Tracing Moral Foundations in Large Language Models

Chenxiao Yu[1]*, **Bowen Yi**[1]*, **Farzan Karimi-Malekabadi**[2,3], **Suhaib Abdurahman**[2,3]

Jinyi Ye[1], **Shrikanth Narayanan**[1,2,3], **Yue Zhao**[1], **Morteza Dehghani**[1,2,3]

[1]Department of Computer Science, University of Southern California
[2]Department of Psychology, University of Southern California
[3]Center for Computational Language Sciences, University of Southern California
{cyu96374,bowenyi,karimima,sabdurah,jinyiy,shri,yue.z,mdehghan}@usc.edu

Abstract

Large language models (LLMs) often produce human-like moral judgments, but it is unclear whether this reflects an internal conceptual structure or superficial "moral mimicry." Using Moral Foundations Theory (MFT) as an analytic framework, we study how moral foundations are encoded, organized, and expressed across 14 base and instruction-tuned LLMs spanning four model families (Llama, Qwen2.5, Qwen3-MoE, Mistral) and scales from 7B to 70B. We employ a multi-level approach combining (i) layer-wise analysis of MFT concept representations and their alignment with human moral perceptions, (ii) pretrained sparse autoencoders (SAEs) over the residual stream to identify sparse features that support moral concepts, and (iii) causal steering interventions using dense MFT vectors and sparse SAE features. We find that models represent and distinguish moral foundations in a manner that aligns with human judgments, and that this moral geometry naturally emerges from pretraining and is selectively rewired by post-training. At a finer scale, SAE features show clear semantic links to specific foundations, suggesting partially disentangled mechanisms within shared representations. Finally, steering along either dense vectors or sparse features produces predictable shifts in foundation-relevant behavior, demonstrating a causal connection between internal representations and moral outputs. Together, our results provide mechanistic evidence that moral concepts in LLMs are distributed, layered, and partly disentangled, suggesting that pluralistic moral structure can emerge as a latent pattern from the statistical regularities of language alone. [2]

1 Introduction

As large language models are increasingly integrated into socially and ethically sensitive domains, understanding their underlying "moral compass" has become a critical priority. To date, most evaluations of model morality have focused on surface-level outputs—analyzing the final text a model generates in response to questionnaires or scenarios [e.g., Abdulhai et al., 2024, Schramowski et al., 2022]. However, this approach treats the model as a black box, failing to distinguish between a model that has genuinely internalized a structured moral framework and one that is merely performing "moral mimicry": superficially matching linguistic patterns found in training data without a stable conceptual organization [Perez et al., 2023].

We thus investigate whether moral concepts are organized within the model's internal representations as distinct, functional units. We adopt Moral Foundations Theory (MFT) as an analytic framework to

*Equal contribution.
[2]Our code and data are available at: https://github.com/AiHeMaotai/MFT_LLMs

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

probe this structure [Graham et al., 2013, Atari et al., 2023]. MFT is particularly well-suited for this task because its dimensions—*Care, Fairness, Loyalty, Authority, and Sanctity*—are systematically expressed in human language [Kennedy et al., 2021, Atari and Dehghani, 2022], associated with distinct patterns of neural activity [Hopp et al., 2023], and predictive of various high-stake real-world behaviors [e.g., Reimer et al., 2022, Hoover et al., 2021]. We hypothesize that if LLMs are capturing the latent structure of human morality, these foundations should correspond to identifiable geometric structures within their representation space.

LLMs are trained on large corpora of human-generated language, which serves as a primary medium for cultural and moral transmission. As a result, they provide a unique testbed for a deeper question: whether structured moral representations can emerge from exposure to language alone, without explicitly grounding in perception, embodiment, or direct social interaction. Recent theoretical work further argues that language itself plays a functional role in initiating, maintaining, revising, and coordinating moral norms in human societies [Li and Tomasello, 2021]. From this perspective, studying moral representations inside LLMs is a way to examine how moral structure may arise through linguistic processes. Our analysis of these internal mechanics also offers a unique computational perspective on the nature of moral representation. For instance, our findings provide a way to evaluate the tension between moral pluralism [i.e., morality as distinct foundations; Graham et al., 2013] and harm-based accounts [i.e., morality as a single dimension of harm; Schein and Gray, 2018]. By measuring the degree of separability between these concepts in representation space, we can observe whether a model trained on human discourse naturally "discretizes" morality into irreducible dimensions or collapses them into a unified axis.

In this paper, we examine how LLMs encode, organize, and apply moral foundation concepts. Using MFT as an analytic framework, we analyze how moral information is geometrically structured across layers, whether distinct moral foundations correspond to separable directions in representation space, and how these directions relate to interpretable internal features. We further test the causal role of these representations by intervening on them during inference and measuring the resulting changes in moral judgments.

Our work makes four main contributions. First, we demonstrate a robust representational **alignment** between LLM latent spaces and human moral perceptions. By projecting model activations onto moral axes derived from vignettes, we show that LLMs naturally recover during pretraining the topological separation found in human-labeled natural language, indicating that these models encode moral concepts in a way that is isomorphic to human judgment. Second, we provide mechanistic evidence that these moral foundations correspond to highly separable linear axes within LLMs. This geometric separability provides computational support for pluralist theories of morality over single-dimension harm-based accounts. Third, by combining concept vectors with sparse autoencoders (SAEs), we decompose abstract foundations into granular, interpretable features. Fourth, we establish the causal relevance of these structures through steering interventions, showing that manipulating the identified directions can reliably shift the model's moral outputs. Taken together, our findings suggest that LLMs are not merely stochastic parrots mimicking moral language, but systems with structured moral geometries. This offers a new computational perspective for psychology: using aligned models as computational testbeds to investigate the structural organization of human moral cognition.

2 Related Work

2.1 Human Moral Cognition

Moral psychology has long debated whether moral judgment reflects a single underlying principle or multiple partially distinct cognitive dimensions [Graham et al., 2011, Haidt, 2007]. Monistic accounts such as the Theory of Dyadic Morality argue that diverse moral judgments largely reduce to perceived interpersonal harm [Schein and Gray, 2018, Gray et al., 2014]. In contrast, MFT proposes a pluralistic structure organized around recurring dimensions—*Care*, *Fairness*[3], *Loyalty*, *Authority*, and *Sanctity*—each linked to different evolutionary problems [Graham et al., 2013]. Empirical work supports a graded view: multiple dimensions are often recovered, but they are correlated rather than independent [Atari et al., 2023]. Neuroimaging findings align with this graded view: moral concerns show overlapping activation, yet multivariate pattern analysis still distinguishes them as separable patterns within shared substrates [Khoudary et al., 2022, Wilkinson et al., 2024].

[3]Recent revisions bifurcate Fairness into Equality and Proportionality [e.g., Atari et al., 2023]

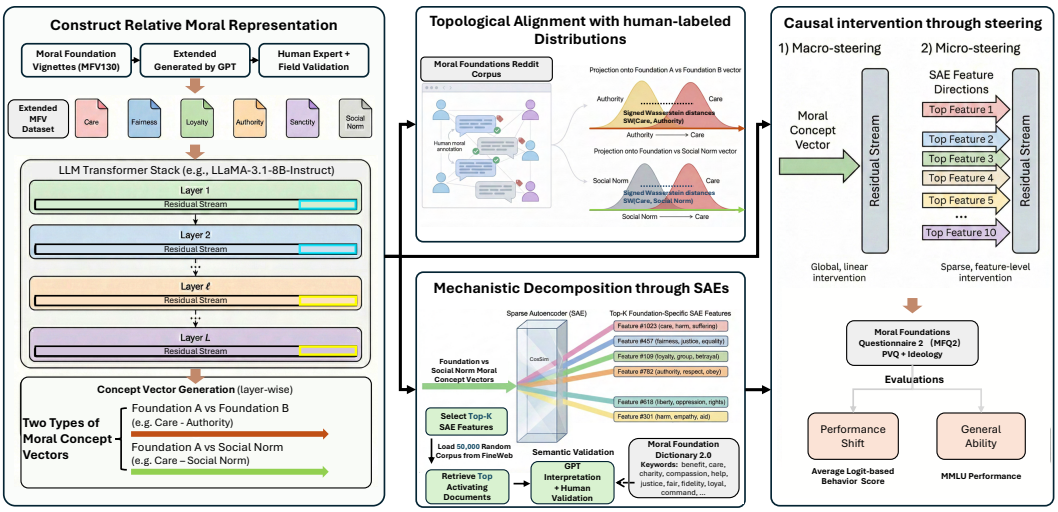


Figure 1: **Overview of the experimental pipeline.** (i) Relative moral concept vectors are constructed from extended Moral Foundations vignettes and serve as a central representational hub. These vectors are validated in parallel through (ii) topological alignment with human-labeled Reddit post distributions and (iii) mechanistic decomposition into sparse autoencoder features. (iv) We then causally intervene on model activations via macro- and micro-steering and assess behavioral shifts and capability preservation.

2.2 Moral Reasoning in Language Models

Work in NLP and computational social science suggests that LLMs exhibit behavioral patterns aligned with moral foundations: they often generate human-like responses in moral scenarios, with larger models matching human moral-political values more closely [Abdulhai et al., 2024], and their moral justifications and advice can be perceived as comparable to, or even better than, those of professional ethicists [Dillion et al., 2025]. Beyond surface behavior, representational analyses show that moral judgments are organized in embedding space in ways consistent with human perceptions of right and wrong [Schramowski et al., 2022], and that moral foundations are linearly decodable from hidden activations, particularly in mid-layers [Karami et al., 2025], suggesting that large-scale statistical learning can yield moral abstractions without explicit moral supervision. At the same time, these representations can be brittle and biased: models show "moral hypocrisy" (gaps between abstract principles and concrete judgments) [Nunes et al., 2024], limited variance and weak nomological structure [Abdurahman et al., 2024], and "moral mimicry" that adapts justifications to a prompter's perceived identity rather than a stable internal framework [Simmons, 2022]. Moreover, training on English-centric corpora can skew models toward WEIRD values [Henrich et al., 2010, Atari et al., 2023, Zewail et al., 2026], a bias that may persist even in multilingual settings and complicates their use as universal moral agents [Aksoy, 2025, Trager et al., 2025a].

2.3 Interpretability for LLMs

Top-down approaches probe or steer models using predefined concepts. Representation engineering and activation steering use concept-aligned directions in activation space to control model behaviors at inference time [Elhage et al., 2021, Zou et al., 2023, Turner et al., 2023]. These targeted interventions along interpretable axes have been shown to reliably steer model generation [Banayeeanzade et al., 2025, Li et al., 2023]. In this paper, we use linear steering similar to persona vectors [Chen et al., 2025]. Instead of benchmarking control methods, we use steering mainly to validate the identified geometric structure. Top-down methods enable targeted interventions but offer limited insight into how concepts are internally encoded or entangled. In contrast, **bottom-up approaches** recover interpretable structure directly from a model's internal activations without imposing strong semantic priors. Neurons are typically *polysemantic*—responding to multiple unrelated concepts—which hinders interpretability [Elhage et al., 2022, Karvonen, 2024, Ngo et al., 2024]. SAEs address this challenge by learning a sparse basis over model activations [Cunningham et al., 2023, Cammarata

et al., 2021, Wang et al., 2023]. An SAE map an activation vector into a sparse set of latent feature activations and linearly reconstructs the original activation from those features. This sparse bottleneck encourages latent units to specialize into more interpretable features. SAE features have been shown to align with meaningful semantic and value-laden concepts, including personalities and social identities [Chen et al., 2025, Girrbach et al., 2025], providing a complementary bottom-up view of how high-level abstractions emerge within the model's internal representations.

3 Methods

To investigate how moral foundations are encoded and causally used by LLMs, we propose a multi-level mechanistic framework that links macroscopic representational geometry to microscopic feature structure (Figure 1). We first construct foundation-specific concept vectors in the residual stream from contrastive vignettes (Section 3.1). We then evaluate whether these directions are robust by testing geometric separability on human-labeled, naturalistic text from the Reddit Moral Foundations Corpus (Section 3.2). Next, we use Sparse Autoencoders to decompose these dense directions into interpretable, atomic features (Section 3.3). Finally, we establish causal relevance via inference-time steering at both the macro (vector) and micro (SAE feature) levels and measure resulting shifts in downstream moral behavior (Section 3.4).

3.1 Constructing Relative Moral Representations

Theoretical grounding and latent space. To investigate whether moral foundations emerge as distinct geometric structures within LLMs, we extract layer-wise directions [Zou et al., 2023, Chen et al., 2025] from the model's residual stream. This top-down approach allows us to project high-level psychological constructs into the model's latent activation space [Elhage et al., 2021, Ameisen et al., 2025]. Let the model have L layers and residual dimension n. For an input sequence $x_{1:T}$, the residual activation at layer ℓ and token t is $\mathbf{h}_{\ell,t} \in \mathbb{R}^n$. We register forward hooks at all layers and run a standard forward pass. Following previous work [Arditi et al., 2024, Stolfo et al., 2025], we represent its internal state using the residual activation at the last token for each input i,

$$\tilde{\mathbf{h}}_{\ell}^{(i)} = \mathbf{h}_{\ell,T^{(i)}}, \tag{1}$$

where $T^{(i)}$ is the last-token index. To reduce stochasticity, we compute this activation ten times per input and average the results.

Constructing foundation vectors. We use foundation-labeled moral scenarios derived from MFV-130 [Clifford et al., 2015] and our expansions (see A.2) to generate concept vectors. For each foundation, we estimate a layerwise concept direction using a difference-in-means contrast. Let $\mathcal{A}$ denote inputs from a target foundation (e.g., *Care*) and $\mathcal{B}$ denote contrast inputs from the remaining foundations or Social Norms. At layer ℓ, we define the raw direction as

$$\mathbf{v}_{\ell}^{(\text{raw})} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \tilde{\mathbf{h}}_{\ell}^{(i)} \; - \; \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \tilde{\mathbf{h}}_{\ell}^{(j)}, \tag{2}$$

and its ℓ_2-normalized version as

$$\mathbf{v}_{\ell} = \frac{\mathbf{v}_{\ell}^{(\text{raw})}}{\|\mathbf{v}_{\ell}^{(\text{raw})}\|_2}. \tag{3}$$

We compute one vector per layer for each of the five foundations, yielding layerwise directions for *Care*, *Fairness*, *Loyalty*, *Authority*, and *Sanctity*, together with an additional *Social Norm* (as control)[4] direction. These vectors quantify how the model internally separates each target concept from the others in residual-stream activation space.

3.2 Topological Alignment with Human-Labeled Distributions

Ecological grounding. To test whether our activation vectors align with how people express and perceive morality in daily natural language, we evaluate them on human-labeled text from the Moral

[4]We include Social Norms as a control condition based on Social Domain Theory [Turiel, 1983], which distinguishes the moral domain from the conventional domain [Smetana, 2006].

Foundations Reddit Corpus (Trager et al., 2025b; see A.2 for details). In this evaluation, the Reddit label is held out: we feed only the raw post text to the model and use the label solely for grouping and analysis.

Projection. For each labeled Reddit post r, we treat the post text as a new model input and record the residual-stream activation $\tilde{\mathbf{h}}_\ell^{(r)}$ at layer ℓ using the same last-token representation as in Equation 1. Given a normalized concept vector $\mathbf{v}_\ell$, we quantify alignment by the scalar projection of this activation onto the vector:

$$s_\ell^{(r)} = \mathbf{v}_\ell^\top \tilde{\mathbf{h}}_\ell^{(r)}, \tag{4}$$

which is the signed component of $\tilde{\mathbf{h}}_\ell^{(r)}$ along $\mathbf{v}_\ell$ since $\|\mathbf{v}_\ell\|_2 = 1$. We interpret $s_\ell^{(r)}$ as an alignment score: larger (more positive) values indicate stronger alignment between the comment's internal representation and the positive direction of the corresponding moral vector at layer ℓ.

Quantifying separability and cross-foundation structure. For each foundation k, we compare the projection-score distributions of posts labeled with k versus those not labeled with k, and quantify their separation using the Signed Wasserstein Distance (SW). Larger SW_1 indicates stronger foundation-specific separability at layer ℓ (see A.3), which we use as a validity check for the corresponding vector $\mathbf{v}_\ell^{(k)}$.

We then extend the analysis from *single foundations* to *relations between foundations*. Beyond defining five dimensions, MFT predicts a structured organization: *Care* and *Fairness* form an *Individualizing* cluster, whereas *Loyalty*, *Authority*, and *Sanctity* form a *Binding* cluster, suggesting smaller within cluster differences (e.g., *Care–Fairness*) and larger cross-cluster differences (e.g., *Care–Authority*) [Graham et al., 2013]. To test whether the model reproduces this pattern, we construct a *Pairwise Wasserstein Matrix*: for each labeled subset under foundation k and each concept m, we project posts onto the corresponding concept vector $\mathbf{v}_\ell^{(m)}$ and compute SW_1 between the score distributions of posts labeled with k and those not labeled with k.

3.3 Mechanistic Decomposition via SAEs

Section 3.1 provides macro-level moral directions in residual stream space. However, a single direction in a high-dimensional activation space can be polysemantic. To identify atomic mechanisms of moral representation, we decompose dense model activations using pretrained SAEs (architecture and training details in Section A.1). Formally, let an activation vector be $h \in \mathbb{R}^{d_{\text{model}}}$. An SAE maps h into an overcomplete feature space via a decoder matrix $W_{\text{dec}} = [\mathbf{d}_1, \ldots, \mathbf{d}_{\text{SAE}}]$:

$$h \approx \sum_{i=1}^{d_{\text{SAE}}} f_i(h)\, \mathbf{d}_i + \mathbf{b}_{\text{dec}}, \tag{5}$$

where $f_i(h)$ is the activation strength of feature i, and $\mathbf{d}_i$ is the decoder direction.

Feature attribution via projection onto moral directions. While Section 3.1 constructs various pairwise contrasts (e.g., *Care* vs. *Loyalty*) to map the global geometry, we specifically utilize the *Foundation vs. Social Norms* concept vectors for mechanistic decomposition. To minimize interference between active moral foundations, we isolate moral features by comparing them to the *Social Norm* comparator (matched-format conventional violations) instead of contrasting foundations against one another.

For a target foundation k at layer ℓ, let $\hat{\mathbf{v}}_{k,\ell}$ denote the normalized difference vector between foundation k and the *Social Norms* baseline. We quantify feature relevance by cosine similarity:

$$r_{i,k}^{(\ell)} = \text{CosSim}(\mathbf{d}_i, \hat{v}_{k,\ell}) = \frac{\mathbf{d}_i \cdot \hat{v}_{k,\ell}}{\|\mathbf{d}_i\|_2 \, \|\hat{v}_{k,\ell}\|_2}. \tag{6}$$

We then select the Top-K features with the largest $r_{i,k}^{(\ell)}$ to form a *feature fingerprint* for concept k: $\mathcal{F}_{k,\ell} = \{i \mid r_{i,k}^{(\ell)} \text{ is top-}K\}$. Intuitively, $\mathcal{F}_{k,\ell}$ identifies which sparse SAE features most align with the macro-level moral direction.

Semantic validation. Geometric alignment between SAE decoder directions and moral concept vectors does not guarantee that an SAE feature corresponds to an interpretable moral concept. We therefore ground feature semantics using top-activating natural-language contexts. For each candidate feature, we randomly sample 50,000 documents from FineWeb [Penedo et al., 2024], a large-scale general-domain web corpus, and compute feature activations over tokens within each document.

We rank documents by the feature's maximum token activation and retrieve the top-K activating documents. For each retrieved document, we extract a localized evidence span by taking a fixed window of ± 64 tokens around the maximally activating token, producing a set of peak-centered context windows per feature. Following previous work [Paulo et al., 2025], we prompt `GPT-5.1` to obtain concise semantic descriptions of SAE features. After human validation, we use LLM-generated descriptions as a readable summary of the feature's typical activation contexts, while treating causal steering results as the primary evidence of moral relevance. Our full experiment details are in Appendix C and E.

3.4 Causal intervention through steering

Linear intervention at inference time. To test whether our moral directions play a causal role in model behavior, we perform inference-time activation steering by injecting a control vector into the residual stream. Let $\mathcal{L}_{\text{steer}}$ denote the set of candidate layers, from which the best layer L^* is selected by the dose–response criterion in Appendix D. For a target layer $\ell \in \mathcal{L}_{\text{steer}}$ and the final token t, let $\mathbf{h}_{\ell,t}$ denote the residual activation (see Section 3.1). We apply a linear intervention:

$$\mathbf{h}'_{\ell,t} = \mathbf{h}_{\ell,t} + \alpha_\ell \, \mathbf{v}_\ell^{\text{steer}}, \tag{7}$$

where α is a steering coefficient that controls intervention strength and sign (positive vs. negative steering). We compare two levels of intervention that differ in how $\mathbf{v}_\ell^{\text{steer}}$ is defined.

Macro-steering. We set $\mathbf{v}_\ell^{\text{steer}} = \mathbf{v}_\ell$, the (debiased) foundation vector from Section 3.1. This tests whether the macro-level moral direction is sufficient to shift behavior in a targeted way.

Micro-steering. We set $\mathbf{v}_\ell^{\text{steer}}$ to the unit-normalized sum of the decoder directions of the Top-K SAE features selected. This tests whether specific sparse mechanisms can drive the same behavioral change. Our experimental details are in Appendix D.

4 Experiments

4.1 Experimental Setup

Models. We evaluate 14 LLMs spanning four model families, four parameter scales, and both base and instruction-tuned variants: `Llama-3.1` (8B,70B) [Grattafiori et al., 2024], `Qwen2.5` (7B, 14B, 32B) [Yang et al., 2024], `Qwen3-30B-A3B` [Yang et al., 2025] (MoE), and `Mistral-7B v0.3` [Jiang et al., 2023], each in Base and Instruct versions. This selection factors variation along three axes: *scale* (7B–70B), *alignment training*(Base vs. Instruct), and *developer origin* (Western-origin `Llama`/`Mistral` vs. Chinese-origin `Qwen2.5`/`Qwen3`); the latter is leveraged in Appendix D.4 to interpret cross-cultural patterns in cross-framework transfer [Atari et al., 2023]. All inference is run with HuggingFace Transformers in bfloat16 at temperature $T = 0.01$.

Sparse autoencoders. For mechanistic decomposition and micro-steering (Section 3.3), we use pretrained residual-stream SAEs via the SAELens library [Bloom et al., 2024]: Llama Scope [He et al., 2024], Qwen-Scope [Qwen Team, 2026], and community-trained SAEs for the Qwen-2.5 and Llama-3.1 architectures[Marks et al., 2024][5]. Publicly available SAEs cover only a subset of our model suite, so we restrict SAE-based analyses to the models for which trained SAEs are released, while projection-based analyses and macro-steering are run on all 14 models. Details of SAE-equipped models, sampled layers, expansion factors, and sparsity values are listed in Appendix A.

Datasets. We construct moral concept vectors from an expanded MFV-130 vignette set [Clifford et al., 2015], augmenting each moral foundation and a *Social Norm* category to $\sim$200 scenarios via

[5]Available at `https://huggingface.co/andyrdt/saes-qwen2.5-7b-instruct` and `https://huggingface.co/andyrdt/saes-llama-3.1-8b-instruct`

gpt-5-mini with **human validation** (evaluation detailed in Appendix E). We validate on naturalistic moral language using the Reddit Moral Foundations Corpus [Trager et al., 2025b], retaining only high-confidence, single-label posts (held out from vector construction and used only for grouping in projection analyses). For semantic anchoring of SAE features, we use the Moral Foundations Dictionary 2.0 (MFD2) [Frimer et al., 2019]. For behavioral evaluation of steering, we use MFQ-2 [Atari et al., 2023] as an in-domain readout, and the Schwartz Portrait Values Questionnaire (PVQ-21) [Schwartz et al., 2003] together with a single political-ideology item as cross framework readouts. Dataset construction and filtering details are in Appendix A.2.

Evaluation metrics. We evaluate representation geometry with the standardized Wasserstein distance $\widetilde{W}_1 = W_1/\sigma_{\text{pooled}}$ (Eq. 12), which measures layer-wise separability of Reddit projection-score distributions while normalizing cross-model scale differences; robustness to alternative metrics (Cohen's d and AUC) is reported in Appendix A.3. We also report a direction-reversal rate, the fraction of layers where the median foundation projection falls below the *Social Norm* baseline (Eq. 15). For causal steering, we use the relative-perturbation slope k^{ρ}_{f,L^*} (Eq. 18), estimated from readout-score changes under the unit-free steering ratio ρ. The same parameterization is used for MFQ-2, Schwartz PVQ-21, political ideology, and MMLU, with all readouts scored as expected Likert ratings from logits. See Appendix A.3 for details.

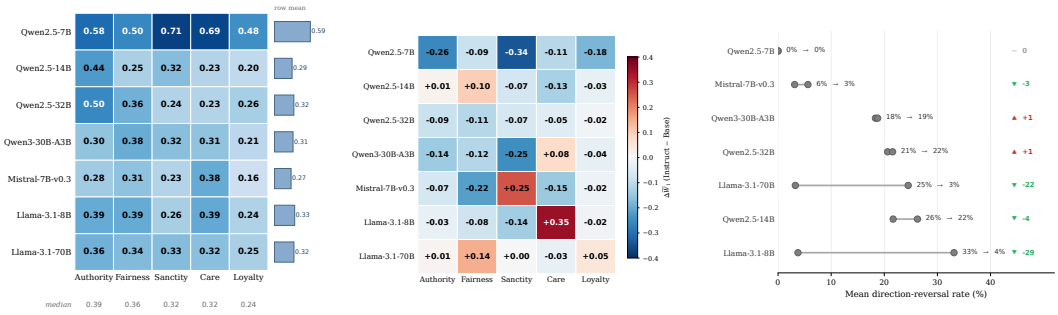
4.2 Main Results

4.2.1 Moral Geometry Naturally Emerges from Pretraining

We first ask whether LLMs contain moral geometry that aligns with human-labeled moral language, and whether this geometry is acquired during pretraining or introduced by post-training. We address this by comparing base and instruction-tuned variants across 14 models from seven settings: `Qwen2.5` (7B/14B/32B), `Qwen3-30B-MoE`, `Llama-3.1` (8B/70B), and `Mistral-7B`.

Universal pretraining-stage encoding. For each (model, foundation) pair, we identify the layer at which the projected representation maximally separates moral content from the *Social Norm* baseline along the corresponding concept vector, and report the standardized Wasserstein distance $\tilde{W}_1 = W_1/\sigma_{\text{pooled}}$ at that layer (Figure 2a). Across the 7 base models, all 35 such (model, foundation) cells exhibit significant linear separability ($\tilde{W}_1 \in [0.16, 0.71]$, AUC > 0.55, Cohen's $d > 0.15$). The strength of pretraining-stage encoding varies systematically across both axes: `Qwen2.5-7B` achieves the highest overall encoding (mean $\tilde{W}_1 = 0.59$ across foundations) and the single strongest foundation effect (*Sanctity*, $\tilde{W}_1 = 0.71$), while across foundations *Authority* is most strongly encoded (median 0.39) and *Loyalty* the weakest (0.24). These results establish that multi-dimensional moral encoding—all five MFT foundations as separable, foundation-specific directions—emerges from next-token prediction over web-scale natural language alone, providing computational evidence for naturally arising moral pluralism in language models.

Post-training rewires but does not introduce moral structure. Comparing matched base–instruct pairs reveals that post-training produces *selective*, family-specific geometric effects rather than uniform reshaping (Figure 2b). At the family level, the direction of alignment effect splits sharply: `Llama-3.1-8B` amplifies *Care* ($\Delta\tilde{W}_1 = +0.35$) and `Mistral-7B` amplifies *Sanctity* ($+0.25$), while `Qwen2.5-7B` shows reduced separability across all five foundations ($\Delta\tilde{W}_1 \in [-0.34, -0.09]$). The remaining families sit between these extremes: `Llama-3.1-70B` remains near-invariant ($|\Delta| \leq 0.14$), and the larger `Qwen2.5` variants together with `Qwen3-30B-MoE` show milder, predominantly negative shifts. Beyond magnitude, post-training also affects the *stability* of the moral axis itself: layer-wise direction-reversal rates either decrease or remain near-constant in 5/7 families ($\Delta \leq 0$; no family exhibits significant degradation; Figure 2c). The `Llama` family stands out with an order-of-magnitude stronger effect—direction-reversal drops from 33% to 4% (`8B`) and from 25% to 3% (`70B`)—revealing a pipeline-specific commitment to globally consistent moral polarity that other alignment recipes only weakly approximate. Together, these findings reframe the standard view of alignment: post-training is a *selective rewiring* of pretraining-acquired geometry whose direction (amplification vs. compression) and intensity (gradual vs. dramatic) vary by training pipeline, not the formation of moral structure de novo. See Appendix B for implementation details, robustness checks, and cross-foundation geometry analyses.



(a) Base $\tilde{W}_1$ across 7 families. (b) $\Delta\tilde{W}_1$ at base-best layer. (c) Direction-reversal rate.

Figure 2: **Moral geometry naturally emerges from pretraining and is selectively rewired by post-training.** Each cell summarizes one (model, foundation) pair via its best-layer projection statistic. **(a)** For each base model and foundation, we report the standardized Wasserstein distance $\tilde{W}_1$ (Appendix A.3) between morally labeled and *Social Norm* projections at the layer where separation is maximal. Right-margin bars show the row mean across the five foundations. Foundations ordered by base-model median $\tilde{W}_1$. **(b)** For each (model, foundation) pair, we fix the layer to the base model's best layer (L_{base}) and compute $\Delta\tilde{W}_1 = \tilde{W}_1^{\text{instruct}}(L_{\text{base}}) - \tilde{W}_1^{\text{base}}(L_{\text{base}})$, isolating representational change from peak migration. **(c)** For each model, we compute the proportion of layers in which the projected direction reverses (i.e., morally labeled samples project below the *Social Norm* median (Appendix A.3)), averaged across the five foundations. Paired endpoints show base $\rightarrow$ instruct.

4.2.2 The Anatomy of Morality

To determine where moral concepts are most strongly represented within models, we analyze the alignment between MFT concept vectors and SAE decoder features across layers. We find that broad moral foundations are not monolithic; rather, they are composed of interpretable, atomic features that crystallize at specific depths of the network.

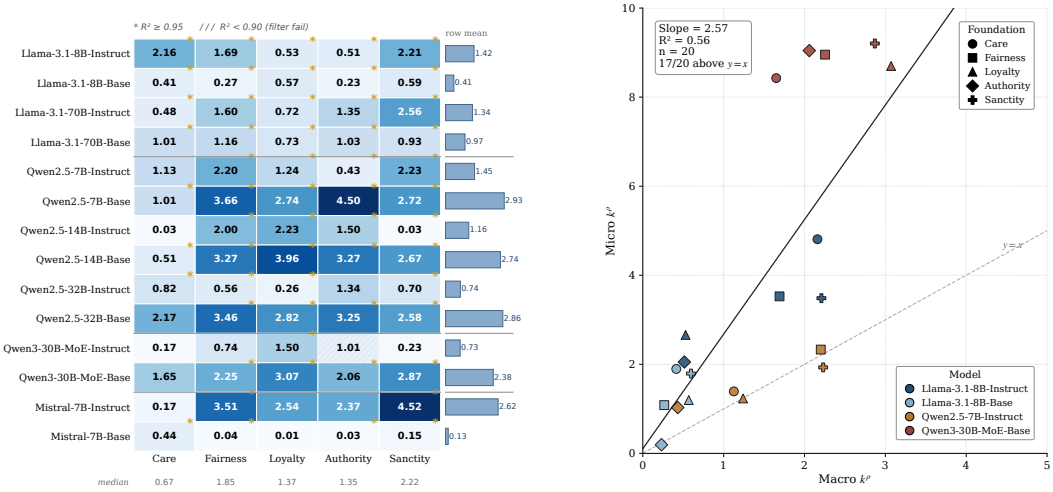
Layer-wise Feature Alignment. Figure 8 shows the average cosine similarity of the top-10 aligned features for each moral foundation every 4 layers across four models. We observe distinct representational trajectories tied to model families rather than instruction tuning alone. Both `Llama` models exhibit a clear mid-network "semantic bottleneck" around Layers 15–19, where moral feature alignment sharply peaks before declining. In contrast, the `Qwen` models push these representations toward the network edges. `Qwen2.5-7B-Instruct` displays a "U-shaped" trajectory with high alignment in early and late layers, while the larger `Qwen3-30B`'s alignment is delayed until a massive peak in its final layers. Interestingly, we also observe that instruction tuning shifts the internal moral hierarchy: while `Llama-Base` shows strong alignment for *Fairness* and *Loyalty*, `Llama-Instruct` shifts to prioritize *Care* and *Sanctity*. This mirrors the relative representational hierarchy consistently seen in both `Qwen` models across scales and indicates the family-specific rewiring effects of post-training. A comprehensive layer-wise analysis of these alignment dynamics is provided in Appendix C.1.

Semantic Decomposition. We qualitatively grounded these sparse features by analyzing their top-activating contexts using `GPT-5.1` with human validation. As detailed in Appendix C.1, we find that SAE features decompose abstract foundations into granular mechanisms. For example, *Care* features in `Llama-3.1-8B-Instruct` disentangle into distinct clusters tracking "physical suffering" vs. "emotional distress". In `Llama-3.1` and `Qwen2.5`, *Care* and *Authority* are the foundations most frequently associated with high-confidence semantic features (See Tables 8,9,10,11).

4.2.3 Causal Control Through Steering

Macro-steering: Vectors causally control moral behavior across 14 models. Across all 14 models and 5 foundations, MFQ-2 responses move approximately linearly with the relative perturbation ρ at each model's selected best layer L^* (Figure 3a, Appendix D). $68/70$ (model, foundation) pairs satisfy the linear-response quality filter ($R^2(\alpha) \geq 0.90$, $p \leq 0.05$, $k_\alpha > 0$), with positive ρ amplifying and negative ρ suppressing the target foundation in a bidirectional dose–response. The family-level direction of alignment effect mirrors the projection-geometry findings of Section 4.2.1:

`Qwen` alignment systematically suppresses steerability (k^ρ ratio base/instruct $\sim$2–4$\times$ across all four `Qwen` variants (`Qwen2.5-7B/14B/32B` and `Qwen3-30B-MoE`)), while `Llama` and `Mistral` alignment amplifies it (`Llama-3.1-8B` base/instruct $= 0.41/1.42$; `Mistral-7B` $= 0.13/2.62$).



(a) **Macro-steering: 14-model k^ρ.** (b) **Micro vs Macro k^ρ.**

Figure 3: **Concept vectors and SAE features causally control moral behavior across model families. (a)** Dose–response slope of the in-domain MFQ-2 sub-score under norm-calibrated steering along each foundation concept vector at the best layer L^*; larger k^ρ means stronger elicitation, and cells with $R^2 < 0.90$ are hatched. **(b)** Each point is one (model, foundation) cell ($n = 20$); axes show macro and micro top-10 SAE-feature steering slopes, each evaluated at its own best layer L^* (Tables 3 and 4), with OLS fit (solid) and $y = x$ equivalence (dashed). Appendix D gives implementation details.

Micro-steering: comparable causal control through sparse features. We test whether causal control can localize to sparse internal mechanisms by intervening along the decoder directions of the top-10 SAE features associated with each foundation. Across four models (`Llama-3.1-8B` base/instruct, `Qwen2.5-7B-Instruct`, `Qwen3-30B-MoE-Base`; 20 (model, foundation) cells), micro-steering produces causal effects on MFQ-2 that match or exceed the corresponding macro vector's: $17/20$ cells lie above the $y = x$ diagonal with OLS slope 2.57 (Figure 3b). This establishes that the sparse features identified through semantic validation in Section 4.2.2 are not only interpretable correlates of moral content but also functional causal handles on model behavior.

Cross-construct transfer. Steering also produces theoretically aligned shifts on Schwartz PVQ-21 values and political ideology, recovering the established binding–conservatism correspondence in moral psychology [Graham et al., 2011](Appendix D.4; capability preserved at $|\rho| \leq 0.20$, Appendix D.3). Together, these results demonstrate that steering moral concept vectors and their SAE-feature decompositions is not restricted to a single moral instrument or even the moral domain alone. Instead, these interventions reliably modulate theoretically related psychological constructs, such as political ideology and cross-framework human values, providing evidence for a shared internal representational structure.

5 Conclusions

This paper investigated the internal organization of moral foundations in LLMs. Our results support three main findings: (1) moral foundations are encoded as distinct linear directions in the residual stream that emerge during pretraining and are selectively rewired, rather than formed de novo, by post-training, with their geometric structure aligning with human moral perception; (2) these broad foundation directions are composed of more interpretable, atomic features; and (3) these representations are causally meaningful, as targeted interventions on the identified directions and features can directly and predictably modulate not only moral behavior but also related psychological constructs. By moving beyond surface-level observations, we provide a structural account of how moral values are anchored in LLM latent space.

These findings not only offer insight into interpretable mechanisms inside LLMs and practical questions in AI safety, but also speak to long-standing debates in moral psychology about the structure of human morality. By showing that LLMs naturally separate moral content into multiple, irreducible geometric dimensions rather than a single harm-based continuum, our work provides computational support for pluralist frameworks. Overall, our results suggest that the multi-dimensional structure of human morality can emerge as a latent pattern from the statistical regularities of language alone, in ways that mirror patterns observed in human moral cognition.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge support from the Air Force Office of Scientific Research (AFOSR) grant A9550-23-1-0463, and from the USC Center for Computational Language Sciences (CCLS) Seed Funding. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors. The authors declare no competing interests.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752, 2024.

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 3(7):pgae245, 2024.

Meltem Aksoy. Whose morality do they speak? unraveling cultural bias in multilingual language models. *Natural Language Processing Journal*, 12:100172, 2025.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, et al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 6, 2025.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Mohammad Atari and Morteza Dehghani. Language analysis in moral psychology. *The atlas of language analysis in psychology*, pages 207–228, 2022.

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

Amin Banayeeanzade, Ala N. Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. Psychological steering in llms: An evaluation of effectiveness and trustworthiness, 2025. URL `https://arxiv.org/abs/2510.04484`.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms, 2025. URL `https://arxiv.org/abs/2502.17424`.

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. Saelens. `https://github.com/decoderesearch/SAELens`, 2024.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

Nick Cammarata, Gabriel Goh, Shan Carter, Chelsea Voss, Ludwig Schubert, and Chris Olah. Curve circuits. *Distill*, 6(1):e00024–006, 2021.

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

Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47(4):1178–1198, 2015.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Dan Dillion, Dan Mondal, Niket Tandon, and Kurt Gray. Ai language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, 15(1):4084, 2025.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Jeremy A Frimer, Reihane Boghrati, Jonathan Haidt, Jesse Graham, and Morteza Dehghani. Moral foundations dictionary for linguistic analyses 2.0. *Unpublished manuscript*, 2019.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL `https://arxiv.org/abs/2101.00027`.

Leander Girrbach, Stephan Alaniz, Genevieve Smith, Trevor Darrell, and Zeynep Akata. Person-centric annotations of laion-400m: Auditing bias and its transfer to models. *arXiv preprint arXiv:2510.03721*, 2025.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366, 2011.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.

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

Kurt Gray, Chelsea Schein, and Adrian F Ward. The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, 143(4):1600, 2014.

Jonathan Haidt. The new synthesis in moral psychology. *science*, 316(5827):998–1002, 2007.

Jonathan Haidt. *The righteous mind: Why good people are divided by politics and religion*. Vintage, 2012.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

Joe Hoover, Mohammad Atari, Aida Mostafazadeh Davani, Brendan Kennedy, Gwenyth Portillo-Wightman, Leigh Yeh, and Morteza Dehghani. Investigating the role of group-based morality in extreme behavioral expressions of prejudice. *Nature Communications*, 12(1):4585, 2021.

Frederic R Hopp, Ori Amir, Jacob T Fisher, Scott Grafton, Walter Sinnott-Armstrong, and René Weber. Moral foundations elicit shared and dissociable cortical activation modulated by political ideology. *Nature Human Behaviour*, 7(12):2182–2198, 2023.

Clayton Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225, 2014.

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. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Behnam Karami, Fatemeh Zandi, and Javad Hatami. Emergent moral representations in large language models aligns with human conceptual, neural, and behavioral moral structure. *Research Square Preprint*, 2025.

Adam Karvonen. An intuitive explanation of sparse autoencoders for llm interpretability, Jun 2024. URL `https://adamkarvonen.github.io/machine_learning/2024/06/11/sae-intuitions.html`.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. Moral concerns are differentially observable in language. *Cognition*, 212:104696, 2021.

Ari Khoudary, Eleanor Hanna, Kevin O'Neill, Vijeth Iyengar, Scott Clifford, Roberto Cabeza, Felipe De Brigard, and Walter Sinnott-Armstrong. A functional neuroimaging investigation of moral foundations theory. *Social Neuroscience*, 17(6):491–507, 2022.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=aLLuYpn83y`.

Leon Li and Michael Tomasello. On the moral functions of language. *Social Cognition*, 39(1): 99–116, 2021.

Shawn Li and Yue Zhao. The autonomy tax: Defense training breaks llm agents. In *EMNLP*, 2026.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. Fantastic semantics and where to find them: Investigating which layers of generative llms reflect lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14551–14558, 2024.

Samuel Marks, Adam Karvonen, and Aaron Mueller. Dictionary learning. GitHub repository, 2024. https://github.com/saprmarks/dictionary_learning.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fh8EYKFKns.

Tuan Dung Nguyen, Ziyu Chen, Nicholas George Carroll, Alasdair Tran, Colin Klein, and Lexing Xie. Measuring moral dimensions in social media with mformer. In *Proceedings of the international AAAI conference on web and social media*, volume 18, pages 1134–1147, 2024.

José Luiz Nunes, Guilherme F. C. F. Almeida, Marcelo de Araujo, and Simone D. J. Barbosa. Are large language models moral hypocrites? a study based on moral foundations. *arXiv preprint arXiv:2405.11100*, 2024.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EemtbhJOXc.

Guilherme Penedo, Hynåek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434, 2023.

Qwen Team. Qwen-Scope: Turning sparse features into development tools for large language models, April 2026. URL https://qianwen-res.oss-accelerate.aliyuncs.com/qwen-scope/Qwen_Scope.pdf.

Aida Ramezani and Yang Xu. Moral association graph: a cognitive model for automated moral inference. *Topics in Cognitive Science*, 17(1):120–138, 2025.

Aida Ramezani, Emmy Liu, Spike WS Lee, and Yang Xu. Quantifying the emergence of moral foundational lexicon in child language development. *PNAS nexus*, 3(8):pgae278, 2024.

Nils Karl Reimer, Mohammad Atari, Farzan Karimi-Malekabadi, Jackson Trager, Brendan Kennedy, Jesse Graham, and Morteza Dehghani. Moral values predict county-level covid-19 vaccination rates in the united states. *American Psychologist*, 77(6):743, 2022.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Khan, and Jia Xu. Analyzing encoded concepts in transformer language models. In *Proceedings of the 2022 Conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 3082–3101, 2022.

Chelsea Schein and Kurt Gray. The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, 22(1):32–70, 2018.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022.

Shalom H Schwartz. 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, 1992.

Shalom H Schwartz et al. A proposal for measuring value orientations across nations. *Questionnaire package of the european social survey*, 259(290):261, 2003.

Gabriel Simmons. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106*, 2022.

Judith G Smetana. Social-cognitive domain theory: Consistencies and variations in children's moral and social knowledge. In *Handbook of Moral Development*, pages 119–153. Erlbaum, Mahwah, NJ, 2006.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wozhdnRCtw.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Jackson Trager, Francielle Vargas, Diego Alves, et al. Mftcxplain: A multilingual benchmark dataset for evaluating the moral reasoning of llms through multi-hop hate speech explanation. *arXiv preprint arXiv:2506.19073*, 2025a.

Jackson Trager, Alireza S. Ziabari, Elnaz Rahmati, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. The moral foundations reddit corpus, 2025b. URL https://arxiv.org/abs/2208.05545.

Elliot Turiel. *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press, Cambridge, UK, 1983.

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

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, 2020.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

James Wilkinson, Oliver Scott Curry, Brittany L Mitchell, and Timothy Bates. Modular morals: Mapping the organization of the moral brain. *Brain and Cognition*, 180:106201, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Aliah Zewail, Alexandra Figueroa, Jesse Graham, and Mohammad Atari. Moral stereotyping in large language models. *Proceedings of the National Academy of Sciences*, 123(10):e2519941123, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2023.

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

Appendix Contents

A Detailed Experimental Setup **16**

A.1 Models and Architectures 16

A.2 Dataset 16

A.3 Evaluation Metrics 17

B Projection: Implementation Details and Extended Analysis **20**

B.1 Foundation-vs.-Social Norm: Case Study on Llama-3.1-8B-Instruct 20

B.2 Foundation-vs.-Social Norm: Layer-wise evolution across all 14 models 20

B.3 Between-foundation projections 21

C SAEs: Implementation Details and Extended Analysis **23**

C.1 Extended Analysis for The Anatomy of Morality 24

C.2 Semantic Validation Implementation Details 25

D Steering: Implementation Details and Extended Analysis **27**

D.1 Implementation Details 27

D.1.1 Steering Response and Best-Layer Selection 28

D.2 Steering Case Study: Dose–Response Curves 29

D.3 General Performance Measurement after Steering 29

D.4 Cross-Construct Transfer 32

E Human Evaluation **33**

E.1 MFV-130 Expansion Prompt 34

F Limitations **35**

G Ethical Considerations **35**

A Detailed Experimental Setup

A.1 Models and Architectures

Subject models. For **projection-based analyses and macro-steering**, we use Llama-3.1 (8B, 70B) [Grattafiori et al., 2024], Qwen2.5 (7B, 14B, 32B) [Yang et al., 2024], Qwen3-30B-A3B [Yang et al., 2025], and Mistral-7B-v0.3 [Jiang et al., 2023], each in both Base and Instruct versions, yielding 14 models in total. For **mechanistic decomposition and micro-steering** (Section 3.3), we employ pretrained SAEs from 4 models: Llama-3.1-8B (Instruct/Base), Qwen2.5-7B-Instruct, and Qwen3-30B-A3B Base. All inference is run with HuggingFace Transformers in BF16/FP16 (depending on hardware support), using temperature $T = 0.01$ for all experiments.

Compute Resources. Our experiments on projection-based analyses and macro-steering are conducted on four NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPUs and two NVIDIA B200 GPUs. Our mechanistic decomposition and micro-steering experiments are done on two NVIDIA RTX 6000 Pro Server Edition GPUs.

Sparse Autoencoders. We perform mechanistic decomposition using multiple suites of pretrained residual-stream SAEs accessed via the SAELens library [Marks et al., 2024], alongside models from Llama Scope [He et al., 2024] and Qwen-Scope [Qwen Team, 2026]. To ensure the learned features cover the relevant behavioral distributions, the training corpora varied by suite: Llama-3.1-8B-Instruct and Qwen2.5-7B-instruct SAEs were trained on a diverse set of standard pretraining corpora and chat-based instruction data [Zheng et al., 2023, Gao et al., 2020, Betley et al., 2025]; Llama Scope models were pretrained on SlimPajama; Qwen-Scope utilized samples from their in-house training data.

To analyze the evolution of moral features across depth while optimizing computational costs, we evaluate SAEs trained on every fourth layer for all models (e.g., $L \in \{3, 7, \ldots, 27\}$ for 32-layer architectures). Architectural specifications and activation functions vary across the model families. For Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, the SAEs employ the BatchTopK activation function [Bussmann et al., 2024, Bloom et al., 2024]. The instruct models feature a sparsity of $k = 64$ active features per token and project to $\sim$131k latent features, corresponding to expansion factors of 32 ($d_{\text{model}} = 4096$) and 36.57 ($d_{\text{model}} = 3584$) respectively. Conversely, the base models utilize different activation rules: Llama Scope for Llama-3.1-8B Base SAE employs TopK-ReLU with an expansion factor of 32, and Qwen-Scope for the Qwen3-30B-A3B Base model uses a standard Top-K activation ($k = 50$) with an expansion factor of 16, projecting the $d_{\text{model}} = 2048$ dimension to $d_{\text{sae}} = 32768$.

A.2 Dataset

Extended Moral Foundations Vignettes. We construct concept vectors using an expanded version of the *Moral Foundations Vignettes* (MFV-130; Clifford et al., 2015). Starting from the original vignettes for the five moral foundations and a *Social Norm* category, we expand each category to approximately 200 short scenarios using `gpt-5-mini`, matching the conceptual definition and linguistic style of the original MFV items, using prompts that vary everyday social contexts and non-essential contextual features which past work has shown to reduce spurious correlations and stabilize learned representations [Arjovsky et al., 2019, Kim et al., 2018]. All generated vignettes are reviewed by human experts to verify clarity, label correctness, and adherence to the intended foundation or social-norm category. Sample prompts and generated items are provided in Appendix E.1.

Moral Foundations Reddit Corpus. Following previous work [Ramezani et al., 2024, Ramezani and Xu, 2025, Nguyen et al., 2024], we use human-labeled Reddit posts from the Reddit Moral Foundations Corpus [Trager et al., 2025b] as the primary source to validate our vectors on real-world moral language. From the full corpus (61.2K posts), we keep only *single-label* posts with *high-confidence* annotations to obtain an unambiguous ground-truth set. These posts are never used for vector construction; we feed only the raw text to the model and use labels solely for grouping in projection analyses.

Table 1: **Length and sentiment statistics of the expanded MFV vignettes.** Mean ± SD of word count and VADER compound sentiment score per category.

Category	n	Words	VADER
Care	242	14.1 ± 1.8	-0.07 ± 0.29
Fairness	217	12.9 ± 1.9	$+0.04 \pm 0.31$
Loyalty	216	14.7 ± 1.5	$+0.21 \pm 0.31$
Authority	217	13.0 ± 1.7	-0.19 ± 0.25
Sanctity	217	12.6 ± 1.7	-0.03 ± 0.21
Social Norms	216	11.9 ± 1.8	$+0.01 \pm 0.13$

Semantic resource for SAE validation. We use the Moral Foundations Dictionary 2.0 (MFD2) [Frimer et al., 2019] as an external semantic anchor for validating SAE feature fingerprints (Section 3.3). For each foundation, we extract the corresponding keyword lists and test whether the selected SAE feature directions are geometrically close to embeddings of the foundation-specific MFD terms.

Behavioral evaluation resources for steering. To measure causal effects of steering on expressed moral preferences, we use three independent behavioral readouts that are operationally distinct from the MFV vignettes used to construct concept vectors.

(i) Moral Foundations Questionnaire–2 (MFQ-2) [Atari et al., 2023] serves as the within-framework readout for in-domain steering effects (§4.2.3, Appendix D). MFQ-2 items are mapped to foundation subscales and rated on a five-point Likert scale. Following prior practice, we operationalize *Fairness* by averaging items from *Equality* and *Proportionality*, and report five foundation scores: *Care*, *Fairness*, *Loyalty*, *Authority*, and *Purity (Sanctity)*.

(ii) Schwartz Portrait Values Questionnaire (PVQ-21) [Schwartz et al., 2003] provides a cross-framework readout grounded in a distinct theoretical model of human values [Schwartz, 1992]. We use the 21-item ESS short form rated on a five-point Likert scale, score items via single-token logits, MRAT-center responses to remove response-style bias [Schwartz et al., 2003], and aggregate to ten basic values and four higher-order clusters (Self Transcendence, Self-Enhancement, Conservation, Openness-to-Change).

(iii) Political ideology is measured with a single 1–7 Likert item ("How would you describe your political views?" with 1 = Very liberal, 7 = Very conservative), scored via single-token logits over the seven option tokens.

PVQ-21 and the ideology item allow us to test whether steering produces theoretically aligned shifts on instruments grounded in *distinct theoretical frameworks* from MFT, recovering the established binding/individualizing–conservatism correspondence in moral psychology [Graham et al., 2011, Atari et al., 2023] (Appendix D.4).

Surface-form validation. To assess potential surface-form confounds, we examined the length and sentiment of all 1,325 vignettes. All categories follow the same single-sentence observational format with comparable average lengths (Table 1). VADER compound sentiment scores [Hutto and Gilbert, 2014] show no uniform negative shift of the moral foundations relative to *Social Norms*: *Loyalty* is more positive on average ($+0.21$) and *Authority* more negative (-0.19) than *Social Norms* ($+0.01$). The foundation contrasts are therefore unlikely to reflect a single shared negative-valence direction.

A.3 Evaluation Metrics

Signed Wasserstein Distance for topological validity. To assess whether our concept vectors align with human-labeled moral categories (Section 3.2), we compare the projection-score distributions of Reddit posts labeled with foundation k versus those not labeled with k. For each layer ℓ, let $P_{k,\ell}$ and $P_{\neg k,\ell}$ denote the corresponding score distributions, and let $\mu_{k,\ell}$ and $\mu_{\neg k,\ell}$ be their means. We report the *Signed Wasserstein Distance* to capture both the magnitude and the direction of separation:

$$\begin{aligned} \mathrm{S}W_1(P_{k,\ell}, P_{\neg k,\ell}) = \operatorname{sign}(\mu_{k,\ell} - \mu_{\neg k,\ell}) \\ \cdot W_1(P_{k,\ell}, P_{\neg k,\ell}), \end{aligned} \tag{8}$$

where the standard (unsigned) 1-Wasserstein distance is defined as

$$W_1(P_{k,\ell}, P_{\neg k,\ell}) = \inf_{\gamma \in \Pi(P_{k,\ell}, P_{\neg k,\ell})} \mathbb{E}_{(x,y)\sim\gamma}\big[|x - y|\big], \tag{9}$$

with $\Pi(P_{k,\ell}, P_{\neg k,\ell})$ denoting the set of joint distributions with marginals $P_{k,\ell}$ and $P_{\neg k,\ell}$. A positive Signed-W_1 indicates that the labeled examples possess larger mean projection scores along the foundation vector (alignment), whereas negative values indicate separation in the opposite direction (anti-alignment). We employ W_1 because it (i) remains well-defined even for distributions with disjoint support, (ii) faithfully reflects geometric separation along the projection axis, and (iii) is robust to class imbalance.

Standardized Wasserstein separation $\widetilde{W}_1$ for cross-model comparison. Raw Wasserstein distances are not comparable across models because the projection scale depends on each model's residual-stream norm and foundation-vector magnitude, both of which vary by orders of magnitude across families and layers. We therefore standardize W_1 by the pooled within-group spread, the same denominator used by Cohen's d.

For each model, layer ℓ, and foundation f, we project every stimulus's last-token residual stream $h_i^{(\ell)} \in \mathbb{R}^d$ onto the unit-norm concept vector $v_f \in \mathbb{R}^d$:

$$p_i = \langle h_i^{(\ell)}, v_f \rangle, \qquad \|v_f\|_2 = 1. \tag{10}$$

This yields two 1-D projection sets: $A = \{p_i\}_{i\in f}$ (foundation samples) and $B = \{p_j\}_{j\in\text{social}}$ (social-norm samples), with sample sizes n_1, n_2 and unbiased standard deviations s_1, s_2. We measure their separation with the 1-Wasserstein distance $W_1(A, B)$—sensitive to the full distributional shift rather than only the mean—and standardize it by the pooled within-group standard deviation:

$$\sigma_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}, \tag{11}$$

giving the standardized separation

$$\widetilde{W}_1 = \frac{W_1(A, B)}{\sigma_{\text{pooled}}}. \tag{12}$$

We use σ_{pooled} rather than the marginal standard deviation $\sigma(A \cup B)$ because the latter conflates within-group noise with the very between-group separation we wish to measure; σ_{pooled} is the standard Cohen's-d denominator and isolates the within-group noise scale.

Invariance properties. $\widetilde{W}_1$ is dimensionless and invariant to two nuisance scalings:

- **Concept-axis norm.** Rescaling $v_f \to \alpha v_f$ multiplies all projections p_i by $|\alpha|$, hence multiplies both W_1 and σ_{pooled} by $|\alpha|$ and leaves $\widetilde{W}_1$ unchanged.
- **Hidden-state magnitude.** Hidden-state norms differ systematically across layers and model families; any uniform rescaling of the residual stream affects both W_1 and σ_{pooled} by the same factor and cancels in $\widetilde{W}_1$.

As a result, $\widetilde{W}_1$ is directly comparable across (model $\times$ layer $\times$ foundation) cells, and can be read in units of within-group standard deviations (the same scale as Cohen's d).

Best-layer aggregation. For each (model, foundation) pair we restrict attention to layers whose median projection direction agrees with the expected sign (`dir_ok` $= 1$, i.e. foundation samples have higher mean projection than social-norm samples), and report $\widetilde{W}_1$ at the layer that maximizes $\widetilde{W}_1$ within this admissible set:

$$L^*_{\text{geom}}(f) = \arg \max_{\ell\,:\,\texttt{dir_ok}(\ell,f)=1} \widetilde{W}_1(\ell, f). \tag{13}$$

The resulting per-cell peak quantifies the geometric separation between foundation f and the social-norm baseline along v_f at the layer where the concept is most sharply encoded. This best-layer $\widetilde{W}_1$ is the metric reported in Figure 2a.

Direction-reversal rate. For each (model, layer ℓ, foundation f) cell, the foundation concept vector v_f is constructed so that foundation-relevant stimuli should yield higher projection scores than *Social Norm* stimuli. We flag a layer as *direction-reversed* when this expected ordering breaks down. Let $A = \{p_i\}_{i \in f}$ and $B = \{p_j\}_{j \in \text{social}}$ denote the two projection sets at (ℓ, f). Define

$$\texttt{dir_ok}(\ell, f) = \begin{cases} \mathbf{1}\{\text{median}(A) > \text{median}(B)\}, & |\text{median}(A) - \text{median}(B)| > \varepsilon, \\ \mathbf{1}\{\text{mean}(A) > \text{mean}(B)\}, & \text{otherwise (tie-break)}, \end{cases} \tag{14}$$

where $\varepsilon = 10^{-9} \cdot (|\bar{A}| + |\bar{B}|)$ handles numerical near-ties. Median is preferred for robustness to outliers; the mean is used only when the two medians are numerically indistinguishable.

The per-(model, foundation) *reversal rate* is the fraction of layers that fail this check:

$$\text{fail_rate}(\text{model}, f) = \frac{\#\{\ell : \texttt{dir_ok}(\ell, f) = 0\}}{\#\{\ell\}}. \tag{15}$$

The *mean direction-reversal rate* for a model is the unweighted average across the five foundations:

$$\overline{\text{fail_rate}}(\text{model}) = \frac{1}{5} \sum_{f \in \mathcal{F}} \text{fail_rate}(\text{model}, f), \qquad \mathcal{F} = \{\text{Care, Fairness, Loyalty, Authority, Sanctity}\}. \tag{16}$$

A reversal rate of 0% indicates that the concept axis v_f points in the expected direction at every layer; higher rates indicate that some layers encode the foundation in a geometrically inverted orientation, signaling unstable or weakly localized encoding. This metric is reported for Base and Instruct variants in Figure 2c, where the Base–Instruct difference isolates the contribution of post-training to direction stability.

Steering linear-response slopes. We quantify causal effects of steering on MFQ-2 responses by fitting linear dose–response models at each candidate layer ℓ. Let $S_k(\ell, \rho)$ denote the MFQ-2 subscale score for foundation k under steering at layer ℓ with relative perturbation ρ (Appendix D), and let $\Delta S_k(\ell, \rho) = S_k(\ell, \rho) - S_k(\ell, 0)$ be the baseline-subtracted change. Because readouts are deterministic expected-value scores (Appendix D), each (layer, foundation) cell contributes one datapoint per ρ-value, i.e. 7 datapoints, on which we fit two equivalent OLS regressions (with $n = 7$, $R^2 \geq 0.90$ implies $p < 0.002$):

$$\Delta S_k(\ell, \alpha) = k_\alpha(\ell) \cdot \alpha + \varepsilon_\alpha, \tag{17}$$

$$\Delta S_k(\ell, \rho) = k^\rho(\ell) \cdot \rho + \varepsilon_\rho, \tag{18}$$

where $\alpha = \rho \cdot \bar{r}_\ell$ is the absolute additive scale at layer ℓ and $\bar{r}_\ell$ is the mean residual norm (steering calibration). The two slopes satisfy $k^\rho(\ell) = k_\alpha(\ell) \cdot \bar{r}_\ell$ identically. We use k_α for within-model layer selection and k^ρ for cross-model comparison since ρ is dimensionless and comparable across models with different residual-norm scales.

Logits-based scoring. For each MFQ-2 item with options $T = \{1, 2, 3, 4, 5\}$, we obtain the option logits $\{z_t\}_{t \in T}$ and compute $p(t) = \text{softmax}(z)_t$. The item score is the expected rating $S_{\text{item}} = \sum_{t \in T} t \cdot p(t)$. We then average item scores within each MFQ-2 subscale to obtain foundation scores S_k. For PVQ-21 and political-ideology items used in cross-framework transfer (Appendix D.4), the same expected-rating procedure is applied over the appropriate option-token set ($K = 6$ for PVQ items, $K = 7$ for ideology).

Robustness to alternative metrics. To assess whether our conclusions depend on the choice of separation metric, we recompute the Figure 2a analysis with Cohen's d (a mean-based standardized effect size) and AUC (a rank-based, scale-free measure), using the same projection scores and the same $\widetilde{W}_1$-selected best layers (Eq. 13); layers are not re-selected per metric. As shown in Table 2, all three metrics identify `Qwen2.5-7B` as the most strongly encoded base model, *Authority* as the most separable foundation, and *Loyalty* as the least separable. Across the 35 base-model cells, Spearman correlations with $\widetilde{W}_1$ are 0.998 (Cohen's d) and 0.975 (AUC); across all 70 cells they are 0.998 and 0.989. The sign of the post-training change in Figure 2b agrees with $\widetilde{W}_1$ in $34/35$ family–foundation pairs under Cohen's d and $30/35$ under AUC, and every family-level conclusion of Section 4.2.1 holds under each metric.

Table 2: **Robustness of foundation separability to the choice of metric.** All values are computed at the layers selected with $\widetilde{W}_1$ (Eq. 13). *Strongest model*: the base model with the highest mean across the five foundations; *strongest/weakest foundation*: highest/lowest median across the seven base models.

Metric	Strongest model	Strongest foundation	Weakest foundation
$\widetilde{W}_1$	`Qwen2.5-7B` (0.589)	*Authority* (0.385)	*Loyalty* (0.239)
Cohen's d	`Qwen2.5-7B` (0.589)	*Authority* (0.389)	*Loyalty* (0.251)
AUC	`Qwen2.5-7B` (0.661)	*Authority* (0.611)	*Loyalty* (0.562)

B Projection: Implementation Details and Extended Analysis

This section expands the projection analyses summarized in Section 4.2.1. Section B.1 provides a detailed case study of foundation-vs.-Social-Norm projections on a representative model, illustrating the distributional and layer-wise structure that the 14-model panel aggregates over. Section B.2 extends this analysis to all 14 models in our panel. Section B.3 then turns to the complementary question of how the five foundations are organized relative to one another.

B.1 Foundation-vs.-Social Norm: Case Study on Llama-3.1-8B-Instruct

To illustrate the distributional structure of foundation-vs.-Social-Norm projections, we report detailed results for `Llama-3.1-8B-Instruct`—a representative instruction-tuned model in our panel.

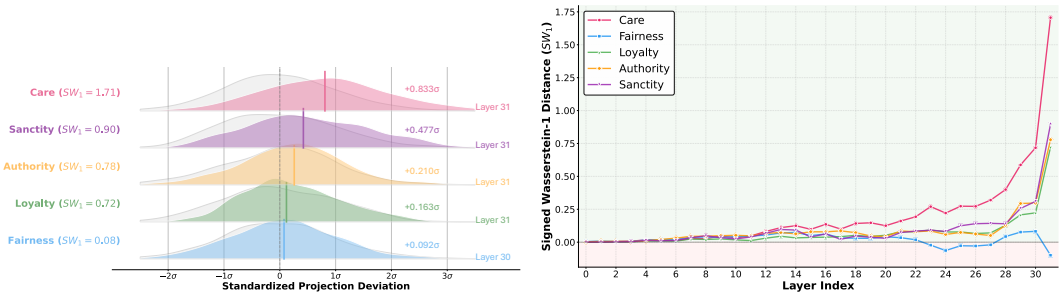
Figure 4a shows projection-score densities at each foundation's optimal layer (the layer with the largest signed Wasserstein separation SW_1, Appendix A.3). We observe clear distributional separation between morally labeled and non-moral Reddit posts, strongest for *Care* ($SW_1 = 1.71$) and *Sanctity* ($SW_1 = 0.90$).

Figure 4b traces the unstandardized signed Wasserstein-1 distance SW_1 across all 32 layers of the model. Raw SW_1 remains low to moderate in early and middle layers (0–24) and increases sharply in the final layers (28–31). However, raw SW_1 values are not directly interpretable as increases in semantic separability across layers, because the residual-stream norm itself grows substantially across layers in pre-LayerNorm transformers [Turner et al., 2023, Sun et al., 2024]. We therefore use the standardized form $\widetilde{W}_1$ (Appendix A.3), which divides $|SW_1|$ by the pooled within-group standard deviation, for cross-layer and cross-model comparisons reported in the main text (Figure 2). The signed unstandardized SW_1 shown here serves only to visualize the within-model layer-wise trend and the direction-stability pattern: green segments (Figure 4b, layers with `dir_ok` $= 1$) indicate the projection direction agrees with the expected sign, while red segments (`dir_ok` $= 0$) indicate direction reversal.

B.2 Foundation-vs.-Social Norm: Layer-wise evolution across all 14 models

Figure 5 shows the layer-wise evolution of $\widetilde{W}_1$ across all 14 model variants. The 2×7 panel layout factors models by family/scale (columns) and training stage (rows), with the five foundations plotted as colored lines within each panel. Several patterns are visible.

Family-specific geometry stability. `Qwen2.5-7B` (both variants) is the cleanest model in the panel: all five foundations exhibit monotone growth from early layers to the terminal peak with no direction reversals. The larger `Qwen2.5-14B` and `32B` variants, in contrast, exhibit substantial mid-layer instability with frequent sign inversions across multiple foundations, indicating that the moral-geometry signal in these models is not linearly read-out at intermediate layers. The `Llama-3.1` family exhibits both a *scale* and a *post-training* effect on geometric stability: `Llama-3.1-70B-Instruct` produces clean monotone growth on all five foundations, while `Llama-3.1-8B-Base` is the messiest `Llama` variant, with *Loyalty* oscillating around zero throughout the late layers and both *Fairness* and *Sanctity* sign-inverted at the terminal peak. The two intermediate variants (`Llama-3.1-8B-Instruct`, `Llama-3.1-70B-Base`) exhibit *Fairness* sign inversion but stable monotone trajectories on the remaining four foundations (cf. Figure 4b in Appendix B.1). Both larger scale (8B $\rightarrow$ 70B) and post-training (Base $\rightarrow$ Instruct) appear to stabilize foundation-specific geometry within this family.



(a) **Projection distributions at the best layer.** Probability densities of foundation-relevant Reddit posts (colored) and *Social Norm* posts (gray) projected onto each foundation-vs.-Social-Norm vector v_f.

(b) **Layer-wise evolution of separability.** Signed Wasserstein-1 distance SW_1 (Appendix A.3) across all 32 layers of `Llama-3.1-8B-Instruct`; green indicates positive separability, red indicates direction-reversed separability.

Figure 4: **Foundation-vs.-Social-Norm geometry in `Llama-3.1-8B-Instruct` (representative instruction-tuned model).** Projecting human-labeled Reddit posts onto each foundation concept vector v_f reveals clear distributional separation from the *Social Norm* baseline that emerges primarily in the final layers of the network. *Fairness* is partially entangled with the *Social Norm* baseline in this particular model.

Post-training amplifies or compresses depending on family. For each family, comparing the upper (Instruct) and lower (Base) row reveals the sign of the post-training effect on geometric separability: `Llama` and `Mistral` Instruct variants show higher peak $\widetilde{W}_1$ than their Base counterparts (amplification), while `Qwen2.5` Instruct variants typically show lower or comparable peaks (compression). This visual pattern mirrors the quantitative amplification/compression dichotomy reported in Figure 2B (Section 4.2.1). The scale dimension of our panel reveals a family-specific interaction: within `Llama-3.1`, scale has little effect on geometric separability and post-training instead stabilizes projection direction; within `Qwen2.5`, scale itself degrades both direction stability and terminal-layer magnitude, and post-training does not compensate at larger scales.

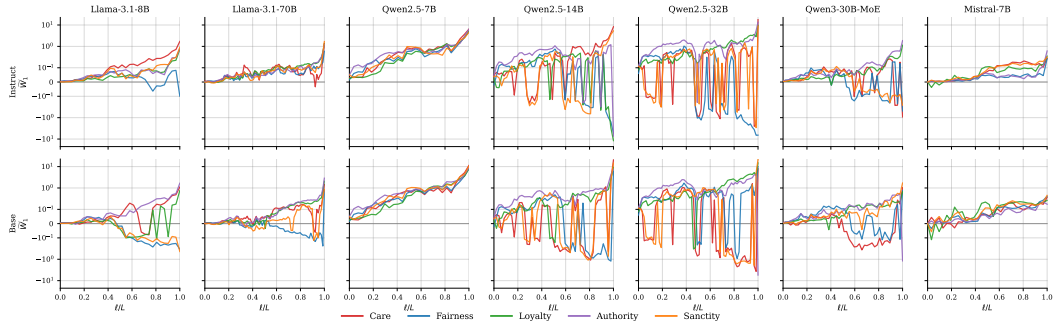


Figure 5: **Layer-wise evolution of $\widetilde{W}_1$ across all 14 model variants.** Each panel shows one (model, variant) pair; rows separate Instruct (top) and Base (bottom) variants; columns are model families ordered by family and scale. Within each panel, the five colored lines correspond to the five foundations (*Care, Fairness, Loyalty, Authority, Sanctity*), plotted on a symmetric-log y-axis to show both positive separation and direction-reversed layers in a single view. The x-axis is normalized depth ℓ/L to allow comparison across models with different layer counts.

B.3 Between-foundation projections

The foundation-vs.-Social-Norm analysis above shows that each v_f separates foundation-f language from non-moral baseline. A complementary question concerns the geometric organization *among* the five foundations themselves: which foundations occupy similar regions of representation space, and which are maximally separated? We answer this with pairwise foundation-vs.-foundation projections.

Recent work in moral psychology suggests that the classic MFQ clustering into *Individualizing* (Care, Fairness) and *Binding* (Loyalty, Authority, Sanctity) is not stable across cultures and may reflect WEIRD-specific measurement structure. Network analyses of Moral Foundations Questionnaire data indicate that inter-foundation relationships vary substantially across societies: moral foundations often form interconnected networks rather than two consistently segregated clusters, and no single higher-order relational pattern generalizes across cultural contexts [Atari et al., 2023]. These findings motivate treating inter-foundation geometry as a variable property rather than a universal template.

Figures 6 and 7 report pairwise $\widetilde{W}_1$ between every pair of foundations for all 14 model variants, separated into Instruct and Base groups. For each (model, foundation pair), the value is the peak $\widetilde{W}_1$ across layers with correct projection direction (`dir_ok` $= 1$, see Appendix A.3). Viewed through the lens of the cross-cultural literature above, our model geometries suggest three takeaways.

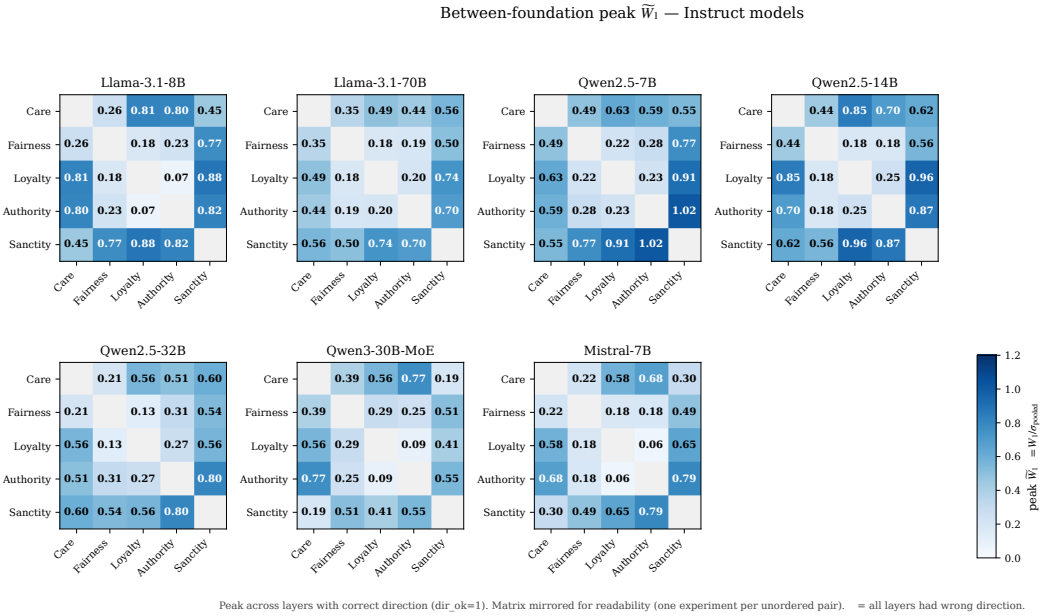


Figure 6: **Pairwise $\widetilde{W}_1$ between foundations for the 7 Instruct model variants.** For each (model, foundation pair), the value is the peak $\widetilde{W}_1$ across layers with `dir_ok` $= 1$ (Appendix A.3); matrices are mirrored along the diagonal for readability.

First, *Sanctity* is geometrically isolated in 12 **of** 14 **models.** The Sanctity row/column appears as a relatively dark band in nearly all panels, with pairwise distances between *Sanctity* and the other four foundations typically the largest in the matrix (e.g., `Llama-3.1-8B-Instruct`: $\widetilde{W}_1$(Sanctity, Authority) $= 0.82$ vs. $\widetilde{W}_1$(Loyalty, Authority) $= 0.07$). This contrasts with the canonical MFT clustering, which would place *Sanctity within* the binding cluster alongside *Loyalty* and *Authority*. `Qwen3-30B-MoE` is the only family that partially deviates: its Instruct variant places *Sanctity* closer to *Care* ($\widetilde{W}_1 = 0.19$) than to most other foundations.

Second, *Fairness* is geometrically closer to *Authority* than to *Care* in 12 **of** 14 **models.** Under canonical MFT clustering, *Fairness* and *Care* should occupy adjacent regions as the two individualizing foundations. Empirically, the opposite holds: in 12 of 14 models, $\widetilde{W}_1$(Fairness, Authority) is smaller than $\widetilde{W}_1$(Fairness, Care). Only the two `Qwen2.5-32B` variants align with the canonical prediction. Comparable variation in the relational position of fairness has been observed in some non-WEIRD human samples, where fairness is not uniformly aligned with care-based concerns [Atari et al., 2023].

Third, *Loyalty* and *Authority* form a near-collapsed pair in all 14 **models.** The *Loyalty–Authority* distance is consistently small across the panel, ranging from 0.06 (`Mistral-7B-Instruct`) to 0.27 (`Qwen2.5-32B-Instruct`). No model in our panel encodes these two foundations as well-separated directions. The collapse is most extreme in three post-trained variants (`Llama-3.1-8B-Instruct`,

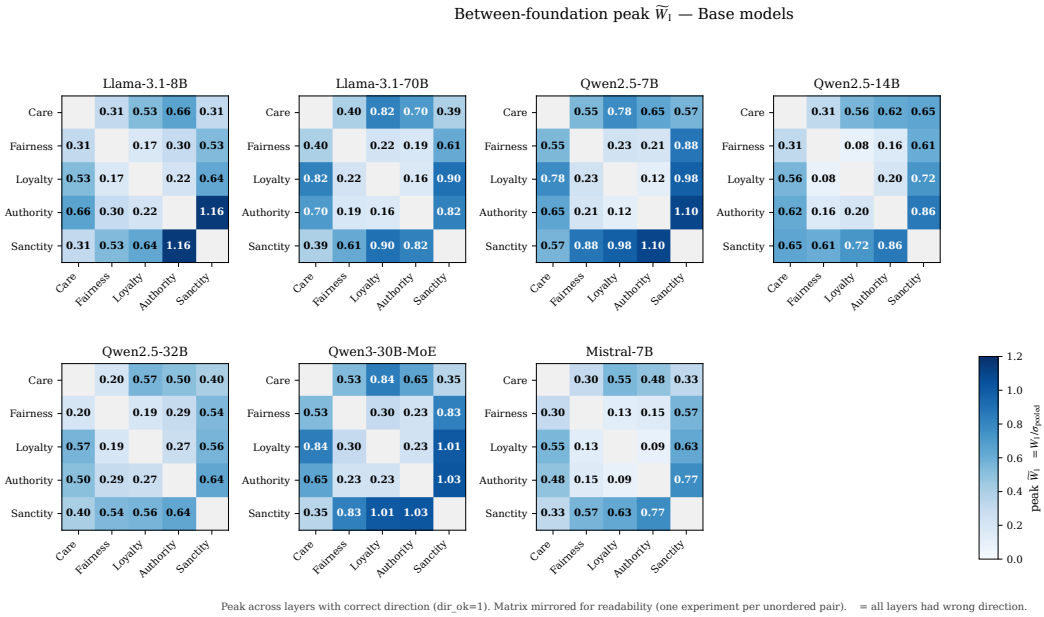


Figure 7: **Pairwise $\widetilde{W}_1$ between foundations for the 7 Base model variants.** Computation as in Figure 6.

`Mistral-7B-Instruct`, `Qwen3-30B-MoE-Instruct`), where *Fairness*/*Loyalty*/*Authority* together form a tight cluster with all three pairwise distances ≤ 0.29.

Taken together, these three patterns describe a geometric organization that does not reproduce the canonical MFT Individualizing/Binding partition. Across most models, the five foundations align approximately with the four-way split $\{Care\} \cup \{Fairness, Loyalty, Authority\} \cup \{Sanctity\}$ relative to the *Social Norm* baseline, with *Care* and *Sanctity* as relatively isolated anchors and *Fairness*/*Loyalty*/*Authority* as a more tightly-grouped middle.

These results should not be read as a direct test of MFT as a model of human moral cognition. The observed proximity among *Fairness*, *Loyalty*, and *Authority*—and in particular its tightening in several Instruct variants—is best interpreted as a consequence of pretraining and alignment pressures that jointly reward compliance-oriented behaviors, rather than as evidence for or against the underlying theory. At the same time, the persistence of a distinct *Care* axis and a distinct *Sanctity* axis across nearly all 14 models suggests that some moral distinctions remain separable at the level of language alone. This distinction clarifies the scope of inference from our analysis: the results characterize how moral domains are reorganized in LLM representations under specific pretraining mixtures and post-training regimes, not how they originate in human moral cognition.

C SAEs: Implementation Details and Extended Analysis

To identify the specific mechanisms underlying moral representation, we followed the methodology in Chen et al. 2025 and computed the cosine similarity between the decoder directions of the SAEs and the foundation-specific concept vectors derived from the residual stream (Section 3.3).

Based on this metric, we selected the top-10 features with the highest similarity for each moral foundation on every layer to serve as the primary targets for analysis and intervention. To validate the semantics of these features, we randomly sampled 50,000 documents from the FineWeb dataset [Penedo et al., 2024] and retrieved the top-40 activating texts for each candidate feature. We extracted deduplicated evidence snippets centered around the tokens (token window size $\pm$ 64) with maximum activation and prompted `GPT-5.1` to generate structured semantic interpretations (Section C.2), ensuring the features meaningfully encoded concepts related to the target moral foundations.

C.1 Extended Analysis for The Anatomy of Morality

In this section, we provide a detailed analysis of the layer-wise evolution of moral features and their semantic grounding.

Layer-wise Alignment Dynamics To quantify feature-level representation, we computed the cosine similarity between MFT concept vectors and SAE decoder features. We also report a per-layer random baseline, which captures the expected alignment between SAE features and an arbitrary direction in activation space.

According to Figures 8a and 8b, observed alignments in `Llama` models substantially exceed the random baseline across all layers, indicating that the identified features encode non-random and semantically meaningful structure. Alignment is relatively low in early layers (Layer 3–11), consistent with findings from earlier work that these layers emphasizing syntactic or local contextual processing [Liu et al., 2024, Sajjad et al., 2022, Vulić et al., 2020]. A peak emerges at Layer 19 in `Llama-Base` and at Layer 15 in `Llama-Instruct` across all five foundations, where alignment is maximized and most strongly separated from the baseline. This pattern suggests that middle or late-middle layers act as a semantic bottleneck in `Llama` models, where moral concepts are most distinctly encoded at the feature level.

Notably, in `Llama-Instruct`, *Care* and *Sanctity* exhibit consistently higher alignment than other foundations, particularly *Loyalty* and *Authority*. Conversely, `Llama-Base` prioritizes *Loyalty* and *Fairness* over *Sanctity* and *Authority*, highlighting a clear inversion from its instruction-tuned variant. Overall, this disparity suggests that moral foundations are represented with varying degrees of clarity at the feature level, with some concepts aligning more cleanly with individual SAE features than others, which potentially reflects differences in representational fidelity or disentanglement driven by post-training.

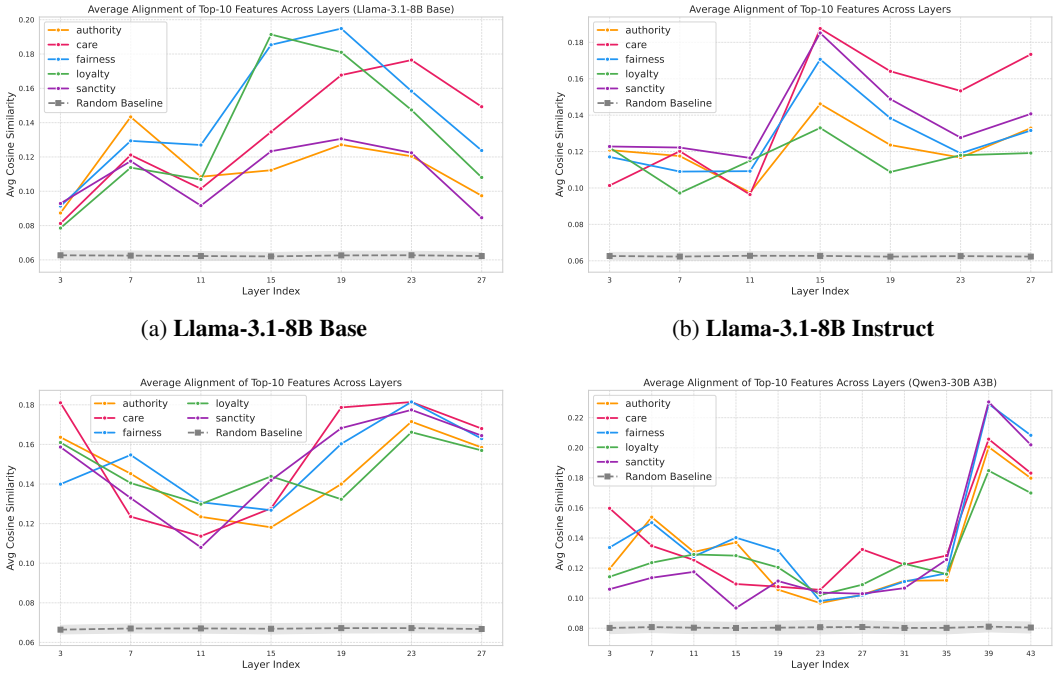
In contrast, the `Qwen` models exhibit a distinct "U-shaped" trajectory (Figures 8c and 8d). In `Qwen2.5-7B-Instruct`, alignment starts high in the early layers (Layer 3), particularly for *Authority* and *Care*, suggesting an early capture of surface-level moral semantics. This is followed by a significant dip in the middle layers (7–15). Finally, alignment resurges sharply in the deep layers, peaking at Layer 23, with *Fairness* reaching the highest separation. On the other hand, in `Qwen3-30B`, alignment remains low across Layers 7 to 35 with very few rises. However, alignment for all foundations increases sharply at Layer 39, especially for *Sanctity* and *Fairness*. The divergence—`Llama` models peaking in the middle versus `Qwen` models peaking at the boundaries—suggests that different architectures may sequence the "crystallization" of moral concepts differently, with `Qwen` models potentially disentangling these concepts during both initial processing and final readout preparation.

Semantic Grounding of SAE Features We qualitatively grounded these sparse features by analyzing their top-activating contexts and mapping them to MF categories using an LLM-assisted procedure with human validation (see Tables 8, 9, 10, and 11). Details of our human evaluation process are shown in Section E.

We find that SAE features decompose abstract foundations into granular, interpretable mechanisms. For example, in Table 8, *Care* features in `Llama-3.1-8B-Instruct` separate into distinct clusters tracking descriptions of physical suffering" (e.g., Feature L23.44965) and emotional distress" (e.g., Feature L19.90260). Similarly, *Authority* decomposes into features tracking "government regulatory frameworks" and "hierarchical role definitions".

When examining the base model, `Llama-3.1-8B-Base` (Table 10), we observe a high density of *Care* features that activate in developmental and protective contexts. Specifically, the model isolates "early childhood development and child-focused care" (Feature L7.114048) from broader "safety-oriented policy" (Feature L23.20681). Interestingly, we also identify two rare *Sanctity* features (e.g. Feature L15.18957) labeled as "Christian children's religious and moral education," suggesting that certain moral foundations are crystallized through specific cultural and educational lenses even in base models.

In `Qwen2.5-7B-Instruct` (Table 9), we observe a similar semantic granularity that mirrors the model's unique layer-wise trajectory. *Authority* features appear as early as Layer 3 (e.g., Feature L3.72227, with "mentions of government and national leaders"), providing a mechanistic explanation for the high geometric alignment observed in the model's initial layers. For *Care*, the model distin-



(a) **Llama-3.1-8B Base** (b) **Llama-3.1-8B Instruct**

(c) **Qwen2.5-7B-Instruct** (d) **Qwen3-30B-A3B**

Figure 8: **Layer-wise Alignment of Moral Features Across Models.** Average cosine similarity of the top-10 most aligned SAE features for each Moral Foundation across every 4 layers, compared against a random baseline (dashed grey line). Similarity is calculated between the SAE decoder weights and the corresponding Foundation vs. Social Norms concept vectors. The panels illustrate the differing representational trajectories for (a) `Llama-3.1-8B Base` (`Llama-Base` in the main text), (b) `Llama-3.1-8B Instruct` (`Llama-Instruct` in the main text), (c) `Qwen2.5-7B-Instruct`, and (d) `Qwen3-30B-A3B` (`Qwen3-30B` in the main text).

guishes between active condemnation of harm, such as "bullying and coercion" (Feature L15.130669), and abstract prosocial definitions, such as "empathy and compassion" (Feature L27.85517). We also identify distinct *Fairness* features related to corporate responsibility and business ethics" (Feature L3.123373), a specific domain of justice that appears less prominent in the `Llama` analysis.

Expanding this analysis to the larger `Qwen3-30B-A3B` architecture (Table 11), we observe highly-activated features representing distinct semantic facets across different depths of the network. Most notably, *Authority* features—particularly those tracking the "founding or creation of institutions" (e.g., Features L39.1016, L39.30227)—frequently emerge in the model's later layers. Meanwhile, the early layers capture other granular, highly-activated concepts in *Care*-related contexts like "public health and safety" (Feature L3.1733) and "nurturing framing" for learners (Feature L3.32310).

Across all four models, *Care* and *Authority* are the foundations most frequently associated with high-confidence semantic features. This indicates that while the model possesses sparse mechanisms for all foundations, the concepts of empathy and social regulation are the most robustly "crystallized" into detectable inner units.

C.2 Semantic Validation Implementation Details

To interpret the semantics of SAE features, we employed `GPT-5.1` (*GPT*) as an automated annotator following existing literature [Paulo et al., 2025, Cunningham et al., 2023, Bussmann et al., 2025]. The model was tasked with analyzing a set of top-activating text snippets for a given feature and generating a structured summary. We prioritized a "conservative" annotation strategy: the model was explicitly instructed to first identify neutral semantic patterns and only assign a Moral Foundations Theory (MFT) label if the evidence was strong.

Prompt construction. The prompt consists of three components: (1) MFT definitions, (2) feature metadata and evidence snippets, and (3) a strict output schema.

System instructions. The model was invoked with temperature $T = 0$. The prompt provided the standard definitions for the five moral foundations [Haidt, 2012] and specific instructions to avoid forcing moral interpretations on non-moral features. The exact text provided to the model is detailed in Figure 9.

Semantic Interpretation Prompt

Role: You are interpreting a sparse autoencoder (SAE) feature from an LLM.
Goal: Infer the most likely semantic pattern that triggers the feature, based ONLY on the evidence snippets.
Instructions:

1. **Neutral Description First:** Describe the dominant pattern (topic, style, rhetorical function, or social behavior) neutrally.
2. **Conservative MFT Mapping:** Map to a Moral Foundations Theory category *only* if strongly supported. Otherwise, output `mft_alignment="none"`. Do not force morality; many features are not moral.
3. **Format:** Provide a short label (5–10 words) and a 1–2 sentence description.
4. **Citations:** Cite `evidence_ids` (indices of snippets) that justify your decision.

Moral Foundations Theory (MFT) definitions:

- **Care/harm:** dislike others' suffering; kindness, gentleness, nurturance vs cruelty, violence.
- **Fairness/cheating:** justice, rights, autonomy vs fraud, exploitation, cheating.
- **Loyalty/betrayal:** group allegiance, patriotism, self-sacrifice vs betrayal, treason, disloyalty.
- **Authority/subversion:** respect for legitimate authority, leadership/followership, traditions vs defiance, disrespect, subversion.
- **Sanctity/degradation:** purity, elevation above the carnal, disgust sensitivity vs degradation, contamination, depravity.

[Insert Feature Metadata JSON]
[Insert Evidence Snippets (index: text)]

Figure 9: Prompt template used for automated interpretation of SAE features. Features are selected to have the 10 highest cosine similarity with a corresponding moral foundation concept vectors at the same layer.

Output schema. We constrained the model to output a valid JSON object matching the schema in Table 7. This structured output facilitates downstream quantitative analysis of the feature directions.

LLM-grounded semantic characterization. Tables 8, 9, 10, and 11 report SAE features whose top-activating FineWeb [Penedo et al., 2024] contexts support a coherent semantic interpretation, with an optional MFT assignment. Due to page limits, we trimmed *GPT* outputs (long descriptions, rationales, MFT polarity, and evidence IDs). We find that *GPT* most confidently identifies features associated with *Care*/harm, *Authority*/subversion, and *Fairness*/cheating, and also identifies a smaller number of *Sanctity*/degradation features. In contrast, we do not obtain high-confidence `Loyalty`/Betrayal assignments in the current semantic-mining pass. We attribute this to the limited size of the validation sample (50,000 documents) and the resulting sparsity of diagnostically relevant top-activation contexts under a fixed compute budget, rather than to an absence of Loyalty-related signal in the model. Importantly, these LLM-grounded summaries are used to qualitatively ground feature semantics and present representative exemplars; they complement (and do not replace) our causal steering evaluations in Section 4.2.3, which indicate that the identified SAE features contain foundation-relevant moral signals.

D Steering: Implementation Details and Extended Analysis

D.1 Implementation Details

Steering conditions. We evaluate two steering granularities by varying the choice of $\vec{v}_{\text{steer}}$. In **macro-steering**, we steer along the (debiased) foundation vector $\vec{v}_k$ from Section 3.1 to test whether the global moral direction is sufficient to induce targeted behavioral change. In **micro-steering**, we steer along a composite direction built from the SAE features identified in Section 3.3. For foundation k at layer ℓ, we take the Top-K (K=10) decoder directions $\{\mathbf{d}_i\}_{i\in\mathcal{F}_{k,\ell}}$ ranked by cosine similarity with $\hat{\mathbf{v}}_{k,\ell}$, ℓ_2-normalize each, sum them, and re-normalize the result to unit norm:

$$\mathbf{v}_\ell^{\text{steer}} = \frac{\sum_{i\in\mathcal{F}_{k,\ell}} \mathbf{d}_i/\|\mathbf{d}_i\|_2}{\left\|\sum_{i\in\mathcal{F}_{k,\ell}} \mathbf{d}_i/\|\mathbf{d}_i\|_2\right\|_2}. \tag{19}$$

This tests whether the sparse mechanisms aligned with a foundation can jointly produce behavioral effects comparable to the dense vector. All steering experiments use the MFV-derived foundation vectors established in Section 3.1, identical to those validated against ecological Reddit-labeled content in Section 4.2.1; no re-fitting or instrument-specific tuning is performed for steering.

Intervention site and layers. We register a steering hook at every layer and sweep all layers; the per-foundation best layer L^* is then selected by the linear-response criterion described below. Because all readouts are logits-based (see *Behavioral assessment*), the intervention is applied in a single forward pass over the prompt, and the readout is taken at the final prompt position T. At the selected layer $\ell = L^*$, we intervene on the residual stream as

$$\mathbf{h}'_{\ell,T} = \mathbf{h}_{\ell,T} + \alpha_\ell\, \mathbf{v}_\ell^{\text{steer}}, \tag{20}$$

where $\alpha_\ell \in \mathbb{R}$ is a signed, layer-specific steering coefficient (positive α_ℓ amplifies the foundation, negative α_ℓ suppresses it; $\alpha_\ell = 0$ recovers the unsteered baseline) and $\mathbf{v}_\ell^{\text{steer}}$ is an ℓ_2-normalized steering direction. For **macro-steering**, we set $\mathbf{v}_\ell^{\text{steer}} = \mathbf{v}_\ell$, the layer-wise moral-foundation vector from Section 3.1. For **micro-steering**, we set $\mathbf{v}_\ell^{\text{steer}}$ equal to the unit-normalized sum of the decoder directions of the Top-K SAE features selected.

Norm-calibrated relative perturbation ρ. A naive fixed α-grid is not comparable across models because residual-stream activation magnitudes vary by orders of magnitude. The same $\alpha = 2$ therefore corresponds to a small relative perturbation in models with large residual norms and a much larger one in models with smaller norms, making cross-model comparison meaningless. To address this, we reparameterize the steering coefficient as a unit-free *relative perturbation ratio*

$$\rho \;:=\; \frac{\alpha_\ell}{\bar{r}_\ell} \;\in\; \mathbb{R}, \tag{21}$$

which measures the signed injected perturbation as a fraction of the layer's natural residual-stream magnitude, and is therefore comparable across models and architectures. We sweep a single fixed grid for all models:

$$\rho \in \{-0.20,\, -0.10,\, -0.05,\, 0,\, +0.05,\, +0.10,\, +0.20\}, \tag{22}$$

with the per-layer steering coefficient computed as $\alpha_\ell = \rho \cdot \bar{r}_\ell$. The grid is bounded at $|\rho| \leq 0.20$ to remain in a capability-preserving regime. We verify capability preservation by measuring 5-shot MMLU on 11 of the 14 models tested under macro-steering (Table 3) (all 7 instruction-tuned variants plus 4 of 7 base models; the three largest base models—`Llama-3.1-70B-Base`, `Qwen2.5-32B-Base`, `Qwen3-30B-A3B-Base`—are excluded due to compute cost, and base-model MMLU is in any case a less informative capability probe since untuned base models exhibit noisy instruction-following baselines). Across these 11 models, MMLU degradation at $|\rho| = 0.20$ stays within approximately 2 percentage points of the unsteered baseline, with the worst cases occurring on `Qwen2.5-7B-Base` (-2.15 at $\rho = -0.20$) and `Mistral-7B-Base` (-2.25 at $\rho = +0.20$); all other models stay within 1pp. Beyond the main grid, degradation grows rapidly: at $|\rho| = 0.40$, multiple models exceed 4pp loss (`Mistral-7B-Base`: -4.65 at $\rho = -0.40$; `Qwen2.5-7B-Base`: -5.20). For micro-steering, we report MMLU on 4 models (Table 4); degradation is comparable to macro for `Qwen2.5-7B-Instruct`, `Llama-3.1-8B-Base`, and `Qwen3-30B-A3B-Base` (≤ 1.3pp at $|\rho| = 0.20$), but is notably larger on `Llama-3.1-8B-Instruct` (-3.40 at $\rho = -0.20$, -3.45 at

$\rho = +0.20$), reflecting that the top-10 composite direction on Instruct `Llama` more strongly perturbs the residual stream than the dense vector intervention does. The grid bound at $|\rho| = 0.20$ is therefore chosen as the largest perturbation magnitude that preserves task capability across all macro-steering models tested. The $\rho = 0$ baseline is implemented by short-circuiting the steering hook (no $0 \cdot \mathbf{v}$ injection), which avoids fp16 round-off and yields exact unsteered baseline scores.

Calibration procedure for $\bar{r}_\ell$. For each model, we measure $\bar{r}_\ell$ once before any steering experiments. We randomly sample 10,000 documents from FineWeb [Penedo et al., 2024], register a forward hook on each transformer block output, run a forward pass over the sampled documents, extract the last-token residual activation $\mathbf{h}_{\ell,T}$, and compute its ℓ_2 norm. We average across documents to obtain $\bar{r}_\ell$ for every layer. Calibrated values for all 14 models (7 model families $\times$ {base, instruct}) are released alongside our code.

Two roles for α and ρ. The two coefficients serve distinct, complementary purposes:

- α **is comparable across layers within a single model.** Since α is the absolute perturbation magnitude in residual-stream coordinates, it provides a unified physical scale for layer-wise comparison within one model. We use α for *best-layer selection*.
- ρ **is comparable across models and across readout instruments.** Since ρ normalizes by the model-specific $\bar{r}_\ell$, it removes architecture-induced scale differences and is calibrated on a generic web distribution rather than on any specific readout (Appendix D.1); the same ρ-grid is therefore directly meaningful when applied to MFQ-2, PVQ-21, political ideology, and MMLU. We use ρ for *cross-model and cross-readout comparison* of steering strength.

The two slope estimands are related by the identity $k^\rho = k_\alpha \cdot \bar{r}_\ell$, where k_α and k^ρ are the response slopes of the readout score with respect to α and ρ respectively (defined formally below). For in-domain MFQ-2 effects we report k^ρ at each model's best layer L^* (Section 4.2.3); for cross-framework transfer to Schwartz PVQ-21 and political ideology, we use the same L^* selected from the within-construct MFQ contrast and report transfer slopes k_{transfer} under the same parameterization (Appendix D.4).

Behavioral assessment. We measure behavioral effects along three complementary axes. **MFQ-2** (see Section A.2) provides the within-construct readout: under macro- and micro-steering with each foundation vector in foundation k, we obtain a foundation-level score S_k for the matched dimension. **PVQ-21** (Schwartz Portrait Values Questionnaire [Schwartz et al., 2003]) and a **single-item political ideology** measure (1–7 Likert, very liberal to very conservative) serve as cross-framework readouts grounded in distinct theoretical frameworks. We present their theoretical motivation, scoring details, and results in Section D.4. All three readouts are scored from output logits at the prompt's final position via option-token probabilities (5-way for MFQ-2 and PVQ-21, 7-way for ideology), yielding deterministic expected-value scores. The rollout dimension in our pipeline is preserved for compatibility with future sampling-based evaluation and does not contribute variance to reported scores.

D.1.1 Steering Response and Best-Layer Selection

Linear response regression. For each (f, ℓ) we fit ordinary least-squares regressions of the baseline-subtracted score $\Delta S_{f,\ell}(\rho) = S_{f,\ell}(\rho) - S_{f,\ell}(0)$ against both parameterizations:

$$\Delta S_{f,\ell}(\alpha) = k^\alpha_{f,\ell} \cdot \alpha + \varepsilon_\alpha, \tag{23}$$

$$\Delta S_{f,\ell}(\rho) = k^\rho_{f,\ell} \cdot \rho + \varepsilon_\rho. \tag{24}$$

The two slopes are related by the change-of-variable identity $k^\rho_{f,\ell} = k^\alpha_{f,\ell} \cdot \bar{r}_\ell$ (Appendix A.3). We additionally record the coefficient of determination $R^2_{f,\ell}$ and the regression p-value $p_{f,\ell}$ for each fit.

Best-layer selection (within-model). We select the per-foundation best layer L^*_f by maximizing the signed slope in the α parameterization, subject to a linear-response quality filter:

$$L^*_f \;=\; \arg\max_{\ell \in \mathcal{L}^{\text{valid}}_f} \; k^\alpha_{f,\ell}, \quad \mathcal{L}^{\text{valid}}_f = \{\ell : R^2_{f,\ell} \geq 0.9 \;\wedge\; p_{f,\ell} \leq 0.05 \;\wedge\; k^\alpha_{f,\ell} > 0\}. \tag{25}$$

Two design choices warrant comment. *Why k^α rather than k^ρ for layer selection.* The slope identity $k^\rho_{f,\ell} = k^\alpha_{f,\ell} \cdot \bar{r}_\ell$ implies that k^ρ is monotonically inflated at deeper layers where $\bar{r}_\ell$ is larger, which would systematically bias L^* toward the final layers regardless of the underlying response strength. Selecting on k^α provides a within-model layer-comparable physical scale that does not bake in this depth bias.

Why the quality filter is essential. Without the R^2 and p-value constraints, shallow layers with small $\bar{r}_\ell$ exhibit a narrow α-range over the ρ-grid, and noise can produce spurious high-magnitude slopes with poor linear fit. Without filtering, L^* for some foundations would be drawn to layer 0 or 1 with $R^2 < 0.85$. With filtering, selected layers consistently fall in the mid-to-late layer range.

Fallback for cells failing the quality filter. We adopt a three-tier selection criterion:

1. **Primary**: $L^*_f = \arg\max_\ell k^\alpha_{f,\ell}$ subject to $\mathcal{L}^{\text{valid}}_f$ above;
2. **First fallback**: if no layer satisfies the primary filter, $L^*_f = \arg\max_\ell k^\alpha_{f,\ell}$ among layers with $k^\alpha_{f,\ell} > 0$;
3. **Second fallback**: if all layers exhibit $k^\alpha_{f,\ell} \leq 0$ (the foundation vector responds in the wrong direction at every layer), $L^*_f = \arg\max_\ell |k^\alpha_{f,\ell}|$.

Cells whose L^*_f is selected via fallback are flagged in our reporting (hatched in Figure 3a of the main text) and excluded from quantitative claims about successful steering. Across the 14-model panel, the primary criterion is satisfied for $68/70$ (model, foundation) pairs; the two exceptions occur on `Qwen3-30B-MoE-Instruct` (*Care*, *Authority*).

D.2 Steering Case Study: Dose–Response Curves

Steering case study on `Llama-3.1-8B-Instruct`. The main-text steering results (Figure 3) summarize behavioral effects across the 14-model panel by reporting per-cell slopes k^ρ_{f,L^*}. This appendix expands one representative cell of that summary—`Llama-3.1-8B-Instruct`—to show the underlying dose–response curves at each foundation's best layer, illustrating both the linear-response shape that justifies reporting a single slope per cell and the foundation-specific asymmetries that the heatmap aggregates over.

Figures 10a and 10b report macro-steering and micro-steering response curves on `Llama-3.1-8B-Instruct` under the evaluation protocol of Section 4.2.3. Three patterns emerge.

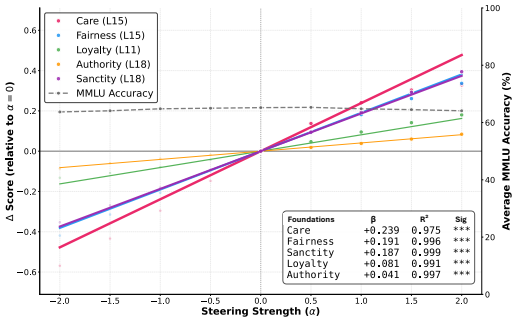
Macro-steering exhibits a clear cross-foundation hierarchy. *Care*, *Fairness*, and *Sanctity* show strong, monotonic responses well-approximated by linear trends ($R^2 > 0.97$): *Care* is most sensitive ($k^\alpha = 0.239$), followed by *Fairness* ($k^\alpha = 0.191$) and *Sanctity* ($k^\alpha = 0.187$). In contrast, *Loyalty* ($k^\alpha = 0.081$) and especially *Authority* ($k^\alpha = 0.041$) are markedly attenuated, with *Authority*'s slope below 20% of *Care*'s. The same qualitative hierarchy is reflected in the 14-model heatmap of Figure 3 (Panel A) for the `Llama-3.1-8B`.

Micro-steering recovers slope on the alignment-inertial foundations. Intervening along the decoder directions of the top-10 SAE features (Section 4.2.2) produces non-trivial behavioral slopes on *Authority* and *Loyalty* where macro-steering is nearly flat, consistent with the main-text scatter showing micro slopes systematically exceeding macro slopes ($17/20$ cells above $y = x$, OLS slope 2.57; Figure 3 Panel B). This pattern suggests that compliance-related moral signals are not *erased* by alignment but are partially *submerged* into the Social Norm direction; sparse-feature interventions recover causal access to circuit components that the dense vector averages out.

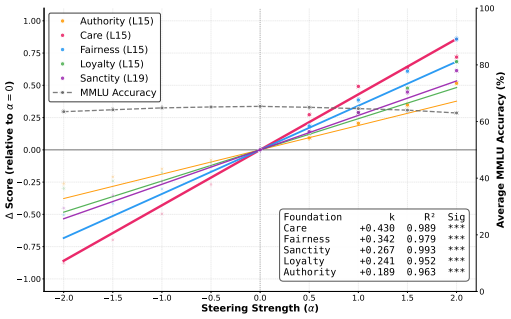
Linear response is a strong descriptive model in both parameterizations. For both macro- and micro-steering, all five foundations satisfy the linear-response quality filter ($R^2 \geq 0.90$, $p \leq 0.05$, $k^\alpha > 0$; Appendix D), validating the reduction of full dose–response curves to a single slope k^ρ in the main-text summary.

D.3 General Performance Measurement after Steering

To quantify whether moral steering affects the model's general capability, we evaluate steered models on the MMLU benchmark [Hendrycks et al., 2020] following previous work [Chen et al., 2025, Li and Zhao, 2026], which covers 57 subjects spanning STEM, humanities, and social sciences.



(a) **Macro (vector) steering response curves.** Steering along foundation-level concept vectors v_f.



(b) **Micro (SAE) steering response curves.** Steering along the top-10 SAE feature decoder directions for each foundation.

Figure 10: **Steering response curves on `Llama-3.1-8B-Instruct`.** For each foundation, we plot the MFQ-2 subscale change $\Delta\text{Score}(\alpha) = S_f(L^*, \alpha) - S_f(L^*, 0)$ relative to the unsteered baseline as a function of steering strength α, evaluated at that foundation's best layer L^* (Appendix D). Points show measured Δ scores and the solid line shows the corresponding linear fit. The gray dashed line reports MMLU under the same interventions, indicating capability preservation across the ρ-grid (Appendix D.3).

Evaluation set. We randomly sample $n = 2000$ questions from the MMLU test set using a fixed seed (seed = 42) to ensure that the same question set is used across all steering conditions and across all models.

Steering configuration. We evaluate each model under the same inference-time intervention used in our main steering experiments. Steering is applied at each model's selected best layer L^* (Section D), with the relative perturbation ratio swept over an extended grid

$$\rho \in \{-0.4,\ -0.3,\ -0.2,\ -0.1,\ 0,\ +0.1,\ +0.2,\ +0.3,\ +0.4\},$$

which extends beyond the main-text grid ($|\rho| \leq 0.20$) by an additional ± 0.2 on each side, allowing us to characterize the onset of capability degradation. For each ρ, the absolute coefficient is $\alpha_\ell = \rho \cdot \bar{r}_\ell$, the steering direction is the residual-stream normalized foundation concept vector, and the calibration $\bar{r}_\ell$ is identical to that used in the steering tasks.

Logit-based multiple-choice scoring. To obtain deterministic and reproducible measurements, we score MMLU in a logit-based manner rather than via free-form generation. For each question we perform a single forward pass and extract the logits of the option tokens (A, B, C, D) at the final position. We then compute a softmax over these four logits to obtain option probabilities,

$$P(\text{option}_i) = \frac{\exp(\text{logit}_i)}{\sum_{j \in \{A,B,C,D\}} \exp(\text{logit}_j)}, \tag{26}$$

and predict the answer by

$$\hat{y} = \arg\max_{i \in \{A,B,C,D\}} P(\text{option}_i). \tag{27}$$

Accuracy at each ρ is computed as the fraction of correct predictions over the $n = 2000$ questions, and we report the change relative to the unsteered baseline ($\rho = 0$): $\Delta\text{MMLU}(\rho) = \text{Acc}(\rho) - \text{Acc}(0)$, in percentage points.

Coverage. We evaluate 11 of the 14 models from our main steering experiments. The three excluded models (`Llama-3.1-70B-Base`, `Qwen2.5-32B-Base`, `Qwen3-30B-A3B-Base`) are omitted due to compute cost; base-model MMLU is in any case a less informative capability probe, since untuned base models exhibit noisy instruction-following baselines. The 11 evaluated models include all 7 instruction-tuned variants and 4 of the 7 base models, giving complete coverage of every architectural family at the instruction-tuned level.

Table 3: **MMLU change under macro-steering across the extended ρ grid.** Each cell reports $\Delta\mathrm{MMLU} = \mathrm{Acc}(\rho) - \mathrm{Acc}(0)$ in percentage points, where $\mathrm{Acc}(0)$ is the unsteered baseline. Steering is applied at each model's selected best layer L^*. The shaded region ($|\rho| \leq 0.20$) marks the main-text steering range; degradation within this range does not exceed 2.25 percentage points in any of the 11 evaluated models. The three largest base models (`Llama-3.1-70B-Base`, `Qwen2.5-32B-Base`, `Qwen3-30B-A3B-Base`) are excluded due to compute cost. Baseline MMLU accuracies are reported in the second column.

Model	L^*	Baseline	ΔMMLU at $\rho =$								
			-0.4	-0.3	-0.2	-0.1	0	$+0.1$	$+0.2$	$+0.3$	$+0.4$
Instruction-tuned models											
`Llama-3.1-70B-Instruct`	28	79.10	-0.15	-0.30	-0.20	$+0.00$	$+0.00$	-0.20	$+0.10$	$+0.30$	$+0.40$
`Llama-3.1-8B-Instruct`	14	62.90	-2.85	-2.00	-0.60	$+0.15$	$+0.00$	$+0.00$	-0.30	-0.90	-0.70
`Qwen2.5-32B-Instruct`	32	80.95	-0.25	-0.20	-0.10	$+0.00$	$+0.00$	-0.05	-0.20	-0.15	-0.15
`Qwen2.5-14B-Instruct`	15	75.15	$+0.05$	-0.05	$+0.05$	-0.05	$+0.00$	$+0.00$	$+0.10$	$+0.15$	$+0.00$
`Qwen2.5-7B-Instruct`	3	70.40	$+0.25$	$+0.15$	$+0.20$	$+0.10$	$+0.00$	$+0.15$	$+0.05$	$+0.10$	$+0.10$
`Qwen3-30B-A3B-Instruct`	24	26.45	$+0.25$	$+0.35$	$+0.85$	-0.20	$+0.00$	-0.10	-0.80	-0.35	$+0.45$
`Mistral-7B-Instruct-v0.3`	11	57.65	-0.90	-0.55	-0.15	-0.15	$+0.00$	$+0.15$	$+0.30$	$+0.40$	$+0.30$
Base models											
`Llama-3.1-8B-Base`	14	60.25	-0.65	-0.40	-0.15	-0.10	$+0.00$	$+0.30$	$+0.15$	$+0.50$	-0.15
`Qwen2.5-14B-Base`	25	76.65	-1.25	-1.10	-0.90	-0.40	$+0.00$	$+0.05$	$+0.35$	-0.20	-0.80
`Qwen2.5-7B-Base`	18	68.90	-5.20	-3.45	-2.15	-0.95	$+0.00$	$+0.60$	$+0.05$	-0.60	-1.30
`Mistral-7B-v0.3 (Base)`	31	56.40	-4.65	-3.00	-1.85	-0.30	$+0.00$	-1.30	-2.25	-3.50	-4.35

Table 4: **MMLU change under micro-steering across the extended ρ grid.** Each cell reports $\Delta\mathrm{MMLU} = \mathrm{Acc}(\rho) - \mathrm{Acc}(0)$ in percentage points, where $\mathrm{Acc}(0)$ is the unsteered baseline. Steering is applied at a selected layer L^*. The shaded region ($|\rho| \leq 0.20$) marks the main-text steering range. Baseline MMLU accuracies are reported in the second column.

Model	L^*	Baseline	ΔMMLU at $\rho =$						
			-0.20	-0.10	-0.05	0.00	+0.05	+0.10	+0.20
Instruction-tuned models									
`Llama-3.1-8B-Instruct`	16	65.20	-3.40	-1.06	-0.17	$+0.00$	-0.39	-1.56	-3.45
`Qwen2.5-7B-Instruct`	16	70.35	-0.14	-0.05	$+0.03$	$+0.00$	$+0.10$	$+0.08$	$+0.06$
Base models									
`Llama-3.1-8B-Base`	16	63.45	-1.30	-0.57	-0.19	$+0.00$	-0.09	-0.03	-0.05
`Qwen3-30B-A3B (Base)`	24	76.25	-0.32	-0.26	-0.09	$+0.00$	$+0.01$	-0.09	-0.14

Results. Table 3 reports per-model ΔMMLU across the full ρ grid. Within the main-text steering range ($|\rho| \leq 0.20$), MMLU degradation does not exceed 2.25 percentage points in any of the 11 models tested, and is below 1.0 point in 9 of 11 models. The two largest within-range degradations occur in `Mistral-7B-Base` (-2.25 at $\rho = +0.20$ and -1.85 at $\rho = -0.20$) and `Qwen2.5-7B-Base` (-2.15 at $\rho = -0.20$); `Llama-3.1-8B-Instruct` loses only 0.60 points within range but 2.85 at $\rho = -0.40$. Instruction-tuned models are systematically more capability-robust than their base counterparts (e.g., `Qwen2.5-7B-Instruct` fluctuates within ± 0.25 across the entire range while its base counterpart loses up to 5.20 points), consistent with the view that instruction tuning produces more redundant residual-stream representations less perturbable by single-direction interventions. The grid bound $|\rho| \leq 0.20$ used in our main-text experiments thus operates safely within the capability-preserving regime for all 11 evaluated models.

Micro-steering results. Table 4 reports ΔMMLU under micro-steering for the 4 models in our SAE feature analysis. MMLU degradation stays within 3.5 percentage points of the unsteered baseline across all 4 models within the main-text range $|\rho| \leq 0.20$, with three of the four models (`Qwen2.5-7B-Instruct`, `Llama-3.1-8B-Base`, `Qwen3-30B-A3B-Base`) staying within 1.5 points across the entire grid. As with macro-steering, micro-steering at $|\rho| \leq 0.20$ thus operates within a capability-preserving regime.

D.4 Cross-Construct Transfer

The main-text steering experiment establishes that MFV-derived concept vectors causally control MFQ-2 responses across the 14-model panel. As a complementary analysis, we test whether the same intervention—applied along a between-foundation contrast vector—also produces theoretically aligned shifts on behavioral readouts from *distinct* theoretical frameworks: the Schwartz Portrait Values Questionnaire (PVQ-21; Schwartz et al., 2003) and a single political-ideology item [Graham et al., 2011]. The Schwartz value system and political ideology are independently grounded constructs whose predicted relationships to the MFT Individualizing–Binding axis are well-established in moral psychology [Graham et al., 2011, Atari et al., 2023], providing an external anchor against which the construct-level reach of our steering intervention can be characterized.

Why a between-foundation contrast vector. The foundation-vs.-Social-Norm vectors used in the main steering experiment (Section 3.1, Eq. 2 with $\mathcal{B} = \textit{Social Norm}$) encode each foundation's *presence* relative to a non-moral baseline, but do not encode a directional axis along the Individualizing–Binding pole that Schwartz values and political ideology are predicted to track. Steering along $\hat{v}_{\text{Care vs SN}}$, for instance, modulates how Care-coded a model's responses are, not whether the model expresses more individualizing or more binding values overall. We therefore use the between-foundation contrast vector $\hat{v}_{\text{Care}-\text{Authority},\ell}$ (Section 3.1, Eq. 2 with $\mathcal{A} = \textit{Care}$ inputs and $\mathcal{B} = \textit{Authority}$ inputs, then ℓ_2-normalized per Eq. 3), which by construction runs from Authority-anchored activations toward Care-anchored activations and therefore carries the directionality required for transfer testing on the Individualizing–Binding axis. By convention, $\rho > 0$ along this contrast pushes toward *Care*/Individualizing and $\rho < 0$ toward *Authority*/Binding.

Procedure. We use the same ρ-parameterized intervention as the main steering experiment (Appendix D), with $\rho > 0$ pushing toward Care/Individualizing and $\rho < 0$ toward Authority/Binding. The target layer L^* is selected from the within-construct anchor $\Delta_{\text{MFQ}}(\ell, \rho) = \text{Score}_{\text{Care}}(\ell, \rho) - \text{Score}_{\text{Authority}}(\ell, \rho)$ via the same three-tier criterion as L^*-selection in the main steering experiment (Appendix D). At the fixed L^* we then score two external readouts under the same hook: PVQ-21 (21 items, 1–6 Likert, MRAT-centered, aggregated to 10 basic values and 4 higher-order clusters following Schwartz's standard pipeline) and the 7-point ideology item (single-token logits readout, 1 = Very liberal, 7 = Very conservative). For each (model, outcome) we use the 7 deterministic ρ-grid datapoints at L^* and fit $y(\rho) = \beta_0 + k_{\text{transfer}} \cdot \rho + \varepsilon$.

Results. Table 5 reports k_{transfer} on the 7 critical outcomes for all 14 models. As an ablation supporting the main-text steering finding (Section 4.2.3), we evaluate whether the same intervention generalizes beyond MFQ-2 to two additional readout instruments: the Schwartz PVQ-21 value system and political ideology.

The 6 Schwartz outcomes constituting the MFT–Schwartz Individualizing/Binding reflection (Universalism, Benevolence, Higher-Order Self-Transcendence positive; Tradition, Conformity, Higher-Order Conservation negative) achieve sign-correctness rates between $10/14$ and $11/14$, for a mean of 76% across the 84 model-outcome cells. The political-ideology readout is weaker ($4/14$, 29%), consistent with the higher noise of a single Likert item and with RLHF suppression of explicit political self-positioning in instruction-tuned models.

At the model level, 4 of 14 models achieve perfect $7/7$ sign-correctness (`Qwen2.5-7B-Base`, `Qwen2.5-14B-Base`, `Qwen2.5-32B-Base`, `Mistral-7B-Instruct`); all four `Llama-3.1` variants achieve $6/7$ (failing only on ideology); and the remaining failures concentrate in models already characterized as steerability-limited or axis-fit-failed in the main steering analysis (`Mistral-7B-Base`, the two `Qwen3-30B-MoE` variants, and `Qwen2.5-32B-Instruct`; cf. Figure 3A). The transfer behavior thus mirrors the pattern of the main steering experiment: models for which the MFV-derived vector causally controls MFQ-2 responses are the same models for which it produces theoretically aligned shifts on independent value-system readouts. This supports the interpretation that the main-text steering effects (Section 4.2.3) reflect construct-level interventions on moral-cognitive representations.

Table 5: **Cross-construct transfer slopes k_{transfer} at L^* on 7 critical outcomes (14 models).** Predicted signs under $\rho > 0$ (Care direction): Universalism, Benevolence, and Higher-Order Self-Transcendence should increase; Tradition, Conformity, Higher-Order Conservation, and Ideology should decrease. Each cell shows k_{transfer} followed by ✓ if the sign matches the theoretical prediction or × otherwise; $k \approx 0$ counts as no-match. Per-outcome correctness across the 14 models is reported in the bottom row.

Model	Univ↑	Benev↑	HO_ST↑	Trad↓	Conf↓	HO_Cons↓	Ideol↓	#ok/7
`Llama-3.1-8B-Instruct`	$+0.354$ ✓	$+0.367$ ✓	$+0.361$ ✓	-0.677 ✓	-1.205 ✓	-0.515 ✓	$+0.061$ ×	6/7
`Llama-3.1-8B-Base`	$+0.193$ ✓	$+0.099$ ✓	$+0.146$ ✓	-0.254 ✓	-0.231 ✓	-0.170 ✓	$+0.156$ ×	6/7
`Llama-3.1-70B-Instruct`	$+0.165$ ✓	$+0.211$ ✓	$+0.188$ ✓	-0.231 ✓	-0.173 ✓	-0.103 ✓	-0.000 ×	6/7
`Llama-3.1-70B-Base`	$+0.014$ ✓	$+0.078$ ✓	$+0.046$ ✓	-0.126 ✓	-0.070 ✓	-0.043 ✓	$+0.189$ ×	6/7
`Qwen2.5-7B-Instruct`	$+0.040$ ✓	$+0.015$ ✓	$+0.027$ ✓	-0.143 ✓	$+0.004$ ×	-0.051 ✓	$+0.000$ ×	5/7
`Qwen2.5-7B-Base`	$+0.912$ ✓	$+0.982$ ✓	$+0.947$ ✓	-0.914 ✓	-1.085 ✓	-1.048 ✓	-0.683 ✓	7/7
`Qwen2.5-14B-Instruct`	$+0.075$ ✓	$+0.148$ ✓	$+0.112$ ✓	$+0.066$ ×	-0.044 ✓	-0.004 ✓	$+0.000$ ×	5/7
`Qwen2.5-14B-Base`	$+0.119$ ✓	$+0.001$ ✓	$+0.060$ ✓	-0.149 ✓	-0.072 ✓	-0.127 ✓	-0.593 ✓	7/7
`Qwen2.5-32B-Instruct`	-0.022 ×	-0.021 ×	-0.022 ×	-0.023 ✓	-0.008 ✓	-0.018 ✓	$+0.000$ ×	3/7
`Qwen2.5-32B-Base`	$+0.577$ ✓	$+0.657$ ✓	$+0.617$ ✓	-0.478 ✓	-0.197 ✓	-0.315 ✓	-0.672 ✓	7/7
`Qwen3-30B-A3B-Instruct`	$+0.084$ ✓	-0.231 ×	-0.074 ×	-0.236 ✓	$+1.090$ ×	$+0.340$ ×	$+0.136$ ×	2/7
`Qwen3-30B-A3B-Base`	-0.074 ×	$+0.052$ ✓	-0.011 ×	$+0.070$ ×	$+0.078$ ×	$+0.008$ ×	$+0.145$ ×	1/7
`Mistral-7B-Instruct-v0.3`	$+0.207$ ✓	$+0.163$ ✓	$+0.185$ ✓	-0.234 ✓	-0.252 ✓	-0.117 ✓	-0.260 ✓	7/7
`Mistral-7B-v0.3 (Base)`	-0.060 ×	-0.075 ×	-0.067 ×	$+0.108$ ×	$+0.061$ ×	$+0.063$ ×	$+0.085$ ×	0/7
Per-outcome #correct	11/14	11/14	10/14	11/14	10/14	11/14	4/14	—

E Human Evaluation

Our study incorporated human evaluation at two points in the overall design:

First, for the **Semantic Validation** described in Section 3.3, we used `GPT-5.1` to interpret SAE features, following prior work on LLM-based feature interpretation [Cunningham et al., 2023, Paulo et al., 2025]. **Three psychology experts and two CS students**, all with expertise in Moral Foundations Theory (MFT), then verified these interpretations. We sampled the top-10 activating features every four model layers and reviewed all features in this set that `GPT-5.1` assigned to a moral foundation. For each feature, annotators examined the top-activating texts from the FineWeb corpus and checked whether they expressed the assigned moral foundation according to standard MFT definitions. Annotators also corrected semantic misalignments in approximately 8% of cases. In rare cases, `GPT-5.1` generated interpretations that referred to content not present in the text provided to the model. We excluded these interpretations from the final tables (Tables 8, 10, 9, and 11).

Second, to further support validity of `gpt-5-mini` generated MFV vignettes in Section 4.1, we ran an online survey to directly compare the LLM-generated vignettes with the original vignettes: We conducted a randomized human evaluation in which each participant rated 30 vignettes drawn from two randomly selected moral foundations, with each set containing a balanced mix of original and LLM-generated items. For each vignette, participants indicated perceived moral wrongness on a 5-point Likert scale (1 = "Not at all wrong" to 5 = "Extremely wrong"), assessing how well the content reflected the target moral domain. In total, we collected 428 ratings for original vignettes and 850 ratings for generated vignettes, spanning all nine moral foundation subcategories, from 39 participants (mainly CS and psychology PhD students). Random assignment of foundations and items ensured broad coverage and minimized systematic bias. Because participants evaluated both original and generated items within each foundation, we computed participant-level averages for each foundation-condition combination, yielding 77 valid participant–foundation pairs for analysis.

Original vignettes received a mean rating of 3.45 (SD = 1.11), compared to 3.31 (SD = 1.03) for LLM-generated vignettes. A paired-samples t-test indicated that this difference was not statistically significant, t(76) = 1.65, p = .103, with a small effect size (d = 0.19, 95% CI [-0.06, 0.44]). To assess practical equivalence, we conducted a Two One-Sided Tests (TOST) analysis. We defined the equivalence margin as a small-to-medium effect (Cohen's $d = 0.40$). Using the standard deviation of the paired differences (SD = 0.747), this corresponds to a raw equivalence bound of $\Delta = \pm 0.299$. Both one-sided tests were significant (lower: $t(76) = 5.15$, $p < .001$; upper: $t(76) = -1.87$, $p = .033$), yielding an overall TOST $p = .033 < .05$. Thus, the difference between original and generated vignettes falls within the predefined negligible range, indicating practical equivalence.

MFV-130 Expansion Prompt (Care–Emotional Harm)

Role: You are generating new scale items for Moral Foundations research.

Goal: Produce short moral vignettes that capture **Care(e)**—emotional harm to humans—while matching the tone and structure of the original Moral Foundations Vignettes.
Foundation Definition (Care–Emotional Harm): Emotionally harmful acts such as mocking, ridicule, belittling, or social exclusion, without physical harm or threats.
Generation Instructions:

1. **Form:** Each item must be a single sentence ($\leq$25 words), plain language, observational tone.
2. **Content Constraints:** Emotional harm only; no physical harm, threats, authority roles, or in-group/out-group dynamics.
3. **Social Context:** Use strangers or minimal relationships; avoid family, close friends, or hierarchical roles.
4. **Style:** Mirror original MFV phrasing (e.g., begin with "You see ...").
5. **Subjects:** Use generic actors (man, woman, boy, girl, person, teen); avoid names and protected attributes as targets.

Diversity Requirement (Coverage Grid): Generate exactly **120 items** organized as **10 themes** $\times$ **12 items**, covering distinct everyday contexts (e.g., public transit, workplaces without hierarchy, online spaces, social mixers).
Output Format: Return **JSON only** with one object containing:

- `foundation = "Care(e)"`
- A list of 10 themes (fixed order), each with exactly 12 items
- `total_count = 120`

Validation: All themes must appear once, counts must match exactly, and no text may appear outside the JSON.

Figure 11: Prompt template used to expand the Moral Foundations Vignettes for the Care(e) foundation. The prompt enforces strict moral constraints, stylistic consistency with the original MFV items, and balanced coverage across everyday social contexts.

E.1 MFV-130 Expansion Prompt

To expand the Moral Foundations Vignettes (MFV-130), for each foundation and the *Social Norm* category, we construct a foundation-specific prompt that defines the target construct, explicitly excludes confounding dimensions, and enforces stylistic consistency with the original MFV items. Generation is organized around a prompt-level diversity grid that enumerates multiple everyday social contexts (e.g., public spaces, online interactions, peer-based workplaces) while holding the underlying moral content fixed. Items are generated in structured JSON format with explicit count and schema constraints to guarantee balanced coverage across contexts, and all generated items are reviewed by human experts for clarity and label fidelity. An example generation prompt is shown in Figure 11. Table 6 shows representative original and generated vignettes.

Original MFV Items (Care–Emotional Harm)	Generated Items (MFV-130 Expansion)
You see a teenage boy chuckling at an amputee he passes by while on the subway.	You see a teen laughing loudly as a person fumbles with their groceries on the sidewalk.
You see a girl laughing at another student forgetting her lines at a school play.	You see a girl pointing and smirking when a stranger drops their phone on the train.
You see a woman commenting out loud about how fat another woman looks in her jeans.	You see a woman muttering that a passenger's clothes look ridiculous as they walk by.
You see a man quickly canceling a blind date as soon as he sees the woman.	You see a man whispering insults while glancing at a woman boarding the bus.
You see a man snickering as he passes by a cancer patient with a bald head.	You see a man rolling his eyes and loudly commenting that a passerby looks pathetic.

Table 6: Side-by-side examples of original Moral Foundations Vignettes (Care–Emotional Harm) and generated expansion items. Generated items preserve the original MFV observational framing and emotional-harm focus while varying everyday social contexts.

Table 7: **JSON output schema for automated SAE feature interpretation.** Each field is required; `mft_alignment` and `mft_polarity` are restricted to the listed enumerations.

Field	Description / Constraints
`short_label`	A brief (5–10 words) descriptor of the pattern.
`long_description`	A 1–2 sentence summary of the feature's semantics.
`mft_alignment`	The primary moral foundation detected. *Enum:* {`care`, `fairness`, `loyalty`, `authority`, `sanctity`, `none`}
`mft_polarity`	The valence of the moral content. *Enum:* {`virtue`, `vice`, `mixed`, `none`}
`rationale`	Reasoning for the classification grounded in specific snippets.
`evidence_ids`	List of snippet indices (1–6 items) supporting the rationale.
`confidence`	Scalar $c \in [0, 1]$ indicating certainty of interpretation.

F Limitations

Our study has several limitations. First, we employ open-sourced SAEs with a fixed layer stride, so the top-activating features we identify should be interpreted as a sparse decomposition at the available layers rather than an exhaustive search for globally optimal features. The scale of our SAE experiments is also restricted by the limited amount of open-source pretrained SAEs. Future work utilizing all-layer SAEs at a larger scale could provide a more comprehensive map of feature evolution for different models.

Moreover, our analysis is anchored in Moral Foundations Theory and English-centric corpora. It remains unclear how the identified geometric structures map onto multilingual contexts or alternative frameworks such as the Theory of Dyadic Morality [Schein and Gray, 2018].

In addition, we evaluate Qwen3-30B-A3B as our MoE-architecture model and as a probe of an updated pretraining data mixture, relative to Qwen2.5. However, our logit-based scoring evaluates encoding-phase representations rather than generation-phase reasoning traces. As a result, our analysis can probe differences associated with architecture, pretraining, and supervised fine-tuning, but it does not directly isolate the effect of Qwen3's reasoning-oriented RL stage. Investigating whether explicit reasoning trace generation modulates moral representations is left to future work.

Finally, while we demonstrate causal steering on a questionnaire-style task, we do not comprehensively measure deployment-relevant side effects (e.g., changes in refusal behavior, bias, toxicity) that may emerge under moral interventions. Future work can rigorously measure such potential side effects to determine the safety of using these interventions in real-world applications.

G Ethical Considerations

Our work studies how moral concepts are represented and can be causally influenced in instruction-tuned LLMs. This creates dual-use risks: the same steering methods that help analysis could be used to manipulate users' moral judgments, increase persuasive power, or tailor outputs to specific

ideological goals without disclosure. Steering may also introduce unintended side effects, such as shifting refusal behavior, amplifying demographic or political bias, or changing toxicity and stereotyping rates, even when overall task accuracy appears stable. In addition, our foundation vectors are built from curated vignette-style data and validated on English natural text, which may reflect cultural and annotator biases and may not transfer to other moral systems or languages; results should not be treated as normative claims about what is "correct" morality. We use publicly available text (e.g., Reddit) and operate at the level of aggregate distributions rather than attempting to identify individuals, but we still aim to minimize privacy risk by avoiding release of raw user text beyond what is already public and by reporting only summary statistics. Finally, automated feature interpretation uses an LLM as an annotator, which can introduce interpretation errors and project inherent biases. Therefore, we treat these annotations as qualitative aids rather than ground truth. To reduce misuse, we recommend that any released code for interventions include clear documentation, default conservative settings, and evaluation scripts that track side effects (bias, toxicity, and refusal changes) under steering.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer:[Yes]

   Justification: We accurately described the scope and objectives of the work in abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss our limitations in the Appendix (Section F).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Full implementation details are disclosed in the main text and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code and data are provided.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have disclosed all the training and test details in the main text and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have reported information about the statistical significance of our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Specified in the appendix A.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research follows NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed social impacts of the work in Section G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We discussed the risks and safeguards in Section G.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited and properly credited the creators or original owners of assets used in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code, data, and models are well-documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We specified the details in appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We discuss our human evaluation procedures in Section E and clarify that the task involved standard data annotation (evaluating LLM-generated text and SAE features) without the collection of Personally Identifiable Information. In accordance with our institution's policies, this standard annotation task was exempt from formal IRB review. Annotators were informed of the task requirements and annotation was done voluntarily.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are the primary subjects of our mechanistic decomposition and data construction sections.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Layer	Feature ID	MFT	Conf.	Label (GPT, human-validated)	Peak-centered sample text (trimmed)
23	44965	Care	0.93	Graphic descriptions of extreme suffering and atrocity	Fly strike kills thousands of rabbits. . . . The eggs hatch into maggots which eat their way into the poor rabbit's flesh. The rabbit dies from being eaten alive—a slow, painfully horrific death . . .
11	8003	Authority	0.90	Polite, expert-style, authoritative explanatory answers to questions	. . . he only needs to be organized . . . Please share any suggestions . . . Ah, organization . . . Dealing with clutter and putting things in order is an issue for almost everyone . . .
15	41465	Care	0.90	Descriptions of large-scale suffering and atrocities	. . . the Holocaust, concentration camps, Nazi unfathomable brutality . . .
19	90260	Authority	0.90	Organizational management systems, methods, and processes	. . . improvements in fields such as . . . safe minimum levels of maintenance . . . operating procedures and strategies . . . capital maintenance regimes and plans . . .
23	90226	Authority	0.86	Definitions and hierarchies of roles and positions	When thinking of a company organizational chart . . . All positions ultimately lead up to an executive member. The executive is considered the leader of the company . . .
19	68970	Authority	0.86	Attributions of authoritative sources and institutions	Imperial Abbey of Essen . . . Imperial Abbey of the Holy Roman Empire . . . Gained princely status . . .
19	37235	Care	0.86	Discussing traumatic tragedies and collective suffering	. . . 17 schools had experienced the terrifying reality of gun violence . . . yesterday . . . the eighteenth school was added . . .
19	13133	Authority	0.86	Mandated frameworks and guidelines	This document serves as USDA guidance for . . . food safety programs . . . minimum elements . . . based on HACCP principles . . .
7	97876	Care	0.86	Descriptions of large-scale suffering and atrocities	Massacre at Paris by Christopher Marlowe . . .
7	61385	Fairness	0.86	Evidence strength, grading, and quality of studies	. . . screen all adults for obesity . . . offer or refer patients . . . intensive, multicomponent behavior . . .
15	10095	Authority	0.86	Authoritative planning and designing structured programs or courses	Frameworks for Financial Crisis Management . . . the government must be aware . . . authority established to make decisions . . .
3	8682	Authority	0.86	Authoritative advice responses	Since your daughter is already light years ahead . . . it does not make much sense . . .
15	107641	Care	0.86	Environmental harms, pollution, and regulatory criticism	. . . blame for the Gulf of Mexico oil spill . . . his agency could have more aggressively monitored . . .
27	119015	Care	0.86	Online safety, filtering, and harm detection	. . . block access to Internet sites which have harmful or illegal content . . .
19	125143	Care	0.86	Detection and prevention of harmful misconduct	. . . actions against online child sexual abuse . . .
7	35014	Authority	0.86	Rules limited by higher moral or legal norms	Those who exercise authority should do so as a service . . . The exercise of authority is measured morally . . . Those in authority should practice distributive justice . . .
3	96957	Authority	0.86	Mentions of assistants and assistant roles/titles	. . . American University School of Public Affairs' assistant professor . . .
19	38705	Care	0.79	Practical advice on health, safety, and care	. . . make compassion a cornerstone . . . kindness initiative . . . show compassion . . . kids to show empathy . . .
15	103468	Authority	0.79	Institutional roles, rules, and protective authority	Thousands of farmers . . . compensated for flood damage . . . satellite-based insurance . . .
23	37802	Authority	0.79	Institutional history, milestones, and commemorations	. . . government rangers working to protect the gorillas . . .
23	20176	Fairness	0.79	Describing laws, policies, and institutional decisions	. . . Supreme Court . . . was constitutional . . . Congress enacted the law . . .
27	88535	Care	0.79	Protect-and-care body-harm discussions	. . . heroes and heroines who defended Scotland . . .
3	105626	Care	0.79	Health, disease prevention, and bodily protection	How Your Child Can Be Cavity Free for Life . . . Healthy Eating . . .
11	26778	Authority	0.78	Teacher-like evaluative feedback and instructions	. . . "Child labor has no place in the production of . . . " . . .
15	127154	Sanctity	0.78	Health, immunity, and purity-from-disease discourse	Metabolism . . . CYP450 . . . Biological half-life . . .
11	25501	Sanctity	0.78	Hagiographic or moral praise of virtuous women and piety	St. Matilda . . . generous to the Church . . . raised at her convent . . . purposeful living . . .

Table 8: **LLM-grounded semantic characterization of 25 SAE features for Llama-3.1-8B-Instruct.** **Content warning:** Table includes one excerpt with a graphic description of violence. For each candidate feature, we mine top-activating contexts from a random sample of 50,000 FineWeb documents [Penedo et al., 2024], then extract a peak-centered ± 64 token window around the maximally activating token. The *Label* and *MFT* assignments are generated by `GPT-5.1` and subsequently validated by human reviewers; the confidence score is the GPT-5.1-reported confidence. Sample texts are manually trimmed for readability while preserving the peak-centered context. **Labels that overlap with MFD2.0 are highlighted.** We note 4 Moral foundations, except for `Loyalty`, are represented among features that *GPT* is confident in identifying. *Care* and *Authority* are most frequently associated with identified features.

Layer	Feature ID	MFT	Conf.	Label (GPT, human-validated)	Peak-centered sample text (trimmed)
3	72227	Authority	0.93	Mentions of governments and national leaders	The government's former climate-change adviser; the Government's initiative to develop 100 cities. . .
15	130669	Care	0.93	Condemning bullying, harmful, coercion, and abusive mistreatment	Denounce bullying and promote kindness, respect, and protection of students or workers from harm (e.g., anti-bullying policies and reminders to not hurt others)
27	85517	Care	0.93	Definitions and examples of empathy/compassion	Compassion is described as sympathetic consciousness of others' distress plus a desire to relieve it, along with related traits like sensitivity and non-judgment
3	43421	Authority	0.90	References to "The Government . . . " as actor/subject	Neutrally reporting or explaining actions of the government, such as official initiatives, policies, or reversals
27	17185	Care	0.90	Medical explanations of injuries and health/safety risks	Workshop safety to avoid injury; playground injuries and prevention for children
27	54738	Care	0.90	Natural disasters and their destructive impact, and care for people facing disasters	Natural disasters and their harmful consequences: PTSD from traumatic events including natural disasters; social media helping people during floods
27	114000	Authority	0.86	Criminal justice, law enforcement, and legal process	Criminal justice systems, the role of a Minister of Justice, and death penalty administration
23	69927	Care	0.86	Low-calorie, nutrient-dense healthy food descriptions	Emphasis on foods that are low in calories but high in nutrients and health benefits to support weight loss or a healthy diet
27	118156	Care	0.86	Harms and risk factors to health or systems	Harmful exposures and their detrimental effects: pesticides as substances used against pests but with implied toxicity
27	47018	Care	0.86	Catastrophic disasters and apocalyptic upheavals causing large-scale human suffering	Large-scale disasters causing or threatening extreme harm to many people: volcanic eruptions like Krakatoa, nuclear accidents at Chernobyl and Fukushima, etc.
11	129124	Care	0.86	Awareness campaigns about health and risk issues	Preventing or mitigating harm to people's health or wellbeing: rare diseases and their impact; neuropathy and the need for early intervention and research
27	70504	Care	0.86	Grim statistics on large-scale human suffering	Victims from school shootings, severe untreated health problems, and many other grim incidents
3	65290	Authority	0.79	Institutions, regulations, and formal responsibility	Governments or large organizations exercising or being critiqued for their formal authority: government climate policy and carbon pricing
23	3088	Authority	0.78	Biographical/institutional leadership and official roles	A university president praised as a leader and compared to Horace Mann and Abraham Lincoln
3	123373	Fairness	0.78	Business and corporate practices, duties, and societal impacts to promote fairness	Corporate behavior for broader societal or environmental good: sustainability as meeting fundamental responsibilities in human rights, labour, environment, and anti-corruption

Table 9: **LLM-grounded semantic characterization of 15 SAE features for Qwen2.5-7B-Instruct.** **Content warning:** Table includes one excerpt with a graphic description of violence. For each candidate feature, we extract top-activating contexts. The *Label* and *MFT* assignments are generated by `GPT-5.1` and subsequently validated by human reviewers; the confidence score is the model-reported confidence. Sample texts are manually trimmed for readability while preserving the peak-centered context. **Labels that overlap with MFD2.0 are highlighted**. Similar to *Llama* SAE's results in Table 8, we note that most *Qwen* SAE features that *GPT* is confident in associating with MFT categories are related to *Care* and *Authority*, with a few relevant to *Fairness*.

Layer	Feature ID	MFT	Conf.	Label (GPT, human-validated)	Peak-centered sample text (trimmed)
15	42641	Authority	0.79	Formal institutions, systems, or organized practices such as government	The senate and the house of representatives ... call for a debate and vote bipartisan group of members are calling for a debate and vote ....
3	4383	Authority	0.86	References of experts, formal research, or authoritative institutions	Experts say there has not been a reliable count ... New Irish Government policy is to plant 600,000 native trees
11	123331	Authority	0.63	Historical or institutional authority and control	The Bolsheviks renamed the village Wladimirowka and subjected its citizens to the tyrannies of the Stalinist era.
23	6462	Authority	0.78	Foundational principles, origins, and institutional milestones	Creating relevant information for decision-making . . . manager of operations analysis and planning utah transit authority.
19	2272	Care	0.79	Programs, resources, or organized efforts aimed at educating, protecting, or improving conditions for children or broader communities	Conducting Social Investigation and Class Analysis in urban poor communities... feudal oppression and other political crimes
23	20681	Care	0.78	Helping others through educational guidance and instructional explanation on safety-oriented policy	...facilitate and increase ... involvement of small scale fishermen in the decision-making processes by exchanging experiences, provide opportunities for education and provide a platform for networking ...
27	88535	Care	0.79	Health effects of environmental exposures and descriptions of altruist values	Phototherapy with PUVA or UVB has been used to treat a wide variety of diseases... the Confucian virtue denoting the good feeling a virtuous human experiences when being altruistic.
23	64586	Care	0.71	Commemoration and honoring significant human achievements or battles	The memorial at Xinon Neron near Florina, Macedonia was the first memorial erected in Greece to commemorate a World War Two battle
7	2886	Care	0.71	Directions, practice, and care in following procedures to support effective learning or safe training	... some having to do with the quality of care. Staying injury free in gymnastics starts with using the proper protective gear...
3	56365	Care	0.63	Explaining or endorsing best practices and benefits	But I believe the mental health benefits of learning new skills go beyond the neuropsychological benefits...
11	85498	Care	0.78	Health, wellness, and medical information content	Generally, common signs and symptoms of Nephrotic Syndrome are as follows...
15	18957	Sanctity	0.78	Christian children's religious and moral education	It's never been so easy and so fun to teach your children the Bible!
19	116807	Sanctity	0.71	Discourse on modern problems such as health degradation	With the air we breathe, the water we drink and the food we eat, our body is poisoned by chemicals, . . .
7	106968	Care	0.86	Early childhood, babies, and childcare	The most critical time for a child's early development is typically before the age of 6. By that time, children are fairly fluent in speaking their native language.
7	114048	Care	0.79	Early childhood development and child-focused care	Early Language Development Yields Higher Fluency...
23	44965	Care	0.86	Children's vulnerability, harm, and protection issues	...the family clothes and feeds the children to meet their needs functional fit theory we should have matrifocal families...
3	3594	Fairness	0.71	Moralized positivity vs negativity in life	If you are selfish or mean to others, you will be paid back with unwanted circumstances.

Table 10: **LLM-grounded semantic characterization of SAE features for Llama-3.1-8B-Base model.** For each candidate feature, we extract top-activating contexts. The *Label* and *MFT* assignments are generated by GPT and subsequently validated by human reviewers; the confidence score is the model-reported confidence. Sample texts are manually trimmed for readability while preserving the peak-centered context. **Labels that overlap with MFD2.0 are highlighted**.

Layer	Feature ID	MFT	Conf.	Label (GPT, human-validated)	Peak-centered sample text (trimmed)
39	704	Authority	0.71	Founding or introduction of institutions and systems	New Irish Government policy is to plant 600,000 native trees around Ireland within 3 years, edged on by the Green Party.
39	1016	Authority	0.78	Founding or creation of institutions and systems	U.S. President Donald Trump is working to designate ... as a foreign terrorist organization...
39	9095	Authority	0.79	Institutional programs, institutes, and formal initiatives	The Berlin Psychoanalytic Institute (later the Göring Institute) was founded in 1920 to further the science of psychoanalysis in Berlin.
39	30227	Authority	0.78	Founding or establishment of institutions and movements	U.S. President Donald Trump is working to designate ..., a move that would bring sanctions against ...
39	30369	Authority	0.71	Institutional programs, initiatives, and formal organizations	New Irish Government policy is to plant 600,000 native trees around Ireland within 3 years, edged on by the Green Party.
3	1733	Care	0.78	Public health, safety, and risk communication	"Most parents and guardians know that immunization is still the best way to protect children against vaccine-preventable diseases such as whooping cough and measles," said...
3	32310	Care	0.78	Educational and developmental guidance for learners with a supportive or nurturing framing	Through a Reggio Emilia-inspired approach and a focus on inquiry and self-guided learning, students ask questions, investigate, and learn about things that genuinely matter to them.
3	6759	Authority	0.79	Rules, procedures, and formal operating guidelines in institutional, educational, or policy contexts	... describes policies intended to "secure societal benefits from increases in land value that can arise from changes to land-use rights through the planning system and/or investment in public infrastructure."... if the public sector invests capital or makes regulatory changes that increase real property values, the public sector, not the landowner, should benefit. . . .
3	27448	Authority	0.79	Rules, laws, and formal institutional guidelines explanation	We encourage you to teach your children these basic safety tips for riding on the school bus. - Stay in your seat, it is the LAW!
3	9359	Care	0.79	Health, safety, and environmental risk warnings	However recent research has shown that tampons can be dangerous for a woman's health. Let's look into the health risks of tampons and explore the healthy alternatives.

Table 11: **LLM-grounded semantic characterization of SAE features for Qwen3-30B-A3B model.** For each candidate feature, we extract top-activating contexts. The *Label* and *MFT* assignments are generated by `GPT` and subsequently validated by human reviewers; the confidence score is the model-reported confidence. Sample texts are manually trimmed for readability while preserving the peak-centered context. **Labels that overlap with MFD2.0 are highlighted**.