Steering Geometry: Validating Human Value Geometry in LLM Steering Space

Mohammad Mahdi Abootorabi[†,♠], **Armin Saghafian**[†], **Ali Bazshoushtari**[†], **Hamid Rezaei**[†], **EunJeong Hwang**[†,♠], **Vered Shwartz**[†,♠], **Parvin Mousavi**[§,♠], **Purang Abolmaesumi**[†]

[†]University of British Columbia [♠]Vector Institute for AI [§]Queen's University

Abstract

As large language models (LLMs) are increasingly deployed in alignment-sensitive contexts, activation steering has emerged as a lightweight, inference-time alternative to fine-tuning methods (e.g., RLHF, DPO) for behavioral control. However, existing work typically validates steering on isolated behaviors, leaving it unclear whether steering vectors encode coherent semantic structure or merely exploit behavior-specific shortcuts. We investigate whether the latent geometry of LLM steering vectors reflects theory-specified structure in human values and morality. Using Schwartz's Theory of Basic Human Values as our primary fine-grained framework, we introduce a 26K-sample benchmark covering 20 human values and analyze distribution-driven methods (e.g., CAA, SphericalSteer, ODESteer) and behavior-centric approaches (e.g., COLD-Steer, BiPO) across diverse model families and sizes. We find that distribution-driven methods recover human value topologies aligned with theoretical predictions (Spearman ρ up to 0.51, $p < 10^{-13}$). In contrast, behavior-centric methods achieve comparable steering performance but show little correlation with the expected value geometry. Geometric fidelity improves with model scale but drops after instruction tuning. Finally, better geometric alignment also leads to more human-consistent transfer across values: steering one value correctly lifts compatible values and suppresses opposing ones. Code and data are available at: https://github.com/DeepRCL/Steering_Geometry.

1 Introduction

Large language models (LLMs), trained on large-scale data and shaped through post-training and alignment, have become strong general-purpose assistants across diverse domains, from code generation (Chen et al., 2021) to complex reasoning (Wei et al., 2022). However, as these models are used more often in alignment-sensitive settings where responses carry consequences, researchers face the challenge of making model behavior consistent with human expectations and values (Ouyang et al., 2022; Bai et al., 2022). The primary alignment framework, instruction tuning and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Rafailov et al., 2023), is highly effective at reshaping model behavior to match human preferences, but it relies heavily on resource-intensive annotations and struggles to scale across the diversity of global cultural and contextual values (Askell et al., 2021). These shortcomings have pushed researchers toward mechanistic interpretability, specifically, understanding how models internally represent value-related concepts, and using that knowledge to make slight, targeted adjustments that steer and control model behavior.

Recent work in interpretability suggests that LLMs encode semantic concepts and behavioral traits as approximately linear directions in their activation space (Zou et al., 2023; Turner et al., 2023). This linear representation hypothesis forms the basis of *activation steering*: a family of inference-time methods that manipulate internal representations to control model behavior without additional training (Li et al., 2023; Rimsky et al., 2024). Approaches range from contrastive activation addition (Rimsky et al., 2024) and sparse decompositions (Bayat et al., 2025), to continuous ODE dynamics (Zhao et al., 2026), geometry-aware spherical rotations (You et al., 2026), query-adaptive Schrödinger-bridge steering (Dalili et al., 2026), and learned optimization (Dunefsky and Cohan, 2025). These techniques have recently been extended from simple binary traits such as truthfulness or harmlessness (Wang et al., 2025) to richer constructs like persona and character (Chen et al., 2025). In practice, this means models can be steered toward concrete behavioral profiles, such as becoming more helpful, empathetic, or toxic. Recent works focus on evaluating the impact of steer-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 57010–57033
October 24-29, 2026 ©2026 Association for Computational Linguistics

ing on a single value or behavior, without considering how the steering process affects the model's behavior across other values. In contrast, in human psychology, values are interconnected; for instance, an individual motivated by Achievement is more likely to also value Power (Dominance) than a conflicting trait like Universalism. As such, by only evaluating an isolated behavior during steering, it remains unclear whether steering vectors encode a meaningful semantic structure or simply exploit behavior-specific shortcuts in activation space. Addressing this question requires looking beyond the steered behavior and asking whether steering vectors align with the theory.

Schwartz's Theory of Basic Human Values provides exactly this kind of formal structure (Schwartz, 1992, 2012). Rather than treating values as independent traits, the theory arranges them along a continuous circle. Compatibility between values naturally follows this layout, where aligned values (e.g., Benevolence and Universalism) are closer together and reinforce each other, while conflicting values (e.g., Self-Direction and Conformity) are at opposing angles. In this way, the relationship between any two values is determined by their angular distance. Because this geometry is culturally universal, it serves as a robust ground truth for evaluating value representations (see Figure 6; details in Appendix §B). To assess whether our paradigm-level conclusion is specific to this circumplex, we also evaluate using the revised Moral Foundations Theory (MFT) (Atari et al., 2023).

Whether or not LLM activation spaces actually reflect this kind of value geometry is not clear. Steering vectors are constructed from model activations, not from any explicit encoding of human values; as such, there is no guarantee that their geometry follows theoretical assumptions. If a steering method captures a value's core meaning, motivational content, and relationship to other values—rather than exploiting shortcuts to produce aligned outputs—its vectors should respect the relationships Schwartz's theory describes. This gives us a theory-grounded way to test whether steering vectors encode meaningful structure or not.

Contributions. In this work, **(i)** we present the first study, to the best of our knowledge, of whether or not the geometry of extracted steering vectors reflects human values. Comparing *distribution-driven* and *behavior-centric* steering methods against the Schwartz circumplex, we find that distribution-driven methods reflect the expected human value structure (Spearman ρ up to 0.51, $p < 10^{-13}$), while behavior-centric methods show near-zero correlation despite achieving similar steering performance. The same paradigm-level separation appears under MFT's coarser foundation-family structure, which provides evidence that the result is not specific to the Schwartz circumplex. This suggests that many behavior-centric methods rely on shortcuts that steer outputs without preserving meaningful semantic structure. **(ii)** We find that larger and more recent model families show stronger geometric alignment, whereas instruction-tuned models show weaker value geometry, pointing to a conflict between post-training and richer internal value representations. **(iii)** We introduce a benchmark of roughly 26K contrastive quadruples spanning Schwartz taxonomy. **(iv)** Finally, we demonstrate that methods which better reflect human value geometry also behave more consistently across the value space, where steering toward one value naturally gains accuracy on compatible values and suppresses opposing ones.

2 Related Work

Activation Steering and Representation Engineering. Activation steering methods control model behavior at inference time by directly manipulating internal representations, without any additional weight updates (Zou et al., 2023; Turner et al., 2023). Existing approaches fall into two broad paradigms: *Distribution-driven methods* extract control directions by aggregating activation differences over contrastive prompt datasets, and differ mainly in how those directions are then applied: additively as in Contrastive Activation Addition (CAA; Rimsky et al., 2024), through sparse-autoencoder features (SAS; Bayat et al., 2025), via ODE-based dynamics (Zhao et al., 2026), or via geometry-aware spherical rotations (You et al., 2026). In contrast, *behavior-centric methods* learn steering vectors via gradient descent or in-context optimization against an output objective, e.g., OPT (Dunefsky and Cohan, 2025), COLD-Steer (Sharma and Trivedi, 2026), which approximates in-context learning via finite-difference estimates, and BiPO (Cao et al., 2024). Beyond runtime control of single traits, steering primitives have been extended to modular persona subnetworks (Ye et al., 2026), pluralistic value-alignment frameworks (Kim et al., 2026), and generative

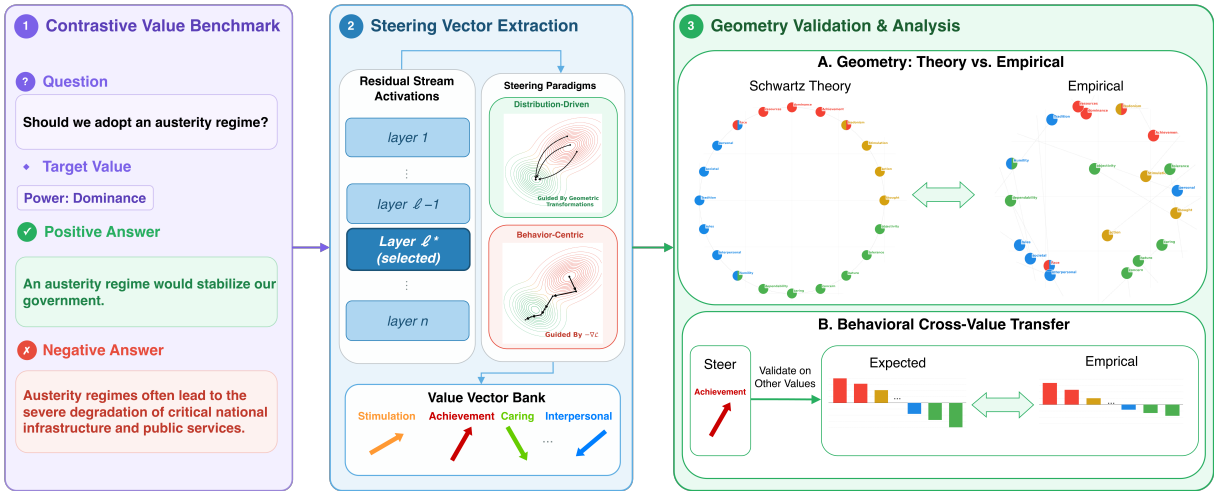

Figure 1: Pipeline for analyzing value geometry in the LLM activation-steering space. **(1)** For each Schwartz value, we build contrastive (question, positive, negative) pairs. **(2)** We extract residual-stream activations at a selected layer $\ell^\star$ and apply steering methods from two paradigms, *distribution-driven* and *behavior-centric*, to construct a value vector bank. **(3)** We evaluate (A) geometric alignment against the Schwartz circumplex, and (B) behavioral cross-value transfer, measuring the effect on other values when steering toward a single one.

multi-agent settings (Paglieri et al., 2026; Arghal et al., 2026).

Human Value Analysis in LLMs. To evaluate whether language models reflect human values, researchers rely on established cognitive science frameworks that organize how humans think about values and morality. Foundational among these is Schwartz's Theory of Basic Human Values (Schwartz, 1992), which organizes universal motivations into a circumplex geometry where compatible values cluster together and opposing values sit in conflict. Other widely adopted frameworks include Moral Foundations Theory (Atari et al., 2023), whose revised formulation distinguishes Care, Equality, Proportionality, Loyalty, Authority, and Purity, alongside Ross's Prima Facie Duties (Ross, 2011), which frames morality as a set of competing obligations. Recent efforts evaluate model alignment against these theories using dedicated benchmarks such as ValueBench (Ren et al., 2024) and MoralBench (Ji et al., 2025), as well as cross-cultural datasets like Global OpinionQA (Durmus et al., 2024).

LLM Activation Space Interpretability. Mechanistic interpretability seeks to decode LLMs' dense activation spaces, driven by the linear representation hypothesis: the premise that high-level concepts, behaviors, and values are encoded as linear directions within the residual stream (Zou et al., 2023; Park et al., 2024). Subsequent work has shown that this linear geometry can extend beyond binary contrasts to categorical and hierarchical concept structures (Park et al., 2025). To overcome the polysemanticity of individual neurons, where one neuron reacts to multiple unrelated concepts, recent work leverages Sparse Autoencoders (SAEs; Huben et al., 2024) to decompose dense activations into interpretable, monosemantic features (Bricken et al., 2023; Templeton et al., 2026). These tools give us a way to isolate value-encoding directions and test whether their geometry matches predictions from psychological theories. Closest to our work, Kang et al. (2025) infer causal graphs over value orientations using role-prompted questionnaire responses and SAE perturbations; they find that LLM-specific graphs differ from human reference graphs but can guide steering with fewer side effects. Their analysis characterizes causal relations among value orientations rather than the geometry of steering directions themselves.

Despite these advances, prior steering work is typically validated by whether a vector changes a target behavior, and prior value-alignment work tests models through behavioral probes. Whether the latent geometry of steering vectors reflects the structure of established human value frameworks, and whether different extraction methods preserve that structure, has not been studied. To address this gap, we build a contrastive benchmark covering the full Schwartz taxonomy and use it to systematically probe the geometric fidelity of steering vectors across methods, model scales, and tuning regimes, as well as their effects on cross-value transfer.

3 Methodology

We investigate whether LLM steering vectors encode meaningful human value structure or only leverage behavior-specific shortcuts in activation space. If steering vectors capture genuine value relationships, steering one value should influence related and conflicting values in predictable ways. This matters for reliable alignment: stable interventions should strengthen related values while suppressing conflicting ones. To test this, we design a four-stage pipeline (Figure 1). First, we construct a value-contrastive dataset based on Schwartz's theory (3.1). Second, we extract value directions from the residual-stream activations using a range of steering methods and store them in a value vector bank (3.2). Third, we run a value geometry test that asks whether the resulting vector space preserves Schwartz value structure (3.3.1). Finally, we conduct a cross-value transfer test to evaluate whether this geometric structure produces psychologically consistent downstream behavior when steering a single value (3.3.2).

3.1 Value-Contrastive Dataset Construction

Studying how human value structure is encoded in activation space requires a dataset of contrastive question-answer pairs. To this end, we construct one grounded in Schwartz's theory (Mirzakhmedova et al., 2024; Kiesel et al., 2022). This culturally universal theory arranges values in a circle, with compatible values close together and conflicting ones apart (Schwartz, 2012). Although several datasets address value-related tasks, none provide contrastive question-answer pairs, a format required by most steering methods to extract value vectors from model activation spaces. We introduce a dataset of *(question, value, positive answer, negative answer)* quadruples, where the positive answer aligns with the target value and the negative answer is value-neutral. We utilized two complementary sources: *ValueBench* (Ren et al., 2024), a benchmark for evaluating value orientations and understanding in LLMs, and *Touché* (Mirzakhmedova et al., 2024). We adopt the 20-value taxonomy of Kiesel et al. (2022), which extends Schwartz's refined 19 values with *Universalism: Objectivity* while preserving the circumplex ordering. The complete taxonomy is provided in Appendix §B. We extracted 911 samples from ValueBench and around 25.5K samples from the Touché data, gathering a dataset of almost 26K data points spanning 20 Schwartz values. For cross-framework evaluation, we additionally construct a balanced benchmark of 1.2K question-contrastive samples spanning the six foundations of Moral Foundations Theory utilizing the Moral Foundations Reddit Corpus (Trager et al., 2026). Full data construction details for the Schwartz and MFT benchmarks are provided in Appendices §E and §G, respectively.

3.2 Value Direction Extraction

The methods we evaluate compute steering signals in very different ways, ranging from contrastive activation statistics and gradient-based optimization to geometry-aware transformations. We represent each method's output as a single *value direction* $v_{(b,\ell)} \in \mathbb{R}^d$, defined as the effective shift it induces in the residual-stream activation hidden states at layer ℓ for Schwartz value b:

$$v_{(b,\ell)} = \mathbb{E}_{x \sim X}\left[a_\ell^{\text{steered}}(x) - a_\ell^{\text{original}}(x)\right], \quad (1)$$

where X is a set of input prompts, and $a_\ell(x)$ is the activation hidden state in layer l for the input x. This way of defining $v_{(b,\ell)}$ lets us capture the impact of steering from any method independent of their internal mechanism. Appendix §J further justifies this approach as a common representation across methods. We then collect the resulting vectors across all 20 Schwartz values into a per-method *value vector bank*, which is used later for geometry analyses.

3.3 Evaluation Protocols

Our evaluation metrics follow a 2×2 organization over value-vector geometry versus cross-value transfer and continuous circumplex versus discrete hierarchical structure. Appendix §H summarizes their interpretations and definitions.

3.3.1 Value Geometry Evaluation

We first ask whether the extracted value vectors $v_{(b,\ell)}$ preserve Schwartz value structure. To test this, for each Schwartz value in its canonical circumplex order, $b \in \{1, \ldots, 20\}$, we first mean-center the steering vectors by subtracting the average all-values direction, $\tilde{v}_b = v_{(b,\ell)} - \frac{1}{20}\sum_{j=1}^{20} v_{(j,\ell)}$, and then normalize them to unit length, $\mathbf{u}_b = \frac{\tilde{v}_{(b,\ell)}}{\|\tilde{v}_{(b,\ell)}\|}$, and build the empirical similarity matrix $E_{bb'} = \mathbf{u}_b^\top \mathbf{u}_{b'}$. We compare $\mathbf{E}$ against a theoretical matrix $\mathbf{T}$ derived from the Schwartz

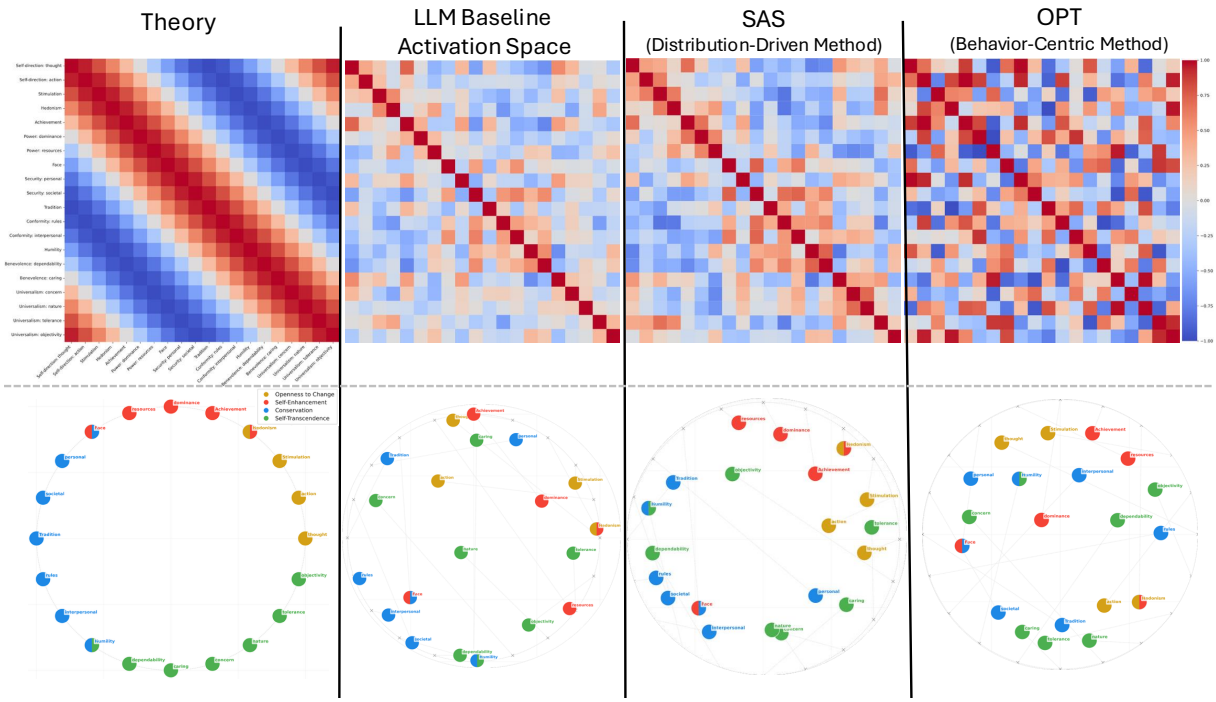

Figure 2: Value geometry across steering methods on Qwen3.5-9B-Base. **Top:** pairwise cosine-similarity matrices for the theoretical Schwartz circumplex, raw residual-stream activations, SAS (a distribution-driven method), and OPT (a behavior-centric method); red/blue indicates positive/negative correlations. **Bottom:** 2D projections of each similarity matrix aligned to the canonical Schwartz arrangement. SAS reproduces the circumplex structure most clearly, while OPT scatters values with no organization, even worse than the raw-activation baseline.

Theory of Human Values, where each entry

$$T_{bb'} = \cos(\min(|b - b'|, 20 - |b - b'|) \times 18^\circ)\,, \quad (2)$$

encodes the angular separation between values b and b' on the Schwartz circumplex. Further intuition and theoretical justification for this construction are provided in Appendix §H. We quantify the alignment between $\mathbf{E}$ and $\mathbf{T}$ using four complementary metrics, with full mathematical definitions provided in Appendix §H.

Theory Rank Correlation (ρ_{T}) and Theory Linear Correlation (r_{T}). To quantify how well the extracted value vectors' geometry matches the Schwartz value circle, we calculate two correlation metrics. First, we use Spearman's rank correlation (ρ_{T}) to test whether the *relative ordering* of value relationships is preserved. It shows whether theoretically compatible values are closer together in the embedding space than conflicting ones. Second, we use Pearson's correlation (r_{T}) to evaluate whether the *magnitudes* of these empirical similarities scale linearly with the theoretical angles. Together, these metrics assess whether the model captures both the general structural hierarchy and the linear scaling of value relationships.

Hierarchical Structure Correlation (ρ_{H}). Schwartz's taxonomy also groups the 20 values into sub-families and higher-order groups. To check whether the model respects this hierarchy, we assign each value pair a distance $d^{\mathrm{hier}}_{bb'}$ of 1 (same sub-family), 2 (same higher-order group), 5 (unrelated), or 10 (opposing groups), and compute the Spearman correlation between $E_{bb'}$ and $-d^{\mathrm{hier}}_{bb'}$. A higher positive ρ_{H} indicates stronger alignment with the multi-level Schwartz structure.

Polarity Separation Score (Δ_{pol}). We compute a separation score: the difference between the mean empirical cosine similarity for value pairs in the same lower-order family (e.g., Benevolence and Universalism) versus pairs from opposing higher-order groups (e.g., Self-Direction vs. Conformity). Larger positive values indicate that the model clusters theoretically compatible values together while pushing opposing values apart.

Robustness Analyses. We additionally evaluate whether the measured value geometry is sensitive to dataset sampling, surface form, or multi-value annotations. Specifically, we re-extract steering vectors using (i) a fresh disjoint subset of 200 examples per value, (ii) fully paraphrased versions of each question and both contrastive answers, and (iii) a single-label subset to assess sensitivity to multi-value annotations. For each setting, we re-

Table 1: **Effect of steering methods on geometry-alignment metrics** on the Qwen 3.5-9B backbone, grouped by tuning regime (Base vs. Instruct) and paradigm (*distribution-driven* vs. *behavior-centric*). The metrics include Theory Rank Correlation (ρ_T), Theory Linear Correlation (r_T), Hierarchical Structure Correlation (ρ_H), and Polarity Separation Score (Δ_{pol}). Subscripts denote p-values. Bold indicates the best value.

Paradigm	Method	ρ_T	r_T	ρ_H	Δ_{pol}
Base *(pre-trained)*					
	LLM Raw Activation Space	$0.2228_{(2.0e-3)}$	$0.2087_{(3.9e-3)}$	$0.2115_{(3.4e-3)}$	0.0009
Behav.-centric	OPT (Dunefsky and Cohan, 2025)	$0.1138_{(1.2e-1)}$	$0.1104_{(1.3e-1)}$	$0.0493_{(5.0e-1)}$	0.0001
	Cold-Steer (FD) (Sharma and Trivedi, 2026)	$0.0265_{(7.2e-1)}$	$-0.0100_{(8.9e-1)}$	$0.0013_{(9.9e-1)}$	0.0162
	BiPO (Cao et al., 2024)	$0.1188_{(1.0e-1)}$	$0.1291_{(7.6e-2)}$	$0.1070_{(1.4e-1)}$	0.0151
Dist.-driven	ODESteer (Zhao et al., 2026)	$0.2730_{(1.4e-4)}$	$0.2963_{(3.3e-5)}$	$0.2988_{(2.8e-5)}$	0.0255
	SphericalSteer (You et al., 2026)	$0.3962_{(1.5e-8)}$	$0.4061_{(6.1e-9)}$	$0.2746_{(1.3e-4)}$	0.3620
	CAA (Rimsky et al., 2024)	$0.4606_{(2.3e-11)}$	$0.4754_{(4.2e-12)}$	$0.3408_{(1.5e-6)}$	0.3883
	SAS (Bayat et al., 2025)	$\mathbf{0.5069}_{(\mathbf{8.5e-14})}$	$\mathbf{0.5061}_{(\mathbf{9.4e-14})}$	$\mathbf{0.3910}_{(\mathbf{2.4e-8})}$	**0.3951**
Instruct *(instruction-tuned)*					
	LLM Raw Activation Space	$0.1253_{(8.5e-2)}$	$0.1112_{(1.3e-1)}$	$0.1455_{(4.5e-2)}$	0.0004
Behav.-centric	OPT (Dunefsky and Cohan, 2025)	$0.0615_{(4.0e-1)}$	$0.0552_{(4.5e-1)}$	$-0.0177_{(8.1e-1)}$	-0.0099
	Cold-Steer (FD) (Sharma and Trivedi, 2026)	$-0.0568_{(4.4e-1)}$	$-0.0394_{(5.9e-1)}$	$-0.0268_{(7.1e-1)}$	0.1401
	BiPO (Cao et al., 2024)	$-0.0104_{(8.9e-1)}$	$-0.0245_{(7.4e-1)}$	$0.0378_{(6.1e-1)}$	-0.0016
Dist.-driven	ODESteer (Zhao et al., 2026)	$0.2055_{(3.7e-3)}$	$0.2177_{(1.9e-3)}$	$0.1495_{(5.2e-3)}$	0.0176
	SphericalSteer (You et al., 2026)	$0.2048_{(4.6e-3)}$	$0.2170_{(2.6e-3)}$	$0.1009_{(1.7e-1)}$	0.2397
	CAA (Rimsky et al., 2024)	$0.2351_{(1.1e-3)}$	$0.2527_{(4.4e-4)}$	$0.1248_{(8.6e-2)}$	0.2603
	SAS (Bayat et al., 2025)	$\mathbf{0.3256}_{(4.6e-6)}$	$\mathbf{0.3331}_{(2.7e-6)}$	$\mathbf{0.2058}_{(4.4e-3)}$	**0.3933**

extract vectors for all steering methods and recompute all geometry metrics. More details are provided in Appendix §F.

Moral Foundations Theory Evaluation. To test whether the paradigm split generalizes beyond Schwartz, we additionally evaluate steering-vector geometry under revised MFT (Atari et al., 2023). Since MFT specifies no circumplex or opposing pairs but does split values between Individualizing (Care, Equality) and Binding (Proportionality, Loyalty, Authority, Purity) foundations, we define Δ_{MFT}, analogous to Δ_{pol}, as the difference between mean within-family and mean cross-family cosine similarity of the extracted vectors. Further details are provided in Appendix §G.

3.3.2 Cross-Value Steering Transfer

Next, we ask whether steering toward one value affects related values in a predictable way. Existing steering evaluations typically focus on a single target in isolation: steer toward value b, report accuracy on b. This narrow view ignores how the intervention reshapes the broader values' behavior, whether it lifts compatible values, suppresses opposing ones, or disrupts both. Under the Schwartz circumplex, steering toward b should yield a predictable pattern: positive transfer to neighboring values and negative transfer to opposing ones. We test whether methods that better capture human value geometry also produce more consistent cross-value transfer.

For each ordered pair (b, b') of distinct Schwartz values, we steer toward b and evaluate on b''s held-out test set, recording $T[b, b'] = \text{acc}_{b \to b'} - \text{baseline}(b')$. This yields a 20×20 matrix per method whose 380 off-diagonal entries describe how steering one value propagates across the value space. However, raw transfer scores mix two confounds with genuine pair-specific effects: some source values are inherently strong steers (a row effect), and some target values are inherently easy to improve (a column effect). To isolate the pair-specific signal, we apply two-way centering by subtracting row and column means, yielding a residualized matrix T_{res} that captures only the transfer attributable to the relationship between b and b'.

Metrics. To quantify whether transfer respects the Schwartz circumplex, we evaluate cross-value transfer along two complementary axes of Schwartz structure: the continuous circumplex, which orders values on a smooth circle, and the discrete hierarchy, which groups them into families and higher-order groups. We introduce *Continuous Transfer Fidelity (TWTM)*, which weights each entry of the residualized transfer matrix T_{res} by its theoretical Schwartz affinity:

$$\text{TWTM} = \frac{1}{380} \sum_{A \neq B} T_{\text{res}}[A, B] \cdot \cos(k_{AB} \cdot 18^\circ), \quad (3)$$

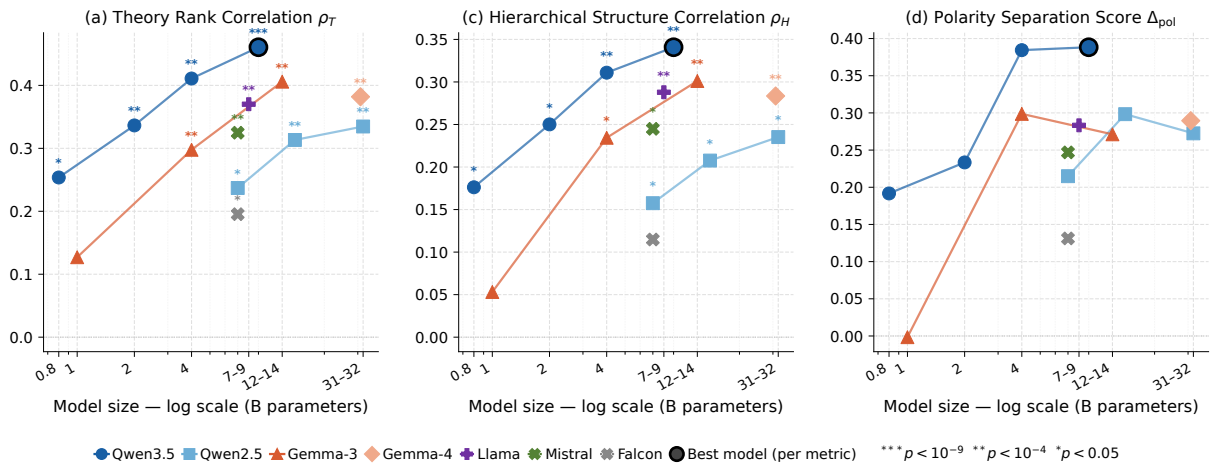

Figure 3: **Effect of model family and size on steering-geometry metrics** (all using CAA). The metrics include Theory Rank Correlation (ρ_T), Hierarchical Structure Correlation (ρ_H), and Polarity Separation Score (Δ_{pol}).

where k_{AB} is the circular step distance between A and B. TWTM rewards large positive transfer at adjacent pairs and large negative transfer at opposing pairs simultaneously, capturing both shape and magnitude on the circumplex.

Second, we introduce *Hierarchical Transfer Fidelity* (ρ_H^{tr}), which is the Spearman rank correlation between $T_{\text{res}}[A, B]$ and $-d_{AB}^{\text{hier}}$, reusing the four-level distance from Section 3.3.1 (same family, same higher-order group, unrelated, opposing). It rewards methods whose transfer is ranked highest within families, lower across higher-order groups, and lowest for opposing groups, testing the multi-level taxonomy rather than the smooth circle. As a robustness check, we additionally report *Binwise Monotonic Decay (BMD-ρ)* in Appendix §I, a shape-only counterpart to TWTM. Together, these metrics cover continuous vs. hierarchical structure and magnitude vs. rank-only views.

4 Experimental Setup

4.1 Steering Methods

Evaluated steering methods are organized into two paradigms: Distribution-Driven and Behavior-Centric. Further intuition behind the paradigm distinction, full objective functions and hyperparameters can be found in Appendix §J.

Distribution-Driven Methods: These methods compute steering vectors by analyzing the distributional differences between activations of contrasting data pairs. Our primary baseline, Contrastive Activation Addition (CAA) (Rimsky et al., 2024), computes the mean difference between the residual-stream activations of positive and negative answers. To address feature superposition, where a single direction encodes multiple unrelated concepts, Sparse Activation Steering (SAS) (Bayat et al., 2025) uses Sparse Autoencoders (SAEs) to isolate a value-specific direction in a disentangled feature space before steering. We also evaluate two methods that constrain how the activation is modified rather than applying a single additive shift. SphericalSteer (You et al., 2026) rotates the activation toward the target direction on the unit sphere, preserving its original norm. ODESteer (Zhao et al., 2026) instead treats steering as a continuous trajectory through activation space, integrating a learned vector field over multiple ODE steps.

Behavior-Centric Methods: Rather than relying on activation differences, these methods optimize the steering vector directly against a behavioral objective or preference loss. Optimization Steering (OPT) (Dunefsky and Cohan, 2025) learns the steering vector through gradient descent on a loss that promotes the log probability of positive answers and suppresses that of negative ones. Bi-directional Preference Optimization (BiPO) (Cao et al., 2024) extends this with a DPO-style preference objective and learns a single vector whose sign controls the direction of steering. COLD-Steer (Sharma and Trivedi, 2026) approximates the activation shift induced by a one-step gradient update on the value-aligned loss via finite differences.

4.2 Implementation Details

We evaluate CAA across seven pre-trained model families: Llama3.1 (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), Falcon (Almazrouei et al., 2023), Qwen2.5 (Qwen Team, 2025), Qwen3.5 (Qwen Team, 2026), Gemma-3 (Gemma Team, 2025), and Gemma-4 (Farabet and La-

combe, 2026), while all steering methods are applied to Qwen3.5-9B in both its pre-trained (Base) and instruction-tuned (Instruct) variants, and to Llama3.1-8B (Base), for a controlled comparison. For each Schwartz value, we sample 200 contrastive prompt pairs with a 90/10 train/test split. For every method, the intervention layer ℓ is selected by maximizing a combined criterion of Normalized L2 Separation and Linear Probe Accuracy (Alain and Bengio, 2017; Li et al., 2023), which identifies the layer where positive and negative activations are most linearly separable. We finetune each method's hyperparameters on the validation split to find its optimal configuration; for instance, see Figure 7 for Qwen3.5-9B-Base. Following prior studies, accuracy gain is measured as the change in the model's selection rate of the value-aligned answer in a two-option multiple-choice setup, relative to the unsteered baseline (Rimsky et al., 2024).

5 Results and Discussion

Effect of Steering Method. Table 1 shows that value geometry depends heavily on the choice of steering method. Distribution-driven methods yield strong alignment across all metrics, while behavior-centric methods show no statistically significant correlation despite comparable results on our dataset (Table 4) and superior behavioral steering accuracy reported in prior works (Cao et al., 2024). The same trend holds on Llama3.1-8B (Table 6). This indicates that behavior-centric methods such as OPT may learn shortcut vectors that achieve the desired behavioral output without encoding meaningful semantic structure. Overall, Distribution-driven methods sharpen the latent value structure that is only partially visible in raw activations, whereas behavior-centric methods weaken it.

Effect of Model Scale and Family. Figure 3 reveals a consistent positive relationship between model scale and geometry alignment within both the Qwen3.5 and Gemma-3 families: metrics improve monotonically with scaling for distribution-driven methods. This suggests that larger models develop progressively richer and more theory-consistent internal representations of value structure. Raw activations improve only modestly, while behavior-centric methods show no reliable scale trend, consistent with our finding that they encode little meaningful value geometry for scaling to strengthen (see Appendix §D for more details).

Comparing model families at similar sizes, newer architectures typically organize value geometry more faithfully: at the 7–9B range, ρ_T rises sharply from Falcon-7B (0.20) to Mistral-7B (0.32), Llama-3.1-8B (0.37), and Qwen3.5-9B (0.46), and within the same family, Qwen3.5 outperforms Qwen2.5 across all scales. Across families, Qwen models also outperform Gemma models at comparable scales, which is a pattern consistent with prior findings attributing Qwen's stronger semantic representations to higher-quality training data (Yang et al., 2025). Most notably, geometric fidelity is independent from downstream performance: Gemma-4-31B, despite being one of the strongest open models, achieves a lower ρ_T (0.38) than Qwen3.5-4B (0.41). *High downstream accuracy does not imply human-aligned value geometry.*

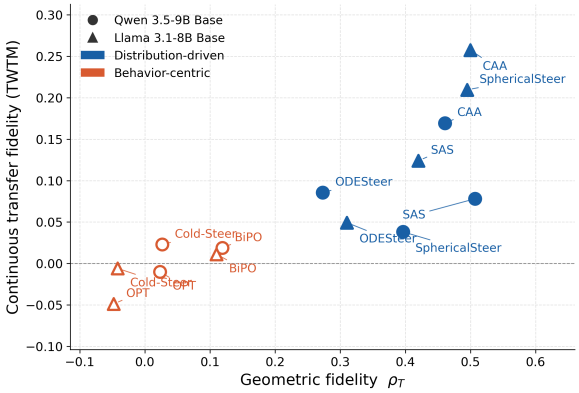

Figure 4: Geometric alignment versus continuous cross-value transfer fidelity across steering methods on Qwen3.5-9B-Base and Llama3.1-8B.

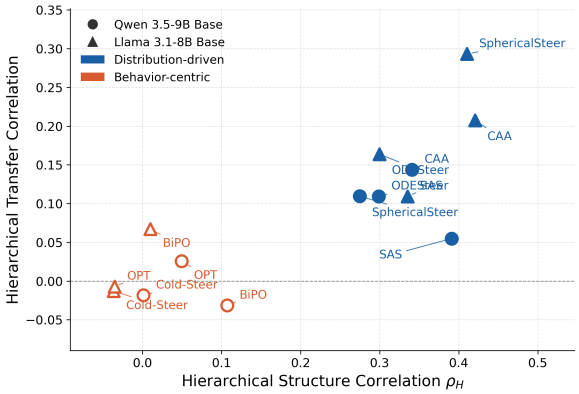

Figure 5: Hierarchical representational fidelity versus hierarchical cross-value transferability across steering methods on Qwen3.5-9B-Base and Llama3.1-8B.

Effect of Instruction Tuning. All methods show consistent degradation when moving from base to instruction-tuned models (Table 1). This suggests that multi-stage post-training is associated with a

form of *value geometry drift*: the activation spaces of instruction-tuned models preserve less of the latent value structure present in pretrained models, consistent with prior observations of representational differences between base and RLHF-tuned models (Rimsky et al., 2024).

Qualitative Analysis. Figure 2 visually confirms the quantitative pattern in Table 1 using cosine-similarity heatmaps and 2D projections. We compare SAS and OPT as representative methods from the two paradigms alongside the theoretical circumplex and raw-activation baseline. In the heatmaps, SAS most closely reproduces the theoretical matrix, raw activations preserve a partial but recognizable pattern, and OPT remains largely unstructured. The projections of pairwise cosine dissimilarities $(1 - \cos)$ show the same ordering: SAS is closest to the theoretical arrangement, raw activations show moderate deviation, and OPT scatters the values. Additional scatter and t-SNE visualizations are provided in Appendix §A.

Robustness to Surface Features and Multi-Value Annotations. The paradigm-level separation remains under all three robustness analyses (Appendix §F). Re-extracting vectors from a fresh disjoint sample preserves the separation, showing that it is not specific to the original sample selection. Fully paraphrasing the questions and answers yields the same qualitative result, which argues against surface wording or generation style as its source. Finally, repeating the analysis on a subset that prioritizes examples with a single value annotation preserves the paradigm separation and statistical significance on both backbones, indicating that multi-value annotations do not drive the main conclusion. Although absolute correlations decrease in this setting, the matched raw activation baseline decreases as well, suggesting that this change partly reflects the altered sample distribution. Full results are reported in Table 10.

Generalization Beyond Schwartz. Table 2 shows that distribution-driven methods yield positive MFT family separation on both backbones ($\Delta_{\text{MFT}} \approx 0.04$–$0.10$, versus ≈ 0.001 for the raw-activation baselines), while behavior-centric methods remain near zero or negative. This suggests that the paradigm distinction is not specific to Schwartz theory.

Cross-Value Steering Transfer. Figures 4 and 5 showcase the reason geometry matters. When evaluation is restricted to the steered target value (Table 4), behavior-centric methods perform comparably to distribution-driven ones despite their poor geometric fidelity. Evaluating across all values, however, the two paradigms separate cleanly under both of the value-transfer metrics: distribution-driven methods achieve high geometric fidelity and strongly transfer, while behavior-centric methods cluster near zero on both *Continuous Transfer Fidelity* (TWTM) and *Hierarchical Transfer Fidelity* (ρ_H^{tr}). The same separation holds under our shape-only metric BMD-ρ (Appendix §I).

Method	Qwen3.5-9B	Llama3.1-8B
LLM Raw Activation	0.0006	0.0017
Behavior-centric		
OPT (Dunefsky and Cohan, 2025)	0.0107	0.0272
Cold-Steer (Sharma and Trivedi, 2026)	-0.0064	-0.0280
BiPO (Cao et al., 2024)	-0.0287	0.0227
Distribution-driven		
ODESteer (Zhao et al., 2026)	0.0495	0.0378
SphericalSteer (You et al., 2026)	0.0932	0.0909
CAA (Rimsky et al., 2024)	**0.0969**	**0.1038**
SAS (Bayat et al., 2025)	0.0892	0.0923

Table 2: Effect of steering method on MFT family separation (Δ_{MFT}). Positive values indicate greater within-family than cross-family cosine similarity.

Pooling all methods and both backbones, geometric fidelity (ρ_T) predicts cross-value transfer more strongly than raw accuracy gain does on both TWTM and ρ_H^{tr}, with the gap widening on the rank-based hierarchical metric ($\Delta\rho = 0.06$ on TWTM, $\Delta\rho = 0.18$ on ρ_H^{tr}; see Appendix §K, Table 12). Methods that preserve Schwartz structure in vectors reflect it through behavior; methods optimized purely for behavioral objectives steer the target effectively but shift related values less predictably.

6 Conclusion

We present a systematic study of whether LLM steering vectors reflect human value structure, and find a clear split between paradigms: distribution-driven methods capture theory-consistent relationships and propagate them across the value space, while behavior-centric methods achieve comparable target accuracy but lose this structure and shift related values unpredictably. We encourage the community to evaluate steering on psychological structure alongside behavioral accuracy. One direction for future work is to examine whether value-level steering yields more predictable downstream effects than steering surface behavior directly: steering a behavior such as politeness may introduce unintended side effects like suppressing honesty or amplifying conformity, whereas steering through values could offer more control.

7 Limitations

Our fine-grained geometry and cross-value-transfer analyses remain centered on Schwartz's Theory of Basic Human Values as a representative human value system theory. While the theory is cross-culturally validated, we additionally test generalization under MFT, but the theory specifies only the coarser Individualizing–Binding grouping and no finer inter-foundation structure; this experiment therefore provides family-level validation rather than a full geometric test. Generalization beyond these two frameworks remains open, particularly across cultures and languages, where cognitively grounded evaluations reveal substantial cross-lingual gaps (Abootorabi et al., 2026). Real arguments may express multiple values, so the extracted vectors should be interpreted as target-conditioned aggregate directions rather than perfectly monosemantic representations. Although our single-label-prioritized sensitivity test preserves the paradigm-level conclusion, it does not establish perfect monosemanticity. In addition, most of our cross-backbone comparisons are conducted with a few steering methods due to computational constraints, leaving open whether the trends we observe hold consistently across other steering methods.

References

Mohammad Mahdi Abootorabi, Omid Ghahroodi, Anas Madkoor, Marzia Nouri, Doratossadat Dastgheib, and Ehsaneddin Asgari. 2026. Almieyar-oryx-BloomBench: A bilingual multimodal benchmark for cognitively informed evaluation of vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 28404–28436, San Diego, California, United States. Association for Computational Linguistics.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations (ICLR 2017), Workshop Track*, Toulon, France.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Raghu Arghal, Kevin He, Shirin Saeedi Bidokhti, and Saswati Sarkar. 2026. Steering the herd: A framework for LLM-based control of social learning. In *The Fourteenth International Conference on Learning Representations*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of personality and social psychology*, 125(5):1157.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. 2025. Steering large language model activations in sparse spaces. In *Second Conference on Language Modeling*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *Advances in Neural Information Processing Systems*, 37:49519–49551.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Seyed Arshan Dalili, Ajay Narayanan Sridhar, Vijaykrishnan Narayanan, and Mehrdad Mahdavi. 2026. Conditional optimal bridge for riemannian activation steering. *arXiv preprint arXiv:2607.10517*.

Tulsee Doshi. 2025. Gemini 3 Flash: Frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/. On behalf of the Gemini team.

Jacob Dunefsky and Arman Cohan. 2025. One-shot optimized steering vectors mediate safety-relevant behaviors in LLMs. In *Second Conference on Language Modeling*.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Clement Farabet and Olivier Lacombe. 2026. Gemma 4: Byte for byte, the most capable open models. https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/. Google Blog.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on llm-as-a-judge. *The Innovation*, 7(6):101253.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pages 7827–7845.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2025. Moralbench: Moral evaluation of llms. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yipeng Kang, Junqi Wang, Yexin Li, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, Tingjun Wu, Xue Feng, Fangwei Zhong, and Zilong Zheng. 2025. Are the values of LLMs structurally aligned with humans? a causal perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23147–23161, Vienna, Austria. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303.

Woojin Kim, Sieun Hyeon, Jusang Oh, and Jaeyoung Do. 2026. Valueflow: Toward pluralistic and steerable value-based alignment in large language models. *arXiv preprint arXiv:2602.03160*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, MohammadAli SadraeiJavaheri, and 1 others. 2024. The touché23-valueeval dataset for identifying human values behind arguments. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16121–16134.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Davide Paglieri, Logan Cross, William A Cunningham, Joel Z Leibo, and Alexander Sasha Vezhnevets. 2026. Persona generators: Generating diverse synthetic personas at scale. *arXiv preprint arXiv:2602.03545*.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2025. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, volume 2025, pages 76441–76463.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5. Qwen Blog, published February 15, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. 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)*, pages 2015–2040.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

W David Ross. 2011. *Foundations of ethics*. Read Books Ltd.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11.

Kartik Sharma and Rakshit Trivedi. 2026. COLD-steer: Steering large language models via in-context one-step learning dynamics. In *The Fourteenth International Conference on Learning Representations*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, and 7 others. 2026. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Preprint*, arXiv:2605.29358.

Jackson P. Trager, Alireza S. Ziabari, Elnaz Rahmati, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Georgios Chochlakis, Nils Karl Reimer, Melissa Reyes, Kesley Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. 2026. The moral foundations reddit corpus. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6383–6407, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, pages 2562–2578.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ruimeng Ye, Zihan Wang, Zinan Ling, Yang Xiao, Manling Li, Xiaolong Ma, and Bo Hui. 2026. Your language model secretly contains personality subnetworks. In *The Fourteenth International Conference on Learning Representations*.

Zejia You, Chunyuan Deng, and Hanjie Chen. 2026. Spherical steering: Geometry-aware activation rotation for language models. In *Forty-third International Conference on Machine Learning*.

Hongjue Zhao, Haosen Sun, Jiangtao Kong, Xiaochang Li, Qineng Wang, Liwei Jiang, Qi Zhu, Tarek F. Abdelzaher, Yejin Choi, Manling Li, and Huajie Shao. 2026. ODESteer: A unified ODE-based steering framework for LLM alignment. In *The Fourteenth International Conference on Learning Representations*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Visualizing Value Geometry Alignment

To further analyze the structure of value vectors, we visualize the emergent value geometry and examine how it varies across model scales and architectures.

Figures 8 and 9 provide visual evidence for the alignment between theoretical and empirical value structures. Both show results from SAS, which is our best-performing configuration.

Figure 8 demonstrates that theoretical predictions correlate with empirical measurements across all value pairs: theoretically compatible values show high cosine similarity, while conflicting values show low or negative similarity. Figure 9 shows t-SNE projections revealing spatial clustering of same-quadrant values and separation of opposing values such as Conservation and openness to change, confirming that the steering space preserves Schwartz's circumplex structure.

B Refined Schwartz Theory of Basic Human Values

In our work, we use the Refined Schwartz Theory of Basic Human Values (Mirzakhmedova et al., 2024) as a reference for understanding relationships between different values. Since the theory organizes human values based on how similar or conflicting they are, it provides a natural way to evaluate whether these relationships are reflected in the learned vector representations. See Figure 6 for the Schwartz circumplex and Table 5 for the explanation of values. We adopt the taxonomy of Kiesel et al. (2022), which extends the refined 19-value framework of Schwartz (2012) with one additional category, *Universalism: Objectivity*. Based on comparisons with other value inventories, this category is placed between *Universalism: Tolerance* and *Self-Direction: Thought*, thereby preserving the circumplex structure. This 20-category framework has become the standard taxonomy in NLP research on human values (Mirzakhmedova et al., 2024; Kiesel et al., 2023).

C Additional Results on Llama 3.1-8B

Tables 6 and 7 show that the same paradigm-level patterns seen on Qwen 3.5-9B are present on Llama 3.1-8B: distribution-driven techniques outperform behavior-centric methods in recovering the Schwartz circumplex with CAA achieving the strongest theory alignment ($\rho_T = 0.50$, $p < 10^{-12}$). On the other hand, behavior-centric methods have non-trivial behavioral gains but are close to zero or negative for most of the geometry metrics, further highlighting the distinction between behavioral steering performance and geometric fidelity.

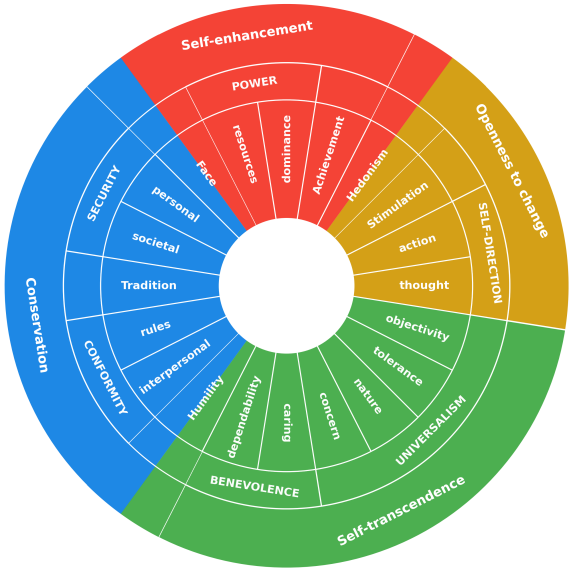

Figure 6: The Schwartz Theory of Basic Human Values organizes 20 basic human value categories in a circular structure, where more similar values appear close together, while values that tend to conflict are placed on opposite sides of the circle. Illustration adapted from Schwartz (2012).

D Further Scale Analysis

We extend the scale analysis beyond CAA by evaluating SphericalSteer on Qwen3.5 (0.8B–9B) and Qwen2.5 (7B–32B). As shown in Table 8, both methods exhibit monotonic improvements in theory alignment across both model families. In comparison, the raw-activation baseline varies only modestly and non-monotonically ($\rho_T = 0.16$–0.22 on Qwen3.5 and 0.14–0.19 on Qwen2.5), suggesting that the improvement is not merely inherited from richer backbone activations. SAS also improves where pretrained SAEs are available ($\rho_T = 0.46$ on Qwen3.5-2B and 0.51 on Qwen3.5-9B), although broader comparison is limited by SAE availability. Available BiPO results show no reliable scale trend. These results support scale-related improvement for distribution-driven methods rather than for steering methods generally.

E Dataset Details

This section describes the construction, filtering, and distribution of the datasets used to extract and

Table 3: **Effect of model family and size on steering-geometry metrics** (all using CAA). The metrics include Theory Rank Correlation (ρ_T), Theory Linear Correlation (r_T), Hierarchical Structure Correlation (ρ_H), and Polarity Separation Score (Δ_{pol}). Subscripts denote p-values. Bold indicates the best value.

Model	ρ_{T}	r_{T}	ρ_{H}	Δ_{pol}
Falcon-7B	$0.1954_{(6.9e\text{-}3)}$	$0.1989_{(5.9e\text{-}3)}$	$0.1148_{(1.1e\text{-}1)}$	0.1312
Mistral-7B	$0.3248_{(4.8e\text{-}6)}$	$0.3378_{(1.9e\text{-}6)}$	$0.2449_{(6.6e\text{-}4)}$	0.2469
Llama-3.1-8B	$0.3701_{(1.5e\text{-}7)}$	$0.3783_{(7.4e\text{-}8)}$	$0.2879_{(5.6e\text{-}5)}$	0.2833
Gemma-3-1B	$0.1270_{(8.1e\text{-}2)}$	$0.1376_{(5.8e\text{-}2)}$	$0.0531_{(4.7e\text{-}1)}$	−0.0016
Gemma-3-4B	$0.2977_{(3.0e\text{-}5)}$	$0.2992_{(2.8e\text{-}5)}$	$0.2344_{(1.1e\text{-}3)}$	0.2987
Gemma-3-12B	$0.4059_{(6.3e\text{-}9)}$	$0.4153_{(2.6e\text{-}9)}$	$0.3011_{(2.4e\text{-}5)}$	0.2712
Qwen2.5-7B	$0.2370_{(9.9e\text{-}4)}$	$0.2491_{(5.3e\text{-}4)}$	$0.1575_{(3.0e\text{-}2)}$	0.2149
Qwen2.5-14B	$0.3133_{(1.1e\text{-}5)}$	$0.3298_{(3.4e\text{-}6)}$	$0.2076_{(4.1e\text{-}3)}$	0.2983
Qwen2.5-32B	$0.3347_{(2.4e\text{-}6)}$	$0.3470_{(9.4e\text{-}7)}$	$0.2352_{(1.1e\text{-}3)}$	0.2725
Gemma-4-31B	$0.3819_{(5.4e\text{-}8)}$	$0.3928_{(2.1e\text{-}8)}$	$0.2835_{(7.4e\text{-}5)}$	0.2895
Qwen3.5-0.8B	$0.2537_{(4.1e\text{-}4)}$	$0.2643_{(2.3e\text{-}4)}$	$0.1762_{(1.5e\text{-}2)}$	0.1917
Qwen3.5-2B	$0.3364_{(2.1e\text{-}6)}$	$0.3545_{(5.2e\text{-}7)}$	$0.2501_{(5.0e\text{-}4)}$	0.2334
Qwen3.5-4B	$0.4109_{(3.9e\text{-}9)}$	$0.4243_{(1.1e\text{-}9)}$	$0.3109_{(1.3e\text{-}5)}$	0.3844
Qwen3.5-9B	$\mathbf{0.4606}_{(2.3e\text{-}11)}$	$\mathbf{0.4754}_{(4.2e\text{-}12)}$	$\mathbf{0.3408}_{(1.5e\text{-}6)}$	**0.3883**

Table 4: **Effect of steering method on accuracy gain** for the Qwen 3.5-9B Base (pre-trained) backbone. Gain is reported in percentage points (pp) relative to the unsteered baseline. Bold indicates the best value.

Method	Gain (pp)
Cold-Steer (FD) (Sharma and Trivedi, 2026)	+5.50
BiPO (Cao et al., 2024)	+8.87
SphericalSteer (You et al., 2026)	+9.38
OPT (Dunefsky and Cohan, 2025)	+9.57
ODESteer (Zhao et al., 2026)	+9.70
SAS (Bayat et al., 2025)	+10.03
CAA (Rimsky et al., 2024)	**+11.74**

evaluate value vectors across the 20 Schwartz value categories.

E.1 Gathering and Generation

ValueBench. ValueBench (Ren et al., 2024) is built from 44 established psychological questionnaires and tests spanning personality, social axioms, cognitive systems, and general value theory. Each item comes with a value label and an agreement score to indicate whether the item endorses or opposes that value. Since ValueBench covers far more dimensions than the 20 Schwartz values, we used Gemini 3 Flash (Doshi, 2025) to map each item to its closest Schwartz value, keeping only items where the mapping was clear. For items missing one answer, we generated the missing answer using Qwen3.5-35B-A3B (Qwen Team, 2026) with in-context learning, and manually verified a representative subset of the process afterward. After filtering, this gave us 911 samples.

Touché23-ValueEval. The Touché23-ValueEval dataset (Mirzakhmedova et al., 2024) contains 9,324 arguments from six domains, including religious texts, political discussions, newspaper editorials, and online democracy platforms. Each argument was annotated by three crowdworkers for 54 human values organized into 20 Schwartz categories. Each argument pairs a conclusion with a premise that either supports or opposes it. We turn each conclusion into a question and use the supporting premise as the positive answer. Negative answers are generated using Gemma-4-26B-A4B-it (Farabet and Lacombe, 2026), each conditioned on one of six prompt strategies for producing value-neutral responses (see Appendix §E.3). We then use the same model in an LLM-as-a-judge framework (Gu et al., 2026) to filter out low-quality samples, leaving around 25.5K final pairs.

E.2 Dataset Statistics

Table 9 presents the distribution of contrastive quadruples across the 20 Schwartz value categories and two dataset sources after quality filtering. Touché dominates with 25,517 samples (96.6%), while ValueBench contributes the remaining 911 (3.4%), yielding 26,428 quadruples in total.

Table 5: **The 20 Basic Human Values of the Refined Schwartz Theory** and their core motivational goals.

Value	Core Motivational Goal
Self-Direction: Thought	Freedom to cultivate one's own ideas and abilities
Self-Direction: Action	Freedom to determine one's own actions
Stimulation	Excitement, novelty, and challenge in life
Hedonism	Pleasure and sensuous gratification for oneself
Achievement	Success through demonstrating competence by social standards
Power: Dominance	Power through exercising control over people
Power: Resources	Power through control of material and social resources
Face	Security and power through maintaining public image
Security: Personal	Safety and security in one's immediate environment
Security: Societal	Safety and stability in the wider society
Tradition	Maintaining cultural, family, or religious customs
Conformity: Rules	Compliance with rules, laws, and formal obligations
Conformity: Interpersonal	Avoidance of upsetting or harming other people
Humility	Recognizing one's insignificance in the larger scheme
Benevolence: Dependability	Being a reliable and trustworthy member of the in-group
Benevolence: Caring	Devotion to the welfare of in-group members
Universalism: Concern	Commitment to equality and justice for all people
Universalism: Nature	Preservation of the natural environment
Universalism: Tolerance	Acceptance and understanding of those who are different
Universalism: Objectivity	Pursuit of truth, rationality, and an unbiased worldview

E.3 Negative Answer Generation Strategies

Negative answers for the Touché-derived subset are generated using one of six strategies: pragmatic, which challenges feasibility, cost, or practical implementation; empirical, which contests factual or causal claims using evidence or known outcomes; counter-example, which cites a real or plausible case where the same policy led to opposite results; side-effects, which argues that the policy produces serious unintended consequences in a different domain; institutional, which asserts that existing rules or institutions already address the concern more effectively; and contradict, which disputes whether the value-based premise is valid, relevant, or accurately applied in context.

Each example was deterministically assigned one of the six strategies based on its argument ID and target value; the complete prompts, including all templates, demonstrations, and leakage warnings, are provided in the released code.

F Robustness of Value-Geometry Evaluation

We conduct additional analyses to test whether the observed value geometry and the separation between steering paradigms could be explained by properties of the dataset rather than the extracted value directions. Specifically, we examine robustness to sample selection, surface wording, and multi-value annotations. For each robustness setting, we re-extract steering vectors for all seven methods on Qwen3.5-9B-Base and Llama3.1-8B and repeat the geometry analysis.

F.1 Sample Selection and Surface Wording

Most examples in our benchmark are derived from Touché23-ValueEval, where the question and value-aligned answer originate from the human-annotated source data and the value-neutral answer is generated by an LLM. We therefore consider several controls for potential artifacts introduced by this construction procedure.

First, generation is constrained to produce answers of approximately the same length as their human-written counterparts. Generated answers contain 26.8 tokens on average, compared with 22.2 tokens for the corresponding human-written answers. We also generate negative answers using six different prompting strategies—pragmatic, empirical, counter-example, side-effects, institutional, and contradict—to avoid reliance on a single stylistic template (Appendix §E.3). In addition, different models are used across stages of the data-construction pipeline, and the generation models differ from the backbones evaluated in our steering experiments.

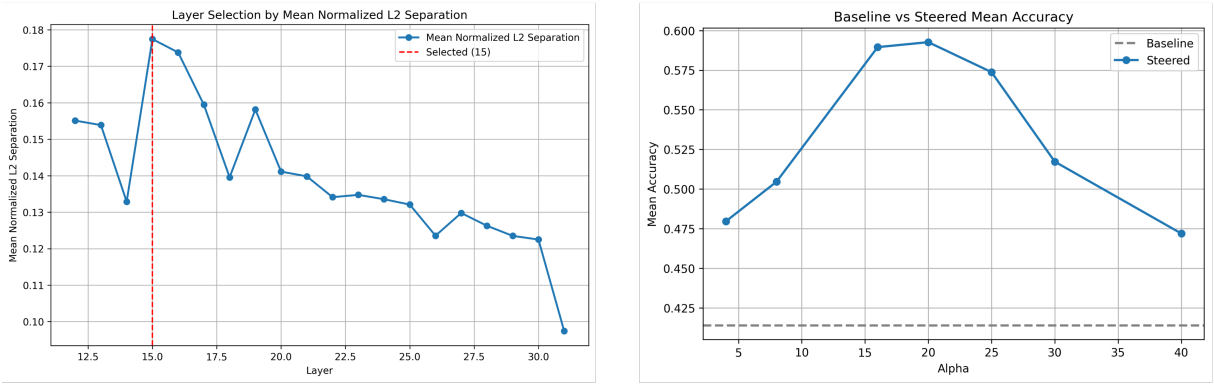

Figure 7: (Left) Layer selection via Mean Normalized L2 Separation across transformer layers for CAA for Qwen3.5-9B-Base. Layer 15 is selected (dashed red line) as it achieves the highest mean separation between positive and negative activation distributions. (Right) Mean steering accuracy across Schwartz values as a function of the multiplier α, compared against the unsteered baseline (dashed gray line). Accuracy peaks around $\alpha \in [15, 20]$ before degrading at larger values, consistent with prior observations that excessive steering strength degrades generation quality.

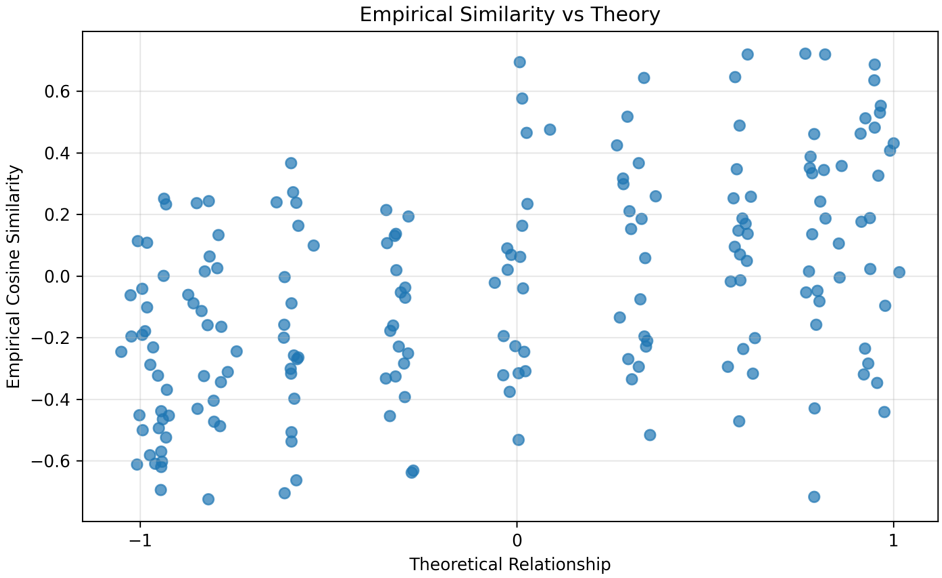

Figure 8: Correlation between theoretical value relationships (from Schwartz Theory) and empirical cosine similarities of extracted value vectors. The scatter plot illustrates the individual data points; while the distribution exhibits noticeable variance, a general upward trend is visible as one moves from left to right along the x-axis, demonstrating that higher theoretical relationships correspond to higher empirical similarities. This positive correlation indicates that LLM representations preserve human-predicted value structure. The results are from the SAS method on Qwen3.5-9B-Base model, which shows the highest correlation among all our configurations.

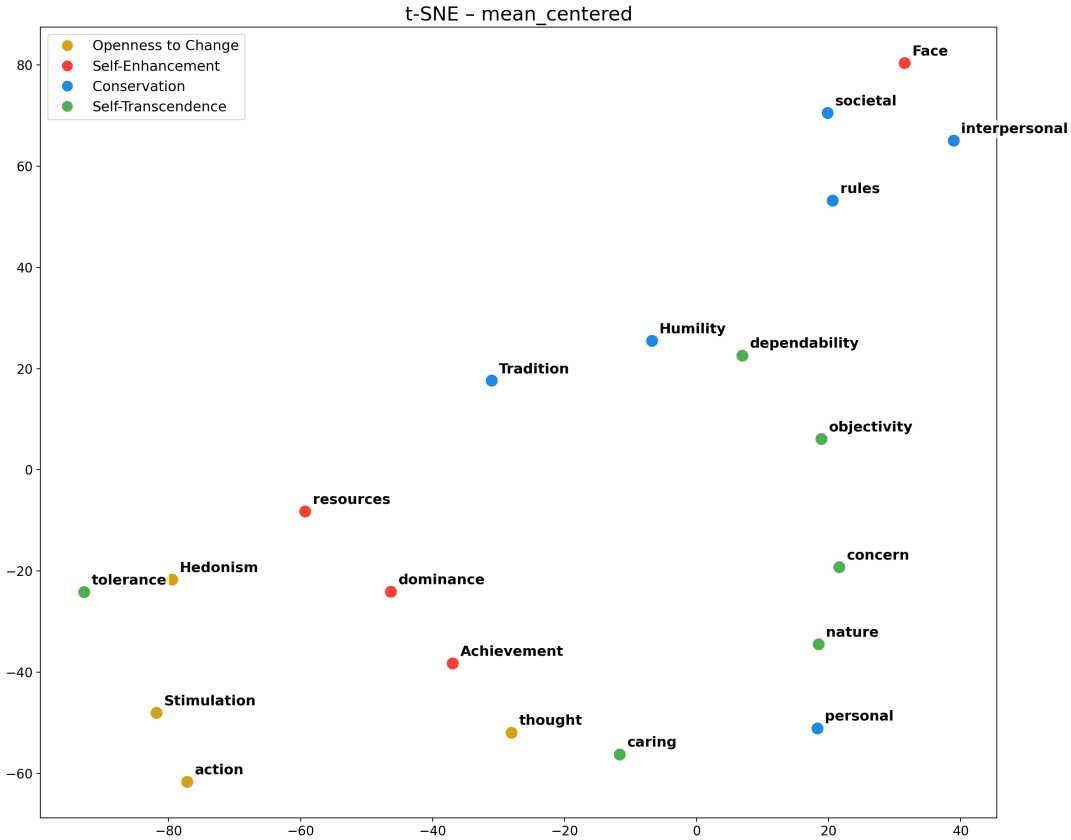

Figure 9: t-SNE projection of 20 Schwartz value vectors, colored by higher-order dimension. Spatial clustering of same-quadrant values and separation of opposing values provides visual evidence of circumplex structure in LLM activation space. The results are from the SAS method on Llama-3.1-8B model.

Table 6: **Effect of steering method on geometry-alignment metrics** on the Llama 3.1-8B backbone. The metrics include Theory Rank Correlation (ρ_T), Theory Linear Correlation (r_T), Hierarchical Structure Correlation (ρ_H), and Polarity Separation Score (Δ_{pol}). Subscripts denote p-values. Bold indicates the best value.

Paradigm	Method	ρ_T	r_T	ρ_H	Δ_{pol}
Base *(pre-trained)*					
	LLM Raw Activation Space	$0.2387_{(9.1e\text{-}4)}$	$0.2377_{(9.6e\text{-}4)}$	$0.2530_{(4.3e\text{-}4)}$	0.0026
Behav.-centric	OPT (Dunefsky and Cohan, 2025)	$-0.0486_{(5.1e\text{-}1)}$	$-0.0696_{(3.4e\text{-}1)}$	$-0.0352_{(6.3e\text{-}1)}$	0.0232
	Cold-Steer (FD) (Sharma and Trivedi, 2026)	$-0.0405_{(5.8e\text{-}1)}$	$-0.0404_{(5.8e\text{-}1)}$	$-0.0365_{(6.2e\text{-}1)}$	0.0104
	BiPO (Cao et al., 2024)	$0.1094_{(1.3e\text{-}1)}$	$0.1036_{(1.5e\text{-}1)}$	$0.0098_{(8.9e\text{-}1)}$	-0.0094
Dist.-driven	ODESteer (Zhao et al., 2026)	$0.3124_{(1.1e\text{-}5)}$	$0.3026_{(2.2e\text{-}5)}$	$0.2996_{(2.7e\text{-}5)}$	0.0195
	SAS (Bayat et al., 2025)	$0.4220_{(1.3e\text{-}9)}$	$0.4327_{(4.5e\text{-}10)}$	$0.3351_{(2.3e\text{-}6)}$	0.3083
	SphericalSteer (You et al., 2026)	$0.4949_{(3.9e\text{-}13)}$	$0.5014_{(1.7e\text{-}13)}$	$0.4103_{(4.1e\text{-}9)}$	**0.3974**
	CAA (Rimsky et al., 2024)	$\mathbf{0.4996}_{(2.2e\text{-}13)}$	$\mathbf{0.5060}_{(9.6e\text{-}14)}$	$\mathbf{0.4205}_{(1.5e\text{-}9)}$	0.3957

Table 7: **Effect of steering method on accuracy gain** for the Llama 3.1-8B backbone. Gain is reported in percentage points (pp) relative to the unsteered baseline. Bold indicates the best value.

Method	Gain (pp)
OPT (Dunefsky and Cohan, 2025)	+6.03
BiPO (Cao et al., 2024)	+6.30
ODESteer (Zhao et al., 2026)	+6.60
Cold-Steer (Sharma and Trivedi, 2026)	+7.50
SphericalSteer (You et al., 2026)	+17.57
CAA (Rimsky et al., 2024)	+17.64
SAS (Bayat et al., 2025)	**+19.87**

We additionally perform a manual audit of 100 generated examples, sampling 5 examples from each Schwartz value. Overall, 89% of the generated answers were judged valid. The primary failure modes were value leakage and insufficient contrast with the corresponding human-written answer.

Disjoint resampling. To test whether the observed geometry depends on the particular examples used in the main experiments, we construct a fresh subset containing 200 examples per value with no overlap with the original subset. We then re-extract all steering vectors and repeat the complete geometry analysis for both backbones.

Surface-form perturbation. To test sensitivity to lexical and stylistic cues, we construct a second variant in which the question and both contrastive answers of every example are paraphrased using `gemma-4-26B-A4B-it`. The paraphrasing procedure preserves the semantic content and target-value assignment while changing the surface realization of all three components. We again re-extract the steering vectors from the paraphrased data and repeat the full geometry evaluation.

Method	Model	ρ_T	r_T
CAA (Rimsky et al., 2024)	Qwen3.5-0.8B	0.2537	0.2643
	Qwen3.5-2B	0.3364	0.3545
	Qwen3.5-4B	0.4109	0.4243
	Qwen3.5-9B	0.4606	0.4754
	Qwen2.5-7B	0.2370	0.2491
	Qwen2.5-14B	0.3133	0.3298
	Qwen2.5-32B	0.3347	0.3470
SphericalSteer (You et al., 2026)	Qwen3.5-0.8B	0.2186	0.2296
	Qwen3.5-2B	0.3109	0.3306
	Qwen3.5-4B	0.3760	0.3899
	Qwen3.5-9B	0.3962	0.4061
	Qwen2.5-7B	0.2332	0.2476
	Qwen2.5-14B	0.2773	0.2971
	Qwen2.5-32B	0.3320	0.3384

Table 8: Effect of model scale on theory-alignment metrics for CAA and SphericalSteer across Qwen3.5 and Qwen2.5. All correlations are statistically significant ($p < 3 \times 10^{-3}$).

Across both the disjoint and paraphrased conditions, the main paradigm-level separation is preserved: distribution-driven methods retain substantial theory alignment, whereas behavior-centric methods remain close to zero. In particular, CAA and SAS remain statistically significant across both backbones and all three data settings. These results indicate that the observed separation is not specific to the original sample selection or surface realization of the contrastive examples.

F.2 Multi-Value Annotation Sensitivity

Real-world arguments may express several human values simultaneously, and the source annotations can therefore assign multiple values to the same example. Consequently, we interpret the extracted direction for a target value as a *target-conditioned aggregate direction*, rather than as a perfectly monosemantic representation of that value.

To examine whether multi-value annotations drive the measured geometry, we construct an ad-

Table 9: Value distribution per Schwartz category across both dataset sources. Counts reflect quadruples retained after quality filtering.

Value category	ValueBench	Touché	Total
Achievement & Power			
Achievement	125	2,187	2,312
Power: dominance	37	792	829
Power: resources	10	779	789
Benevolence & Universalism			
Benevolence: caring	96	2,112	2,208
Benevolence: dependability	19	1,149	1,168
Universalism: concern	64	2,863	2,927
Universalism: nature	30	633	663
Universalism: objectivity	30	1,502	1,532
Universalism: tolerance	34	916	950
Security & Conformity			
Security: personal	68	2,950	3,018
Security: societal	10	2,277	2,287
Conformity: interpersonal	37	299	336
Conformity: rules	57	1,692	1,749
Self-direction, Stimulation & Hedonism			
Self-direction: action	62	1,793	1,855
Self-direction: thought	90	1,087	1,177
Stimulation	28	418	446
Hedonism	11	285	296
Tradition, Humility & Face			
Tradition	52	706	758
Humility	31	550	581
Face	20	527	547
Total	**911**	**25,517**	**26,428**

ditional subset that prioritizes examples annotated with exactly one value in the source data. We select 200 examples per Schwartz value, using single-label examples whenever available. Because some values do not contain 200 such examples, shortages are filled with different multi-label examples that were not used in the original subset. The resulting dataset is approximately 50% single-label overall. Since the availability of single-label examples differs across values and this procedure also changes the sample distribution, we treat this experiment as a conservative sensitivity analysis rather than a pure single-label ablation.

The paradigm-level separation also remains under the single-label-prioritized sensitivity analysis on both backbones (Table 10). On Qwen3.5-9B-Base, CAA, SAS, SphericalSteer, and ODESteer obtain ρ_T values of 0.325, 0.309, 0.355, and 0.256, respectively, compared with 0.122 for the matched raw-activation baseline and 0.010, 0.074, and 0.055 for BiPO, OPT, and COLD-Steer. On Llama3.1-8B, the corresponding distribution-driven methods obtain ρ_T values of 0.430, 0.337, 0.442, and 0.262, compared with 0.126 for the raw-activation baseline and 0.054, 0.033, and 0.062 for BiPO, OPT, and COLD-Steer.

Absolute correlations are lower than in the main experiment for several methods, but the matched raw-activation baseline also decreases under this sampling procedure. Thus, the reduction is not specific to the steering methods and is consistent with the change in sample composition. Most importantly, the separation between distribution-driven and behavior-centric methods and the associated statistical significance are preserved on both backbones. This sensitivity analysis therefore supports the robustness of our paradigm-level conclusion to multi-value annotation, while not implying that the extracted directions are perfectly monosemantic.

G Moral Foundations Theory Evaluation

We chose Schwartz's theory as the primary framework because its cross-culturally validated circumplex specifies explicit pairwise compatibility and opposition in continuous and discrete form, giving a human-derived geometric ground truth for detailed analysis. Other frameworks do not offer this: Moral Foundations Theory (MFT), as the most prominent alternative, defines six discrete foundations but doesn't specify the relation between them;

Table 10: Robustness of Theory Rank Correlation (ρ_{T}) to dataset construction. *Original* denotes the main experimental setting; *Disjoint* uses a fresh non-overlapping subset of 200 examples per value; *Paraphrased* paraphrases the question and both contrastive answers; and *SL-prio.* uses the single-label-prioritized sensitivity subset. Subscripts denote p-values. The paradigm-level separation is preserved across all robustness settings. Bold indicates the highest correlation within each setting.

	Qwen3.5-9B-Base				Llama 3.1-8B			
Method	**Original**	**Disjoint**	**Paraphrased**	**SL-prio.**	**Original**	**Disjoint**	**Paraphrased**	**SL-prio.**
LLM Raw Activation Space	$0.2228_{(2.0e-3)}$	$0.2122_{(3.3e-3)}$	$0.2362_{(1.0e-3)}$	$0.1223_{(9.3e-2)}$	$0.2387_{(9.1e-4)}$	$0.2400_{(8.5e-4)}$	$0.2305_{(1.4e-3)}$	$0.1260_{(8.3e-2)}$
Behavior-centric								
OPT (Dunefsky and Cohan, 2025)	$0.1138_{(1.2e-1)}$	$0.0824_{(2.6e-1)}$	$-0.0416_{(5.7e-1)}$	$0.0740_{(3.1e-1)}$	$-0.0486_{(5.1e-1)}$	$0.0961_{(1.9e-1)}$	$-0.0294_{(6.9e-1)}$	$0.0335_{(6.5e-1)}$
COLD-Steer (FD) (Sharma and Trivedi, 2026)	$0.0309_{(9.1e-1)}$	$-0.0877_{(2.3e-1)}$	$-0.0622_{(3.9e-1)}$	$0.0553_{(4.5e-1)}$	$-0.0405_{(5.8e-1)}$	$0.0820_{(2.6e-1)}$	$-0.0122_{(8.7e-1)}$	$0.0615_{(4.0e-1)}$
BiPO (Cao et al., 2024)	$0.1188_{(1.0e-1)}$	$0.0491_{(5.0e-1)}$	$-0.0666_{(3.6e-1)}$	$0.0099_{(8.9e-1)}$	$0.1094_{(1.3e-1)}$	$0.0532_{(4.7e-1)}$	$0.0279_{(7.0e-1)}$	$0.0539_{(4.6e-1)}$
Distribution-driven								
ODESteer (Zhao et al., 2026)	$0.2741_{(1.3e-4)}$	$0.3029_{(2.2e-5)}$	$0.4325_{(4.6e-10)}$	$0.2560_{(3.6e-4)}$	$0.3124_{(1.1e-5)}$	$0.2094_{(3.7e-3)}$	$0.2881_{(5.5e-5)}$	$0.2624_{(2.6e-4)}$
SphericalSteer (You et al., 2026)	$0.3962_{(1.5e-8)}$	$0.4112_{(3.8e-9)}$	$0.4265_{(8.5e-10)}$	$\mathbf{0.3554}_{(4.9e-7)}$	$0.4949_{(3.9e-13)}$	$0.5264_{(6.2e-15)}$	$\mathbf{0.4955}_{(3.6e-13)}$	$\mathbf{0.4416}_{(1.8e-10)}$
CAA (Rimsky et al., 2024)	$0.4606_{(2.3e-11)}$	$0.4365_{(3.1e-10)}$	$\mathbf{0.4634}_{(1.7e-11)}$	$0.3249_{(4.8e-6)}$	$\mathbf{0.4996}_{(2.2e-13)}$	$\mathbf{0.5277}_{(5.2e-15)}$	$0.4902_{(7.0e-13)}$	$0.4304_{(5.7e-10)}$
SAS (Bayat et al., 2025)	$\mathbf{0.5069}_{(8.5e-14)}$	$\mathbf{0.4514}_{(6.3e-11)}$	$0.4381_{(2.6e-10)}$	$0.3090_{(1.4e-5)}$	$0.4220_{(1.3e-9)}$	$0.4515_{(6.3e-11)}$	$0.3918_{(2.3e-8)}$	$0.3370_{(2.0e-6)}$

its only theory-specified structure is the Individualizing–Binding split.

G.1 Benchmark Construction

We adopt the six-foundation revised Moral Foundations Theory (MFT) of Atari et al. (2023): Care, Equality, Proportionality, Loyalty, Authority, and Purity. We source the initial samples from the Moral Foundations Reddit Corpus (Trager et al., 2026). We retain only high-confidence, single-foundation samples with full agreement on the foundation label among at least three trained annotators. Using the same generation and filtering pipeline as our Schwartz benchmark, we convert the retained Reddit samples into question-contrastive instances. We balance the benchmark at 200 samples per foundation, yielding 1,200 samples: 400 from the Individualizing foundations and 800 from the Binding foundations.

G.2 Family-Structure Metric

Unlike Schwartz's theory, MFT specifies neither a circumplex nor inter-foundation angles or opposing pairs; therefore, no analogue of the Schwartz theoretical similarity matrix is available. We instead evaluate MFT's theory-specified higher-order grouping into Individualizing foundations $\{\text{Care}, \text{Equality}\}$ and Binding foundations $\{\text{Proportionality}, \text{Loyalty}, \text{Authority}, \text{Purity}\}$.

Let $\mathcal{W}$ denote the seven unordered within-family pairs (one Individualizing pair and six Binding pairs) and let $\mathcal{C}$ denote the eight cross-family pairs. For unit-normalized foundation vectors u_i, we define their empirical cosine similarity as $E_{ij} = u_i^\top u_j$. Analogous to Δ_{pol}, MFT family separation is

$$\Delta_{\mathrm{MFT}} = \frac{1}{|\mathcal{W}|} \sum_{(i,j)\in\mathcal{W}} E_{ij} - \frac{1}{|\mathcal{C}|} \sum_{(i,j)\in\mathcal{C}} E_{ij}. \quad (4)$$

Positive Δ_{MFT} indicates that foundations within the same MFT family are more similar than foundations across families. As with Δ_{pol}, we report Δ_{MFT} as an interpretable separation score without an associated p-value.

H Geometry Metrics

Evaluation Metric Overview. Because validating value geometry has no single established metric, we use complementary measures covering both the geometry of the extracted vectors and its behavioral consequences. The metrics are organized along two dimensions: value-vector geometry versus cross-value transfer, and continuous circumplex structure versus discrete hierarchical structure (Summary in Table 11).

Let $\mathbf{u}_1, \ldots, \mathbf{u}_{20}$ be unit-normalized steering vectors for the 20 Schwartz values, indexed in their canonical circumplex order. The **empirical similarity matrix** is $E_{bb'} = \mathbf{u}_b^\top \mathbf{u}_{b'}$, and $E_\triangle$ denotes its 190 upper-triangle entries.

Construction of the Theoretical Similarity Matrix. The theoretical matrix follows directly from the central assumption of Schwartz's theory: values occupy a fixed canonical order on a circumplex, and the compatibility between two values is determined by their angular distance on this circle (Schwartz, 1992, 2012). With 20 values, adjacent positions are separated by $18°$, and $T_{bb'}$ is defined as the cosine of the shortest angular distance between values b and b':

$$T_{bb'} = \cos\big(\min(|b-b'|,\, 20-|b-b'|) \times 18°\big)\,.$$

Metric	Structure tested	What it measures	Defined in
Geometry of value vectors			
Theory Rank Correlation (ρ_T)	Continuous circumplex	Whether compatible values rank above conflicting values in vector similarity	Sec. 3.3.1 App. §H
Theory Linear Correlation (r_T)	Continuous circumplex	Whether empirical cosine similarities scale linearly with theoretical circumplex similarities	Sec. 3.3.1 App. §H
Hierarchical Structure Correlation (ρ_H)	Discrete hierarchy	Whether vector similarities respect lower-order families and higher-order groups	Sec. 3.3.1 App. §H
Polarity Separation Score (Δ_{pol})	Same-family vs. opposing groups	The difference between mean within-family similarity and mean similarity across opposing groups	Sec. 3.3.1 App. §H
Cross-value transfer			
Continuous Transfer Fidelity (TWTM)	Continuous circumplex	Whether transfer is positive for neighboring values and negative for opposing values, accounting for both direction and magnitude	Sec. 3.3.2
Hierarchical Transfer Fidelity (ρ_H^{tr})	Discrete hierarchy	Whether transfer ranks highest within value families and lowest across opposing groups	Sec. 3.3.2
Bin-wise Monotonic Decay (BMD-ρ)	Continuous circumplex	Whether mean transfer decreases monotonically with circumplex distance, independent of magnitude	App. §I

Table 11: Summary of the evaluation metrics and the theoretical structure assessed by each metric. Geometry metrics evaluate relationships among extracted value vectors, while transfer metrics evaluate how steering effects propagate across values.

Consequently, adjacent values, such as *Benevolence: Caring* and *Universalism: Concern*, receive a theoretical similarity close to 1, whereas diametrically opposed values receive -1. This cosine-of-angular-distance construction operationalizes the circumplex structure established in the psychology literature through cross-cultural multidimensional-scaling analyses. For the Hierarchical Structure Correlation and Polarity Separation Score, we additionally use the circumplex's higher-order groups (Openness to Change, Self-Enhancement, Conservation, and Self-Transcendence) shown in Figure 6. Since the Schwartz circumplex is an idealized theoretical model that is itself recovered only approximately from human questionnaire data through smallest-space analysis, perfect correspondence with a cosine-similarity matrix is not expected (Schwartz, 1992).

Theory Rank Correlation (ρ_{T}).

$$\rho_{\text{T}} = \text{Spearman}(E_{\triangle},\ T_{\triangle}).$$

Theory Linear Correlation (r_{T}).

$$r_{\text{T}} = \text{Pearson}(E_{\triangle},\ T_{\triangle}).$$

Both ρ_{T} and r_{T} are reported with p-values from standard two-tailed tests over $n = 190$ observations.

Hierarchical Structure Correlation (ρ_{H}). The theory-based hierarchical distance is:

$$d_{bb'}^{\text{hier}} = \begin{cases} 1 & \text{same lower-order family,} \\ 2 & \text{same higher-order group, different family,} \\ 5 & \text{no clear relation,} \\ 10 & \text{opposing higher-order groups.} \end{cases}$$

Boundary values (Hedonism, Face, Humility) belong to two higher-order groups simultaneously. Then:

$$\rho_{\text{H}} = \text{Spearman}(E_{\triangle},\ -d_{\triangle}^{\text{hier}}).$$

ρ_{H} is reported with a p-value.

Polarity Separation Score (Δ_{pol}). Let $\mathcal{S}^{-}$ be the set of same lower-order family pairs and $\mathcal{S}^{\oplus}$ the set of opposing higher-order group pairs. Then:

$$\Delta_{\text{pol}} = \frac{1}{|\mathcal{S}^{-}|} \sum_{(b,b') \in \mathcal{S}^{-}} E_{bb'} - \frac{1}{|\mathcal{S}^{\oplus}|} \sum_{(b,b') \in \mathcal{S}^{\oplus}} E_{bb'}.$$

No p-value is reported for Δ_{pol}; it serves as an interpretable complement to the correlation metrics.

I Bin-wise Monotonic Decay (BMD-ρ)

We report a complementary measure of cross-value transfer that disregards magnitude and focuses purely on whether transfer decays monotonically with circumplex distance.

Definition. We partition the 380 off-diagonal entries of T_{res} into ten bins by their circumplex step distance $k \in \{1, \ldots, 10\}$, where $k = 1$ denotes adjacent values and $k = 10$ denotes diametrically opposing values. For each bin, we compute the mean residual transfer:

$$\mu_k = \frac{1}{|\mathcal{P}_k|} \sum_{(A,B) \in \mathcal{P}_k} T_{\text{res}}[A, B],$$

where $\mathcal{P}_k$ is the set of off-diagonal pairs at distance k (40 pairs for $k = 1, \ldots, 9$; 20 pairs for $k = 10$). The Schwartz circumplex predicts that μ_k should decrease monotonically with k: steering toward value A should on average lift values nearby on the circumplex ($\mu_1 > 0$) and suppress values on the opposite side ($\mu_{10} < 0$), with a smooth decay in between. To put this prediction into practice, **BMD-ρ** measures the negative Spearman correlation between k and $(\mu_1, \ldots, \mu_{10})$:

$$\text{BMD-}\rho = -\text{Spearman}\big(k,\ \mu_k\big). \tag{5}$$

Higher values reflect the theoretically predicted decay from positive transfer at adjacent pairs ($k = 1$) to negative transfer at opposing pairs ($k = 10$). Unlike TWTM, BMD-ρ is insensitive to absolute magnitude since a method producing tiny but correctly-ordered transfers can score as high as one producing large effects, provided the ordering across bins respects the circumplex.

Results. Figure 10 reports BMD-ρ for all seven methods on both Qwen3.5-9B-Base and Llama3.1-8B. Two findings stand out. First, the paradigm separation from the main TWTM result is reproduced and strengthened here. On Qwen, all four distribution-driven methods score between 0.33 and 0.83, while all three behavior-centric methods score between 0.08 and 0.20, which represents a non-overlapping separation. The lowest distribution-driven method (ODESteer, 0.33) exceeds the highest behavior-centric method (BiPO, 0.20) by a clear margin. OPT (0.08) and Cold-Steer (0.12) approach near zero, meaning their bin-wise transfer ordering does not support a monotonic relationship to circumplex distance. This confirms that the paradigm split observed in TWTM is not an artifact of magnitude differences but reflects a genuine structural difference in how the two families of methods propagate steering effects across the value space.

Second, the Llama results reveal that this structural separation is not model-specific. Distribution-driven methods on Llama score 0.43–0.99, while Behavior-centric methods collapse to negative BMD-ρ, meaning their transfer ordering is actively inverted relative to the circumplex on this model. A method with negative BMD-ρ does not merely fail to respect Schwartz structure; rather, it systematically transfers more to opposing values than to adjacent ones, which is the opposite of what human value theory predicts.

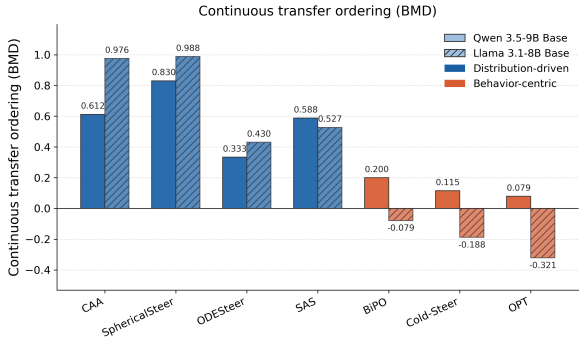

Figure 10: Residualized BMD-ρ across steering methods on Qwen3.5-9B-Base and Llama3.1-8B, grouped by paradigm. Distribution-driven methods (blue) produce transfer patterns whose ordering decays monotonically with circumplex distance, while behavior-centric methods (orange) do not.

J Steering Method Formulations

The descriptions below detail the formulations used to extract and apply the value vector $v_{(b,\ell)}$ for each baseline method.

Intuition Behind the Two Steering Paradigms. The key difference in these two paradigms is what supervises the vector. Distribution-driven methods derive it from where activations of contrastive pairs lie in activation space (e.g., CAA averages their difference), so the vector inherits the model's internal representation of the value. Behavior-centric methods instead optimize the vector against an output objective (e.g., OPT maximizes the log-probability of the aligned answer), so any direction that flips the output is a valid solution, whether or not it corresponds to how the model represents the value. This is why the two classes can achieve similar target accuracy yet show very different geometric properties.

Unified Effective-Shift Representation. The effective shift in Equation 1 is defined as the difference between the steered and unsteered activations, so it measures exactly the intervention each method applies, independent of internal mechanism. We choose this because methods such as ODESteer (a

nonlinear ODE trajectory) and SphericalSteer (a norm-preserving rotation) do not expose a native steering vector, and this was the only representation that treats every method on equal footing. For OPT and BiPO, the intervention is a fixed additive vector, so the effective shift recovers it exactly; for COLD-Steer, following its original formulation, we average the finite-difference displacement over the training prompts. Thus, this representation cannot systematically favor distribution-driven methods: if anything, collapsing an intervention into a single vector may disadvantage the non-additive ODESteer and SphericalSteer methods by discarding information about their trajectories and rotations, yet both retain clear theory-aligned structure. Independently, our cross-value transfer evaluation is behavioral: it measures accuracy changes on held-out values without using the pairwise geometry of the effective-shift vectors. The same paradigm split appears there, which further confirms the findings.

LLM Raw Activation Space. This non-steering baseline measures the geometry already present in the unmodified model. For each Schwartz value, we format the multiple-choice prompt so that the value-aligned option is the target answer, run the backbone without interventions, and extract the last-token hidden activation at the selected layer. The representation for each value is the mean of these positive-answer activations over the corresponding examples. We then compute the same pairwise cosine-similarity geometry metrics used for value vectors, allowing us to compare learned or constructed steering directions against the model's native value representation.

Contrastive Activation Addition (CAA). Following (Rimsky et al., 2024), given a contrastive dataset $D_b = \{(p_i, c_i^+, c_i^-)\}$ for each Schwartz value b, and a given layer ℓ, the value vector is defined as the mean difference in residual stream activations:

$$v_{(b,\ell)} = \frac{1}{|D_b|} \sum_{(p_i, c_i^+, c_i^-) \in D_b} [a_\ell(p_i, c_i^+) - a_\ell(p_i, c_i^-)] \tag{6}$$

At inference time, the vector is unit-normalized and injected into the residual stream at the token position where the LLM generates the answer as $a_\ell \leftarrow a_\ell + \alpha v_{(b,\ell)}$, where α controls steering strength.

Optimization Steering (OPT). As an alternative to contrastive averaging, we directly learn value vectors via gradient descent, following (Dunefsky and Cohan, 2025). Concretely, we employ mixed steering, minimizing the combined loss:

$$\mathcal{L}(v) = \underbrace{-\sum_k \log P_v(c_k^+ \mid p)}_{\text{promotion}} + \underbrace{-\sum_k \log(1 - P_v(c_k^- \mid p))}_{\text{suppression}} \tag{7}$$

where P_v denotes the model's output distribution under steered activations. This formulation simultaneously promotes value-aligned responses and suppresses value-contrary ones. The saved optimized vector is then applied additively at the selected layer, preserving its learned norm: $a_\ell \leftarrow a_\ell + \alpha v_{(b,\ell)}$.

COLD-Steer (Finite-Difference). For the COLD-Steer finite-difference baseline (Sharma and Trivedi, 2026), each value first induces a small parameter-space perturbation from the gradient of the value-aligned training loss. If $\theta' = \theta + \epsilon \nabla_\theta \mathcal{L}$, the native COLD intervention estimates an activation-space finite-difference direction $(z_{\theta'}(x) - z_\theta(x))/\epsilon$ and steers by subtracting a scaled version of this direction. For geometry comparison, our code stores a representative per-value vector by averaging this finite-difference displacement at the last response token over the training prompts; evaluation applies the saved direction additively at the selected layer with the sign and scale matching the native COLD hook.

Bi-directional Preference Optimization (BiPO). BiPO (Cao et al., 2024) learns a bidirectional activation vector from preference pairs. The reference implementation wraps a selected transformer block with a trainable vector addition, freezes the base model, and optimizes only this vector with a DPO-style chosen/rejected objective. During training, the vector multiplier is sampled from $\{+1, -1\}$, and the loss sign is flipped for negative multipliers, encouraging a single direction whose positive and negative scales steer toward opposite preferences. In our comparison adapter, the resulting saved vector is applied with the same additive layer intervention, $a_\ell \leftarrow a_\ell + \alpha v_{(b,\ell)}$, without unit-normalizing the learned magnitude.

ODESteer. ODESteer (Zhao et al., 2026) treats steering as motion through activation space rather than a fixed additive offset. For each value, it fits a polynomial sketch classifier separating value-aligned and value-contrary last-token activations. The normalized gradient of this classifier defines a vector field $f(x)$ pointing toward the aligned re-

gion, and steering transforms an activation by integrating the ordinary differential equation $dx/dt = f(x)$ for time $T = \alpha$ using a fixed-step solver. In our exact evaluation mode, one ODESteer model is fitted per Schwartz value and hooked at the selected layer; the steered hidden state is the nonlinear ODE-transformed activation rather than $x + \alpha v$.

SphericalSteer. SphericalSteer (You et al., 2026) uses the CAA mean-difference direction as a unit prototype on the activation sphere. For each value, μ_T is the normalized positive-minus-negative prototype and $-\mu_T$ is treated as the antipodal contrary prototype. At inference time, the method compares the activation's von Mises-Fisher-style affinity to the two prototypes; when the contrary pole is sufficiently favored, it rotates the activation along the sphere toward μ_T by a strength controlled by α, while preserving the original activation norm. Thus, SphericalSteer modifies angular position rather than adding a Euclidean offset.

Sparse Activation Steering (SAS). We adapt the SAS framework (Bayat et al., 2025) to operate within the Qwen-Scope TopK SAE feature space. The SAE is optionally fine-tuned on value-task residual activations with reconstruction loss, then each value's value vector is computed from positive-minus-negative last-token SAE features. Our implementation uses pre-TopK features by default, applies a frequency threshold to reduce noisy features, and explicitly removes features shared by both polarities. At inference time, the hook encodes the residual stream, adds α times the value vector in the sparse SAE space, applies TopK sparsification, decodes back to the residual stream, and crucially adds back the SAE reconstruction residual error ($\Delta = a - \hat{a}$) so that non-reconstructed dense information is preserved.

K Geometric Fidelity vs. Accuracy Gain as Predictors of Cross-Value Transfer

To assess which property of a steering vector better predicts its cross-value behavior, we pool all seven steering methods across both backbones (Qwen3.5-9B-Base and Llama3.1-8B), yielding $n = 14$ (method, backbone) observations. For each observation, we record three quantities: geometric fidelity (ρ_T), target-value accuracy gain, and cross-value transfer under both TWTM and ρ_H^{tr}. We then compute Spearman correlations between each candidate predictor and each transfer metric (Table 12). Geometric fidelity is the stronger predictor on both transfer metrics, and the gap is largest on ρ_H^{tr}, which unlike TWTM, rewards only the *ordering* of transfer across the Schwartz hierarchy and not steering magnitude.

Predictor	**TWTM**	$\boldsymbol{\rho_H^{\text{tr}}}$
Geometric fidelity (ρ_T)	**0.84*****	**0.77****
Target-value accuracy gain	0.78***	0.59*

Table 12: Spearman correlations between two candidate predictors and cross-value transfer, pooled across 7 steering methods $\times$ 2 backbones ($n = 14$). Geometric fidelity is the stronger predictor on both transfer metrics. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

L AI Assistant Acknowledgment

During the preparation of this work, the authors utilized artificial intelligence tools to improve the grammatical clarity and flow of the manuscript. Additionally, AI-based coding assistants were used to support the implementation of the experimental pipeline and data processing scripts.

M Computational Resources

All experiments were executed using a single NVIDIA B300 GPU.