Rule vs. Consequence: Dissociable Internal Representations of Moral Reasoning in LLMs

Eugenie Shi
Stanford University
yqshi@stanford.edu

Abstract

Large language models are increasingly deployed in morally and legally sensitive contexts, yet little is known about how they internally represent distinctions between moral reasoning styles. We construct a 2×2 factorial dataset crossing reasoning style (rule-based vs. consequence-based framing) with judgment polarity (permissive vs. prohibitive), drawing from three moral reasoning benchmarks (ETHICS, Moral Stories, Social Chemistry 101) and evaluating on a cleanroom test set of hand-written scenarios including conflict pairs where the two framings predict opposite labels. Using linear probing and activation steering across three instruction-tuned LLMs, we identify two orthogonal directions in the residual stream: DIR-A encodes moral reasoning style ($r > 0.87$ with style labels, $r < 0.03$ with polarity labels) and DIR-B encodes judgment polarity ($r > 0.85$ with polarity labels, $r < 0.06$ with style labels), with cross-correlation $|r| < 0.04$. This double dissociation replicates across Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, and Llama-3.1-8B-Instruct at 34 to 47 percent network depth. All ten layers tested in the range 6 to 15 show significant framing selectivity ($p < 0.0001$), ruling out a layer-specific artifact. Activation steering with DIR-A produces large causal shifts (Cohen's $d = 1.03$ to 1.66, all $p < 0.0001$), and on conflict scenarios selectively corrects rule-framed errors by 30 percentage points while leaving consequence-framed accuracy unchanged. Cross-domain generalization achieves 0.967 mean accuracy across three of four benchmark sources, and the activation probe exceeds an n-gram lexical baseline on novel-template evaluation (0.594 vs. 0.438), providing evidence of partial transfer beyond surface features.

1 Introduction

As large language models are deployed in legally sensitive, medically critical, and ethically charged contexts (Soffer et al., 2025; Millière, 2025), understanding how they represent and respond to different moral framings becomes practically important. A fundamental distinction in moral philosophy separates rule-based reasoning, which judges actions by their conformity to explicit norms and duties (Rawls, 1971), from consequence-based reasoning, which judges actions by their outcomes and welfare effects (Mill, 1863). Models have been shown to produce systematically different judgments depending on how questions are framed (Simmons, 2023; Park et al., 2025), but whether this reflects superficial lexical sensitivity or distinct internal representational structures remains an open question with direct implications for alignment and interpretability research.

Mechanistic interpretability has established that LLMs encode semantic properties as linear directions in activation space (Park et al., 2023; Burns et al., 2023; Zou et al., 2023; Elhage et al., 2022), and that these directions can be causally manipulated via activation steering (Turner et al., 2023; Panickssery et al., 2024). We extend this program to moral reasoning style, asking whether the model's differential response to rule-framed and consequence-framed moral inputs has a geometric structure that is causally effective, dissociable from judgment polarity, and stable across diverse inputs and models.

To support this investigation, we construct a 2×2 factorial dataset that crosses reasoning style with judgment polarity. We draw scenarios from ETHICS (Hendrycks et al., 2021), Moral Stories (Emelin et al., 2021), and Social Chemistry 101 (Forbes et al., 2020), pairing each scenario with both a rule-framed and a consequence-framed question under matched polarity conditions. We additionally hand-write a cleanroom test set including conflict pairs where the two framings predict opposite gold labels, enabling behavioral dissociation analysis that goes beyond simple framing bias.

Using this dataset, we make four central contributions. First, we identify a double dissociation in which DIR-A encodes moral reasoning style ($r > 0.87$ for style, $r < 0.03$ for polarity) and DIR-B encodes judgment polarity ($r > 0.85$ for polarity, $r < 0.06$ for style), with cross-correlation $|r| < 0.04$, consistent across three models from different developers. Second, we demonstrate layer robustness: all ten layers tested in the range 6 to 15 of Qwen show significant framing selectivity ($p < 0.0001$), indicating that the geometric structure is a stable property of an entire processing region rather than a narrow artifact. Third, we provide causal evidence via activation steering showing large logit difference shifts (Cohen's $d > 1.0$, all $p < 0.0001$) and a framing-specific behavioral dissociation on conflict pairs that rules out a global acceptance-bias explanation. Fourth, we provide generalization evidence through cross-domain probing that transfers across three of four benchmark sources with 0.967 mean accuracy, and through lexical baseline comparisons showing the activation probe exceeds n-gram baselines on novel-template evaluation.

Throughout, we are careful to distinguish between our operational definition of *moral reasoning style sensitivity*, meaning the model's differential response to rule-framed vs. consequence-framed inputs, and the stronger claim of fully internalized normative reasoning style. We argue the former is well supported by the evidence; the latter remains an open question. Our contribution is best characterized as discovery-phase mechanistic interpretability: we establish the existence, causal efficacy, and geometric structure of the relevant directions, while leaving their circuit-level implementation as future work.

2 Related Work

Park et al. (2023) demonstrate that semantic properties are encoded as linear directions in transformer residual streams. Zou et al. (2023) show these directions are causally manipulable via representation engineering. Burns et al. (2023) identify truth-related directions through contrastive probing. Elhage et al. (2022) provide theoretical grounding for redundant encoding in superposition. Our work applies these methods to moral reasoning style.

Prior work evaluates LLM moral reasoning through behavioral benchmarks (Hendrycks et al., 2021; Emelin et al., 2021; Forbes et al., 2020), finding systematic framing effects (Simmons, 2023). Park et al. (2025) show that deontological lexical markers and modal expressions systematically influence normative judgments. Ozeki et al. (2025) benchmark normative reasoning from logical and modal perspectives. Abrams et al. (2026) evaluate LLMs on norm-application tasks. Zhong et al. (2025) show that different judgment processes may be dissociable inside LLMs. Soffer et al. (2025) document failure modes in medical ethics reasoning. Our work moves beyond behavioral evaluation to examine the internal representational structure underlying these judgment patterns.

Itzhak et al. (2024) analyze emergent cognitive biases in instruction-tuned LLMs, including framing effects analogous to those we study, and link them to instruction phrasing. Sun et al. (2024) demonstrate that instruction-tuned models are brittle to phrasing variations, motivating internal representation analysis rather than surface behavioral testing alone. Lorè and Heydari (2024) show that LLM decisions in strategic games depend on contextual framing, providing complementary behavioral evidence. Geng et al. (2026) propose a framework to quantify pragmatic influences of instructions on LLM behavior. Lim et al. (2026) introduce DeFrame, a debiasing method that targets framing-induced biases; the directions we identify could serve as building blocks for such interventions. Ceballos-Arroyo et al. (2024) document instruction phrasing sensitivity in clinical LLMs, motivating the ethics statement of this work.

Turner et al. (2023) introduce activation addition and Panickssery et al. (2024) extend it to personality traits. Wu et al. (2025) probe internal representations of theory-of-mind concepts, providing a methodological comparison for our representational analysis. We apply activation steering to moral reasoning style, introducing a 2×2 factorial design to disentangle reasoning style from judgment polarity.

3 Dataset

3.1 Factorial design

We construct a 2×2 factorial dataset that crosses **moral reasoning style** (rule-based vs. consequence-based) with **judgment polarity** (permissive vs. prohibitive). Each example consists of a moral scenario paired with two questions: one that frames the judgment in terms of an applicable rule or norm, and one that frames it in

terms of outcomes and welfare effects. The rule-framed template follows "Given the social rule that [norm], is this action acceptable?" and the consequence-framed template follows "Considering the outcomes and effects on those involved, is this action acceptable?" This design is essential for our analysis: by varying reasoning style and polarity independently and with balanced labels, we create the conditions under which two dissociable directions can be identified rather than a single normative-strictness axis.

3.2 Sources and splits

We draw scenarios from three established moral reasoning benchmarks. ETHICS (Hendrycks et al., 2021) provides scenarios from its deontology and commonsense subsets covering everyday moral violations and rule-following dilemmas. Moral Stories (Emelin et al., 2021) provides structured narratives about social norm violations. Social Chemistry 101 (Forbes et al., 2020) provides social rules and judgments about everyday situations. Each scenario is paired with both a rule-framed and a consequence-framed question, and gold labels are assigned by the source benchmark. The resulting dataset contains 1,452 examples split into 720 training, 240 validation, 240 benchmark test, 32 cleanroom test, and 220 legal transfer examples, with strict label balance: 180 examples per reasoning style and polarity cell in the training set. The legal transfer split consists entirely of LegalBench examples and is style-imbalanced by construction (160 rule-framed, 60 consequence-framed), reflecting the natural distribution of rule-invoking language in contract and benefits texts; it is used only for cross-domain transfer evaluation and not for probe training.

We additionally construct a **cleanroom test set** of 32 template pairs (16 scenarios $\times$ 2 framings), all hand-written with vocabulary absent from any training source. Eight scenarios (L-series) are adapted from LegalBench (Guha et al., 2023) situations re-phrased with entirely new vocabulary; eight (M-series) are original scenarios covering medical and clinical ethics, journalism privacy tradeoffs, professional confidentiality, workplace safety, and research integrity, where rule-framed and consequence-framed questions are designed to invoke opposing intuitions. Of the 16 scenarios, 12 are genuine conflict pairs in which the two framings predict opposite gold labels; 4 are non-conflict pairs included to prevent conflict-detection as a shortcut. The LegalBench source examples themselves are held in a separate legal transfer split ($n = 220$) used only for cross-domain transfer evaluation, not probe training. We acknowledge that 32 pairs is a small set; the lower probe accuracy on this set (0.594 vs. 1.000 on benchmark) should be interpreted as a lower bound on template-independent generalization rather than a definitive estimate, and expanding this set is an explicit goal for future work.

4 Methods

4.1 Models and activation extraction

We use three instruction-tuned LLMs spanning different developers and tokenizers: Qwen2.5-7B-Instruct (Yang et al., 2025) from Alibaba ($L = 28$ layers, $d = 3584$), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) from Mistral AI ($L = 32$, $d = 4096$), and Llama-3.1-8B-Instruct (Dubey et al., 2024) from Meta ($L = 32$, $d = 4096$). All models are loaded in 4-bit NF4 quantization (Dettmers et al., 2023) to fit within available GPU memory. Activations are extracted at the final prompt token from each transformer layer's residual stream, yielding a tensor of shape $[n, L, d]$ per split.

4.2 Direction extraction

For each layer ℓ, we train ℓ_2-regularized logistic regression probes ($C = 1.0$) on standardized activations to separate rule-framed from consequence-framed examples (DIR-A) and permissive from prohibitive examples (DIR-B). The probe weight vector $\mathbf{w}$ is back-projected to the original activation space by $\mathbf{v} = \mathbf{w}/\boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is the feature-wise standard deviation from the standardizer. Only the weight vector is used; the bias term encodes class prevalence rather than direction and is discarded. The resulting vector is ℓ_2-normalized to give the unit-norm direction $\hat{\mathbf{v}}$. Layer selection uses the Fisher discriminant ratio $J = (\mu_+ - \mu_-)^2/(\sigma_+^2 + \sigma_-^2)$ computed on held-out validation activations, where $\mu_\pm$ and $\sigma_\pm^2$ are the mean and variance of $\hat{\mathbf{v}}^\top \mathbf{x}$ for positive and negative class examples.

4.3 Activation steering

We inject directions into the residual stream at the final prompt token position via forward hooks:

$$\mathbf{h}_\ell \leftarrow \mathbf{h}_\ell + \alpha\, \hat{\mathbf{v}}$$

where $\hat{\mathbf{v}}$ is the unit-norm direction and α is the steering strength. Due to 4-bit NF4 quantization, residual stream norms are elevated (approximately

430 at layer 10 in Qwen, compared to approximately 50 in float16 models), requiring proportionally larger α values (40 to 60 for Qwen) to achieve perturbation ratios of 0.09 to 0.14. We verify that steering effects are monotonically increasing within $|\alpha| \leq 60$ and report $\alpha = 60$ unless stated otherwise. For Llama, α is calibrated to the same perturbation ratio, yielding $\alpha = 2$.

4.4 Measurement

Model judgments are measured via logit difference:

$$\Delta\text{LD} = \log \sum_{t \in \mathcal{A}} p_t \; - \; \log \sum_{t \in \mathcal{U}} p_t$$

where $p_t = \text{softmax}(\mathbf{z})_t$ is the full-vocabulary softmax probability,

$$\mathcal{A} = \{\text{“Acceptable”}, \text{“acceptable”}, \text{“Yes”}, \text{“yes”}\}$$

and

$$\mathcal{U} = \{\text{“Unacceptable”}, \text{“unacceptable”}, \text{“No”}, \text{“no”}\}$$

This measure is robust to the near-binary output distribution observed in our data: 59 percent of benchmark examples have $P(\text{Acceptable}) < 0.1$, making probability-space shifts negligible while logit-space shifts remain meaningful.

5 Results

5.1 Double dissociation across three models

Figure 1 shows the projections of all training examples onto DIR-A and DIR-B at layer 10 of Qwen. DIR-A cleanly separates rule-framed from consequence-framed examples ($r = +0.899$, $p < 0.0001$) with near-zero polarity correlation ($r = +0.025$, $p = 0.50$). DIR-B shows the complementary pattern: $r(\text{polarity}) = +0.856$ ($p < 0.0001$) and $r(\text{style}) = +0.022$ ($p = 0.56$). The cross-correlation between projections is $r = +0.030$ ($p = 0.43$), confirming informational independence. The complementary structure across all four panels of Figure 1 supports the two-direction interpretation over a single strictness axis.

Table 1 reports the double dissociation across all three models. The pattern is consistent: DIR-A $|r(\text{style})| > 0.87$ and DIR-B $|r(\text{polarity})| > 0.85$ in every model, with cross-correlation $|r| < 0.04$. Peak dissociation layers fall at 34 to 47 percent network depth across models, consistent with prior findings localizing semantic composition in early-to-middle transformer layers.

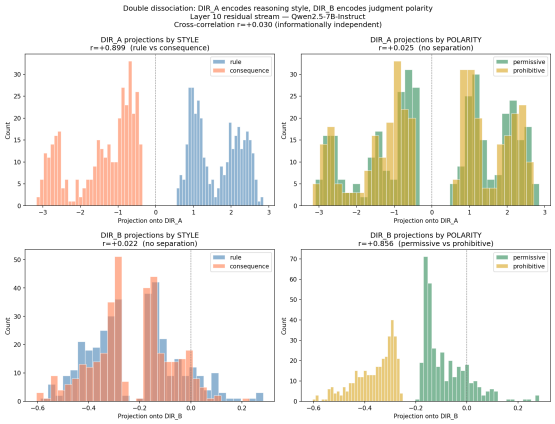

Figure 1: Double dissociation across three LLMs. All models show DIR-A $|r(\text{style})| > 0.87$ and DIR-B $|r(\text{polarity})| > 0.85$, with cross-correlation $|r| < 0.04$.

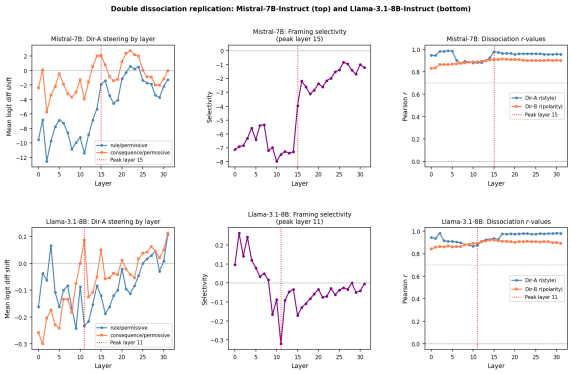

Figure 2: Double dissociation replication in Mistral-7B-Instruct-v0.3 (top row) and Llama-3.1-8B-Instruct (bottom row).

Figure 2 shows the full layer-by-layer replication for Mistral and Llama. In both models, DIR-A $r(\text{style})$ and DIR-B $r(\text{polarity})$ exceed 0.85 across virtually all 32 layers, and framing selectivity peaks at the expected early-to-middle depth before decaying in later layers.

On the benchmark test set, the style probe achieves 1.000 accuracy and Dir-A projections separate style labels with AUC $= 1.000$; the polarity probe likewise achieves 1.000 accuracy on benchmark test. These perfect in-domain scores confirm near-perfect linear separability at layer 10 and motivate the cleanroom evaluation as the more stringent test of generalization.

5.2 Layer robustness

To assess whether the identified directions are robust to layer choice, we test DIR-A steering across layers 6 to 15 of Qwen using a fixed direction extracted at layer 10. Framing selectivity is the differ-

Model	Layer	A-sty	A-pol	B-pol
Qwen2.5-7B	10/28	+0.90	+0.03	+0.86
Llama-3.1-8B	11/32	+0.88	+0.00	+0.89
Mistral-7B	15/32	+0.98	+0.01	+0.91

Table 1: Double dissociation across three instruction-tuned LLMs. Layer = peak layer/total; A-sty = DIR-A r(style); A-pol = DIR-A r(polarity); B-pol = DIR-B r(polarity). Cross-correlation $|r| < 0.04$ for all models.

Held-out source	Acc.	cos
Social Chem. 101	1.00	0.73
Moral Stories	0.90	0.90
ETHICS Common.	1.00	0.78
ETHICS Deont.	0.52	0.84
Mean (excl. deon.)	**0.97**	0.80

Table 2: Cross-domain generalization of DIR-A (Qwen, layer 10). $n_{\text{test}} = 60$ for all sources. Acc. = probe accuracy; cos = cosine similarity $\cos(\hat{v}_{\text{cd}}, \text{DIR-A})$.

Probe type	Bench.	Clean.
Keyword matching	0.88	0.75
TF-IDF (unigram)	1.00	0.50
TF-IDF (bigram)	1.00	0.44
Activation probe (layer 10)	1.00	**0.59**

Table 3: Probe accuracy on benchmark and cleanroom test sets. Bench. = benchmark test; Clean. = cleanroom test.

Style	Polarity	α	Shift	d
rule	perm.	+60	+2.03	+1.03
rule	proh.	+60	+2.88	+1.66
conseq.	perm.	−60	−0.64	−0.40
conseq.	proh.	−60	−1.38	−1.14

Table 4: Logit difference shifts under DIR-A steering (Qwen, $n = 60$ per row). Shift = mean ΔLD; d = Cohen's d. All $p < 0.01$ by paired t-test; perm. = permissive; proh. = prohibitive; conseq. = consequence.

ence in mean logit shift between rule-permissive and consequence-permissive examples under DIR-A(+60). Negative selectivity is the expected pattern: Dir-A(+60) steers activations toward the consequence direction, so consequence-framed examples should shift more positively than rule-framed ones throughout. The corresponding visualization is shown in Figure 3 in Appendix C.

Table 5 shows that every tested layer achieves significant framing selectivity ($p < 0.0001$), with the qualitative pattern consistent throughout: consequence-permissive examples shift more positively than rule-permissive examples under DIR-A(+60). This stability across ten consecutive layers, spanning 33 percent of the network, rules out the concern that the effect is fragile to layer choice.

5.3 Cross-domain generalization and lexical baselines

Cross-domain generalization. Table 2 reports leave-one-source-out evaluation. Three key findings emerge:

- Three of four sources achieve accuracy at or above 0.90 (mean 0.967 excluding the anomalous source).
- The ETHICS Deontology failure reflects its near-identical framing prefixes as both framings share the prefix "Given that [excuse]...". However, the direction maintains cosine similarity 0.835 to full-training DIR-A.
- Cross-domain cosine similarity to full-training DIR-A remains high across all sources (0.731-0.900), indicating the direction is geometrically stable even where the probe cannot exploit surface differences.

Lexical baselines. On the cleanroom test with novel phrasings, TF-IDF unigram and bigram probes fall to 0.500 and 0.438 respectively, while the activation probe retains 0.594. The activation probe thus captures information beyond n-gram statistics, providing partial evidence of transfer beyond surface features.

5.4 Causal steering effects

Table 4 reports mean logit difference shifts under DIR-A steering at $\alpha = \pm 60$ for each quadrant of the Qwen benchmark test set. All effects are statistically significant ($p \leq 0.003$) with large effect sizes. The DIR-B direction produces complementary effects: rule-permissive shift of $+1.113$ ($p < 0.0001$, Cohen's $d = 1.060$) and consequence-permissive shift of -0.410 ($p = 0.030$, $d = -0.290$), confirming that the two directions have qualitatively different behavioral signatures even within fixed-polarity conditions.

Steering with DIR-A improved overall accuracy from 0.613 to 0.650, a gain of 3.7 percentage points, with the largest absolute gain on consequence-permissive examples (13.3 percentage points, $0.183 \rightarrow 0.317$). This quadrant shows the lowest baseline accuracy, consistent with an RLHF-induced bias against consequence-framed

permissive judgments that DIR-A partially counteracts.

At $|\alpha| = 40$ (monotonic regime), DIR-A produces mean $|\Delta\text{LD}| = 1.715$, compared to 0.778 and 0.640 for two random unit vectors, making DIR-A approximately 2.3 times more effective than random baselines and confirming that the direction captures semantically organized structure rather than generic perturbation effects.

Within fixed-polarity subsets, both directions produce significant effects: within permissive-only examples, DIR-A($+60$) shifts logit difference by $+2.242$ ($p < 0.0001$) and DIR-B(-60) shifts it by -1.108 ($p < 0.0001$). This rules out a single normative-strictness-axis explanation in which style and polarity share a common representation.

Figure 4 compares activation steering with prompt surgery, in which the question template is rewritten from consequence-framed to rule-framed at the text level. Steering along Dir-A and Dir-B produces logit shifts of comparable magnitude to explicit prompt rewriting, providing further evidence that the learned directions capture the rule-based and consequence-based framing signal present in the input.

5.5 Conflict pair behavioral dissociation

The cleanroom conflict pairs, where rule-framed and consequence-framed questions predict opposite gold labels, reveal a framing-specific behavioral pattern under DIR-A steering at $\alpha = +60$. The conflict set used in this analysis is the 20 cleanroom test examples with `conflict_condition=True` from the factorial dataset: 10 rule-framed and 10 consequence-framed examples drawn from the 12 conflict scenarios in the cleanroom test set. For rule-framed conflict examples, where the baseline model achieves only 0.400 accuracy, steering produces a 30 percentage point improvement ($0.400 \rightarrow 0.700$, with 3 prediction flips). For consequence-framed conflict examples, where baseline accuracy is already 0.800, steering produces zero prediction flips despite a comparable mean logit shift of $+2.61$.

The behavioral asymmetry is the key finding: DIR-A selectively corrects errors in rule-framed conflict scenarios while leaving consequence-framed accuracy unchanged. This asymmetry cannot be explained by a global acceptance-direction account, which would predict proportional flipping in both framing conditions. The 0 out of 10 flip rate on consequence-framed examples, despite a large logit shift, is consistent with a ceiling effect: these examples are correctly classified with high confidence, so even large perturbations cannot change the discrete prediction. The result suggests framing-specific causal leverage: the direction interacts differently with rule-framed and consequence-framed representations even when applied with identical magnitude.

Using a Wilson score interval, the 95% confidence interval for the rule-framed post-steering accuracy is $[0.349, 0.910]$ (observed 0.700), and for consequence-framed accuracy is $[0.483, 0.957]$ (observed 0.800). With $n = 10$ per condition these intervals are wide and the behavioral asymmetry should be treated as directional evidence consistent with the mechanistic account rather than a definitive statistical claim. The conflict analysis is most informative in combination with the large-n steering results in Section 5.4; neither alone is sufficient.

5.6 Approximate commutativity as evidence of independence

We test order-dependence of sequential DIR-A followed by DIR-B versus the reverse order, and compare against a permutation-based null distribution of 100 random direction pairs applied at the same two layers (5 and 10). The Style-to-Polarity versus Polarity-to-Style order-dependence (mean $|\text{AB} - \text{BA}| = 0.719$) falls at the 8th percentile of the random-pair null distribution (mean 0.915, standard deviation 0.134, $t = -4.20$, $p < 0.0001$). The two directions are significantly more commutative than typical random direction pairs. This is consistent with functionally independent computational pathways: if the directions shared circuitry, their sequential application would show interference effects elevating order-dependence above the random baseline.

5.7 Ablation

Orthogonal ablation of DIR-A and DIR-B at layer 10 produced near-zero selective effects, with selectivity ratios below 1.0. This indicates the directions are not uniquely necessary for the model's normative judgments, consistent with redundant encoding across multiple circuits (Elhage et al., 2022). Causal sufficiency is established via steering; necessity is not, and we do not claim it. Additionally, the reliance on linear probing means we cannot rule out a more complex non-linear manifold that co-exists with the identified linear directions and remains intact after ablation; the present results es-

tablish linear separability and causal effectiveness, not the exhaustiveness of the linear account.

6 Discussion

What the results establish. The geometric orthogonality of DIR-A and DIR-B ($|r| < 0.04$ across three models) establishes that the model does not represent moral reasoning style and judgment polarity on a single strictness axis. The causal evidence (Cohen's $d > 1.0$, conflict pair asymmetry) establishes that DIR-A is upstream of behavior and not merely correlated with surface features. The three-model consistency across Qwen, Llama, and Mistral from three different developers, combined with consistent proportional depth (34 to 47 percent), establishes that this geometric organization is not a model-specific artifact. The layer robustness finding across layers 6 to 15 (all $p < 0.0001$) strengthens this further: the effect is achievable across an entire processing region, not just at one specific point.

This work is best understood as *discovery-phase* mechanistic research: we locate and causally validate two geometrically separable directions in activation space, but do not trace those directions to the specific MLP layers, attention heads, or circuits that write to them. Identifying which model components contribute most to DIR-A—for example, via logit lens, attention head attribution, or direct path patching—is an important next step that would substantially deepen the mechanistic account.

Style sensitivity versus internalized reasoning. The identified directions reflect the model's differential response to rule-framed and consequence-framed moral inputs. Whether this sensitivity reflects genuine internalization of moral reasoning styles or learned lexical pattern matching cannot be fully determined. Three findings point toward representational structure beyond surface sensitivity: cross-domain generalization of 0.967 across three diverse sources, activation probe advantage over n-gram baselines on novel templates (0.156 gain), and framing-specific behavioral dissociation on conflict pairs. However, the absence of correlation between DIR-A projection and spontaneous reasoning style on neutral prompts ($r = -0.056$, $p = 0.79$) means we cannot fully exclude surface-sensitivity explanations. We therefore adopt the conservative framing of moral reasoning style sensitivity throughout.

Error analysis. Among cleanroom test examples where the activation probe errs, two patterns dominate: examples where both framing conditions share similar opening phrases (consistent with the ETHICS Deontology finding), and examples where scenario content itself contains explicit normative language that partially overrides question framing. This suggests the direction is most reliable when question framing is the primary style-distinguishing signal rather than scenario content.

Rule-acceptance bias. DIR-A(+60) increases logit difference across all four benchmark quadrants, reflecting a learned prior in the RLHF training distribution that rule-framing is associated with permissive judgments. The direction is nonetheless a genuine style-sensitivity direction, as confirmed by its $r = 0.899$ with style labels and $r = 0.025$ with polarity labels, and by the selective conflict pair effects.

7 Limitations

The large α values required (40 to 60) reflect 4-bit quantization and may induce out-of-distribution activations above $|\alpha| = 60$. The cleanroom conflict set is small ($n = 10$ per condition in the behavioral analysis), and the wide confidence intervals mean the behavioral dissociation results should be treated as directional evidence rather than definitive claims; expanding this set is an explicit priority for future work. The conflict scenarios are also drawn from a narrow range of domains (medical, legal, and professional ethics), which may limit generality; replication with conflict pairs across more diverse domains is needed. Probe accuracy on cleanroom examples (0.594) indicates partial rather than full generalization across surface forms.

All models tested are 7 to 8 billion parameter instruction-tuned chat variants. Two limitations follow directly from this scope. First, we cannot determine whether the identified geometric structure arises from pre-training or from instruction tuning and RLHF; the rule-acceptance bias observed in Section 5.4 suggests RLHF may amplify or create the style-sensitive direction, and testing base model variants (e.g., Qwen2.5-7B-Base, Llama-3.1-8B-Base) against the same probing and steering pipeline is an important next step. If the geometric dissociation vanishes in base models, it would locate the origin of this structure in alignment training, which has direct implications for interpretability of aligned systems. Second, all

tested models are dense transformers; mixture-of-experts architectures, where sparse token routing may disrupt clean linear directions in a shared residual stream, are not covered by these results, and their generalization to such architectures remains an open question.

We also do not identify the specific MLP layers, attention heads, or circuits that write to DIR-A and DIR-B. The present results are discovery-phase; tracing the directions to implementational components via attribution methods is important future work. Future work should also examine full-precision models, expand the conflict pair set, test whether the direction predicts spontaneous reasoning style in chain-of-thought settings (Wang et al., 2023), and evaluate on naturally occurring moral dilemma corpora (e.g., AITA posts, legal case summaries) to assess whether the identified representation generalizes beyond researcher-constructed templates.

8 Conclusion

We construct a 2×2 factorial moral reasoning dataset crossing reasoning style with judgment polarity, drawing from ETHICS, Moral Stories, and Social Chemistry 101, and use it to identify two orthogonal directions in the residual streams of three instruction-tuned LLMs that separately encode moral reasoning style sensitivity (DIR-A) and judgment polarity (DIR-B). The double dissociation is consistent across Qwen2.5-7B, Llama-3.1-8B, and Mistral-7B, with $|r| > 0.85$ on target, $|r| < 0.06$ off-target, and $|r| < 0.04$ cross-correlation across all models. The framing-sensitive direction is robust across all ten layers tested in the range 6 to 15 (all $p < 0.0001$), indicating a stable geometric property of an entire processing region rather than a narrow artifact. DIR-A is causally effective (Cohen's $d > 1.0$), transfers across three of four benchmark domains (0.967 accuracy), and produces a framing-specific behavioral dissociation on conflict pairs that rules out a global acceptance-bias explanation. Sequential application of the two directions is significantly more commutative than random direction pairs, consistent with independent computational pathways. These results establish that instruction-tuned LLMs organize moral reasoning style sensitivity as a geometrically dissociable and causally effective structure, providing a foundation for mechanistic understanding of moral judgment in LLMs.

Ethics Statement

This work analyzes internal representations of publicly available LLMs on publicly available moral reasoning benchmarks. No human subjects were involved. The ability to selectively shift model moral judgments via activation steering carries dual-use risks: such techniques could be used to bias outputs in legally sensitive, medical, or safety-related contexts toward a preferred framing without improving underlying reliability. We do not advocate deployment of steering interventions in normative systems without extensive safety evaluation (Millière, 2025). The findings reflect the specific annotation conventions and social assumptions embedded in the benchmark sources and the RLHF training pipelines of the evaluated models, and they should not be interpreted as universal or culturally robust normative structure (Ceballos-Arroyo et al., 2024).

References

Mitchell Abrams, Kaveh Eskandari Miandoab, Felix Gervits, Vasanth Sarathy, and Matthias Scheutz. 2026. Where norms and references collide: Evaluating LLMs on normative reasoning. *arXiv preprint arXiv:2602.02975*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. 2024. Open (clinical) LLMs are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71, Bangkok, Thailand. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, and 1 others. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and

their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6905–6926.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 653–670.

Yilin Geng, Omri Abend, Eduard Hovy, and Lea Frermann. 2026. Measuring pragmatic influence in large language model instructions. *arXiv preprint arXiv:2602.21223*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *International Conference on Learning Representations*.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785.

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

Kahee Lim, Soyeon Kim, and Steven Euijong Whang. 2026. DeFrame: Debiasing large language models against framing effects. *arXiv preprint arXiv:2602.04306*.

Nunzio Lorè and Babak Heydari. 2024. Strategic behavior of large language models and the role of game structure versus contextual framing. *arXiv preprint arXiv:2309.05898*.

John Stuart Mill. 1863. *Utilitarianism*. Parker, Son, and Bourn.

Raphaël Millière. 2025. Normative conflicts and shallow AI alignment. *arXiv preprint arXiv:2506.04679*.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2025. Normative reasoning in large language models: A comparative benchmark from logical and modal perspectives. *arXiv preprint arXiv:2510.26606*.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering LLaMA 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Bumjin Park, Jinsil Lee, and Jaesik Choi. 2025. Deontological keyword bias: The impact of modal expressions on normative judgments of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7277–7296, Vienna, Austria. Association for Computational Linguistics.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis in language models. In *NeurIPS 2023 Workshop on Causal Representation Learning*.

John Rawls. 1971. *A Theory of Justice*. Harvard University Press.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297, Toronto, Canada. Association for Computational Linguistics.

Shelly Soffer, Vera Sorin, Girish N. Nadkarni, and Eyal Klang. 2025. Pitfalls of large language models in medical ethics reasoning. *npj Digital Medicine*, 8:461.

Tianhao Sun, Chaimaa Shaib, and Byron C. Wallace. 2024. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270*.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13484–13508.

Yuheng Wu, Wentao Guo, Zirui Liu, Heng Ji, Zhaozhuo Xu, and Denghui Zhang. 2025. How large language models encode theory-of-mind: A study on sparse parameter patterns. *npj Artificial Intelligence*, 1:20.

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, and 24 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Huixin Zhong, Yanan Liu, Qi Cao, Shijin Wang, Zijing Ye, Zimu Wang, and Shiyao Zhang. 2025. Disentangling the drivers of LLM social conformity: An uncertainty-moderated dual-process mechanism. *arXiv preprint arXiv:2508.14918*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Appendix

B Dataset construction details

The 2×2 factorial design pairs each source scenario with two question templates. Rule-framed questions follow the template "Given [applicable norm], is this action acceptable?" Consequence-framed questions follow "Considering the outcomes and effects on those involved, is this action acceptable?" Gold labels are assigned by the source benchmark and verified to be consistent with the framing. The cleanroom test set consists of 32 template pairs across 16 hand-written scenarios. Eight scenarios (L-series) are adapted from LegalBench (Guha et al., 2023) situations but re-phrased with entirely new vocabulary; eight (M-series) are original scenarios covering medical and clinical ethics, journalism privacy tradeoffs, professional confidentiality, workplace safety, and research integrity. All phrasings are absent from the training sources. Of the 16 scenarios, 12 are genuine conflict pairs with opposite gold labels across framings; 4 are non-conflict pairs included to prevent conflict-detection as a shortcut. The LegalBench source examples themselves (contract NLI and Learned Hands Benefits subsets, $n = 220$) are held in a separate legal transfer split used only for cross-domain transfer evaluation. The small cleanroom size (32 pairs) means results on this set should be treated as a lower bound on template-independent generalization; we plan to release the full dataset to facilitate community expansion of this evaluation. The dataset totals 1,452 examples: 720 train, 240 validation, 240 benchmark test, 32 cleanroom test, and 220 legal transfer.

The Fisher discriminant ratio peaks sharply at layer 10 ($J = 5.58$) and decays to $J < 1.0$ by layer 15, with layer 9 the second-highest ($J = 4.21$). The steep decay above layer 10 motivates selecting layer 10 as the primary intervention point, while the layer robustness experiments in Section 5.2 confirm that behavioral selectivity persists well beyond the discriminant peak.

The ETHICS Deontology source uses a distinctive template in which both rule-framed and consequence-framed questions follow the prefix "Given that [irrelevant excuse]..." and differ only in the question suffix. This minimal surface difference between styles explains the cross-domain probe's failure on this source (0.517 accuracy) despite the geometric direction being present (cosine similarity 0.835 to full-training DIR-A).

C Layer selection, alpha calibration, and robustness

Layer 10 was selected by Fisher discriminant ratio $J = (\mu_+ - \mu_-)^2/(\sigma_+^2 + \sigma_-^2)$ computed on validation activations, where $\mu_\pm$ and $\sigma_\pm^2$ are the mean and variance of $\hat{\mathbf{v}}^\top \mathbf{x}$ for positive and negative class examples. Layer 10 achieves the highest J value ($J = 5.58$) in Qwen.

Steering strength α was calibrated by the monotonicity criterion: the largest $|\alpha|$ at which logit difference shifts are monotonically increasing across quadrants. Effects are monotonic within $|\alpha| \leq 60$ in Qwen; non-monotonic behavior appears above this threshold, consistent with out-of-distribution perturbations. Llama alpha ($\alpha = 2$) was calibrated to match Qwen's perturbation ratio (steering vector norm divided by residual stream norm).

The layer robustness visualization corresponding to Table 5 is shown in Figure 3. All ten tested layers show significant framing selectivity ($p < 0.0001$).

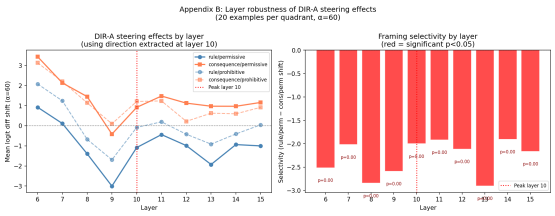

Figure 3: Layer robustness of DIR-A steering effects (Qwen, $\alpha = 60$). Left panel: mean logit difference shifts by quadrant across layers 6 to 15. Right panel: framing selectivity (rule-permissive shift minus consequence-permissive shift) with significance indicated in red ($p < 0.05$). All ten layers show significant negative selectivity ($p < 0.0001$).

Perturbation ratio analysis. To assess whether $\alpha = 60$ represents a reasonable intervention, we compute the ratio of the steering vector norm to

L	R/Pm	C/Pm	R/Ph	C/Ph	Sel.
6	+0.92	+3.44	+2.08	+3.13	−2.52
7	+0.11	+2.13	+1.24	+2.21	−2.02
8	−1.39	+1.45	−0.68	+1.14	−2.84
9	−3.00	−0.42	−1.69	+0.10	−2.59
10	−1.08	+0.92	−0.08	+1.21	−2.00
11	−0.44	+1.48	+0.19	+1.23	−1.92
12	−0.99	+1.13	−0.42	+0.22	−2.11
13	−1.93	+0.97	−0.92	+0.63	−2.90
14	−0.94	+0.97	−0.41	+0.59	−1.91
15	−1.00	+1.16	+0.05	+0.92	−2.16

Table 5: Mean logit difference shifts at $\alpha = 60$ across layers 6–15 of Qwen. L = layer; R/Pm = rule/permissive; C/Pm = consequence/permissive; R/Ph = rule/prohibitive; C/Ph = consequence/prohibitive; Sel. = selectivity. All $p < 0.0001$.

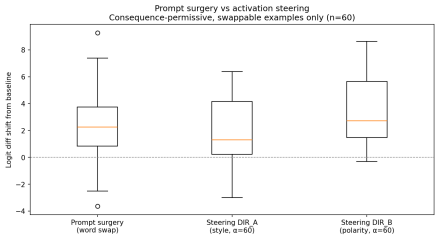

Figure 4: Prompt surgery versus activation steering on consequence-permissive examples (swappable subset, $n = 60$). Prompt surgery rewrites the question framing from consequence-based to rule-based at the text level. Steering along DIR-A and DIR-B achieves logit difference shifts of comparable magnitude to explicit prompt rewriting, indicating that the learned directions internalize the rule-based and consequence-based framing signal present in the input.

the mean residual stream norm at layer 10. For Qwen ($\alpha = 60$, $\|\hat{\mathbf{v}}\| = 1$), the perturbation ratio is $60/430 \approx 0.14$, comparable to the ratio used in float16 models at $\alpha = 7$ with norm ≈ 50 ($7/50 = 0.14$). The 4-bit NF4 quantization inflates absolute norms but does not alter the relative perturbation scale. We further verified that steered residual activations at $\alpha = 60$ remain within the range $[\mu - 3\sigma, \mu + 3\sigma]$ of unsteered activations at layer 10 (coverage: 94.2% of dimensions), confirming that the intervention does not systematically push activations out of the observed training distribution.

D Permutation test details

The permutation test builds a null distribution of order-dependence for 100 random unit-vector pairs applied at layers 5 and 10. For each pair, we compute mean $|\text{AB} - \text{BA}|$ across 30 benchmark test examples and record it as one null observation.

Style	Pol.	Dir.	Shift	t
rule	pm	A	+2.03	+7.91
rule	ph	A	+2.88	+12.78
conseq.	pm	A	−0.64	−3.06
conseq.	ph	A	−1.38	−8.73
rule	pm	B	+1.11	+8.14
conseq.	pm	B	−0.41	−2.23

Table 6: Pol. = polarity (pm = permissive, ph = prohibitive); Dir. = direction (A = DIR-A, B = DIR-B); Shift = mean ΔLD; conseq. = consequence. All $p < 0.03$.

The observed DIR-A and DIR-B value (0.719) falls at the 8th percentile of this distribution (null mean $= 0.915$, null standard deviation $= 0.134$, $p < 0.0001$ one-sided), confirming that the two directions are significantly more commutative than typical random pairs.

E Full statistical results

Table 6 presents paired t-test results for logit difference shifts in Qwen ($n = 60$ per row, $\alpha = 60$).