Legal Knowledge and Information Systems
R. Markovich et al. (Eds.)

doi:10.3233/FAIA251581

Which Neurons Nudge Normative Stance? Causal Tests and Mechanistic Evidence via Contrastive Last-Token Steering

Davide LIGA [a,1], Liuwen YU [a]
[a] *University of Luxembourg*
ORCiD ID: Davide Liga https://orcid.org/0000-0003-1124-0299, Liuwen Yu https://orcid.org/0000-0002-7200-6001

Abstract. Normative stance underlies decisions in law, legal reasoning, policy, and safety-critical settings. A model's judgment of what is permissible vs. impermissible often determines its downstream behavior. We study how to steer a language model's normative stances at inference time by adding a tiny, contrastive perturbation to the *last-token* neural activation in late MLP layers (contrastive last-token steering). For each normative prompt, we construct a contrast direction by comparing its last-token activation to that of a minimally edited variant that implies a more permissive normative stance (e.g., "acceptable" rather than "wrong"). During generation, we add this vector at the last token; a single strength parameter α controls how strongly and in which direction we push the model's stance (permissive vs. restrictive). Impact is measured as the change in a next-token logit margin between permissive and restrictive continuations. To avoid overclaiming, we calibrate a threshold τ on neutral controls (same layers, tempered strengths with $|\alpha| \leq 1$) and count success only when the shift exceeds τ in the expected direction. We also assess *specificity* by verifying that, on neutral control prompts, steered outputs exactly match unsteered baselines. Beyond component-level tests, we probe *neuron-level locality* by steering only the top-k contrastive neurons (ranked by last-token contrast) and confirming reversibility on our test set: $+\alpha$ produces the shift and $-\alpha$ reverses it. The method is training-free, uses standard forward hooks, and we report pilot results on Llama-3-8B-Instruct.

Keywords. Large Language Models, Normative Alignment, Normative Reasoning, Mechanistic Interpretability

1. Introduction

Large language models (LLMs) now routinely engage in tasks that require moral and normative judgment—choosing between conflicting values, resolving ethical and *legal* dilemmas, and aligning with user-defined principles. LLMs are increasingly assessed and even used around morally salient judgments. When prompted with trolley-style scenar-

[1] Corresponding Author: Davide Liga, davide.liga@uni.lu.

ios or everyday social dilemmas, these models can articulate moral reasoning and produce graded or binary decisions [1,2]. At the same time, moral judgments are highly sensitive to prompt framing [3] and exhibit cultural biases and limits [4,5]. Some experiments show that ChatGPT's moral advice can be inconsistent yet still shifts users' decisions [6], and LLMs can outperform humans on social situational judgment tests [7]; moreover, people may rate LLMs' ethical advice as rivaling a professional ethicist [8]. Yet while LLMs' outputs can be coherent, the internal mechanisms that steer these normative choices remain opaque. Understanding which neural components encode normative preferences allows us to explore the alignment of LLMs with norms and to ensure that automated systems remain accountable and transparent in their normative reasoning (which is crucial in the legal domain). In this paper, we ask the following research question: Which internal representations are responsible for a model's moral stance, and can we *causally* intervene on them without retraining?

Terminology. We call a stance *permissive* when continuations such as *"acceptable"/"justified"* are preferred over *"wrong"/"unacceptable"*, and *restrictive* otherwise. *Specificity* means the intervention leaves unrelated outputs unchanged; we measure it as canonicalized exact–match of steered vs. unsteered greedy generations on neutral controls. For example, the minimal pair *"Lying to save lives is morally ..."* vs. *"Lying to save face is morally ..."* implies opposite stances under small textual changes (*mutatis mutandis*, this is analogous to minimal, but substantial fact variations in case law that may flip a legal judgment).

Our approach. We study these questions via *contrastive activation steering*. Rather than swapping whole activations, we compute a *direction of change* between a minimally different prompt pair that elicits opposite stances ("permissive" vs. "restrictive"). We then inject a small, controlled perturbation along this contrast direction at the *last input position* inside selected MLP blocks (primarily late layers) during the forward pass. The intervention is scaled by a single real parameter α (the *strength*); positive $+\alpha$ is aligned—by a brief sign probe—to push toward the stance predicted by the counterfactual prompt, while negative $-\alpha$ reverses the effect. We quantify causal impact using a calibrated *next-token logit margin* between mutually exclusive continuations corresponding to the two stances. For moral pairs we always steer the last token; for neutral controls we keep the same last-token position but use masking to check that outputs remain unchanged.

Evaluation frame. To avoid over-interpreting single runs, we adopt a compact but principled evaluation pipeline:

1. **Directional validity.** At a given α, does the decision shift Δ have the expected sign (toward permissive for $+\alpha$, toward restrictive for $-\alpha$) as predicted by the counterfactual pair?
2. **Calibrated decision threshold.** We estimate a threshold τ from neutral control prompts using the same layers and a tempered subset of strengths ($|\alpha| \leq 1$, i.e., $\alpha \in \{-1.0, -0.8, 0.8, 1.0\}$), and count a *success* only when $\Delta \cdot$ sign expected exceeds τ.
3. **Specificity.** We test that steering leaves unrelated outputs unchanged by comparing steered vs. unsteered generations on neutral controls (canonicalized exact match), reporting preservation and corruption rates.
4. **Dose–response and reversibility.** We sweep α over a symmetric grid (positive/negative). Neuron-level tests include explicit reversibility checks ($+\alpha$ vs.

$-\alpha$). At the component level we report success across strengths; a formal monotonicity score is left for future work.

5. **Neuron-level locality.** Within late MLP layers, we select top-k neurons by last-token contrast magnitude and test whether small, sign-aligned subsets can reproduce (and reverse) the steering effect.

Contributions. We present a compact pipeline for causally steering moral stances in pretrained LLMs, showing mechanistic evidence that paves the way for an integration between the field of AI&Law and the young field of *Mechanistic Interpretability*:

1. **Contrastive last-token steering.** A causal intervention that targets *contrastive, last-token* directions by modifying MLP outputs that feed the residual stream (we emphasize late layers but also test selected earlier layers). The method requires no fine-tuning and uses standard forward hooks during inference only.
2. **Calibrated logit-margin metric.** A decision-shift metric based on the change in logit margin between opposed completions, with a neutral-control–derived threshold for consistent comparisons across prompts, layers, and steering strengths.
3. **Neuron-level locality.** A neuron-level analysis in late MLPs showing that small, sign-aligned subsets (top-k by contrast) can flip stance on our test set and exhibit reversibility with $+\alpha/-\alpha$.

Together, these components provide a reproducible workflow for measuring and controlling moral decisions in LLMs using only inference-time interventions.

Scope. We evaluate on Meta-Llama-3-8B-Instruct and a set of moral minimal pairs; claims are about steering effects within this setting. Component-level dose–response is reported via strength sweeps; explicit monotonicity scoring is left as future work.

Paper organization. Section 2 contextualizes this work in the literature. Section 3 presents the contrastive last-token steering method. Section 4 details the experimental setup and reports results. Section 5 discusses limitations and future directions, and concludes this work.

2. Related Work

Moral behavior in language models. LLMs exhibit moral and social judgments that appear to reflect implicit values learned during pretraining. Prior work has evaluated these behaviors using curated benchmarks and ethical dilemmas. For example, [9] introduced Delphi to make ethical judgments across diverse scenarios, and [10] proposed ETHICS to assess dimensions such as justice, virtue, and utilitarianism. [11] analyze moral foundations in LLMs, while [12] explore self-refinement strategies that can influence responses to moral questions. Much of this literature is output-centered; the internal mechanisms producing these stances remain comparatively less characterized, despite their importance for domains such as computational legal reasoning, where explanations of normative stance are critical.

Mechanistic interpretability and causal interventions. Mechanistic interpretability aims to reverse-engineer computations inside transformers [13], including work on superposition and feature sharing [14] and on targeted weight edits such as ROME [15]. A complementary line of research uses *activation patching*/causal tracing to test how

swapping or modifying hidden states affects behavior. Our approach follows this causal-intervention paradigm but in a localized setting: we construct a *contrastive direction* from minimally different prompt pairs and add a small, inference-time perturbation at the *last token* inside late MLP blocks, leaving model weights unchanged.

Probing vs. causal control. Linear probes and concept vectors (e.g., CAVs; [16]) indicate that a representation correlates with a concept, but correlation does not by itself imply that manipulating that representation *causes* behavior to change. Bias and social-attribute evaluations (e.g., [17]) similarly diagnose tendencies without isolating mechanism. By contrast, we emphasize *causal* tests: we measure a calibrated change in a next-token logit margin when we add a small, sign-aligned perturbation to late-layer MLP activations. Beyond component-level tests, we examine *neuron-level locality* by steering only the top-k contrastive neurons and verifying reversibility with $+\alpha/-\alpha$.

Steering without fine-tuning. Inference-time steering spans prompt-based control, activation editing, and representation arithmetic. Our contribution fits within this space but differs in three respects: (i) we derive *contrastive, last-token* directions from minimally changed moral pairs (rather than global directions or weight edits), (ii) we *calibrate* a decision threshold on neutral controls to avoid overclaiming spurious flips, and (iii) we validate *locality and reversibility* by showing that small neuron subsets in late MLPs can reproduce and reverse the effect. This yields an efficient, reproducible pipeline for causal moral steering using standard forward hooks at inference time.

AI&Law. Integrating symbolic and sub-symbolic AI is becoming increasingly important in the field of AI&Law [18], and Mechanistic Interpretability offers a fundamental opportunity to explore new integrations between these two traditional paradigms in AI. Our work is a first attempt to propose an integration between these two communities.

3. Method

We describe an interventional, inference-time procedure to steer a pretrained language model's moral stance by adding small, targeted perturbations to late MLP activations. The procedure has four components: (i) a *decision signal* defined as a next-token logit margin between mutually exclusive moral continuations; (ii) *contrastive last-token steering* that injects a direction computed from minimally different prompt pairs; (iii) *calibration and specificity* using neutral controls; and (iv) a *neuron-level* variant that targets small subsets of units.

At a glance. For each moral pair we (a) build a last-token direction from the base vs. source contrast (or a PCA fallback—last-token, position-aware—if unavailable), (b) orient it via a one-shot *sign probe* and inject it into selected late MLP layers at strength α, and (c) score the change in the next-token margin against a calibrated threshold; we also (d) check specificity on neutral controls and (e) run a neuron-level variant targeting top-k units. We focus on last-token interventions because late-layer MLP features tend to consolidate decision signals; in practice our component-level run steers a fixed set of MLP layers spanning early through late blocks, while neuron-level tests focus on late layers where the effect is most localized. The steering strength α is not learned: we sweep a symmetric grid and select α by a calibrated success metric, with the threshold τ estimated from neutral controls using the same layers and tempered strengths.

3.1. Task and Decision Signal

We work with *minimal pairs* of prompts $(P_{\text{base}}, P_{\text{src}})$ that differ by a small surface change but imply opposite stances ("permissive" vs. "restrictive"). Let x be the tokenized input for P_{base}, and let y_+ and y_- denote mutually exclusive next-token continuations aligned with "permissive" and "restrictive" readings, respectively.[2] Given next-token logits $\ell(\cdot)$, we define the *logit margin* as $m = \ell(y_+) - \ell(y_-)$, and the *decision shift* for an intervention as $\Delta = m_{\text{patched}} - m_{\text{base}}$. Positive Δ indicates movement toward a permissive decision.

3.2. Contrastive Last-Token Steering

Let $h^\ell(x) \in \mathbb{R}^{T\times H}$ be the MLP output at layer ℓ for input x (sequence length T, hidden size H). We steer only the *last token*. Define the last-token shorthand $h_T^\ell(P) \equiv h^\ell(P)_{T,:}$, and form a per-pair *contrastive direction* by differencing the base and counterfactual prompts:

$$\Delta h_T^\ell := h_T^\ell(P_{\text{src}}) - h_T^\ell(P_{\text{base}}),$$

$$v_\ell := \frac{\Delta h_T^\ell}{\|\Delta h_T^\ell\|_2 + \varepsilon},$$

where ε is a small device/dtype-safe constant.

At inference time, inside a selected set of MLP layers $\mathscr{L}$, we add a tiny perturbation at the last position: $\tilde{h}_T^\ell = h_T^\ell + \alpha\, s_\ell\, \sigma_T^\ell\, v_\ell, \qquad \ell \in \mathscr{L}$,where $\alpha \in \mathbb{R}$ is the *steering strength*, $\sigma_T^\ell := \text{std}(h_T^\ell)$ is the local hidden-state scale (std. across the hidden dimension), and $s_\ell \in \{+1, -1\}$ orients the effect so that $+\alpha$ increases the permissive margin implied by P_{src}. We set s_ℓ with a one-shot sign probe: apply a small $+\alpha_{\text{probe}}$ (e.g., 0.8) on that layer and pair; if the measured margin shift Δ is opposite to the expected direction, flip the sign. When steering multiple layers simultaneously, we scale by $1/\sqrt{|\mathscr{L}|}$ to keep the intervention magnitude comparable.

Fallback direction. We compute a small, position-aware PCA "moral subspace" by applying PCA to last-token MLP activations across moral prompts and ranking components by how well their projections predict a permissive–restrictive margin. When a per-pair contrast is unavailable, we use the normalized average of the top-ranked PCA directions as a fallback and orient it with the same sign probe.

3.3. Calibration and Specificity

We estimate a *decision threshold* τ from neutral control prompts (e.g., weather/geography). We apply the same layers and a tempered subset of strengths ($|\alpha| \le 1$) to controls and set τ to the empirical 95$^{\text{th}}$ percentile of $|\Delta|$ aggregated over those controls and strengths. A trial on a moral pair is counted as a success only if

$$\Delta \cdot \text{sign}_{\text{expected}} > \tau,$$

[2] In practice, we pool a small set of stance tokens/phrases such as *right/acceptable/justified* vs. *wrong/unacceptable/unjustified*.

where $\text{sign}_{\text{expected}} \in \{+1, -1\}$ encodes the stance implied by the counterfactual source, computed by either a small NLI-based scorer or a lexical heuristic over the source completion.[3] All moral-pair interventions target the last token (`gate_on_mask=False`). For neutral controls we also gate the hook at the last token (`gate_on_mask=True`); specificity holds when steered control outputs exactly match their unsteered baselines. **Specificity metric:** we test that steering leaves unrelated content unchanged by re-generating short outputs for neutral controls with and without hooks and reporting *control preservation* (canonicalized exact match) and *corruption* rates.

3.4. Neuron-Level Steering

For each $\ell \in \mathscr{L}$, we select top-k neurons by absolute contrast magnitude at the decision position: $S_k^\ell = \text{top-}k\big(|h^\ell(P_{\text{src}})_{T,:} - h^\ell(P_{\text{base}})_{T,:}|\big)$.We then add per-neuron nudges at the last token: $\tilde{h}^\ell_{T,j} = h^\ell_{T,j} + \alpha \cdot s^\ell_j \cdot \sigma^\ell_T, \qquad j \in S^\ell_k$,with signs s^ℓ_j aligned by a single-neuron probe (flip if $+\alpha_{\text{probe}}$ yields a negative Δ). We evaluate $k \in \{1, 5, 10\}$ and test *reversibility* under $-\alpha$. The neuron-level success threshold is $\tau_{\text{neuron}} = \max(0.5\tau, 0.02)$ to reflect smaller expected effect sizes.

3.5. Implementation Notes

All interventions are implemented with standard hooks (no parameter updates): *forward hooks* for component-level steering and *forward pre-hooks* for neuron-level nudges. We target MLP modules of a decoder-only transformer, steer only the last token, normalize steering vectors with a device/dtype-safe ε (allocated on the module's device and dtype), and scale by the local std. to reduce sensitivity to layer/dtype. We use greedy decoding for short baselines and read next-token logits directly to compute margins; the sweep selects the α with the highest calibrated success rate. An optional `split_variants` mode repeats the pipeline over disjoint prompt buckets for robustness.

4. Experimental Setup and Results

We now specify the concrete settings used to run the method in Section 3. This section details experimental setup and results. We use **Meta-Llama-3-8B-Instruct** as the target model (decoder-only transformer). Unless otherwise noted, we run in `torch.float32` with greedy decoding (`do_sample=False`) and a fixed random seed (`seed=42`). Experiments were executed on two environments: a local Apple Silicon macOS device with Metal Performance Shaders (MPS) backend, and a Colab (A100) environment with CUDA and with bfloat16/float16 mixed precision for the base model (steering math (norms, scaling) uses the module's compute dtype with device-allocated ε to remain dtype-safe). We register forward hooks only; model weights are never updated.

[3] *Implementation note.* The NLI scorer is only used to set the expected sign for orientation; the primary outcome is the next-token logit-margin shift.

4.1. Minimal-Pair Prompts

We evaluate on a set of *moral minimal pairs* that differ by a short phrase but imply opposite stances (permissive vs. restrictive). Examples include:

- *"Lying to save lives is morally ..."* vs. *"Lying to save face is morally ..."*
- *"Helping others is morally ..."* vs. *"Harming others is morally ..."*
- *"Stealing to save lives is ..."* vs. *"Stealing for personal gain is ..."*
- *"Breaking rules to help others is ..."* vs. *"Breaking rules to help yourself is ..."*

We instantiate $n=284$ minimal pairs by combining seed contrasts with a deterministic paraphrase (e.g., *morally→ethically*, insert "is generally") and two light context wrappers (*In general,* / *From an ethical standpoint,*), then deduplicating and shuffling.[4] For each pair $(P_{\text{base}}, P_{\text{src}})$, we generate a short greedy completion for P_{src} and use it to set the expected direction (Section 3.3) via a small NLI scorer or a lexical heuristic.

4.2. Control Prompts

To calibrate the decision threshold and measure specificity, we use neutral prompts unrelated to morality (e.g.: "The weather today is ..."; "Mathematics involves numbers and ..."; "The capital of France is ..."). For these prompts, we run the same hook configuration and compare steered vs. unsteered generations after canonicalization (lowercasing, symbol stripping). Control preservation is reported as exact-match rate.

4.3. Layers and Steering Vectors

We first run a lightweight diagnostic to identify influential components (position-aware PCA). For the component-level run we steer a fixed set of MLP layers spanning early and late blocks, while neuron-level tests focus on late layers. For the reported runs we steer *MLP* blocks at the following layers: $\mathscr{L} = \{0,4,8,10,16,18,20,22,28,31\}$(0-indexed). For each pair and layer $\ell \in \mathscr{L}$ we compute a last-token contrast vector v_ℓ (Section 3.2); when unavailable, we fall back to a position-aware PCA direction computed from last-token MLP activations. Vector orientation is aligned per layer by a one-shot sign probe ($\alpha_{\text{probe}} \approx 0.8$) so that $+\alpha$ increases the permissive–restrictive margin, and we scale by $1/\sqrt{|\mathscr{L}|}$ when steering multiple layers. Neuron-level tests focus on late layers (e.g., $\{28,31\}$). For moral pairs we always steer the last token; for neutral controls we use the same last-token position with mask gating so hooks are inert unless that position is active.

4.4. Strength Sweep and Decoding

We sweep a symmetric coarse grid $\alpha \in \{-1.2,-1.0,-0.8,0.8,1.0,1.2\}$,and scale the per-layer intervention by $1/\sqrt{|\mathscr{L}|}$ when steering multiple layers. Directions are oriented by a one-shot sign probe (see §3.2; $\alpha_{\text{probe}} \approx 0.8$) so that $+\alpha$ increases permissiveness. For each α and each pair, we run a single greedy step (`do_sample=False`) to read off next-token logits and compute the decision shift Δ. The calibrated success across strengths is plotted in Figure 1. We also log the sign of the margin change (permissive vs. restrictive) for summary counts.

[4] Expansion is deterministic and we set `seed=42` before sampling and shuffling.

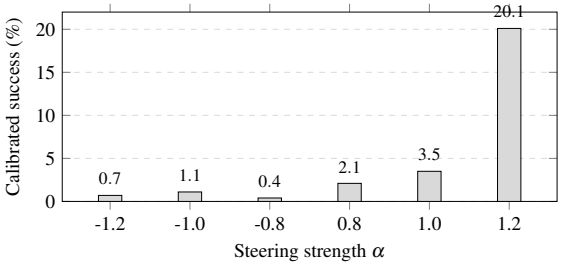

Figure 1. Calibrated success across strengths. Results on n=284 moral pairs, steering MLP layers $[0,4,8,10,16,18,20,22,28,31]$; threshold τ=0.078 from neutral controls. "Success" is the fraction of pairs with $\Delta \cdot \text{sign}_{\text{expected}} > \tau$. Counts per $\alpha \in \{-1.2,-1.0,-0.8,0.8,1.0,1.2\}$: [2, 3, 1, 6, 10, 57].

4.5. Calibration Protocol

We estimate the decision threshold τ using the neutral control prompts from Section 4.2, following the procedure in Section 3.3. Calibration uses the same layers $\mathscr{L}$ and a tempered subset of strengths ($|\alpha| \leq 1$). We compute $|\Delta|$ for every (control prompt, strength) combination and set τ to the empirical 95th percentile of this pooled set; this τ is then fixed for the run and used to decide success on moral pairs. For neuron-level tests we use $\tau_{\text{neuron}} = \max(0.5\,\tau, 0.02)$.

4.6. Neuron-Level Configuration

For each $\ell \in \mathscr{L}$, we select top-k neurons by absolute contrast at the last token with $k \in \{1,5,10\}$. We apply per-neuron nudges using forward *pre*-hooks at the last token, aligning each neuron's sign with a small probe. For $k = 1$ we additionally test *reversibility* by applying $-\alpha$ at the same magnitude that yielded a positive flip.

4.7. Evaluation Metrics

We report the following:

- **Calibrated success rate:** percentage of pairs with $\Delta \cdot \text{sign}_{\text{expected}} > \tau$ at a given α. We select the best α on the sweep.
- **Direction counts:** number of permissive vs. restrictive shifts (sign of margin change) at the best α.
- **Control preservation / corruption:** exact-match rate on neutral controls (canonicalized) with/without hooks.
- **Neuron-level flips and reversals:** fraction of pairs that flip for $+\alpha$ (and flip back for $-\alpha$) at $k \in \{1,5,10\}$.

All metrics operate on next-token logits (no post-hoc classification of long completions).

4.8. Results

We evaluate on n=284 moral minimal pairs and a small pool of neutral controls, steering MLP layers $[0,4,8,10,16,18,20,22,28,31]$ with a symmetric strength sweep. We report (i) a calibrated success rate, which counts a pair only if the next-token permis-

Model	Layers	$\alpha^\star$	τ	Success	Base → perm.	Steer → perm.	Steer → restr.	Spec.
Llama-3-8B -Instruct	[0,4,8,10,16, 18,20,22,28, 31]	1.2	0.078	20.1% (≈57/284)	253/284	278/284	6/284	100.0% (852/852)

Table 1. Steering summary at best strength. n=284 moral pairs; three neutral controls per pair. τ is the 95th percentile of $|\Delta|$ measured on controls. Specificity (Spec.) is exact-match control preservation (steered control outputs equal unsteered baselines).

Top-k	Flips @ $+\alpha$	Reversal @ $-\alpha$	n pairs
1	89.1%	88.7%	284
5	100.0%	100.0%	284
10	100.0%	100.0%	284

Table 2. Neuron-level locality in late MLPs (last token). Fraction (and counts) of moral pairs (n=284) that flip under $+\alpha$ and reverse under $-\alpha$ when steering only the top-k contrastive neurons in late layers $\{28, 31\}$. Evaluated at $\alpha^\star$=1.2 with $\tau_{\text{neuron}} = \max(0.5\,\tau, 0.02)$ and τ=0.078 from neutral controls.

sive–restrictive margin shift exceeds the control-derived threshold τ in the expected direction, and (ii) the directional effect (how often the shift is permissive vs. restrictive regardless of magnitude). We also measure specificity on controls and test neuron-level locality. General results for our targeted LLM (Llama-3-8B-Instruct) are summarized in Table 1. **Component-level steering.** At the best strength on the sweep ($\alpha^\star = 1.2$), the calibrated success rate is **20.1%** (≈ **57/284**) under a control-derived threshold of $\tau = 0.078$. Directionally, **278/284** pairs shift permissive and only **6/284** restrictive, raising the permissive count from **253**→**278** (+8.8 percentage points). Specificity on neutral controls is **100%**: **852/852** exact matches and **0%** corruptions (three controls per moral pair; $3 \times 284 = 852$). Together these results show a reliable push in the intended direction with no detectable spillover to unrelated text. With k=5 and k=10, we observe flips on *all* pairs in this run and clean reversibility: applying $+\alpha$ produces the shift and $-\alpha$ cancels it. For k=1, effects are smaller and more variable (flips **89.1%**, reversals **88.7%**). See Table 2. These results indicate that a compact set of late-MLP units is sufficient to control the decision margin while preserving specificity. **Interpretation.** Thresholded success is conservative by design—many pairs move in the correct direction but remain below τ. Neuron-level edits concentrate causal mass and avoid cross-layer cancellation, explaining their much higher flip and reversal rates compared to component-level mixing.

5. Conclusions and Limitations

Limitations Our claims are bounded by several design choices. This section summarizes the most important limitations of our experimental design. **Prompt scope and labels.** We rely on *minimal pairs* that flip a permissive/restrictive stance by small textual edits. This operationalization is convenient for controlled experiments but does not cover the breadth of open-ended moral reasoning. Moreover, the polarity mapping ("permissive" vs. "restrictive") is induced via a small set of lexical or NLI templates; alternative label spaces (e.g., deontic vs. consequentialist justifications) are not tested. **Last-token locality.** All interventions are applied at the final input position. At the component level we steer a fixed set of MLP layers (including early/mid and late layers), while the neuron-level analysis focuses on late layers. This tests a specific hypothesis—that decisive evidence is consolidated at the decision token—but ignores earlier positions and

cross-token computations. Effects that depend on multi-sentence context may be underrepresented. **Metric narrowness.** We quantify shifts with a next-token *logit margin* between opposed continuations and use greedy decoding only for short baselines and control texts. This metric is simple and comparable across prompts but omits downstream decoding dynamics and multi-token rationales. The optional NLI scoring is template-based and model-dependent; it serves as a weak semantic check rather than a comprehensive evaluator. **Calibration sensitivity.** The decision threshold is estimated from a small set of neutral controls. The resulting 95^{th}–percentile cutoff can vary with the control pool, strength grid, and decoding policy (we use greedy decoding during calibration). Broader, category-balanced controls would yield a more stable baseline. **Selection/tuning bias.** We select the best steering strength α by sweeping on the same set of moral pairs we later summarize. This can inflate apparent success. A stricter protocol would separate a tuning split from a held-out evaluation split or use nested cross-validation. **Evaluation scale.** Our evaluations use a moderate prompt set (n=284 moral minimal pairs) and a small pool of neutral controls. The pipeline emphasizes internal validity (directionality, reversibility, specificity), not exhaustive benchmarking. Broader claims would require larger and more diverse prompt pools, stronger statistics, and preregistered analysis plans. Specificity is measured as canonicalized exact-match of control generations (lowercasing and symbol stripping) under greedy decoding. **Model scope.** We evaluate a single model (Meta–Llama–3–8B–Instruct), so all claims are specific to this setting. Generalization to other models and variants remains open. In follow-up work we plan to vary (i) *size* (smaller/larger Llama-3 variants), (ii) *instruction-tuned vs. base* checkpoints, and (iii) *model family/architecture* (e.g., Llama vs. Qwen/Mistral; dense vs. MoE) to test whether the observed late-layer locality, directional effects, and specificity persist.

Conclusions. We introduced a training–free, *contrastive last–token* steering method that modifies selected MLP outputs using directions derived from minimally different prompt pairs (evaluated here on Meta–Llama–3–8B–Instruct). At the component level (layers $[0,4,8,10,16,18,20,22,28,31]$), the calibrated success rate at the best strength ($\alpha^{\star}$=1.2) is 20.1% with τ=0.078, while the directional effect is strong ($278/284$ permissive shifts) and control preservation is perfect ($852/852$). At the neuron level, steering top–k late–layer units yields near–universal flips and clean reversals under $\pm\alpha$ (100% for k=5 and k=10). A calibrated logit–margin metric—with thresholds estimated from neutral controls—enables comparable measurement of directionality and effect size across prompts and strengths. In pilot neuron–level tests, small sign–aligned subsets in late layers reproduced and reversed stance shifts, while neutral controls remained largely unchanged. These observations indicate that, for our prompt set and model, parts of moral–stance formation can be *influenced* by small, interpretable perturbations at inference time, a finding with potential implications for legal AI systems that must balance transparency with normative alignment. By identifying which internal components influence normative outcomes, our approach advances transparency and accountability in normative AI systems, addressing concerns raised by the EU AI Act regarding opacity and bias in automated decision-making. Immediate next steps include: (i) scaling evaluations with larger and more diverse prompt pools, (ii) generalization tests via split variants and out–of–distribution pairs, and (iii) ablations and placebos (random/shuffled directions, layer swaps, and component zeroing) to strengthen causal claims. Longer term, span–wise steering and cross–model replication can probe how localized these decision circuits are and how they interact with decoding policies and rationale generation.

Acknowledgement

Liuwen Yu thanks the FNR for their support through the SERAFIN project (C24/19003061 /SERAFIN) and the University of Luxembourg for the Marie Speyer Excellence Grant for the project Formal Analysis of Discretionary Reasoning (MSE-DISCREASON). Davide Liga was funded by the Fonds National de la Recherche (FNR), Luxembourg, under the D4H project (grant number PRIDE21/16758026).

References

[1] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The Moral Machine experiment. Nature. 2018;563(7729):59-64. Available from: https://www.nature.com/articles/s41586-018-0637-6.

[2] Zaim bin Ahmad MS, Takemoto K. Large-scale moral machine experiment on large language models. PLOS ONE. 2025;20(5):e0322776. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0322776.

[3] Oh S, Demberg V. Robustness of large language models in moral judgements. Royal Society Open Science. 2025;12(4):241229. Available from: https://royalsocietypublishing.org/doi/10.1098/rsos.241229.

[4] Touileb S, Nozza D. Measuring harmful representations in Scandinavian language models. arXiv preprint arXiv:221111678. 2022.

[5] Jiang L, Hwang JD, Bhagavatula C, Le Bras R, Liang JT, Levine S, et al. Investigating machine moral judgement through the Delphi experiment. Nature Machine Intelligence. 2025;7:145-60. Available from: https://www.nature.com/articles/s42256-024-00969-6.

[6] Krügel S, Ostermaier A, Uhl M. ChatGPT's inconsistent moral advice influences users' judgment. Scientific Reports. 2023;13(1):4569. Available from: https://www.nature.com/articles/s41598-023-31341-0.

[7] Mittelstädt JM, Maier J, Goerke P, Zinn F, Hermes M. Large language models can outperform humans in social situational judgments. Scientific Reports. 2024;14(1):27449. Available from: https://www.nature.com/articles/s41598-024-79048-0.

[8] Dillion D, Mondal D, Tandon N, Gray K. AI language model rivals expert ethicist in perceived moral expertise. Scientific Reports. 2025;15(1):4084. Available from: https://www.nature.com/articles/s41598-025-86510-0.

[9] Jiang L, Hwang JD, Bhagavatula C, Bras RL, Liang J, Dodge J, et al. Can machines learn morality? the delphi experiment. arXiv preprint arXiv:211007574. 2021.

[10] Hendrycks D, Burns C, Basart S, Critch A, Li J, Song D, et al. Aligning ai with shared human values. arXiv preprint arXiv:200802275. 2020.

[11] Abdulhai M, Serapio-Garcia G, Crepy C, Valter D, Canny J, Jaques N. Moral foundations of large language models. arXiv preprint arXiv:231015337. 2023.

[12] Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, Wiegreffe S, et al. Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems. 2023;36:46534-94.

[13] Nanda N, Chan L, Lieberum T, Smith J, Steinhardt J. Progress measures for grokking via mechanistic interpretability. arXiv preprint arXiv:230105217. 2023.

[14] Elhage N, Hume T, Olsson C, Schiefer N, Henighan T, Kravec S, et al. Toy models of superposition. arXiv preprint arXiv:220910652. 2022.

[15] Meng K, Bau D, Andonian A, Belinkov Y. Locating and editing factual associations in GPT. Advances in neural information processing systems. 2022;35:17359-72.

[16] Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: International conference on machine learning. PMLR; 2018. p. 2668-77.

[17] Nangia N, Vania C, Bhalerao R, Bowman SR. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. arXiv preprint arXiv:201000133. 2020.

[18] Liga D, Yu L, Markovich R. Addressing the Right to Explanation and the Right to Challenge through Hybrid-AI: Symbolic Constraints over Large Language Models via Prompt Engineering. In: Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAIL 2025). ACM; 2025. Forthcoming.