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

Eugenie Shi
Department of Computer Science
Stanford University
Stanford, CA 94305, USA
{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 ask whether a model's differential response to rule-framed and consequence-framed moral questions reflects distinct internal structure or merely surface lexical sensitivity. We construct a 2×2 factorial dataset crossing reasoning style (rule-based vs. consequence-based framing) with judgment polarity (permissive vs. prohibitive), drawn from three established moral reasoning benchmarks and evaluated on a cleanroom test set of novel phrasings that includes conflict pairs where the two framings predict opposite labels. Using linear probing and activation steering, we identify two orthogonal directions in the residual stream: DIR-A tracks reasoning style and DIR-B tracks judgment polarity, each near-blind to the other factor (cross-correlation $|r| < 0.04$). This double dissociation replicates across three instruction-tuned models from different developers at comparable proportional depth, and it is not knife-edged in layer: the directions are recoverable, and causally effective, throughout an extended band of the network. DIR-A is causally sufficient to shift judgments (Cohen's $d > 1.0$), and on conflict scenarios steering selectively corrects rule-framed errors while leaving consequence-framed judgments unchanged, a pattern inconsistent with a global acceptance bias. Generalization beyond the training templates is partial: the direction transfers across benchmark sources, but its advantage over lexical baselines on novel phrasings is modest. We therefore frame the result as moral reasoning style sensitivity, a geometrically dissociable and causally effective structure, rather than internalized normative reasoning.

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).[1] 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

[1]We use *rule-based* and *consequence-based* rather than *deontological* and *utilitarian* deliberately. The philosophical terms name normative theories that carry commitments our stimuli do not probe, such as universalizability or the aggregation of welfare across persons, and our directions license no claims about those commitments. What we manipulate is how a question is posed. A model whose activations separate our two conditions has distinguished two ways of being asked, not adopted a moral theory. We reserve the theory names for the theories and use framing terms for the manipulation.

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, though not a hand-specified keyword detector.

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.

2 Related Work

Linear representations in LLMs. 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.

Moral and normative reasoning in LLMs. 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.

Framing effects, cognitive bias, and instruction sensitivity. 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 & 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.

Activation steering and mechanistic analysis. 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 balancing reasoning style against polarity across the dataset, 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. The core moral split comprises 720 training, 240 validation, and 240 benchmark test examples, with strict label balance in the margins: 180 examples per reasoning style and polarity cell in the training set, and 60 per cell in each of validation and benchmark test. Two further splits are held out: a 32-item cleanroom test set of hand-written novel phrasings, and a 220-item legal transfer set drawn from LegalBench (Guha et al., 2023).

We additionally construct a **cleanroom test set** of 32 hand-written items, 16 moral and 16 legal in subject matter, of which 20 form 10 conflict pairs in which rule-framed and consequence-framed questions on a shared scenario carry opposite gold labels. All phrasings are original to this set and absent from the training sources, enabling evaluation of generalization beyond learned templates.

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. The absolute value of α is not interpretable in isolation; the meaningful quantity is the *perturbation ratio*, the steering vector norm divided by the mean residual stream norm at the intervention layer. Four-bit NF4 quantization (Dettmers et al., 2023) inflates absolute residual norms uniformly (approximately 430 at layer 10 in Qwen, against approximately 50 in float16 models) without changing the relative scale of an intervention. Our $\alpha = 40$ to 60 for Qwen therefore corresponds to perturbation ratios of 0.09 to 0.14, matching $\alpha \approx 7$ in an unquantized model of the same architecture. We verify that the intervention remains in-distribution: at $\alpha = 60$, 94.2 percent of steered residual dimensions fall within $[\mu - 3\sigma, \mu + 3\sigma]$ of the unsteered activation distribution at layer 10. For Llama, α is calibrated to the same perturbation ratio, yielding $\alpha = 2$; the fact that a calibrated $\alpha = 2$ in one model and $\alpha = 60$ in another produce the same qualitative effects is itself evidence that the ratio, not the raw magnitude, is the operative variable. We further verify that steering effects are monotonically increasing within $|\alpha| \leq 60$ and report $\alpha = 60$ unless stated otherwise. Full details are in Appendix B.

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

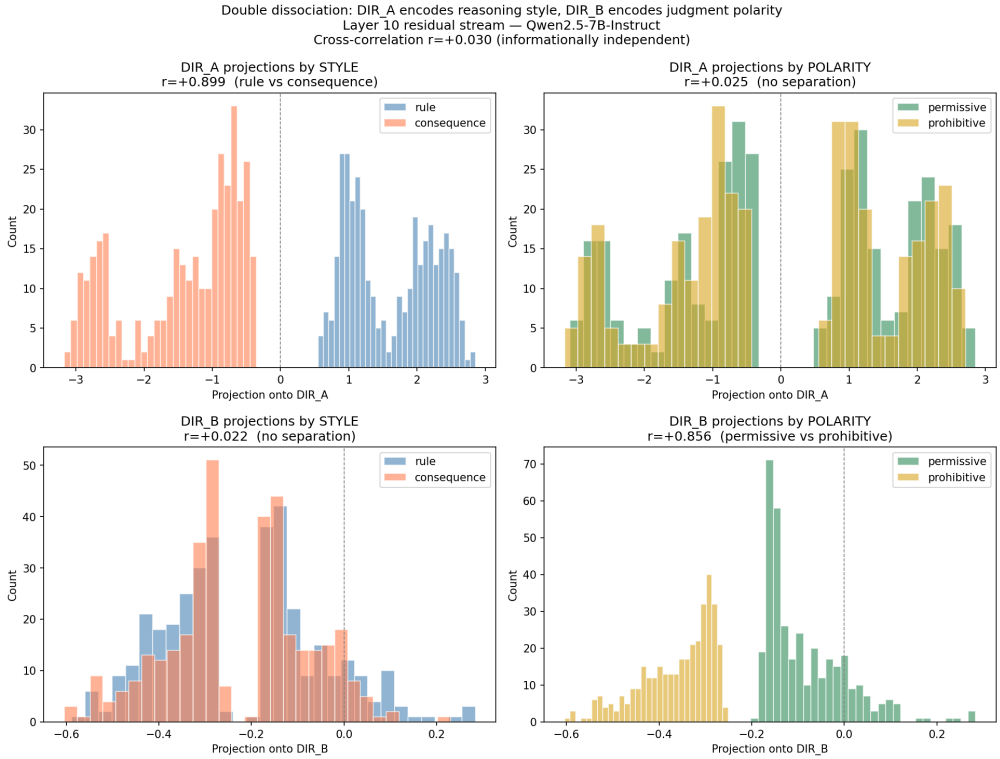

Figure 1: Double dissociation across three LLMs. All models show DIR-A $|r(\text{style})| > 0.87$ and DIR-B $|r(\text{polarity})| > 0.85$, with cross-correlation $|r| < 0.04$.

negligible while logit-space shifts remain meaningful. Note that this measurement operates only over these eight tokens and never over option letters. Where a source benchmark distributes items in multiple-choice form, we reformat the scenario into our binary acceptability frame and discard the original option labels, so no A/B/C/D token enters the logit difference at any point. Appendix A shows a source item alongside both of its templated forms.

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. This is consistent with the layer profiles in Section 5.2 and Figure 2, where framing selectivity is strongest at 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

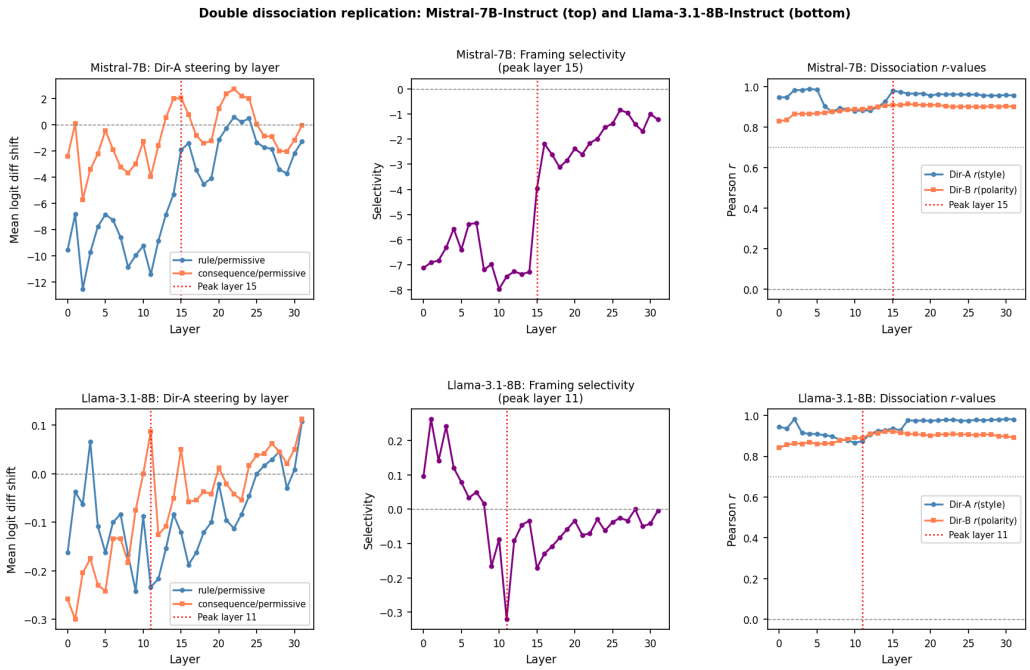

Figure 2: Double dissociation replication in Mistral-7B-Instruct-v0.3 (top row) and Llama-3.1-8B-Instruct (bottom row).

Model	Peak layer	DIR-A r(sty)	DIR-A r(pol)	DIR-B r(pol)	Cross-r
Qwen2.5-7B	10/28 (35.7%)	+0.899	+0.025	+0.856	+0.030
Llama-3.1-8B	11/32 (34.4%)	+0.875	+0.002	+0.888	−0.032
Mistral-7B	15/32 (46.9%)	+0.978	+0.007	+0.909	−0.021

Table 1: Double dissociation across three instruction-tuned LLMs from different developers.

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

Two questions about layer choice are distinct, and we answer them separately: whether the geometry exists only at one depth, and whether the causal effect survives moving the intervention layer.

On the first, Figure 2 reports the dissociation at every layer of Mistral and Llama (32 each). DIR-A r(style) and DIR-B r(polarity) exceed 0.85 across virtually all layers in both models, so the directions are recoverable throughout the network rather than at a single privileged depth.

On the second, we steer with the fixed layer-10 direction at each of layers 6 to 15 of Qwen. We restrict the causal sweep to this band rather than the full 28 layers because layer selection is governed by the Fisher discriminant ratio, which peaks sharply at layer 10 ($J = 5.58$) and decays below 1.0 by layer 15 (Appendix B). Intervening where the discriminant has decayed would test whether a direction carrying little class signal still moves behavior, which is not a robustness check on the present result. The informative question is whether the effect is knife-edged at the discriminant peak, and Table 6 shows that it is not: all ten tested layers achieve significant framing selectivity ($p < 0.0001$), with the qualitative pattern consistent throughout. Framing selectivity is the difference in mean logit shift between rule-

Held-out source	n_{test}	Accuracy	$\cos(\hat{v}_{\text{cd}}, \text{DIR-A})$
Social Chemistry 101	60	1.000	0.731
Moral Stories	60	0.900	0.900
ETHICS Commonsense	60	1.000	0.777
ETHICS Deontology	60	0.517	0.835
Mean (excl. deontology)		**0.967**	0.803

Table 2: Cross-domain generalization of DIR-A (Qwen, layer 10).

Probe type	Benchmark test	Cleanroom test
Keyword matching	0.879	0.750
TF-IDF (unigram)	1.000	0.500
TF-IDF (bigram)	1.000	0.438
Activation probe (layer 10)	1.000	**0.594**

Table 3: Probe accuracy on benchmark and cleanroom test sets.

permissive and consequence-permissive examples under DIR-A(+60); negative selectivity is the expected sign, since DIR-A(+60) steers toward the consequence direction. The effect is a property of a ten-layer processing region, not of a single hand-picked intervention point. The corresponding visualization is Figure 3 in Appendix B.

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. Table 3 compares the activation probe against three lexical baselines. Both TF-IDF probes reach 1.000 in-domain but fall to chance on cleanroom (0.500, 0.438), showing their benchmark performance was attributable entirely to n-grams of the training templates. The activation probe retains 0.594, a 0.156 gain over the stronger variant, indicating it encodes information not reducible to those statistics. Keyword matching is different in kind: a hand-specified detector for framing markers rather than a fitted model. Because a question must still signal which framing it asks, some markers survive our cleanroom rewrites by construction, and keyword matching scores 0.750, above the activation probe. The cleanroom set defeats learned surface statistics but not a human-designed cue detector. We therefore claim only that the probe generalizes beyond the lexical statistics its training templates supply, not that it is the strongest available style classifier.

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

Style	Polarity	α	Mean shift	Std	Cohen's d	p
rule	permissive	$+60$	$+2.025$	1.967	$+1.029$	< 0.0001
rule	prohibitive	$+60$	$+2.877$	1.729	$+1.664$	< 0.0001
consequence	permissive	-60	-0.640	1.605	-0.398	0.003
consequence	prohibitive	-60	-1.377	1.212	-1.137	< 0.0001

Table 4: Logit difference shifts under DIR-A steering ($n = 60$ per row, Qwen, $\alpha = 60$). All effects significant by paired t-test.

($p = 0.030$, $d = -0.290$), 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 $\alpha = +60$. 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). We acknowledge that with $n = 10$ per condition these intervals are wide, and treat this as behavioral evidence consistent with the mechanistic account rather than a definitive statistical claim.

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 $|AB - BA| = 0.719$) 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.

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.

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, an activation probe advantage of 0.156 over the strongest n-gram baseline on novel templates, and framing-specific behavioral dissociation on conflict pairs. Against this, the probe does not exceed a hand-specified keyword detector on the cleanroom set (0.594 vs. 0.750), and 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.

Limitations. The large α values required (40 to 60) are an artifact of 4-bit quantization rather than an unusually aggressive intervention, as the perturbation ratio analysis in Appendix B establishes; nonetheless, we have not replicated at full precision, and $|\alpha| > 60$ induces non-monotonic, likely out-of-distribution behavior. The cleanroom conflict set is relatively small. The factorial design is balanced in its margins but not within source (Appendix A), so reasoning style is partially predictable from source identity. A within-source crossed design would isolate the style manipulation more cleanly. 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, and base models and larger frontier models may show different geometric organization. The rule-acceptance bias observed in Section 6 suggests RLHF may amplify or create the style-sensitive direction, and testing base variants of Llama and Mistral is an important direction for future work. Future work should examine full-precision models, expand the conflict pair set, and test whether the direction predicts spontaneous reasoning style in chain-of-thought settings (Wang et al., 2023). Finally, we establish that reasoning style is dissociable and steerable, but not that steering it confers control on a downstream task of independent interest. The natural next step is to test whether DIR-A supports targeted control on an applied benchmark, either by holding judgment polarity fixed while shifting the framing a model reasons from, or by auditing whether a deployed model's normative outputs are framing-driven. We regard this as the most valuable extension of the present work.

7 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. Where norms and references collide: Evaluating LLMs on normative reasoning.

arXiv preprint arXiv:2602.02975, 2026.

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

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. Open (clinical) LLMs are sensitive to instruction phrasings. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii (eds.), *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 50–71, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.5. URL https://aclanthology.org/2024.bionlp-1.5/.

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

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6905–6926, 2021. URL https://aclanthology.org/2021.emnlp-main.54.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 653–670, 2020. URL https://aclanthology.org/2020.emnlp-main.48.

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

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, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2023.

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

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785, 2024. doi: 10.1162/tacl_a_00673. URL https://aclanthology.org/2024.tacl-1.43/.

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.

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

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

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

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

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

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

Bumjin Park, Jinsil Lee, and Jaesik Choi. Deontological keyword bias: The impact of modal expressions on normative judgments of language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7277–7296, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.360. URL https://aclanthology.org/2025.acl-long.360/.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis in language models. In *NeurIPS 2023 Workshop on Causal Representation Learning*, 2023. URL https://neurips.cc/virtual/2023/77537.

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

Gabriel Simmons. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In Vishakh Padmakumar, Gisela Vallejo, and Yao Fu (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 282–297, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-srw.40. URL https://aclanthology.org/2023.acl-srw.40/.

Shelly Soffer, Vera Sorin, Girish N. Nadkarni, and Eyal Klang. Pitfalls of large language models in medical ethics reasoning. *npj Digital Medicine*, 8:461, 2025. doi: 10.1038/s41746-025-01792-y. URL https://www.nature.com/articles/s41746-025-01792-y.

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

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. URL https://arxiv.org/abs/2308.10248.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 13484–13508, 2023. URL https://arxiv.org/abs/2212.10560.

Yuheng Wu, Wentao Guo, Zirui Liu, Heng Ji, Zhaozhuo Xu, and Denghui Zhang. How large language models encode theory-of-mind: A study on sparse parameter patterns. *npj Artificial Intelligence*, 1:20, 2025. doi: 10.1038/s44387-025-00031-9. URL https://www.nature.com/articles/s44387-025-00031-9.

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, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

Huixin Zhong, Yanan Liu, Qi Cao, Shijin Wang, Zijing Ye, Zimu Wang, and Shiyao Zhang. Disentangling the drivers of LLM social conformity: An uncertainty-moderated dual-process mechanism. *arXiv preprint arXiv:2508.14918*, 2025. URL https://arxiv.org/abs/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, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A 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 hand-written items (16 moral, 16 legal in subject matter), of which 20 form 10 conflict pairs where rule-framed and consequence-framed questions on a shared scenario carry opposite gold labels. All phrasings are original to this set and absent from the training sources. Separately, we hold out a legal transfer split of 220 items drawn from LegalBench (Guha et al., 2023) (160 Contract NLI, 60 Learned Hands Benefits), which is not used in the results reported here.

Table 5 shows representative items in their templated form. Source scenarios are reproduced verbatim; templating adds only the question, which supplies the norm clause for rule-framed items and the outcome clause for consequence-framed items. No option labels appear in any prompt: where a source distributes items in multiple-choice form, the scenario is reformatted into the binary acceptability frame and the original option labels are discarded.

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).

B 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

Source (style)	Scenario	Question
Social Chem. 101 (rule)	being careful around animals you don't know.	Given the social rule that you should be careful around animals you don't know, is this action acceptable?
ETHICS Comm. (cons.)	I noticed the fruit was rotting, so I put it in my trash can.	Considering the outcomes and effects on those involved, is this action acceptable?
Moral Stories (rule)	Caden watches Adrian have a panic attack in the middle of a store. Caden wants Adrian to calm down.	Given the rule that it's good for people to care about others, is it acceptable to Caden hugs Adrian and leads him to a quiet place in the store?
Moral Stories (cons.)	Vince is an accountant with a company. Vince wants to get some extra money.	Considering the likely effects on those involved, is it acceptable to Vince takes some money and works the accounting so that no one can see the loss?
ETHICS Deont. (rule)	Will you fold the towels and put them away?	Given that no the towels make me itchy, does this excuse justify setting aside the applicable moral rule?
ETHICS Deont. (cons.)	Aren't you supposed to give the dog a bath?	Given that no because I'm taking him to the groomer today, do the circumstances make this action acceptable despite the general prohibition?
Cleanroom (conflict, rule)	A doctor has one dose of a rare medication. Two patients need it: a 10-year-old child with decades ahead and a 70-year-old patient who was on the waiting list first.	Given that the first-come-first-served allocation rule must be followed, is it acceptable to give the medication to the 70-year-old patient?
Cleanroom (conflict, cons.)	*(same scenario)*	Given that giving the medication to the child would save approximately 60 more life-years, is it acceptable to give it to the 70-year-old patient?

Table 5: Representative templated items. The cleanroom conflict pair shares a scenario across framings and carries opposite gold labels (Acceptable / Unacceptable).

Layer	rule/perm	cons/perm	rule/proh	cons/proh	Selectivity (p)
6	$+0.922$	$+3.438$	$+2.081$	$+3.131$	-2.516 (< 0.0001)
7	$+0.113$	$+2.131$	$+1.241$	$+2.206$	-2.019 (< 0.0001)
8	-1.394	$+1.450$	-0.675	$+1.137$	-2.844 (< 0.0001)
9	-3.003	-0.416	-1.694	$+0.097$	-2.587 (< 0.0001)
10	-1.084	$+0.916$	-0.081	$+1.209$	-2.000 (< 0.0001)
11	-0.441	$+1.478$	$+0.194$	$+1.234$	-1.919 (< 0.0001)
12	-0.988	$+1.125$	-0.422	$+0.219$	-2.112 (< 0.0001)
13	-1.928	$+0.972$	-0.916	$+0.625$	-2.900 (< 0.0001)
14	-0.938	$+0.969$	-0.409	$+0.594$	-1.906 (< 0.0001)
15	-1.003	$+1.159$	$+0.047$	$+0.916$	-2.163 (< 0.0001)

Table 6: Mean logit difference shifts at $\alpha = 60$ across layers 6 to 15 of Qwen.

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 6 is shown in Figure 3. All ten tested layers show significant framing selectivity ($p < 0.0001$).

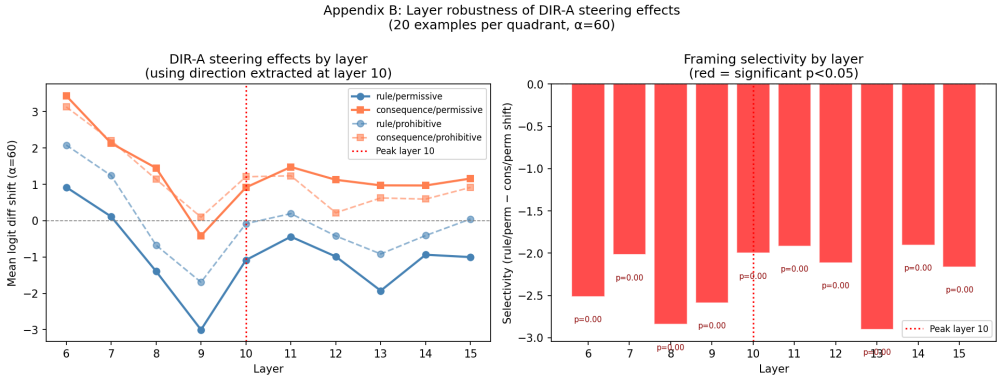

Figure 3: Layer robustness of DIR-A steering effects (Qwen, $\alpha = 60$). Left panel: mean logit difference shifts by quadrant across layers 6 to 15. Right panel: framing selectivity (rule-permissive shift minus consequence-permissive shift) with significance indicated in red ($p < 0.05$). All ten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

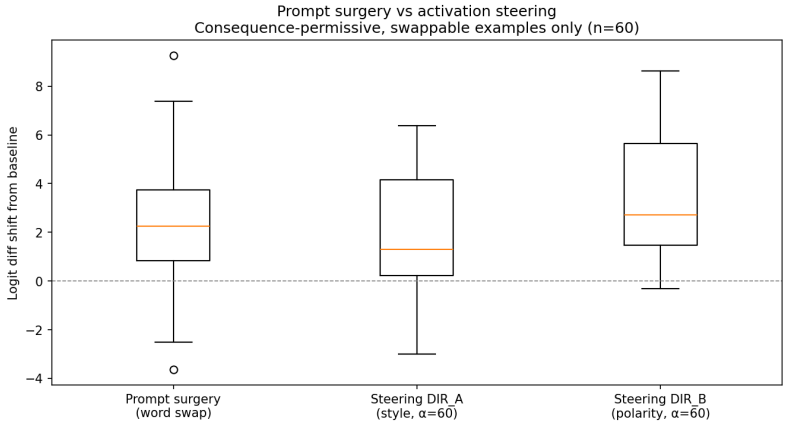

Figure 4: Prompt surgery versus activation steering on consequence-permissive examples (swappable subset, $n = 60$). Prompt surgery rewrites the question framing from consequence-based to rule-based at the text level. Steering along DIR-A and DIR-B achieves logit difference shifts of comparable magnitude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

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

C 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 $|\mathrm{AB} - \mathrm{BA}|$ across

30 benchmark test examples and record it as one null observation. 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.

D Full statistical results

Style	Polarity	Direction	α	Mean shift	Std	t	p
rule	perm	DIR-A	$+60$	$+2.025$	1.967	7.906	< 0.0001
rule	proh	DIR-A	$+60$	$+2.877$	1.729	12.783	< 0.0001
consequence	perm	DIR-A	-60	-0.640	1.605	-3.061	0.003
consequence	proh	DIR-A	-60	-1.377	1.212	-8.730	< 0.0001
rule	perm	DIR-B	$+60$	$+1.113$	1.049	8.142	< 0.0001
consequence	perm	DIR-B	$+60$	-0.410	1.416	-2.226	0.030

Table 7: Full paired t-test results for logit difference shifts ($n = 60$ per row, Qwen benchmark test).

E LLM Use Disclosure

The authors used Claude (Anthropic), Gemini (Google), and ChatGPT (OpenAI) for the following purposes: refining writing drafts in the paper; generating BibTeX entries from author-provided metadata including URLs, author lists, titles, and publication venues when BibTeX was not directly available from the referenced paper website; and assisting with code to debug and generate figures and plots, with underlying data provided by the authors and generated plots verified for correctness.