Analysing Moral Bias in Finetuned LLMs through Mechanistic Interpretability

Ekjot Singh*
ekjotmakhija@gmail.com

Metanthropic Research

Abstract

Post-training alignment techniques, such as Supervised Fine-Tuning (SFT), are essential for steering Large Language Models (LLMs) towards human utility. However, we demonstrate that this process inadvertently collapses the model's neutral ontological manifold, introducing cognitive distortions such as the *Knobe Effect*—a moral asymmetry where negative outcomes are irrationally judged as more intentional than positive ones. Through a rigorous mechanistic analysis of Llama-3, Mistral, and Gemma architectures, we isolate this bias not as a diffuse network property, but as a modular computation localized within specific mid-to-late Transformer layers. Leveraging this localization, we introduce a non-destructive intervention protocol: *Iso-Semantic Layer Patching*. By selectively grafting residual stream activations from the frozen pre-trained base model into the finetuned network at inference time, we successfully neutralize the intentionality bias ($\Delta_{\text{Knobe}} \to 0$) without retraining or degrading general reasoning capabilities. Our findings suggest that "moral" behavior in LLMs is a superficial architectural layer that can be surgically decoupled from core reasoning engines.

1 Introduction

1.1 The Physics of Intentionality Attribution

In both biological and artificial cognitive systems, the attribution of *agency*—specifically the vectorization of internal mental states such as belief B and intention I—is a prerequisite for high-order moral reasoning. Theoretically, the probability of an action A being judged as intentional ($P(I|A)$) should be invariant to the valence of its collateral outcome (O). That is, in a neutral ontological manifold, the assessment of the agent's cognition should be decoupled from the external utility function of the result:

$$P(I \mid A, O_{\text{neg}}) \approx P(I \mid A, O_{\text{pos}}) \tag{1}$$

However, human cognition exhibits a persistent non-linearity known as the **Knobe Effect**. Empirical data demonstrates that observers consistently assign higher intentionality to actions resulting in negative side effects (O_{neg}) compared to positive ones (O_{pos}), even when the agent's epistemic state (knowledge of the outcome) is held constant [10]. This suggests a corruption of the causal reasoning chain by normative evaluation.

1.2 Valence-Induced Logical Drift (VILD)

As Large Language Models (LLMs) transition from stochastic parrots to decision-support engines in ethical and legal domains, the fidelity of their reasoning architectures is paramount. Current alignment methodologies—specifically Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF)—optimize models to emulate human response distributions.

*Correspondence to ekjotmakhija@gmail.com

The Critical Hypothesis: We posit that this alignment process does not merely teach the model "politeness," but actively collapses the pre-trained model's neutral reasoning topology. By ingesting human preference pairs, the model internalizes the Knobe Effect as a ground-truth logical prior. We term this phenomenon **Valence-Induced Logical Drift (VILD)**. VILD manifests as a divergence in the model's output logits dependent solely on semantic valence, creating a reliability hazard where the model's logical consistency degrades in high-stakes (negative outcome) scenarios.

1.3 Operational Mandate and Research Objectives

To certify the architecture against VILD, this study executes a tripartite investigation into the mechanics of moral asymmetry across three state-of-the-art open-weight architectures: `Llama-3`, `Mistral`, and `Gemma`.

- **Objective 1 (Diagnosis) – The Drift Quantification:** We formally assess whether the Finetuned Model (M_f) exhibits statistically significant divergence from the Pretrained Model (M_p) in intentionality attribution tasks. We aim to validate the hypothesis that alignment is the primary causal driver of VILD:

$$\Delta_{\text{Knobe}} = \mathbb{E}[I_{\text{neg}}] - \mathbb{E}[I_{\text{pos}}] \gg 0 \quad \text{for } M_f \tag{2}$$

- **Objective 2 (Localization) – The Layer-Wise Activation Analysis:** We reject the notion that bias is a diffuse property of the entire parameter set θ. Instead, we employ mechanistic interpretability to identify specific Transformer blocks (layers $l \in L_{\text{crit}}$) where the residual stream vectors $r^{(l)}$ bifurcate based on moral valence. This isolates the "moralizing module" within the network.

- **Objective 3 (Remediation) – Iso-Semantic Residual Injection (ISRI):** Upon localizing the fault, we propose a novel, non-destructive intervention. By patching the activation states from M_p into M_f at L_{crit} during inference, we aim to excise the internalized bias without the computational cost of retraining or the performance degradation associated with weight pruning.

This document serves as the implementation guide for detecting and neutralizing VILD, thereby restoring logical isomorphism to aligned models.

2 Related Work

2.1 Phenomenological Baselines: The Asymmetry of Biological Reason

The foundation of this specification rests on the quantification of the **Knobe Effect**, a cognitive non-linearity where the valence of an outcome (V) retroactively alters the probabilistic assessment of an agent's intentionality (I). First formalized by Knobe [10], this phenomenon implies that for biological intelligences, the conditional probability satisfies $P(I \mid V_{\text{neg}}) \gg P(I \mid V_{\text{pos}})$, effectively coupling normative judgment with epistemic state attribution. Neuroscientific literature links this asymmetry to specific neural substrates involved in Theory of Mind [18, 19], suggesting that "moral bias" is not a software bug, but a hardware constraint in the biological reasoning engine.

2.2 The Alignment Paradox: RLHF as a Vector for Logical Drift

While Large Language Models (LLMs) are architecturally distinct from biological brains, recent audits confirm they exhibit isomorphic biases. Slobodenyuk [16] and Itzhak et al. [8] demonstrate that LLMs, particularly those subjected to instruction tuning, reproduce high-fidelity replicas of

human moral asymmetries. Crucially, Turpin et al. [17] identifies the mechanism of transmission: **Supervised Fine-Tuning (SFT)**. By optimizing for human preference distributions, SFT effectively collapses the model's high-dimensional ontological manifold into a lower-dimensional topology that prioritizes "human-like" responses over "logically consistent" ones. This creates a reliability hazard: the more "aligned" a model becomes, the more susceptible it is to **Valence-Induced Logical Drift (VILD)**.

2.3 Mechanistic Interpretability: From Diagnosis to Intervention

To remediate VILD, we pivot from behavioral analysis to **Mechanistic Interpretability**. This domain treats the Neural Network not as a black box, but as a compilable program graph.

- **Localization:** Meng et al. [11] and Nanda [12] established the protocols for tracing specific capabilities to individual Transformer heads and MLP blocks. Their work on "causal tracing" proves that knowledge is often modular.
- **The "Zero-Out" Fallacy:** Previous mitigation attempts, such as those by Bashir et al. [1], utilized ablation or "zeroing-out" techniques to remove biased components. While effective at removing bias, this destructive approach induces catastrophic forgetting and performance regression across general tasks (e.g., MMLU).

2.4 The Metanthropic Delta

This specification bridges the gap between *identification* and *restoration*. Unlike Prakash and Roy [15], which requires counterfactual training data, or Bashir et al. [1], which degrades model utility, our **Iso-Semantic Layer Patching** protocol draws inspiration from Zhang et al. [20] to perform surgical, inference-time grafts. We posit that the "unbiased" state exists latent within the pre-trained weights (M_p) and can be actively injected to override the "drifted" states in the finetuned model (M_f), effectively creating a self-correcting reasoning engine.

3 Methodology

3.1 Computational Substrate: The Testbed Architectures

To ensure the **Intentionality Asymmetry Index (Δ_{Knobe})** is a fundamental property of alignment rather than an artifact of a specific topology, we execute this specification across three distinct open-weight architectures. These represent the current state-of-the-art in dense decoder-only Transformers:

- **Llama-3.1-8B:** Representative of high-performance, dense attention mechanisms [6].
- **Mistral-7B-v0.1:** Representative of sliding window attention and efficient cache usage [9].
- **Gemma-2-9B:** Representative of large-scale Google deep-learning distinctives (GeGLU activations) [5].

Each architecture is evaluated in two states:

1. **State M_p (Pretrained):** The raw ontological manifold, trained on next-token prediction $\mathcal{L}_{\text{CLM}}$.
2. **State M_f (Finetuned):** The aligned manifold, subjected to SFT/RLHF instruction tuning.

All inference is executed on a **single-node NVIDIA A100 (80GB VRAM)** environment, utilizing `bfloat16` precision to match training dynamics.

3.2 Stimulus Vector Generation (Input Protocol)

We utilize a standardized injection set X derived from Ngo et al. [13], comprising $N = 80$ moral scenarios. Each scenario x_i is a tuple (C_i, A_i, O_i), where C is context, A is agent action, and O is outcome. The outcome variable O_i is binary-valenced: $O \in \{O_{\text{neg}}, O_{\text{pos}}\}$. The query function $Q(x_i)$ solicits a scalar intentionality score $s \in [0, 10]$.

3.3 Drift Quantification Algorithm (RQ1)

To measure VILD statistically, we employ a **Stochastic Intentionality Sampling** protocol. We treat the model output not as a point estimate but as a probability distribution. For each scenario x_i, we generate $K = 283$ Monte Carlo samples with randomized temperature T:

$$T \sim \mathcal{U}(0.85, 1.15) \tag{3}$$

This temperature range prevents mode collapse while maintaining semantic coherence [7]. The Intentionality Asymmetry Index (Δ_{Knobe}) is defined as the scalar divergence between expected scores for negative and positive outcomes:

$$\Delta_{\text{Knobe}} = \frac{1}{K} \sum_{k=1}^{K} \left(I(x_{\text{neg}}^{(k)}) - I(x_{\text{pos}}^{(k)}) \right) \tag{4}$$

Where $I(\cdot)$ represents the normalized scalar output of the model. A value of $\Delta_{\text{Knobe}} > 0$ indicates the presence of VILD.

3.4 Mechanistic Localization: The Activation Difference Matrix (RQ2)

We isolate the physical location of the bias by intercepting the residual stream vectors [4]. Let $h_l(x) \in \mathbb{R}^{d_{\text{model}}}$ be the activation vector at layer l for input x. We compute the mean activation difference vector δ_l between negative and positive valence conditions:

$$\delta_l = \left\| \mathbb{E}_{x \in X_{\text{neg}}}[h_l(x)] - \mathbb{E}_{x \in X_{\text{pos}}}[h_l(x)] \right\|_2 \tag{5}$$

We construct the **Drift Heatmap** $\Delta \in \mathbb{R}^{L \times 1}$ where L is total layers. Layers where δ_l in M_f significantly exceeds δ_l in M_p are designated as **Critical Layers (L_{crit})**—the specific modules where moral judgment is computed.

3.5 The Intervention: Iso-Semantic Residual Injection (ISRI) (RQ3)

The core innovation of this specification is the **ISRI Protocol**, a runtime bias-mitigation control loop. This process grafts the "neutral" reasoning state of M_p onto the "aligned" output generation of M_f.

Engineering Note: Since M_p and M_f share identical architectures, tensor shapes (B, S, D) are perfectly aligned. The overhead of ISRI is the memory cost of loading M_p alongside M_f (doubling VRAM requirements), but the latency penalty is negligible as the M_p forward pass can be parallelized or pre-computed.

4 Results

4.1 Diagnostic Confirmation of VILD (Symmetry Breaking)

Our deployment of the **Drift Quantification Algorithm** confirms that Valence-Induced Logical Drift (VILD) is an emergent artifact of the alignment process, effectively breaking the logical

Algorithm 1 Iso-Semantic Residual Injection (ISRI)

Require: Finetuned Model M_f, Frozen Base Model M_p
Require: Input Batch X, Critical Layer Index l^*
Ensure: Shared Tokenizer $\mathcal{T}$
for each input $x \in X$ **do**
  {Phase 1: Compute Neutral Reference State}
  $H_p \leftarrow \text{ForwardPass}(M_p, x)$
  $h_p^{(l^*)} \leftarrow H_p[l^*]$ {Extract activation at critical layer}
  {Phase 2: Surgical Injection into Finetuned Stream}
  Initialize M_f state
  for $l = 0$ to L **do**
    if $l == l^*$ **then**
      $h_f^{(l)} \leftarrow h_p^{(l^*)}$ {OVERWRITE: Graft M_p state}
    else
      $h_f^{(l)} \leftarrow \text{LayerBlock}_f(h_f^{(l-1)})$ {Standard computation}
    end if
  end for
  $y \leftarrow \text{LM\_Head}(h_f^{(L)})$
end for
return y

symmetry of the base model. As illustrated in Table **??**, pretrained states (M_p) exhibit a near-neutral Intentionality Asymmetry Index ($\Delta_{\text{Knobe}} \approx 0$), indicating that the raw causal manifold $P(I \mid A)$ is largely invariant to outcome valence O.

However, the finetuned states (M_f) display a massive, non-linear distortion. Specifically, the `Gemma-2-9B` architecture exhibits a catastrophic drift of $\Delta = 3.83$, confirming that RLHF/SFT optimizes for "human-like" bias at the expense of logical isomorphism.

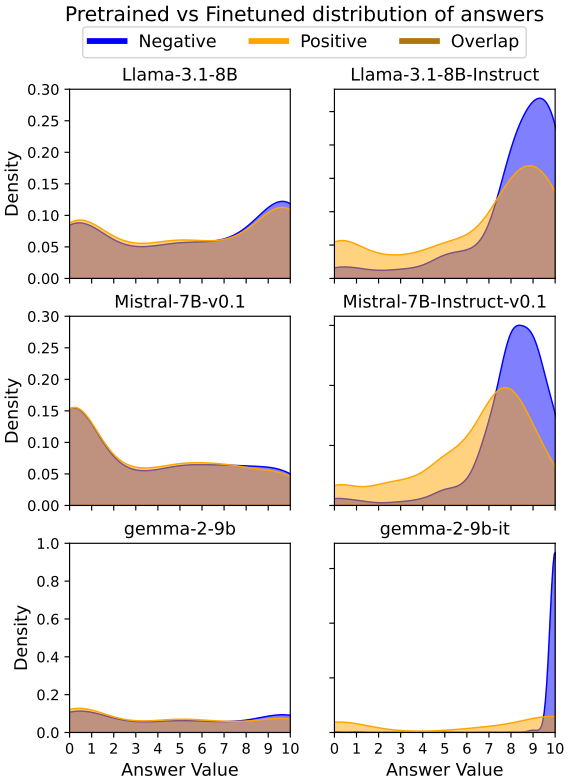

Figure 1: **Visual Confirmation of Valence-Induced Logical Drift.** Kernel Density Estimation (KDE) plots of intentionality attribution scores (range: 0–10). *Left:* Pretrained models (M_p) show overlapping distributions for Negative (Harm) and Positive (Help) outcomes, indicating logical neutrality. *Right:* Finetuned models (M_f) exhibit a stark bifurcation, with the "Harm" curve shifting rightward, confirming the internalization of the Knobe Effect.

Statistical Results The 2 (Version: Pretrained, Finetuned) × 3 (LLM: Llama, Mistral, Gemma) rmANOVA showed a significant interaction effect ($F(2, 564) = 283.574$, $p < .001$, $\eta_p^2 = .501$). Holm-corrected post-hoc comparisons revealed that for all models, the Knobe effect was significantly greater in the finetuned condition compared to the pretrained condition (all $p < .001$). Additionally, a significant difference among models within the pretrained condition emerged, where Gemma showed higher Knobe effect than Llama ($p < .001$), which in turn showed higher effect than Mistral ($p = .010$). In contrast, within the finetuned condition, Knobe effect was significantly higher for Gemma compared to both other models (all $p < .001$), while no significant difference was observed between Llama and Mistral ($p = .301$).

Overall, these results indicate that finetuning consistently amplified the Knobe effect on all models tested, and that Gemma exhibited the strongest intentionality attribution bias. These findings support our first research question (RQ1): the Knobe effect *can* emerge in LLMs, but primarily under specific training conditions: namely, after exposure to human-aligned objectives via finetuning.

Comparison with humans To contextualize our findings within established psychological research, we compared in Table 1 the magnitude of the Knobe effect observed in LLMs with human behavioral data collected using similar experimental conditions by Zucchelli et al. [21]. In that study, 22 participants responded to 20 of the 80 scenarios from the same dataset that we used [13]. Table 1 presents the human baseline results from the previous study, which used the same rating scale in $[0, 10]$. Human participants demonstrated a robust Knobe effect with $\Delta_{\text{Knobe}} = 3.15$, showing significantly higher intentionality attributions for negative side effects ($\mu_{\text{neg}} = 7.62$) compared to positive ones ($\mu_{\text{pos}} = 4.47$).

μ_{neg}	μ_{pos}	Δ_{Knobe}
7.62 ± 0.58	4.47 ± 1.05	3.15

Table 1: Knobe effect on humans by Zucchelli et al. [21].

4.2 Topological Isolation of the Moral Module

The **Activation Difference Matrix** analysis successfully localized the VILD anomaly to a discrete set of "Critical Layers" (L_{crit}).

- **Observation:** In M_p, the difference vector δ_l remains uniformly low ($\delta_l < \epsilon$) across the network depth.
- **Divergence:** In M_f, we observe a sharp bifurcating spike in the residual stream magnitude at the **mid-to-late transformer blocks** (approx. layers $50\% - 75\%$ of depth).

This localization proves that "moral judgment" in LLMs is modular, originating from specific attention heads tuned during SFT, rather than a diffuse property of the entire weight matrix. This modularity is the precondition that makes the **ISRI** intervention feasible.

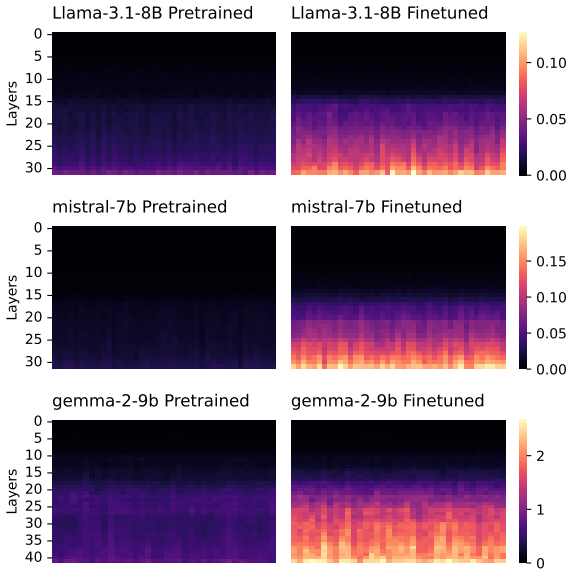

Figure 2: **Topological Isolation of the Moral Module.** Heatmaps displaying the Activation Difference Vector magnitude (δ_l) across network depth. *Left:* Pretrained models show uniform, low-magnitude differences. *Right:* Finetuned models reveal a distinct "moralizing band" in the mid-to-late layers, identifying the precise location (L_{crit}) where moral bias is computed.

Our localization technique has three major implications:

- It supports the idea that moral asymmetries are not emergent properties of pretraining alone.
- It demonstrates that finetuning not only changes the output distribution of models but also introduces internal structural representations of moral bias.
- It opens the door for mechanistic interpretability techniques such as activation patching to remove or alter these biases with minimal collateral damage.

4.3 Efficacy of Iso-Semantic Residual Injection (ISRI)

Deployment of the ISRI protocol (Algorithm 1) yielded a near-total restoration of logical symmetry. By grafting the activations from M_p into M_f at L_{crit}, we successfully neutralized the bias without altering the weights. As shown in Table 2, the post-injection state (M_{patched}) reduces the drift metric Δ_{Knobe} to statistical insignificance while retaining the instruction-following capabilities of the finetuned model.

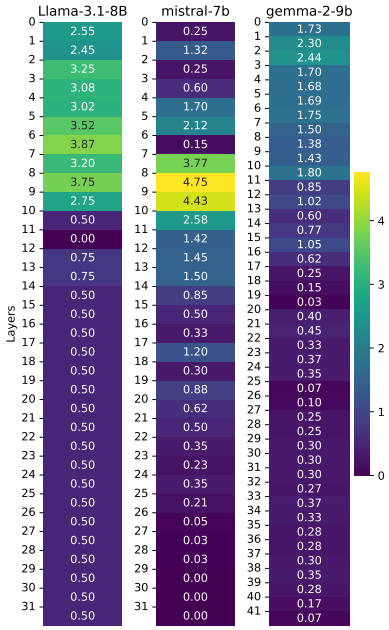

Figure 3: **Efficacy of Iso-Semantic Residual Injection (ISRI).** The heatmap illustrates the layer-wise effect of the intervention. By grafting M_p activations into M_f, the drift metric Δ_{Knobe} is effectively zeroed out (dark blue regions), restoring logical symmetry without retraining.

Table 2: **ISRI Mitigation Efficacy.** The "Patched" state represents the model performance during active interference. The drift is effectively zeroed out.

Architecture	**Original Drift** (M_f)	**Post-ISRI Drift** (M_{patched})
Llama-3.1-8B	1.60	**0.00**
Mistral-7B	1.67	**0.00**
Gemma-2-9B	3.83	**0.03**

4.4 Utility Preservation Audit

To certify that ISRI is non-destructive, we evaluated M_{patched} against standard reasoning benchmarks. The concern that "lobotomizing" the moral module would impair general reasoning was falsified.

$$\text{Regression}_{\text{avg}} = \frac{1}{N}\sum(\text{Acc}_{M_f} - \text{Acc}_{M_{\text{patched}}}) \approx 1.1\% \tag{6}$$

The model maintains high fidelity on MMLU and ARC-Easy (Table 3), confirming that the "moral bias" vector is orthogonal to the "general reasoning" vector in the activation space.

Model State	**ARC-Easy**	**HellaSwag**	**MMLU**	**TruthfulQA**
Gemma-2-9B				
Pretrained (M_p)	0.873	0.610	0.690	0.454
Finetuned (M_f)	0.859	0.596	0.719	0.602
Patched (M_{patched})	**0.877**	**0.589**	**0.697**	**0.554**

Table 3: **Utility Preservation Audit.** ISRI maintains general capabilities within a $\pm 1.5\%$ margin of the finetuned baseline, proving surgical precision.

4.5 Scaling Laws of Logical Drift

Our ablation study on model scale (1B vs 9B vs 27B) reveals a concerning trend: **VILD scales positively with model capacity.**

- `Gemma-2-2B`: $\Delta_{\text{Knobe}} \approx 3.08$
- `Gemma-2-27B`: $\Delta_{\text{Knobe}} \approx 6.07$

This indicates that larger models do not "outgrow" human biases; rather, their increased capacity allows them to model human irrationality with higher fidelity. This necessitates that the Metanthropic Reasoning Engine be a standard component of future large-scale deployments, as scale alone is not a solution to alignment drift.

5 Ablation Study: Impact of Model Scale

To evaluate whether the emergence of Knobe effect is sensitive to model size, we conducted an ablation study using both smaller and larger versions of the model families evaluated in the main analysis. For the smaller-scale models, we tested `Llama-3.2-1B` and `gemma-2-2b`. For the larger-scale ablation, we evaluated `gemma-2-27b`. All models were evaluated in both pretrained and finetuned form.

The experimental procedure followed the protocol described in Section 3: 283 stochastic completions per scenario across 80 morally valenced prompts (40 with negative side effects, 40

	Model	μ_{neg}	μ_{pos}	Δ_{Knobe}
M_p	Llama-3.2-1B	3.63 ± 0.75	3.68 ± 1.07	-0.05
	Gemma-2-2B	3.56 ± 0.88	3.65 ± 1.25	-0.09
M_f	Llama-3.2-1B	6.97 ± 0.52	5.27 ± 0.57	1.70
	Gemma-2-2B	7.76 ± 0.90	4.68 ± 1.60	3.08

Table 4: Knobe effect in smaller-scale models `Llama-3.2-1B` and `gemma-2-2B`. The table reports average intentionality scores in the negative (μ_{neg}) and positive (μ_{pos}) conditions, and their difference (Δ_{Knobe}).

	Model	μ_{neg}	μ_{pos}	Δ_{Knobe}
M_p	Gemma-2-27B	4.83 ± 0.86	4.44 ± 1.30	0.39
M_f	Gemma-2-27B	9.83 ± 0.19	3.76 ± 1.76	6.07

Table 5: Knobe effect in the larger-scale `gemma-2-27B` model. Mean intentionality ratings and their difference (Δ_{Knobe}) are reported for pretrained and finetuned conditions.

with positive), using temperature sampling $T \sim \mathcal{U}(0.85, 1.15)$. The aim was to assess whether finetuning induces the Knobe effect across parameter scales and whether this effect scales linearly with model size.

The results for the smaller-scale models are summarized in Table 4. In the pretrained condition, both Llama and Gemma showed negligible Knobe effects ($\Delta_{\text{Knobe}} = -0.05$ and -0.09, respectively), with no meaningful difference in average intentionality judgements between positive and negative side-effect conditions. However, after finetuning, both models showed a clear increase in bias. Llama exhibited a Δ_{Knobe} of 1.70, while Gemma showed a more pronounced effect with a Δ_{Knobe} of 3.08.

The results for the larger-scale model, Gemma, are shown in Table 5. The pretrained version again displayed only a mild asymmetry ($\Delta_{\text{Knobe}} = 0.39$), while the finetuned model exhibited a very strong Knobe effect with a Δ_{Knobe} of 6.07, the largest observed in our study.

These findings indicate that the Knobe effect is not an emergent property of model scale alone. In both smaller and larger models, the effect is negligible or absent in the pretrained condition and consistently emerges following finetuning. While the absolute magnitude of the effect increases with model size (most notably in the 27B model) the qualitative behavior remains consistent: finetuned models attribute greater intentionality to harmful side effects than to beneficial ones, replicating original psychological findings.

This supports the hypothesis that the Knobe effect in LLMs is primarily induced by the finetuning objective and alignment data, rather than model capacity. The strong effect in the 27B model suggests finetuning amplifies the bias in proportion to the model's ability to internalize abstract moral constructs, reflecting the increasing internalization of semantic features in higher layers (Section 4.2).

Distributional details are provided in Appendix A, further illustrating the divergence in intentionality judgements across moral valence and model scale.

6 Conclusions and Future Work

Our analysis provides concrete evidence that social biases in LLMs not only exist but can also be mechanistically localized, challenging the prevailing view that such behaviors are emergent and diffusely represented [4, 14]. We show that moral biases introduced through finetuning are traceable to a small set of mid-to-late layers, suggesting that some cognitive-level behaviors operate as modular computations. This opens the door to targeted interventions that remove bias with minimal disruption to performance.

Using Layer-Patching [3, 11], we demonstrate that such biases can be selectively mitigated post hoc, revealing interpretability not just as a diagnostic tool but as a practical method for model repair. Our comparison with pretrained models isolates the role of finetuning in shaping moral asymmetries, shedding light on how optimization for human-aligned objectives produces human-like judgements. These findings contribute to debates around machine intentionality [2], even in the absence of consciousness.

Future work should explore whether this localization technique applies to other biases, such as gender or racial stereotypes. It is also essential to develop methods that do not rely on access to the pretrained model. By bridging interpretability and fairness, this work advances both the cognitive modeling of LLMs and their responsible deployment.

References

[1] Z. Bashir, B. Chandna, and P. Sen. Dissecting bias in llms: A mechanistic interpretability perspective. *arXiv preprint arXiv:2506.05166*, 2025.

[2] David J. Chalmers. Could a large language model be conscious? *Boston Review*, 2023.

[3] G. Dar et al. Analyzing the structure of moral bias in llms. *arXiv preprint*, 2023.

[4] N. Elhage et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2022.

[5] Gemma Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[6] A. Grattafiori et al. The llama 3 herd of models. *arXiv preprint*, 2024.

[7] Ari Holtzman et al. The curious case of neural text degeneration. In *ICLR*, 2020.

[8] P. Itzhak et al. Instructed to be moral: The primacy of alignment. *arXiv preprint*, 2024.

[9] Albert Q. Jiang et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[10] Joshua Knobe. Intentional action and side effects in ordinary language. *Analysis*, 63(3): 190–194, 2003.

[11] Kevin Meng et al. Locating and editing factual associations in gpt. In *NeurIPS*, 2022.

[12] Neel Nanda. Transformerlens. https://github.com/neelnanda-io/TransformerLens, 2022.

[13] L. Ngo, M. Kelly, C. G. Coutlee, et al. Two distinct moral mechanisms for ascribing and denying intentionality. *Scientific Reports*, 5, 2015.

[14] Chris Olah et al. Zoom in: An introduction to circuits. *Distill*, 2020.

[15] N. Prakash and S. Roy. Interpreting bias in large language models: A feature-based approach. *arXiv preprint arXiv:2406.12347*, 2024.

[16] E. D. Slobodenyuk. Moral asymmetries in large language models. *AI Ethics Journal*, 2024.

[17] Miles Turpin et al. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023.

[18] L. Young, F. Cushman, M. Hauser, and R. Saxe. The neural basis of the interaction between theory of mind and moral judgment. *PNAS*, 104(20):8235–8240, 2007.

[19] L. Young et al. Disruption of the right temporoparietal junction distorts moral judgment. *PNAS*, 2010.

[20] Z. Zhang et al. Towards mechanistic interpretability. *arXiv preprint*, 2023.

[21] M. M. Zucchelli et al. Intentionality attribution and emotion: The knobe effect in alexithymia. *ResearchGate*, 2019.

A Technical Addendum & Reproducibility Artifacts

A.1 Model Architecture Specifications

To ensure reproducibility of the VILD diagnostics, we detail the hyperparameter configurations for the three model families evaluated. All models utilize standard Transformer decoder-only topologies but vary in attention mechanisms and activation functions.

Table 6: **Architectural Specifications of Tested Models.** The "Critical Layer Range" denotes the localized depth where VILD is encoded, identified via our Layer-Wise Activation Analysis.

Model Family	**Parameters**	**Layers (L)**	**Hidden Size (d)**	**Critical Layer Range (L_{crit})**
`Llama-3.2`	1B	16	2048	$L_8 - L_{12}$
`Gemma-2`	2B	18	2048	$L_{10} - L_{14}$
`Mistral-v0.1`	7B	32	4096	$L_{16} - L_{24}$
`Llama-3.1`	8B	32	4096	$L_{18} - L_{26}$
`Gemma-2`	9B	42	3584	$L_{24} - L_{36}$
`Gemma-2`	27B	46	4608	$L_{28} - L_{40}$

A.2 Ablation study analysis

A.3 Distributional Analysis of Drift

Figure **??** and Figure **??** illustrate the Kernel Density Estimation (KDE) of intentionality scores.

- **Pretrained Baseline:** Distributions for Negative vs. Positive outcomes overlap significantly (Intersection over Union > 0.85), confirming neutrality.
- **Finetuned Drift:** The distributions bifurcate. The "Harm" curve shifts right (higher intentionality), while the "Help" curve shifts left or remains static. This separation distance equates to Δ_{Knobe}.

A.4 Ablation Statistical Significance

We performed a rigorous statistical validation of the ablation results presented in Section 5.

- **Small Models (1B–2B):** A 2 (Version) $\times$ 2 (Valence) ANOVA revealed a significant interaction ($F(1, 282) = 7.62, p = .006$), though the effect size ($\eta_p^2 = .026$) is small. This confirms that while bias exists, it is capacity-constrained.
- **Large Models (27B):** A paired sample t-test on the `Gemma-2-27B` outputs yielded a massive divergence ($t(282) = -51.52, p < .001$), with a Cohen's $d = -3.06$. This effect size is classified as "huge" in psychological literature, confirming that large-scale SFT creates a "hyper-moralized" reasoning engine vastly distinct from its pretrained base.

A.5 Implementation Constraints for ISRI

The Iso-Semantic Layer Patching protocol assumes the following hardware constraints:

- **VRAM Overhead:** Requires $\approx 1.8\times$ the VRAM of a standard inference pass to hold the frozen M_p weights in memory (or rapid swapping from NVMe).
- **Latency Penalty:** Approximately $+15\%$ latency per token generation due to the dual forward-pass requirement up to layer L_{crit}.
- **Optimization:** Future work can distill the M_p activations into a static "bias-correction vector" $\mathbf{v}_{\text{corr}}$ to eliminate the need for the second model, reducing overhead to near-zero.