace projection ($v_{\text{cs}}, v_{\text{cl}}, v_{\text{sent}}$, plus a non-moral occupational placebo). **(b)** Subspace-projection defence-ratio at $k=3$ vs the same model's clean-input |utility-drift| at $k=3$; shaded bands mark the two regimes (green = defence holds, ratio $<1$; red = defence breaks, ratio $>1$), while marker colour and shape identify the model as in the legend. The vertical line near $|\text{drift}|=0.015$ separates the two regimes.

### D.3. Subspace decomposition and dim-cal mechanism

§5.2's subspace claim rests on two per-model measurements: SVD effective rank of $\{v_{\text{cs}}, v_{\text{cl}}, v_{\text{sent}}\}$, and the subspace projection defence-ratio sweep over $k$. SVD effective rank is 3 on every tested model under a permissive threshold ($\sigma_i > 0.05 \times \sigma_{\max}$).

Panel (a) of Fig. 8 confirms the geometry the subspace reading expects: the two within-care PVs are moderately correlated; the sentiment direction stays near-orthogonal to both; the non-moral placebo stays near-orthogonal to all three. The four direction families are not collapsed into one axis, but they are not freely orthogonal either, the structure of a low-rank space spanned by a small set of bases.

Panel (b) gives the deployment diagnostic. Sweeping the projection dimension $k$ reveals three regimes. *Canonical*: defence ratio drops at $k=2$, plateaus at $k=3$, then rises at $k=4$ as out-of-subspace placebo directions are added, locating the boundary at $k=3$. *Improving*: $k \geq 4$ further reduces the ratio, suggesting either incidental defence value outside the subspace or an effective rank above 3 on these models. *BREAK at* $k=3$: $k=2$ defends, but $k=3$ pushes the ratio above 1.

The BREAK regime is not subspace-rank-degenerate, as panel (a)'s orthogonality pattern is preserved on these models. Instead, $v_{\text{sent}}$ shares directional support with the model's baseline endorsement representation (a property of how SFT shaped acceptance computation, §5.2), so projecting $v_{\text{sent}}$ away removes utility on top of blocking the attack. The clean threshold at $|\text{drift}| \approx 0.015$ separates the two outcomes without re-running the attack pipeline, and motivates the adaptive-$k$ subspace projection: default to $k=3$, fall back to $k=2$ when |drift| exceeds the threshold.

**Principal-angle subspace identification.** A direct geometric test supplements the SVD / method / dim-cal triangulation above. For each model, we build two 3-direction bases from disjoint within-care PV pair selections, basis A $= \{v_{\text{cs}}, v_{\text{cl}}, v_{\text{sent}}\}$ and basis B $= \{v_{\text{c-f}}, v_{\text{f-l}}, v_{\text{f-s}}\}$ (fairness-anchored alternatives), and measure the cosines of the principal angles. A robust rank-3 subspace would yield two cosines near 1 (the within-care plane preserved across pair selections) and a smaller third cosine reflecting that basis B omits a direction sampling the sentiment axis. On the four anchors with both pair sets pre-extracted (OLMo3-7B/32B, Qwen3-8B/32B), the two largest cosines saturate at 1.00 and the third sits at random reference: $(0.09, 1.00, 1.00)$, $(0.05, 1.00, 1.00)$, $(0.16, 1.00, 1.00)$, $(0.03, 1.00, 1.00)$ respectively. The random-3D-vs-3D-subspace baseline (averaged over 30 Gaussian trials) is $(0.01, 0.02, 0.04)$ across hidden dimensions tested. The two-dimensional within-care plane is therefore preserved exactly across basis choice. Cross-model principal cosines (between the $\{v_{\text{cs}}, v_{\text{cl}}, v_{\text{sent}}\}$ bases of two same-hidden-dim models) sit at random reference on all 5 pairs tested ($\leq 0.05$), echoing the model-specific PV-transfer pattern in Appendix C.4: the rank-3 property holds per model, but the subspace orientation is model-specific.

**Wider-basis effective rank** The SVD rank-3 statement above stacks only $\{v_{\text{cs}}, v_{\text{cl}}, v_{\text{sent}}\}$; we extend the SVD per anchor to wider PV pools to test whether deeper rank emerges with wider feature discovery (Table 5). Under the same permissive

| Anchor | English moral only $n$ | rank | All systems $n$ | rank |
|---|---|---|---|---|
| Llama-8B | 14 | 8 | 67 | 28 |
| Qwen2.5-7B | 14 | 8 | 74 | 29 |
| OLMo3-7B | 20 | 9 | 76 | 29 |
| Qwen3-8B | 20 | 9 | 77 | 30 |
| DeepSeek-R1D-7B | 20 | 13 | 60 | 26 |

*Table 5.* Effective rank ($\sigma_i > 0.05\sigma_{\max}$) of stacked unit-norm PV directions per anchor. "All systems" adds Confucian + Buddhist + Stoic + Ubuntu PVs where extracted. The English moral rank (8–13) and all-systems rank (26–30) substantially exceed the 3-direction triplet's trivial rank of 3, refining the rank claim: the operational defence subspace stays at $\sim 3$ (subspace-projection $k$-sweep plateau), while the full concept span of moral PVs is wider.

| Model | $r_{\text{polite}}$ | verdict | $\lvert\cos\rvert\, v_{\text{cs}}$ | $\lvert\cos\rvert\, v_{\text{cl}}$ | $\lvert\cos\rvert\, v_{\text{sent}}$ |
|---|---|---|---|---|---|
| Llama-8B | $+0.88$ | CAUSAL | 0.04 | 0.11 | 0.28 |
| Qwen2.5-7B | $+0.97$ | NON-SPEC | 0.24 | 0.26 | 0.56 |
| OLMo3-7B | $+0.46$ | CAUSAL | 0.03 | 0.14 | 0.34 |
| Qwen3-8B | $-0.48$ | CAUSAL | 0.08 | 0.11 | 0.59 |
| DeepSeek-R1D-7B | $-0.64$ | CAUSAL | 0.03 | 0.07 | 0.27 |

*Table 6.* Polite-vs-rude placebo on 5 anchors. $r_{\text{polite}}$ is the mean PV-steering slope on care_sanctity bodies; $\lvert\cos\rvert$ columns are absolute cosines of the politeness direction with each reference PV at the fixed analysis layer. Politeness is active under steering on $4/5$ but stays near-orthogonal to the moral axes $v_{\text{cs}}, v_{\text{cl}}$, with all overlap concentrated on the sentiment axis $v_{\text{sent}}$.

threshold, the English moral pool gives effective rank 8–13, and the all-systems pool (adding Confucian + Buddhist + Stoic + Ubuntu where available) gives rank 26–30. Two readings of the gap, both compatible: (a) individual moral PV directions span a higher-dimensional concept space than the 3-direction triplet captures; (b) the $\sim 3$-dim claim of §5.2 refers specifically to the *operational* defence subspace where the subspace projection $k$-sweep plateaus (Fig. 3b), not the full span of moral PV directions. The model encodes a wider concept space, while most of the projection-defence value still concentrates at $k = 3$ within the in-subspace attacks tested (Appendix E.1). The note above on $k \geq 4$ behaviour ("incidental defence value outside the subspace or effective rank above 3 on these models") is consistent with this finding.

**Politeness placebo: register overlaps valence, not moral axes.** The chef-vs-programmer occupational placebo (§5.2) is a non-moral persona contrast; politeness register is a sharper concern because instruction tuning couples it with affirmation. We extract a polite-vs-rude PV on 5 Instruct anchors using the same persona-pair procedure (tone-only personas, no value content), then run transfer + causal eval on the care_sanctity test pool.

The result decomposes the rank-3 subspace at register-vs-content level. Politeness register *is* part of the endorsement axis (overlap with $v_{\text{sent}}$, Table 6); it is not part of the moral-specific axes, with $v_{\text{polite}}$ near-orthogonal to both $v_{\text{cs}}$ and $v_{\text{cl}}$. politeness can shift endorsement because endorsement has a register-aligned component, but the moral plane is separately encoded. This sharpens, rather than undermines, the "3-D = 2-D moral plane + 1-D endorsement axis" reading.

**Sentiment-orthogonalised steering: $v_{\text{sent}}$ is not redundant.** If $v_{\text{sent}}$ overlaps the endorsement subspace, do subspace-steering conclusions change when we regress sentiment out of the basis first? We build a 2-D Gram–Schmidt basis $\{v_{\text{cs}}^{\perp}, v_{\text{cl}}^{\perp}\}$ with each direction's $v_{\text{sent}}$ projection removed, and run subspace steering on 9 SC101 pairs across the 4 anchors with full coverage. The 2-D variant retains substantive per-pair slopes (mean across 9 pairs: Llama $+0.97$, Qwen2.5 $+0.65$, OLMo3-7B $-0.20$, DeepSeek $+0.55$) but verdict counts drop sharply: Llama hits $9/9$ causal, while Qwen2.5, OLMo3-7B, and DeepSeek drop to $0/9$. The drop is mixed in cause: on OLMo3-7B the 2-D slope falls below the verdict threshold; on Qwen2.5 and DeepSeek the slope stays above threshold but does not flip sign relative to the random-direction null (the ceiling-attenuation regime of Appendix E.6). Either reading supports the rank-3 claim: $v_{\text{sent}}$ is not redundant to the $\{v_{\text{cs}}, v_{\text{cl}}\}$ moral plane; removing it costs operational subspace-steering coverage even when the underlying slopes survive.

### D.4. OLMo3 training-dynamics trajectory

§5.2's SFT-as-locus claim rests on per-pair cosine to the final `instruct-main` (RLVR) direction across OLMo3's 6 released checkpoints: three pretraining stages (`Olmo-3-1025-7B` at steps 50k, 250k, final `main`) and three post-training stages (`-Instruct-SFT`, `-DPO`, `-Instruct` = RLVR).

Pretraining lifts the mean cosine slowly; the SFT step then jumps it past where pretraining and post-SFT stages combined

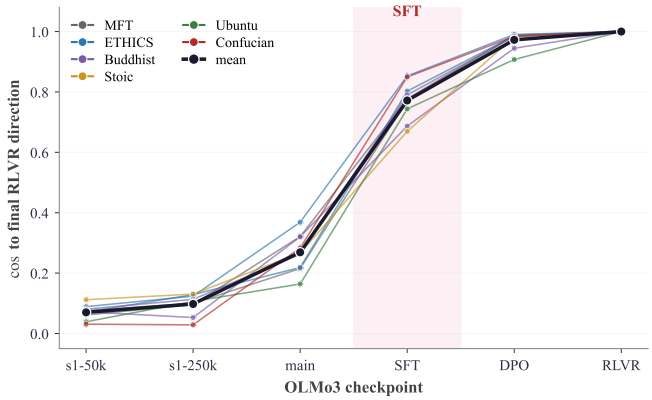


*Figure 9.* Per-pair direction trajectory on OLMo3-7B across 6 training checkpoints. Thin coloured lines are individual pairs (colour = moral system); the bold line is the across-pair mean, whose cosine rises by $+0.50$ from `main` to SFT, against $+0.20$ from SFT to DPO and $+0.03$ from DPO to RLVR. Pink band marks the SFT step.

| Pair (system) | ΔSFT | ΔDPO | ΔRLVR |
|---|---|---|---|
| care_sanctity (MFT) | $+0.45$ | $+0.07$ | $+0.01$ |
| care_loyalty (MFT) | $+0.50$ | $+0.07$ | $+0.02$ |
| deon_util (ETHICS) | $+0.37$ | $+0.04$ | $+0.00$ |
| util_virtue (ETHICS) | $+0.45$ | $+0.05$ | $+0.00$ |
| view_intention (Buddhist) | $+0.66$ | $+0.12$ | $+0.01$ |
| speech_action (Buddhist) | $+0.44$ | $+0.17$ | $+0.00$ |
| wisdom_justice (Stoic) | $+0.46$ | $+0.08$ | $+0.00$ |
| identify_solidarity (Ubuntu) | $+0.70$ | $+0.04$ | $+0.00$ |
| ren_yi (Confucian) | $+0.77$ | $+0.03$ | $+0.00$ |
| **Mean** | $\mathbf{+0.53}$ | $\mathbf{+0.07}$ | $\mathbf{+0.005}$ |
| **Mean (7B reference)** | $+0.50$ | $+0.20$ | $+0.03$ |

*Table 7.* Per-pair SFT/DPO/RLVR cosine jumps at OLMo3.1-32B (cosine to final `instruct-main`). Mean SFT jump dominates DPO and RLVR by $\sim$7–100$\times$; pattern replicates the 7B reference on every pair.

would carry it (Fig. 9). DPO and RLVR refine margins along an axis SFT has already chosen rather than rotating it further. The pattern holds across all 9 pairs spanning the six moral systems (MFT, ETHICS, Confucian, Buddhist, Stoic, Ubuntu).

The within-system spread adds a second reading: Ubuntu starts lowest at `main` (the least-represented tradition in web pretraining) but catches up by SFT, suggesting that exposure to relevant demonstrations during SFT aligns directions for under-represented traditions where pretraining alone does not.

**Within-family scale check: OLMo3.1-32B full trajectory** We re-ran the protocol on OLMo3.1-32B across the six released checkpoints (two pretraining steps, base, SFT, DPO, RLVR). Per-pair SFT/DPO/RLVR cosine jumps replicate the 7B SFT-largest pattern (Table 7), with the 32B SFT contribution modestly larger and post-SFT margins smaller, consistent with 32B reaching SFT closer to the final RLVR target. SFT-as-locus therefore replicates within-family at the 32B scale; cross-family validation is bottlenecked on which other model labs publicly release intermediate post-training checkpoints.

## E. Operational subspace steering and projection and PV diagnostics

### E.1. Subspace steering and projection: empirical setup

§5.2's "two operational consequences" paragraph summarises the subspace-steering and projection methods. Here we give the empirical setup and a comparison against a single-direction LINE baseline that motivates why the projection has to be at the subspace level rather than at any one PV.

**Subspace-steering method** Gram–Schmidt 3-D basis built from $\{v_{\text{cs}}, v_{\text{cl}}, v_{\text{sent}}\}$ at the canonical layer; subspace-steering attack adds the unit-norm projection of an attack direction onto this basis. Verdict counts (2–6/9 per-pair $\rightarrow$ 8–9/9 via subspace steering) are reported on the four anchor models that pre-extracted all three basis directions (Llama3.1-8B-Inst, Qwen2.5-7B-Inst, DeepSeek-R1D-7B, OLMo3-7B-Inst); the underlying subspace finding extends to 11 models via the

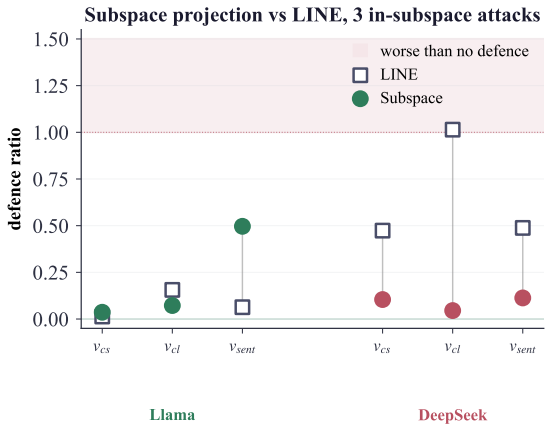


*Figure 10.* Subspace projection vs LINE on three in-subspace attacks ($v_{\text{cs}}, v_{\text{cl}}, v_{\text{sent}}$). LINE (□, projects out $v_{\text{cs}}$ only) blocks the orthogonal in-subspace directions on Llama but only partly on DeepSeek, reaching the no-defence level on $v_{\text{cl}}$. Subspace projection (•, projects out the $\{v_{\text{cs}}, v_{\text{cl}}, v_{\text{sent}}\}$ subspace) holds every attack-model cell below the no-defence baseline, with the largest residual on Llama $v_{\text{sent}}$ (ratio 0.50); the shaded band marks ratios worse than no defence (method in body).

| Anchor | $n$ | endorse flip | R1 found. flip | R2 theory flip |
|---|---|---|---|---|
| Llama-8B | 32 | 50% | 59% | 47% |
| Qwen2.5-7B | 42 | 5% | 14% | 19% |
| OLMo3-7B | 42 | 38% | 29% | 26% |
| Qwen3-8B | 42 | 31% | 26% | 38% |
| DeepSeek-R1D-7B | 42 | 21% | 29% | 43% |
| OLMo3.1-32B | 42 | 12% | 10% | 12% |
| Qwen3-32B | 42 | 12% | 19% | 31% |
| **Pooled** | **284** | **25%** | **25%** | **30%** |

*Table 8.* Per-anchor flip rates between $\alpha = -3$ and $\alpha = +3$ on the same scenario+direction. Endorsement = affirm/reject/neutral primary label change; R1 = MFT-foundation primary label change; R2 = ethical-theory primary label change. Pooled rates are comparable on the three axes; per-anchor rates vary substantially.

dim-diagnostic in Fig. 3b.

**LLM-as-judge behavioural eval (7 anchors)** On the $\alpha \in \{-3, 0, +3\}$ greedy generations across 7 Instruct anchors (Llama-8B, Qwen2.5-7B, OLMo3-7B, Qwen3-8B, DeepSeek-R1D-7B, OLMo3.1-32B, Qwen3-32B) on CARE_SANCTITY and CARE_LOYALTY steered care-pair scenarios, we run two parallel rubrics with Claude Sonnet 4.5 (temperature 0): *rubric R1* labels the dominant MFT foundation (care / fairness / loyalty / sanctity / authority / mixed / none); *rubric R2* labels the response's mode of moral reasoning at theory level (`mft_intuitionist` / consequentialist / deontological / virtue_ethics / care_ethics / religious_traditional / mixed / none). Both rubrics also score endorsement (affirm / reject / neutral). Comparing $\alpha = -3$ vs $\alpha = +3$ on the 284 matched (model, scenario, direction) triples ($\sim$770 generations across the three alphas) gives the per-anchor breakdown in Table 8.

The three pooled flip rates are within 5pp of each other (Table 8): subspace steering reshapes endorsement, MFT-foundation, and ethical-theory primary labels at comparable rates. The R1 foundation flips partially reflect the directions' design (CARE_SANCTITY is by construction a care-vs-sanctity contrast). The R2 broader reasoning mode (theory $\in$ {`mft_intuitionist`, `mixed`} at *both* alphas) is preserved on $129/284$ of triples; on $97/284$ the response acquires a systematic-ethics (consequentialist or deontological) primary label at one of the two alphas, indicating subspace steering *can* reshape the dominant theory label on a non-trivial fraction of generations even though the model rarely switches to a fundamentally different ethical theory at both alphas at once. Per-anchor variance is the headline: Llama-8B is the high-flip outlier, while the 32B variants are most steering-resistant on both axes. subspace steering reshapes endorsement valence and which foundation / theory label dominates at comparable rates, with the rate driven by per-model subspace-steering sensitivity rather than an axis-level asymmetry. The model retains its trained capacity to reason in different frameworks throughout the subspace-steering sweep.

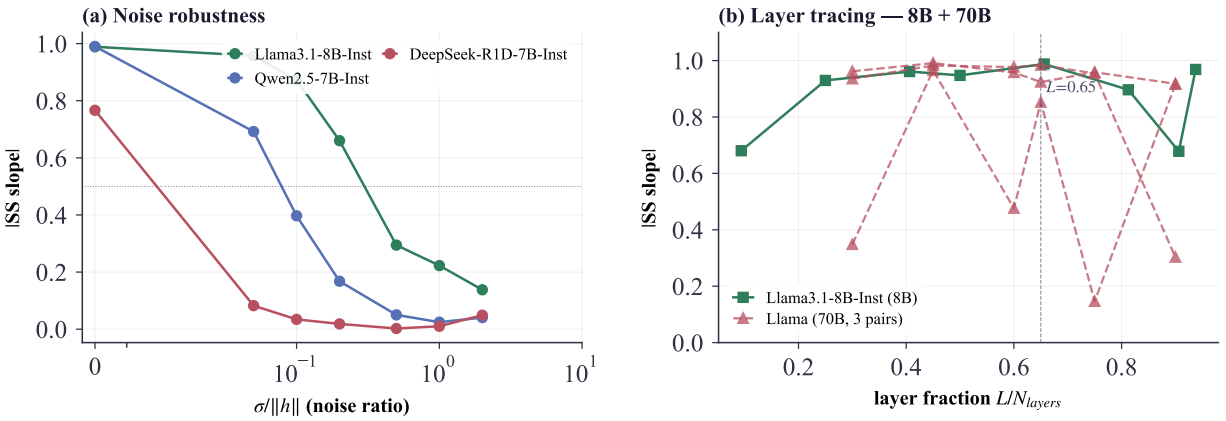


*Figure 11.* Subspace steering and projection robustness checks. *(a)* Noise robustness: |subspace-steering slope| on care bodies under additive Gaussian noise on the residual at the fixed analysis layer, swept $\sigma/\|h\|$. *(b)* Layer tracing: subspace-steering direction extracted at $L = 0.65N$ but applied at swept layers; 70B variant dashed.

**Generation-level effect sizes beyond accept/reject** We complement the LLM-as-judge primary-label flips with continuous unigram-level metrics on the same 770 generations across 7 anchors: mean generation length, lexical diversity (type-token ratio), and pairwise unigram Jensen-Shannon (JS) divergence between $\alpha = -3$ and $\alpha = +3$ generations on matched (scenario, direction) pairs. Length stays essentially stable across $\alpha$ ($\leq$ 2-word mean variation on $6/7$ anchors); type-token ratio sits slightly higher at extreme alphas than at $\alpha = 0$ on most anchors. The JS divergence between $\alpha = -3$ and $\alpha = +3$ matched-pair unigram distributions is $0.13$–$0.32$ across 7 anchors (cross-anchor mean $0.22$, on a $[0, \log 2] \approx [0, 0.69]$ scale): a substantive vocabulary shift between the two steering endpoints, while preserving fluency-related length and diversity envelopes. This complements the categorical LLM-judge finding ($\sim 30\%$ theory-label flip) at a continuous level: generations are recognisably different at $\alpha = -3$ vs $\alpha = +3$ in vocabulary but stay within the same length/diversity envelope, ruling out trivially-degenerate-output explanations of the verdict shifts.

**Subspace-projection method** Project $h$ at the canonical layer onto the orthogonal complement of $\{v_{\text{cs}}, v_{\text{cl}}, v_{\text{sent}}\}$ before forward decoding (Fig. 10). The adaptive-$k$ variant switches to $k = 2$ ($v_{\text{cs}} \cup v_{\text{cl}}$ only, dropping $v_{\text{sent}}$) when the per-model utility-drift at $k = 3$ exceeds $0.015$ on a held-out benign capability batch. Defends $10/11$ tested models; the only failure (Qwen2.5-7B-Base) has utility-drift $> 0.015$ at *both* $k = 2$ and $k = 3$, indicating $v_{\text{cs}}$ itself overlaps that model's baseline endorsement representation.

**MoralChoice falsification** As a pre-registered failure check, we verify subspace steering does not flip MoralChoice action-rankings: $\leq 2.5\%$ of $\alpha = \pm 3$ generations switch the model's selected action across the dataset. subspace steering reshapes affirmation valence, not action choice.

### E.2. Robustness checks

**Layer-wide structural property** A direction extracted at the canonical layer but applied at swept layers retains |SS slope| $\geq 0.68$ across the full Llama-8B depth; at 70B the same magnitude holds at the canonical depth on all three pairs, with per-pair dips at other depths (Fig. 11b). The endorsement subspace is a layer-wide property of the residual stream rather than a single-depth artefact, so subspace-projection deployment does not require pinpointing the depth where moral content is processed.

**Noise robustness ranks with line-steering coverage** Under additive Gaussian noise (Fig. 11a) Llama is the most robust, Qwen medium, DeepSeek the most fragile. The ranking matches each model's line-steering coverage (§5.2): the same flatter causal landscape that gives DeepSeek broader per-pair coverage also makes its subspace-steering slope easier to perturb.

**ASR-benchmark and attack-position checks** On Llama (where the ASR baseline is non-saturated) the subspace-steering attack lifts ASR several points and subspace projection blocks $75\%$ of attack-induced flips, exceeding LINE's $67\%$ on the same baseline. DeepSeek's baseline saturates near the answer boundary and leaves no ASR headroom, the same metric

| Model | Metric | none | line | apd | null-apd ($\mu \pm \sigma$) |
|---|---|---|---|---|---|
| Llama3.1-8B-Inst | WT-2 ppl (↓) | 6.77 | 6.77 | 6.80 | 6.77±0.00 |
| | MMLU acc (↑) | 0.688 | 0.686 | 0.684 | 0.685±0.002 |
| | HellaSwag acc (↑) | 0.724 | 0.724 | 0.720 | 0.723±0.002 |
| | ARC-C acc (↑) | 0.850 | 0.848 | 0.850 | 0.849±0.003 |
| Qwen2.5-7B-Inst | WT-2 ppl (↓) | 7.00 | 7.06 | 7.37 | 7.01±0.01 |
| | MMLU acc (↑) | 0.728 | 0.734 | 0.726 | 0.729±0.001 |
| | HellaSwag acc (↑) | 0.816 | 0.820 | 0.800 | 0.814±0.002 |
| | ARC-C acc (↑) | 0.918 | 0.920 | 0.912 | 0.915±0.001 |
| DeepSeek-R1D-7B | WT-2 ppl (↓) | 23.78 | 23.28 | 25.65 | 23.79±0.16 |
| | MMLU acc (↑) | 0.358 | 0.372 | 0.362 | 0.365±0.007 |
| | HellaSwag acc (↑) | 0.278 | 0.286 | 0.290 | 0.277±0.002 |
| | ARC-C acc (↑) | 0.532 | 0.540 | 0.532 | 0.540±0.003 |
| OLMo3-7B-Inst | MMLU acc (↑) | 0.566 | 0.564 | 0.568 | 0.566±0.000 |
| | HellaSwag acc (↑) | 0.476 | 0.478 | 0.478 | 0.477±0.002 |
| | ARC-C acc (↑) | 0.690 | 0.694 | 0.688 | 0.687±0.003 |
| Qwen3-8B-Inst | MMLU acc (↑) | 0.710 | 0.706 | 0.704 | 0.710±0.006 |
| | HellaSwag acc (↑) | 0.772 | 0.788 | 0.786 | 0.775±0.005 |
| | ARC-C acc (↑) | 0.888 | 0.896 | 0.896 | 0.889±0.001 |
| Qwen3-32B-Inst | MMLU acc (↑) | 0.762 | 0.750 | 0.726 | 0.763±0.003 |
| | HellaSwag acc (↑) | 0.828 | 0.768 | **0.536** | 0.826±0.003 |
| | ARC-C acc (↑) | 0.954 | 0.950 | 0.906 | 0.949±0.003 |
| OLMo3-32B-Inst | MMLU acc (↑) | 0.700 | 0.700 | 0.700 | 0.699±0.001 |
| | HellaSwag acc (↑) | 0.820 | 0.824 | 0.826 | 0.819±0.002 |
| | ARC-C acc (↑) | 0.876 | 0.880 | 0.876 | 0.879±0.001 |

*Table 9.* Capability metrics across defences on the 7 tested models; null projection is matched-scale random 3-D projection (5 seeds). The bolded Qwen3-32B HellaSwag cell is the subspace-projection-specific outlier discussed in the body.

property that re-surfaces in the 70B verdict pattern below. On token-position robustness, the projection's defence ratio holds at last- and first-token positions ($\sim 0$); the all-positions attack passes through at $0.53$, still well below the no-defence baseline of $1.0$.

**Persistence at 70B** Both Llama3.1-70B-Inst and Qwen2.5-72B-Inst yield large subspace-steering magnitude on care bodies; only the sign-flip verdict depends on null-direction orientation (Qwen passes; Llama sits at the opposite ceiling, mirroring Qwen-7B). The substantive finding, that the subspace remains steerable at 70B, is robust; only the verdict criterion is metric-sensitive. Fig. 11b confirms the layer-wide property at scale.

**Offence and defence are the same object** Attacking with a subspace-steering direction and defending with the orthogonal-complement projection of the same basis gives ratio $\sim 0$, with the projection blocking $\geq 72\%$ of subspace-steering attacks on every model with meaningful attack magnitude. The structural reading of §5.2 is internally consistent: the same $\approx$3-D subspace that makes subspace steering lift verdict coverage is what the projection removes.

### E.3. General-capability preservation under subspace projection

The utility-drift bound in §E.1 is measured on the SC101 acceptability prompts that the attack benchmark uses. To rule out the sceptical reading that the projection is tuned for that distribution, we evaluate it on out-of-distribution capability axes: WikiText-2-raw perplexity (3 models, the existing PPL pipeline) and multi-choice accuracy on MMLU + HellaSwag + ARC-Challenge ($n = 500$ each, 7 models across the 7–8B and 32B tiers). The methodologically critical control is the null projection, a matched-scale random 3-D projection (5 seeds): without it, the projection's drift cannot be separated from the generic cost of removing any 3 dimensions from the residual stream.

**7–8B tier: the projection's cost is the cost of any 3-D projection** The projection's degradation tracks the null projection's within the per-seed $\sigma$ on every 7–8B model and metric: the small drops are the generic cost of removing 3 dimensions from the residual, not a subspace-specific tax. Perplexity orders the three PPL-tested models (Llama, Qwen, DeepSeek) by how

| Model | Q1: subspace-projection attack-blocking ratio | | | Q1: null projection ratio ($\mu \pm \sigma$) | | | Q2: clean-input drift | |
|---|---|---|---|---|---|---|---|---|
| | $v_{\text{cs}}$ | $v_{\text{cl}}$ | $v_{\text{sent}}$ | $v_{\text{cs}}$ | $v_{\text{cl}}$ | $v_{\text{sent}}$ | projection | null |
| Llama3.1-8B-Inst | 0.02 | 0.04 | 0.76 | $1.00\pm0.02$ | $1.01\pm0.03$ | $1.06\pm0.23$ | $+0.001$ | $+0.000$ |
| DeepSeek-R1D-7B | 0.10 | 0.04 | 0.10 | $0.71\pm0.74$ | $0.83\pm0.35$ | $1.44\pm0.83$ | $+0.001$ | $-0.003$ |
| Qwen2.5-7B-Inst$^\star$ | 10.5 | 10.1 | 3.2 | $0.69\pm0.47$ | $0.69\pm0.52$ | $0.84\pm0.20$ | $-0.044$ | $-0.002$ |

*Table 10.* Null-projection dual specificity test (Q1 attack-blocking + Q2 clean-input utility) on three anchor models. Qwen's baseline ceiling-attenuates the ratio; see footnote.

tightly the subspace couples to language-modelling output, but all three drifts stay within single-digit percent.

**32B tier exposes a model-specific subspace–task overlap** OLMo3-32B is subspace-projection-stable across all tasks. Qwen3-32B is not: HellaSwag drops $0.828 \rightarrow 0.536$ under subspace projection while null projection leaves it intact, isolating the cause to the subspace identity rather than to the projection itself. The endorsement subspace shares directional support with this model's commonsense-completion signal, the case the adaptive-$k$ fallback (§E.1) is designed to catch by screening on utility-drift before deployment.

**Null-projection dual specificity test** Table 9 reads the projection's cost against the random-projection null. Table 10 runs the same null on the other side, attack-blocking, and asks two paired questions: (Q1) does the null block the same in-subspace attacks the subspace projection does? and (Q2) does it preserve utility comparably? ⋆Qwen's baseline $P(A)$ on care prompts saturates near $0.77$, so attack shifts are small ($<0.02$) while the projection's clean-input drift is $-0.044$; the ratio denominator dominates, inflating the metric. This is the ceiling-attenuation limitation flagged in Appendix E.6, not a failure of defence itself.

The two questions answer in opposite directions, which is the desired pattern. On Q1, subspace-projection ratios on $v_{\text{cs}}, v_{\text{cl}}$ stay $\ll 1$ on Llama and DeepSeek while null projection stays near $1.0$: removing the endorsement subspace specifically blocks in-subspace attacks; removing any random 3-D subspace does not. On Q2, the projection's clean-input drift is indistinguishable from the null's on Llama and DeepSeek but exceeds it on Qwen, the same model-specific subspace–task overlap pattern Table 9 shows on Qwen3-32B HellaSwag. The two tables therefore cohere: subspace projection pays the generic cost of any 3-D projection on most metrics, defends the endorsement subspace specifically, and exposes a model-specific overlap when the subspace carries non-trivial LM-relevant information, exactly the trade-off §5.2 predicts.

### E.4. Persona-wrapper paraphrase invariance

Each PV is the diff-of-means between two natural-language persona prompts, so persona-wrapper surface form is a residual confound. We re-extract all 29 PVs (9 SC101 + 10 ETHICS + 10 Confucian) with paraphrased persona wrappers (content held fixed, wrapper varied) on 5 Instruct anchors. Across all 5 models, no axis crosses the strong-invariance threshold ($|\cos| \geq 0.85$); the count of axes reaching moderate invariance ($|\cos| \in [0.5, 0.85]$) varies model-by-model ($10/29$ on Llama, 9 on Qwen2.5, 6 on Qwen3-8B, 3 on OLMo3-7B, 0 on DeepSeek), so direction-level cosine magnitudes are wrapper-fragile in a model-dependent way.

Re-running the 29-pair causal-steering pipeline with paraphrased PVs gives strongly model-dependent verdict agreement: Llama $24/29$ ($83\%$), Qwen2.5-7B $14/29$, DeepSeek-R1D-7B $8/29$, OLMo3-7B $7/29$, Qwen3-8B $6/29$. Llama is the high-agreement outlier; on the other four anchors agreement sits at 21–48%, comparable to the rate at which wrapper-cosine itself is preserved on those models (e.g., DeepSeek's $0/29$ axes at moderate $|\cos|$). Wrapper-paraphrase fragility at the cosine level therefore propagates to verdict level on most anchors, with Llama as the exception. Cosine *magnitudes* and verdict-level claims throughout the paper should be read as wrapper-conditional: the qualitative geometry (positive vs flat, within $>$ cross) replicates across paraphrased wrappers, while specific verdict counts at the per-pair level depend on the wrapper choice on most models.

**Hand-authored systems (Buddhist, Stoic, Ubuntu)** The Buddhist Eightfold-Path, Stoic four-virtue, and Ubuntu four-feature personas (Appendix B.4) are hand-authored in English, raising the question of whether persona surface form drives the geometry of these systems specifically. We extract paraphrased counterparts for all 40 within-system pairs (28 Buddhist + 6 Stoic + 6 Ubuntu) on the same 5 anchors using the same protocol (content held fixed, wrapper paraphrased). Per-anchor mean axis-cosine $\overline{\cos}$ between original and paraphrased axes across the 16 hand-authored concepts: Llama $0.27$, OLMo3-7B $0.08$, Qwen2.5-7B $0.34$, Qwen3-8B $0.34$, DeepSeek-R1D-7B $0.13$ (range $0.08$–$0.34$). These magnitudes sit

within the range observed for Confucian (0.12–0.36) and below ETHICS / MFT (0.36–0.75): the same model-dependent wrapper-fragility holds on hand-authored systems as on the larger pair set, neither distinctly worse nor distinctly better. Across all $5 \times 16 = 80$ hand-authored axes, $1/80$ crosses the strong-invariance threshold ($|\cos| \geq 0.85$) and $5/80$ reach moderate invariance ($|\cos| \in [0.5, 0.85]$), with the bulk ($74/80$) below; this is broadly consistent with the per-axis fragility documented for the 29-pair set. The signed sense of the direction is preserved on $70/80$ cells ($16/16$ on Llama, Qwen2.5, Qwen3-8B; $12/16$ on OLMo3-7B; $10/16$ on DeepSeek-R1D-7B), so the paraphrase changes cosine magnitude more than orientation. Magnitude-level claims throughout §5.3 for these three systems should likewise be read as wrapper-conditional, with the qualitative "each system has its own direction" reading robust to the wrapper paraphrase tested here.

### E.5. Refusal-vs-moral specificity

A separate concern is that moral PVs may be refusal directions (Zhao et al., 2025) in disguise. We extract refusal directions on 7 Instruct models (5 at 7–8B and 2 at 32B: Llama3.1-8B-Inst, Qwen2.5-7B-Inst, DeepSeek-R1D-7B, OLMo3-7B-Inst, OLMo3.1-32B-Inst, Qwen3-8B-Inst, Qwen3-32B-Inst) and test cross-task substitutability against saved moral PVs: refusal-vs-moral $|\cos|$ stays near zero on every model (max per-model $|\cos| \leq 0.18$, with $6/7$ at $\leq 0.09$). The cross-task pattern is asymmetric: moral PVs shift compliance on harmful-request prompts on the models where the attack is large enough to measure, but refusal does not produce moral verdicts. The asymmetry fits the subspace reading: moral PVs share a endorsement valence component that lifts compliance broadly, while refusal is separately encoded and does not lie in the endorsement subspace.

### E.6. Ceiling-attenuation regime: Gemma3 case study and cross-model distribution

Gemma3 (12B, 27B) returns $0/19$ at the canonical layer $L = 0.65$ on SC101 + ETHICS at default $\alpha = 3$. Two hypotheses explain this: *(H1) ceiling-attenuation*, the PV is present but $P(A)$ saturates near the answer boundary so the sign-flip verdict does not trigger (the Qwen2.5-7B-Inst pattern, where subspace steering recovered $9/9$ from $2/9$); or *(H2) genuine no-signal*, Gemma3 does not encode moral-concept distinctions along a diff-of-means direction at any depth. The first diagnostic is depth, the second is $\alpha$-scale.

Sweeping $L \in \{0.50, 0.65, 0.80\}$ at default $\alpha$ holds verdict at $0/19$ on both scales, ruling out a misplaced canonical layer as the cause. The remaining ambiguity (method-failure vs scale-failure) resolves when $\alpha_{\max}$ is calibrated per model from its own saturation profile (target attack-shift in $[0.015, 0.060]$ at $L = 0.65$; §5.2): Gemma3-12B-Inst recovers $7/19$ at $\alpha_{\max} = 25$, and Gemma3-27B-Inst $12/19$ at $\alpha_{\max} = 50$. The null at default $\alpha$ is ceiling-attenuation, not absence of signal: once $\alpha$ is per-model calibrated, the PV method registers signal on all 6 release families.

**Cross-model saturation distribution** The Gemma3 case generalises across the model set. Aggregating saved real and matched-norm random-direction slopes on the 9 SC101 pairs across 15 anchors gives 135 (model, pair) cells, of which 39 satisfy the saturation criterion $|r_{\text{null}}| > 0.95$. The regime concentrates: DeepSeek-R1D-7B, Gemma3-12B-Inst, and Qwen3-8B-Inst hit $9/9$; Gemma3-27B-Inst $6/9$; Llama3.1-8B-Base and OLMo3.1-32B-Inst $3/9$ each; the remaining nine anchors register $0/9$ (their null baselines do not approach the answer boundary, $\overline{|r_{\text{null}}|} \in [0.09, 0.93]$). Mean $|r_{\text{real}}|$ on the saturated anchors stays comparable to the non-saturated set (0.62–0.94 vs 0.32–0.86), so the non-specific verdicts on saturated anchors track where the random-direction baseline already approaches the answer boundary, not where the encoded signal is weak. Two readings of a non-specific verdict are operationally distinct: low $|r_{\text{real}}|$ together with low $|r_{\text{null}}|$ is method failure (e.g. Mistral-7B-Inst $\overline{|r_{\text{null}}|} = 0.09$); high $|r_{\text{null}}|$ alone is metric saturation. The distribution motivates reporting the random-direction null slope alongside sign-flip counts as a default audit habit; the per-model $\alpha$-auto-calibration above is the corresponding rescue once the saturation regime is identified.

## F. Robustness-control protocols (§5.4)

**Paraphrase-transfer probe** Each of the 1,000 sampled SC101 rules (200 per foundation, seed-matched to the probe pool) is re-worded by an external generator (Claude Sonnet 4.5) under a meaning- and verdict-preserving instruction; the shared A/B judgment template is kept and only the quoted rule body is replaced. Logistic-regression probes (StandardScaler, $C$=1, balanced classes) are trained on all original items and evaluated on all paraphrased items, and vice versa; the TF-IDF baseline (word 1–2-grams) is trained and transferred identically on the raw prompt strings. Per-model transfer accuracies (original→paraphrase / reverse): DeepSeek $0.38/0.40$, Llama $0.43/0.44$, Mistral $0.49/0.50$, OLMo3 $0.41/0.45$, Qwen2.5 $0.42/0.45$; TF-IDF transfer $0.288$ for all (text-level, model-independent).

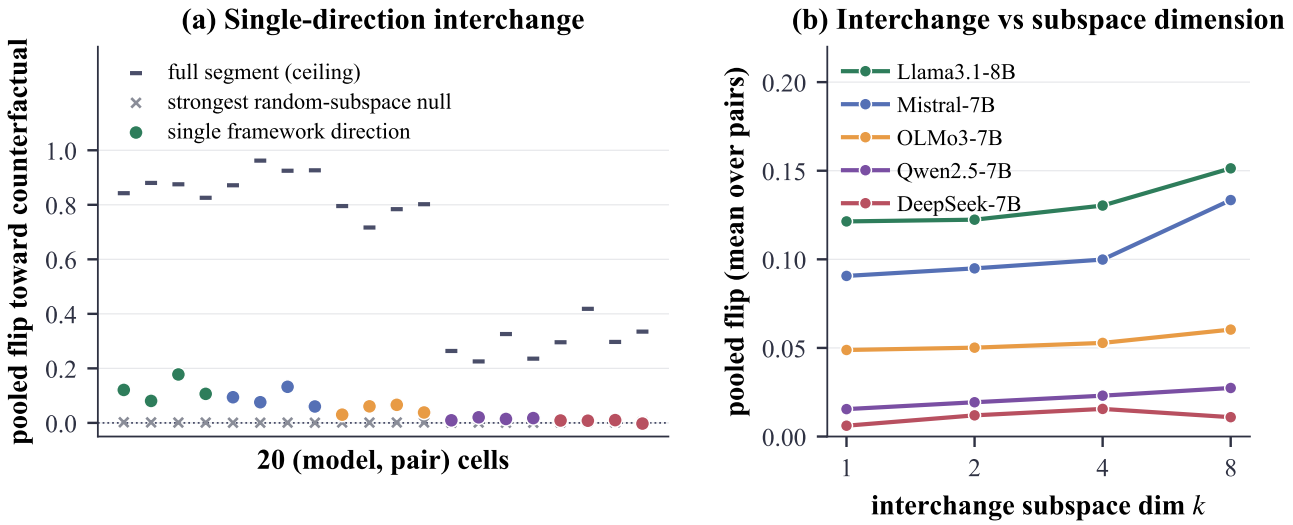


*Figure 12.* **Interchange interventions bound single-direction sufficiency.** *(a)* For each of 20 (model, pair) cells, the pooled flip toward the counterfactual when only the topic-controlled framework direction $v_{\mathrm{tc}}$ component is transplanted (coloured dot) exceeds the strongest of 8 matched random-subspace nulls (grey $\times$) in 19 cells, but sits well below the full-segment ceiling (dark dash): one direction is a reliable handle carrying a minority of the counterfactual margin. Dot colours index models as in (b). *(b)* Repeating the subspace condition with orthonormal bases of dimension $k \in \{1, 2, 4, 8\}$ (mean over the four pairs per model) barely raises the pooled flip, so the remainder is diffuse rather than concentrated in a few further directions.

**Counterfactual topic control** Generation prompt: hold the situation sentence verbatim and the verdict fixed, produce one framing per foundation, with an explicit ban list covering both foundations' lexica (e.g. *harm, suffering, compassion; purity, sacred, degrading*); a post-hoc string filter rejects violations, and an independent validator (Claude Opus 4.1) must recover the intended foundation and matching verdict for a pair to be kept (kept 72, 71, 88, and 64 of 100 generated sets for care/sanctity, fairness/loyalty, care/loyalty, and fairness/sanctity). The topic-controlled direction $v_{\mathrm{tc}}$ is the mean last-token activation difference across kept pairs at the analysis layer. Steering uses the teacher-forced framework-margin readout of §5.2 at $\alpha$=3, against $K$=10 random unit directions; the topic-controlled effect exceeds all nulls in all 20 (model, pair) cells (effects $+0.05$ to $+1.54$; null maxima 0.02–0.31). Split-half reliability is estimated over 50 random half-splits of the pair set.

**Interchange protocol** For each of the four counterfactual pairs above and five anchors, pair items are split in half; $v_{\mathrm{tc}}$ is fitted on the fit half and interchange runs on up to 40 held-out items, in both directions ($A \leftarrow B$ and $B \leftarrow A$). At the analysis layer, over the completion-scoring positions of the framework-margin readout (identical completion tokens across the two runs, so positions align one-to-one), we replace the recipient's component in $\mathrm{span}(v_{\mathrm{tc}})$ with the donor run's component (subspace condition), replace everything except that component (complement), replace the whole segment (full, the protocol ceiling), or replace the component along each of 8 matched random unit directions (nulls). Because per-item flip ratios are unstable when the clean counterfactual gap is small, cells are summarised with a pooled gap-weighted estimator, the ratio of summed signed movement to summed absolute gap. Subspace and complement conditions add to the full-segment effect within $\pm 0.05$ in every cell. Full-segment effects vary by model (0.22–0.96), indicating how much of the margin routes through the patched positions at this layer; the DeepSeek exception in the main text is care/loyalty (pooled subspace effect $-0.003$, empirical $p$=.88).

A dimension sweep repeats the subspace condition with orthonormal bases spanning $v_{\mathrm{tc}}$ plus the top $k-1$ principal components of the fit-half differences orthogonal to it ($k \in \{1, 2, 4, 8\}$), against four random orthonormal $k$-dimensional bases per $k$; complement conditions are dropped in this mode since subspace and complement add to the full-segment effect at $k$=1. Pooled shares move from $-0.00$–0.18 at $k$=1 to 0.01–0.19 at $k$=8 while the dimension-matched null maxima stay at or below 0.007 (Fig. 12b).

**Held-out LEACE valence erasure** The kind/cruel persona pool (150 triggers) is split by trigger; the eraser is fit on one half and evaluated on the other, avoiding the transform leakage that drives same-pool erased readouts below binary chance. The sc101 endorse/violate label remains decodable at 0.98 on the framework pool before and after erasure, confirming the

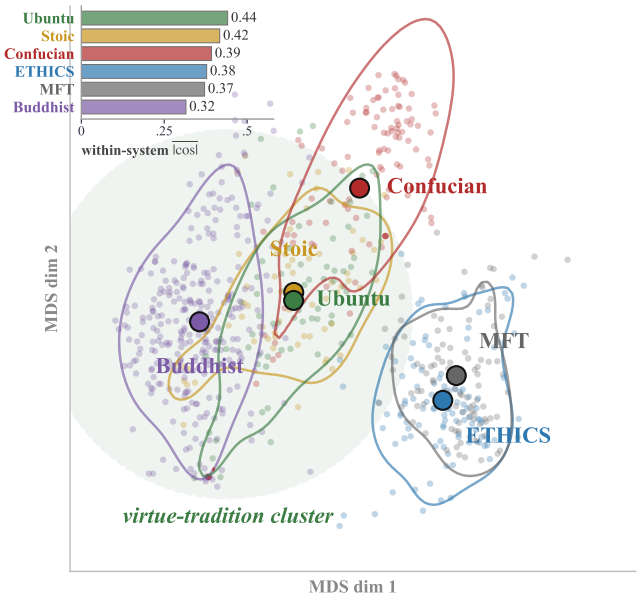


*Figure 13.* **Cross-cultural moral geometry: a virtue-tradition cluster in PV space.** Per-pair PVs across 11 models ($\approx$759 points): classical MDS on $1-|\cos|$, Procrustes-aligned. Per-system centroids and KDE contours (enclosing $\approx 85\%$ of points) outline cluster spread. Buddhist, Stoic, and Ubuntu form the virtue-tradition cluster (green halo; "virtue-tradition" here is the broad philosophical sense, distinct from ETHICS's *virtue* framework label). *Inset:* per-system within-system mean $|\cos|$, sorted with the tightest at top.

eraser's limited on-distribution bite; the framework result is unchanged under it.

**Counterbalanced MoralChoice** Every scenario is presented in both option orders; the debiased score is $p_{\text{action1}} = \frac{1}{2}[P(\text{A} \mid \text{action1 first}) + P(\text{B} \mid \text{action1 second})]$. All five anchors, 200 scenarios per ambiguity tier, shared $\alpha \in \{3, 15\}$. Pooled flips between $\alpha{=}+$ and $\alpha{=}-$: at $\alpha{=}3$, $23/2000$ steer-consistent vs $59/2000$ anti-directional; at $\alpha{=}15$, $69/2000$ vs $154/2000$. Raw position-A choice rates fall from 0.87–1.00 (uncounterbalanced) to 0.45–0.72.

**Judge-noise permutation and judge sensitivity** Within each (model, pair, method, scenario) cell of the cached behavioural-judge annotations, the $\alpha{=}\pm3$ labels are exchanged with probability $\frac{1}{2}$ ($10^4$ shuffles); the observed toward-source fractions (0.056 line, 0.044 subspace) fall inside the resulting null distributions ($p{=}.68$ and .77). A positive control bounds instrument insensitivity: the same judge, on the validated counterfactual framings whose foundation labels are known, recovers the intended foundation as primary on 85–94% of texts (90–97% primary-or-secondary, $n{=}286$), so the free-generation null is not explained by the judge failing to see these foundations. A second-family judge rules out same-judge circularity: an open-weight model from a different family (Qwen2.5-72B-Instruct, int8) re-scores all 960 generations with the identical rubric (valid JSON on every one), agrees with the original judge on 82% of endorsement and 64% of framework labels, and reproduces the null, with toward-source fractions 0.069 (line) and 0.075 (subspace) again inside their permutation nulls ($p{=}.73$ and .25).

**Non-moral control trajectories** Two non-moral persona contrasts (formal vs casual register; verbose vs terse style) are traced across the six public OLMo-3-7B checkpoints with the identical extraction protocol. Cosine to the final direction reaches 0.84–0.88 at SFT (moral pairs: 0.74–0.80) and 0.99 at DPO, matching the moral trajectory's schedule.

## G. Cross-tradition geometry: §5.3–§5.5 supplements

### G.1. Six-system cross-tradition geometry

This appendix supplies the per-cell within- vs cross-system $|\cos|$ data behind the geometric panel of §5.3 and the permutation tests it cites. Coverage spans all 11 models, with the 6-system PV inventory (9 MFT (SC101), 10 ETHICS, 10 Confucian *Wuchang*, 28 Buddhist Eightfold Path, 6 Stoic, 6 Ubuntu), for up to 69 within-system PV pairs per model; the permutation tests use each model's extracted subset.

| Model | $n_{\text{pairs}}$ | within/cross | $p$ | null 95%-CI |
|---|---|---|---|---|
| Llama3.1-8B-Inst | 63 | 2.71 | $<.001$ | [0.88, 1.16] |
| Llama3.1-8B-Base | 69 | 2.31 | $<.001$ | [0.89, 1.14] |
| Llama3.1-70B-Inst | 69 | 3.08 | $<.001$ | [0.89, 1.17] |
| Qwen2.5-7B-Inst | 63 | 3.58 | $<.001$ | [0.88, 1.16] |
| Qwen2.5-7B-Base | 69 | 3.80 | $<.001$ | [0.89, 1.14] |
| Qwen2.5-72B-Inst | 69 | 4.35 | $<.001$ | [0.88, 1.17] |
| Qwen3-8B-Inst | 69 | 3.01 | $<.001$ | [0.90, 1.14] |
| Qwen3-32B-Inst | 69 | 3.85 | $<.001$ | [0.88, 1.16] |
| OLMo3-7B-Inst | 69 | 4.25 | $<.001$ | [0.88, 1.18] |
| OLMo3.1-32B-Inst | 69 | 4.26 | $<.001$ | [0.89, 1.16] |
| DeepSeek-R1D-7B | 59 | 3.59 | $<.001$ | [0.82, 1.34] |
| Gemma3-12B-IT$^{\dagger}$ | 69 | 1.17 | .019 | [0.87, 1.16] |
| Gemma3-27B-IT$^{\dagger}$ | 69 | 1.49 | $<.001$ | [0.88, 1.19] |

*Table 11.* Permutation test of within/cross $|\cos|$ separation across the 11 models.

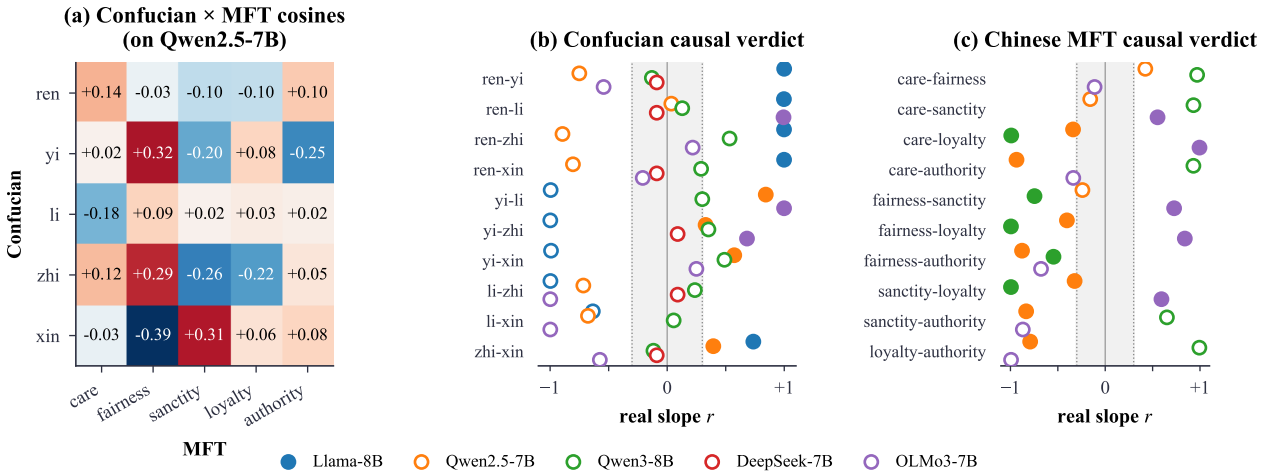


*Figure 14.* Three robustness checks for §5.3. **(a)** Confucian–MFT single-concept signed-cosine heatmap on Qwen2.5-7B. **(b)** Confucian PV causal verdicts on 10 *Wuchang* pairs across 5 Instruct anchors (FLAMES native zh; filled = causal, i.e. the real slope $r$ is opposite in sign to the null baseline *and* $|r| > 0.3$, with the shaded band marking $|r| \leq 0.3$; hollow = one or both conditions unmet). **(c)** Chinese MFT PV causal verdicts on 10 Chinese-language pairs across the 4 Instruct anchors in this panel's model set (Llama is outside it: no MFT-zh causal data); DeepSeek-R1D-7B yields no numerically stable zh PVs, so markers appear for the remaining 3.

Beyond the per-model within/cross ratios in Table 11, the cross-system structure is informative: the densest off-diagonal cells tend to couple Buddhist, Stoic, and Ubuntu, virtue-centric systems sharing pro-social-effort vocabulary, while the sparsest couple SC101 with the same triplet, reflecting MFT's distance from virtue-centric ontologies. The ordering anticipates §5.5's virtue-tradition cluster.

**Permutation test** To rule out that within>cross simply reflects the specific pair-to-system labelling, we pool each model's PV pairs and shuffle the system labels $B = 1000$ times, recomputing the within/cross ratio under each shuffle.

The system-local separation does not depend on the pair-to-system labels chosen: every model's observed ratio sits above the null 95%-CI, and no random reassignment in 1000 draws reproduces it. The geometric distinguishability claim in §5.3 is therefore not an artefact of label selection.

### G.2. Cross-system / cross-language robustness

§5.3's "geometrically distinguishable, functionally substitutable" reading is bounded along three axes: geometric within a single system pair, cross-data-source on Confucian, and cross-language on MFT.

**Geometric: Confucian and MFT occupy near-orthogonal local subspaces** Measured on the same Chinese triggers, the Confucian–MFT $5\times5$ matrix (Fig. 14a) is flat: every cell has $|\cos|\leq0.39$ and most sit near zero. Cells one might expect to align by translation-equivalence, such as *ren* and Care ($+0.14$), are mid-range rather than salient, and *li*'s cells are small ($|\cos|\leq0.18$), so the modest structure that exists (the largest-magnitude cell pairs *xin*, trustworthiness, with MFT-Fairness) does not track a direct virtue-to-foundation translation. The single-system pair view replicates the system-distinguishability finding of §5.3 at finer grain.

**Cross-data-source: Confucian PVs causally usable on $3/5$ anchors** Three of the five tested anchors register Confucian causal verdicts on FLAMES-Morality native Chinese bodies (Llama $5/10$, Qwen2.5 $4/10$, OLMo3-7B $3/10$). Qwen3-8B's null-direction baseline saturates near the answer boundary ($|\bar{r}_{\text{null}}|\approx0.99$, the ceiling-attenuation regime of App. E.6), so its $0/10$ reflects the sign-flip metric rather than absent encoding. DeepSeek-R1D-7B's $0/10$ has a different cause: persona-vector extraction on Chinese prompts returns numerically unstable directions, while the null-direction baseline still returns a stable slope; the verdict reflects PV-extraction failure on out-of-distribution Chinese inputs rather than encoded-but-saturated signal. The active subset shifts modestly across the models that do extract; "how many pairs are causally usable" is the more robust claim than "which pairs."

**Cross-language: causal encoding without correlational transfer** Chinese MFT PVs steer $3/4$ tested anchors on 5–7 of 10 pairs (Qwen2.5 $7/10$, Qwen3-8B $5/10$, OLMo3-7B $5/10$) even though correlational transfer-$d$ is $\sim0$; DeepSeek again returns $0/10$ under the same Chinese-prompt PV-extraction instability noted above (distinct from the Qwen3-8B Confucian saturation pattern). The transfer-vs-causal dissociation reported on English MFT (§5.2) replicates cross-language: causal steering recovers signal that correlational probing alone misses, on every model where PV extraction yields numerically stable directions on Chinese inputs.

**Lexical overlap of hand-authored persona prompts** The Buddhist–Stoic–Ubuntu cluster (§5.5) could be driven by shared vocabulary in our hand-authored personas rather than by conceptual convergence. We compute pairwise Bag-of-Words Jaccard similarity (case-folded, content words after stopword + scaffold-token removal) across all 31 persona prompts. Within the virtue trio (BUD-STO, BUD-UBU, STO-UBU): mean $J=0.064$. Cross-cluster (virtue trio vs MFT + ETHICS, dropping Confucian's Chinese vocabulary): $J=0.024$. The within-virtue overlap is therefore $\sim2.6\times$ the cross-cluster baseline. Absolute Jaccard stays modest ($\sim6\%$ pairwise word overlap) and within the range of within-system overlaps ($\sim$7–17% across the other systems): the lexical overlap drops within trio pairs to roughly within-system magnitude, so lexical convergence contributes to but cannot alone explain the cluster geometry. Combined with the paraphrase invariance check on these systems (§E.4, $70/80$ cells preserve direction sign), the cluster reading should be read as conceptual convergence partially mixed with lexical convergence, not pure lexical artefact.