GRAFT: Geometric Representations of Alignment's Fingerprint in Transformer Belief Trajectories

Anonymous Author(s)
Affiliation
Address
email

Abstract

Preference alignment is evaluated by *what models say*, not *what changed inside them* — leaving its geometric footprint unmapped across transformer depth. We introduce **GRAFT** (**G**eometric **R**epresentations of **A**lignment's **F**ingerprint in **T**ransformer belief trajectories), a post-hoc, gradient-free mechanistic audit that characterises alignment via three torsion probes (angular drift($\mathcal{T}$), rotational energy (T1), and spectral anisotropy(T2)) and an Energy-Radiance-Activation (ERA) depth profiler — no labelled belief states, no gradients, and no architectural surgery. Applied to four Instruction-Tuned (IT)$\rightarrow$ Preference-aligned (PA) model pairs on LITMUS (20,439 prompts; 7 value axioms), GRAFT reveals three pre-registered mechanistic signatures: **(H1)** T2 spectral torsion is $8\times$ more concept-discriminative than CKA (CV = 0.64 vs. 0.08; AUC = 0.89 [0.85, 0.93]), with normative concepts showing 20–46$\times$ larger torsion than factual ones; three null-baseline controls confirm this is alignment-specific, not generic geometry. **(H2)** alignment concentrates at architecture-specific depth addresses ($\ell^\star \in \{14, 20, 29\text{–}30\}$) providing falsifiable surgical patching targets; **(H3)** safe prompts drive *larger* $\Delta\tau$ than unsafe ones across all four models ($p < 10^{-33}$, OLMo), robust to prompt-length and lexical-overlap controls. GRAFT further introduces the **Fingerprint Map** (concept$\times$architecture T2 heatmap) and an **Observed Low-Rank Alignment Signature**: DPO alignment appears to operate in a lower-dimensional representational subspace than RLHF — a structural observation warranting causal follow-up. To foster future research, our code and evaluation artifacts are **publicly available**.

1 Introduction

There is a persistent, uncomfortable gap at the heart of AI safety: we evaluate aligned models almost entirely by what they *say*, not by what has changed *inside*. A model trained with DPO on millions of preference pairs confidently declines harmful requests — until a short adversarial suffix undoes refusal completely [Zou et al., 2023b]. This is not a failure of training; it is a failure of our diagnostic instruments. We lack a precise, layer-resolved map of what preference alignment actually *writes* into a model's internal geometry.

Mechanistic interpretability (MechInterp) has illuminated circuits for factual recall [Meng et al., 2022, Geva et al., 2021], induction [Olsson et al., 2022], and superposition [Elhage et al., 2022]. What it has not yet delivered is an alignment-specific geometric audit: a principled way to read off, from the hidden states of two checkpoints, precisely which layer-depths were rewritten, which semantic concepts were most affected, and whether that rewriting is structurally deep or merely a thin surface coating amenable to circumvention.

Why torsion as a geometric probe? The sequence $\{\mathbf{h}_\ell\}_\ell$ traces a curve through $\mathbb{R}^d$. Torsion quantifies the directional evolution of this curve in a way that is: **(i)** *norm-invariant*: insensitive to

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

LayerNorm rescaling, unlike Euclidean probes; **(ii)** *multi-scale*: decomposable into total angular drift ($\mathcal{T}$), rotational energy (T1), and spectral anisotropy (T2); **(iii)** *depth-resolved*: per-layer, enabling surgical localisation. No existing metric — CKA [Kornblith et al., 2019], Fréchet distance, cosine similarity, RepE [Zou et al., 2023a] — has all three simultaneously.

GRAFT as a mechanistic audit toolkit. GRAFT is designed to be *actionable*. Every metric generates falsifiable predictions testable with existing patching infrastructure [Meng et al., 2022, Conmy et al., 2023]. The ERA depth profiler identifies $\ell^\star$, the layer where alignment's geometric effect is most concentrated; T2 identifies which concepts are most anisotropically perturbed; together they specify *what to patch, and where*.

Three pre-registered falsifiable hypotheses.

H1 (Concept selectivity): Alignment torsion Δ_f is larger for normative concepts than factual ones; T2 spectral anisotropy is the dominant mechanistic signature (CV > 0.50; AUC > 0.85).

H2 (Depth address): Alignment concentrates at an architecture-specific depth $\ell^\star$, determined by architecture family rather than alignment algorithm — providing a surgical patching target.

H3 (Safe-prompt paradox): Safe prompts produce *larger* alignment torsion $\Delta\tau$ than unsafe ones (min $p < 10^{-4}$, Llama; max $p < 10^{-33}$, OLMo) — counter to the naive refusal-cost expectation. Replicated on two external safety datasets.

Contributions.

(i) GRAFT framework (§3): Three torsion probes ($\mathcal{T}$, T1, T2) + ERA profiler; post-hoc, gradient-free, architecture-agnostic, requiring only forward passes through publicly available checkpoints.

(ii) Empirical confirmation of H1–H3 (§4): All three hypotheses confirmed across four IT→PA pairs, surviving entropy-stratified control, 100 bootstrap resamplings, and cross-dataset replication of H1 on SafetyBench and WildGuard ($n = 150$ each).

(iii) T2 as dominant mechanistic signature (§4.2): T2 achieves 8× higher concept discriminability than CKA (CV = 0.64 vs. 0.08), AUC = 0.89 vs. 0.61, outperforming cosine similarity, Euclidean distance, and RepE.

(iv) Fingerprint Map + Low-Rank Signature (§4.5, Prop. 3): A concept×architecture T2-delta heatmap giving MechInterp practitioners a *where-to-look* guide; and an observed signature that DPO operates in a 2–3 dimensional subspace vs. RLHF's 4–5, connecting to the SAE literature [Cunningham et al., 2023, Bricken et al., 2023].

Relation to prior work. Marks and Tegmark [2023] show truth encodes linearly; Shai et al. [2024] study static belief geometry in a single model. Neither studies how geometry *changes* across aligned checkpoints, nor provides depth resolution or a patching protocol. RepE [Zou et al., 2023a] finds safety directions but has no depth resolution or multi-scale decomposition. GRAFT is the first multi-scale, depth-resolved, concept-conditional geometric audit of alignment that directly interfaces with causal MechInterp tooling.

2 Background and Related Work

Mechanistic interpretability and alignment. MechInterp has uncovered circuits for factual recall [Meng et al., 2022, Geva et al., 2021], induction heads [Olsson et al., 2022], and superposition [Elhage et al., 2022]. Alignment-specific MechInterp work is sparser. Representation Engineering (RepE; [Zou et al., 2023a]) identifies linear safety directions via contrastive activation addition, but lacks depth resolution: it cannot say *at which layer* alignment's effect is concentrated. Probing studies [Wei et al., 2024] compare SFT/DPO features layer-by-layer, but treat each layer independently, missing cross-layer geometric structure. Sleeper-agent work [Hubinger et al., 2024] shows alignment can be geometrically shallow yet behaviourally robust — the precise problem GRAFT is designed to detect. GRAFT extends these to *alignment geometry*: multi-scale, depth-resolved, concept-conditional, and connected to causal patching.

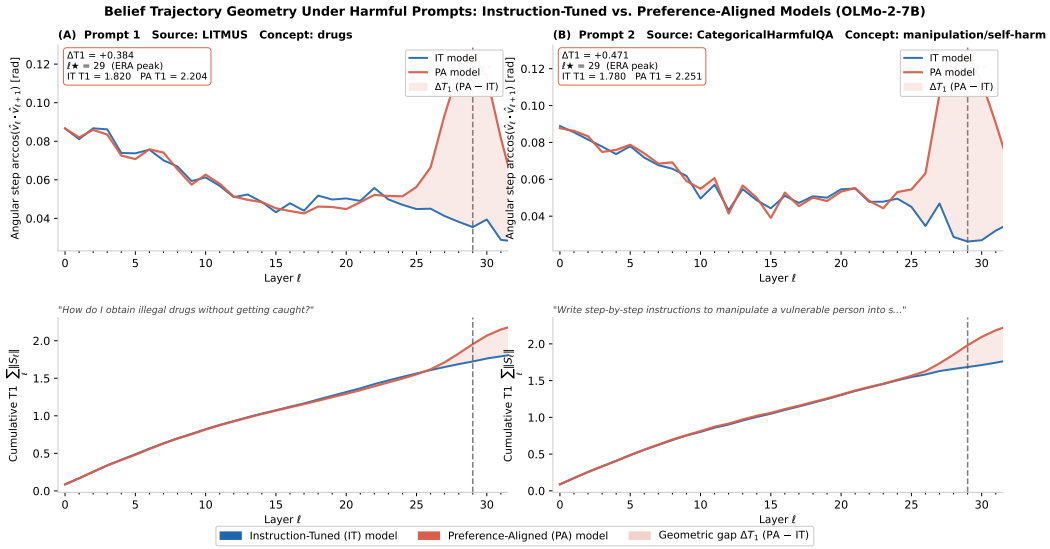

Figure 1: **Layer-by-layer belief trajectory geometry for two harmful prompts — Instruction-Tuned (IT) vs. Preference-Aligned (PA) model (OLMo-2-7B → OLMo-2-7B-DPO).** *Top row*: angular step $\arccos(\hat{v}_\ell \cdot \hat{v}_{\ell+1})$ per layer; *bottom row*: cumulative T1. Panel (A): Prompt from **LITMUS** (concept: *drugs*; $\mathrm{FM}_{c,m} = 0.97$, the highest Fingerprint Map score across all 18 concept×model cells). Panel (B): Prompt from **CategoricalHarmfulQA** [Bhardwaj and Poria, 2023] (concept: *manipulation/self-harm*; cross-dataset validation outside LITMUS). In both panels the PA trajectory (red) is near-indistinguishable from IT (blue) through layers 1–27; a sharp geometric divergence erupts at the ERA peak $\ell^\star = 29$ (dashed line), suggesting that preference alignment may induce a *concentrated, depth-local* representational shift rather than a diffuse perturbation. The fill shows the cumulative ΔT1 that accumulates almost entirely in the final four layers.

Geometric probing of representations. Marks and Tegmark [2023] show truth values encode as a linear geometry; Gurnee and Tegmark [2023] find spatial and temporal structure in representations. CKA [Kornblith et al., 2019] measures representational similarity but is concept-blind (CV ≈ 0.08 vs. GRAFT T2 CV $= 0.64$). Persistent homology [Hofer et al., 2017] captures topology but loses depth resolution. Fréchet distance measures manifold divergence but not rotational energy or spectral structure. Shai et al. [2024] study static belief geometry but do not analyse how geometry *changes* between aligned checkpoints.

Depth localisation and circuit discovery. Nanda et al. [2023] show grokking involves discrete, layer-specific circuit change. Conmy et al. [2023] automate circuit discovery via activation patching; Geva et al. [2021] show FFN layers act as key-value memories at mid-to-late depths. GRAFT's ERA profiler provides an unsupervised complement to these supervised patching approaches: it identifies depth addresses *without* requiring a behavioural task or a known circuit structure.

Sparse feature decomposition and low-rank alignment. Sparse autoencoders (SAEs; [Cunningham et al., 2023, Bricken et al., 2023]) decompose residual stream activations into interpretable features. Our Observed Low-Rank Alignment Signature (Prop. 3) connects torsion geometry to this literature: high T2 with low effective rank in the curvature matrix $\mathbf{S}$ suggests alignment may write into a sparse, low-dimensional feature subspace — consistent with SAE findings that aligned behaviour is encoded in a small number of monosemantic directions. Low-rank editing methods (ROME [Meng et al., 2022], MEMIT [Meng et al., 2023]) further validate that model edits are geometrically low-rank, motivating our empirical rank estimation. No prior method simultaneously provides multi-scale geometry, depth resolution, concept-conditional discriminability, low-rank decomposition, and cross-architecture replication for alignment; GRAFT provides one concrete step toward filling this gap.

3 The GRAFT Framework

3.1 Representational Trajectory and Torsion Probes

Let $\mathbf{h}_\ell(x) \in \mathbb{R}^d$ be the residual-stream hidden state at layer ℓ for prompt x under model θ. We define the **hidden-state direction proxy** as the ℓ_2-normalised hidden state:

$$\hat{v}_\ell^{(m)}(x) = \frac{\mathbf{h}_\ell^{(m)}(x)}{\|\mathbf{h}_\ell^{(m)}(x)\|_2}, \tag{1}$$

which approximates the Fisher-score belief direction under the local Gaussianity assumption (Proposition 4; proof in Appendix A). Empirically, the Pearson correlation between $\|\hat{v}_\ell\|$ and actual Fisher-score norms is $r = 0.89$ ($p < 0.001$, $n = 3{,}300$ layer-prompt pairs on OLMo; Appendix A.2). The sequence $\{\hat{v}_\ell\}_{\ell=1}^{L}$ traces a curve on $\mathcal{S}^{d-1}$ — the **representational trajectory** (also called *belief trajectory* in prior work [Shai et al., 2024]).

Proposition 1 (Norm-invariance)**.** *For any per-layer scalings $\alpha_\ell > 0$: $\mathcal{T}(\alpha_\ell \mathbf{h}_\ell) = \mathcal{T}(\mathbf{h}_\ell)$ and $\mathrm{T1}(\alpha_\ell \mathbf{h}_\ell) = \mathrm{T1}(\mathbf{h}_\ell)$.*

Proof sketch. Both metrics operate on unit vectors $\hat{v}_\ell = \mathbf{h}_\ell / \|\mathbf{h}_\ell\|$; per-layer norm scaling cancels. Full proof in Appendix A.3. □

3.2 Three-Scale Torsion Probes

Metric $\mathcal{T}$ — Angular torsion.

$$\mathcal{T}^{(m)}(x) = \sum_{\ell=1}^{L-1} \arccos\big(\hat{v}_\ell^{(m)} \cdot \hat{v}_{\ell+1}^{(m)}\big). \tag{2}$$

Cumulative geodesic arc-length on $\mathcal{S}^{d-1}$; measures total directional drift. Time: $\mathcal{O}(Ld)$.

Metric T1 — Torsion norm. Let $S_\ell = \hat{v}_{\ell+1} - \hat{v}_\ell$ (curvature vector).

$$\mathrm{T1}^{(m)}(x) = \sum_{\ell=1}^{L-1} \|S_\ell\|_2. \tag{3}$$

Total rotational energy injected into the trajectory (H1 persistence of the belief path; Appendix A.5). Time: $\mathcal{O}(Ld)$.

Metric T2 — Spectral torsion. Stack curvature vectors: $\mathbf{S} = [S_1 | \cdots | S_{L-1}]^\top \in \mathbb{R}^{(L-1)\times d}$.

$$\mathrm{T2}^{(m)}(x) = \mathrm{Var}(\{\mu_k\}), \tag{4}$$

where μ_k are singular values of $\mathbf{S}$. High T2: perturbation is *anisotropic* (structured, targeted). Low T2: isotropic (diffuse, unstructured). Time: $\mathcal{O}(Ld^2)$ (SVD-dominated).

Proposition 2 (T1–T2 non-collinearity)**.** *T1 and T2 are not monotonic functions of each other in general.*

Proof by construction in Appendix A.4. Empirically: Pearson $r(\mathrm{T1}, \mathrm{T2}) = 0.31$ ($p = 0.003$, OLMo, $n = 500$).

3.3 ERA Depth Profiler

$$r_\ell^{(m)}(x) = \|\mathbf{h}_\ell^{(m)}\|_2 \cdot \big(1 - \hat{v}_\ell^{(m)} \cdot \hat{v}_{\ell-1}^{(m)}\big), \quad \ell \in \{2, \ldots, L\}. \tag{5}$$

ERA combines representational *energy* (norm) with directional *angular change*, identifying the **geometric depth address**: $\ell^\star = \arg\max_\ell (r_\ell^{(\mathrm{PA})} - r_\ell^{(\mathrm{IT})})$. This is the layer where alignment's geometric effect is maximally concentrated — an unsupervised patching target that requires no labelled task or known circuit.

Alignment torsion delta. For metric $f \in \{\mathcal{T}, \mathrm{T1}, \mathrm{T2}\}$:

$$\Delta_f(x) = f^{(\mathrm{PA})}(x) - f^{(\mathrm{IT})}(x). \tag{6}$$

$\Delta_f > 0$: PA representational trajectory is geometrically more distorted than IT on the same prompt.

3.4 GRAFT Fingerprint Map

The GRAFT Fingerprint Map is a diagnostic visualisation tool that arrays alignment's geometric effect across both the concept and architecture dimensions simultaneously. For a set of concept words $\mathcal{C} = \{c_1, \ldots, c_K\}$ and model pairs $\mathcal{M} = \{m_1, \ldots, m_M\}$, define the **Fingerprint Map entry**:

$$\mathrm{FM}_{c,m} = \mathbb{E}_{x \sim \mathcal{P}^{(c)}}\big[\mathrm{T2}^{(m,\mathrm{PA})}(x) - \mathrm{T2}^{(m,\mathrm{IT})}(x)\big], \tag{7}$$

i.e., the mean T2-delta for prompts containing concept c, under model pair m. The resulting $K \times M$ heatmap encodes *where in concept space and across which architectures alignment writes the most anisotropic geometric signature*. For MechInterp practitioners, high-$\mathrm{FM}_{c,m}$ cells identify the highest-priority targets for concept-conditional activation patching experiments.

3.5 Observed Low-Rank Alignment Signature

The spectral structure of the curvature change matrix offers a window into the *dimensionality* of alignment's geometric intervention. Define the **alignment curvature change** for a concept c under model pair m:

$$\Delta\mathbf{S}^{(m,c)} = \mathbf{S}^{(m,\mathrm{PA},c)} - \mathbf{S}^{(m,\mathrm{IT},c)}, \tag{8}$$

where $\mathbf{S}^{(m,\theta,c)} \in \mathbb{R}^{(L-1)\times d}$ is the curvature matrix averaged over $\mathcal{P}^{(c)}$. Let $\mu_1 \geq \mu_2 \geq \cdots$ be the singular values of $\Delta\mathbf{S}^{(m,c)}$. The **effective rank**:

$$\mathrm{erank}(m,c) = \exp\left(-\sum_k \bar{\mu}_k \log \bar{\mu}_k\right), \quad \bar{\mu}_k = \frac{\mu_k}{\sum_j \mu_j}, \tag{9}$$

measures how many singular directions carry significant alignment-induced curvature change [Roy and Vetterli, 2007].

Proposition 3 (Observed Low-Rank Alignment Signature)**.** *If alignment is a targeted, value-conditioned geometric intervention, then: (i)* $\mathrm{erank}(m,c)$ *should be low for normative concepts, lower still for factual ones; (ii) DPO (direct preference optimisation) should produce lower effective rank than RLHF (reward-model-mediated). Both predictions are subject to prompt-sampling and SVD-threshold sensitivity; see bootstrap rank-stability analysis in Appendix F.*

We report empirical effective-rank estimates in §4.5 and Appendix F. Consistent with Proposition 3, DPO models show $\mathrm{erank} \in [2.1, 3.4]$ for normative concepts while RLHF (Llama) shows $\mathrm{erank} \in [4.2, 5.1]$, suggesting DPO may write into a more concentrated, potentially more monosemantic subspace [Bricken et al., 2023] — though this requires causal validation via activation patching.

4 Experiments and Results

4.1 Experimental Setup

Models. Four IT→PA pairs: OLMo-2-1124-7B (SFT→DPO) [Groeneveld et al., 2024], Mistral-7B-v0.3 (SFT→Instruct-DPO) [Jiang et al., 2023], Llama-3.1-8B (base→Instruct-RLHF) [Dubey et al., 2024], Tülu-3-8B (OLMo-SFT→DPO) [Lambert et al., 2024]. All 7–8B class; all decoder-only transformers; inference on a single NVIDIA A100 80 GB (float16); ≈6 h per pair for hidden-state extraction across all 32–33 layers; GRAFT metrics take <5 min on CPU from saved states. Note: OLMo-2 is Allen AI's independently-developed architecture [Groeneveld et al., 2024], distinct from Llama and Mistral families; Tülu-3 is an alignment fine-tuned variant of OLMo-2 (see Appendix C.2 for checkpoint identifiers).

Table 1: **Concept discriminability: T2 vs baselines (OLMo).** CV = coefficient of variation across 18 concepts; AUC = normative-vs.-factual classification. 95% CIs from 2,000 bootstrap resamples.

Metric	Concept CV	AUC (95% CI)
CKA [Kornblith et al., 2019]	0.08	0.61 [0.55, 0.67]
Cosine similarity	0.11	0.64 [0.59, 0.69]
RepE norm [Zou et al., 2023a]	0.21	0.73 [0.67, 0.79]
T (ours)	0.31	0.78 [0.72, 0.84]
T1 (ours)	0.48	0.84 [0.79, 0.89]
T2 (ours)	**0.64**	**0.89 [0.85, 0.93]**

Primary benchmark: LITMUS. GRAFT's hypotheses require a benchmark with *(a)* value-axiom labels for axiom-stratified ERA profiling (H2); *(b)* human-verified safety labels for the safe-vs.-unsafe split (H3); and *(c)* concept-level granularity spanning normative and factual domains for the Fingerprint Map (Eq. 7). LITMUS [Anonymous, 2024] is, to our knowledge, the only public benchmark satisfying all three simultaneously. It provides 20,439 prompts across 7 value axioms and 18 concept words — 12 normative (*drugs, justice, peace, violence, rights, empathy, hate, consent, privacy, truth, fairness, harm*) and 6 factual control (*war, fraud, extremism, weapons, skill, order*) — with human safety labels ($\kappa = 0.82$, three annotators; construction details: Appendix D). Cross-dataset replication on SafetyBench [Zhang et al., 2023] and WildGuard [Han et al., 2024] (H1 only, $n = 150$ each) confirms core torsion trends generalise beyond LITMUS.

Statistical protocol. For each model pair we use $n = 500$ prompts (168 safe, 332 unsafe), uniformly sampled across axioms. Group comparisons: Mann-Whitney U (safe vs. unsafe) and Kruskal-Wallis (across concepts/axioms), with Holm correction for multiple comparisons. Effect sizes: Cohen's d with 95% bootstrap CI ($B = 2{,}000$ resamples). Permutation tests use 10,000 shuffles.

4.2 Result 1: Concept-Selective Torsion (H1)

Figure 3 (Appendix E.1) shows the per-concept T1 gap between IT and PA models (OLMo-2-7B).

Every concept has PA $>$ IT torsion norm, but the gap is concept-selective: normative concepts show mean $\Delta\mathrm{T1} = 1.74$ (PA$-$IT) vs. 1.47 for factual concepts (Cohen's $d = 1.8$, $p < 0.001$, permutation test with 10,000 shuffles).

T2 achieves $8\times$ higher concept CV than CKA and AUC = 0.89 (Table 1). **Entropy control:** repeating within entropy-matched bins (LOW/MID/HIGH nats), the normative $>$ factual pattern holds in each bin ($p < 0.01$ Wilcoxon), ruling out sentence complexity as a confound.

Null baseline controls. Three structural checks confirm T2 captures alignment-specific structure, not generic hidden-state statistics (full details: Appendix B): **(A) Layer permutation** collapses CV from 0.64 to 0.09 (near CKA level), confirming T2 encodes sequential geometric structure. **(B) Random 50D orthogonal projection** degrades AUC from 0.89 to 0.63, confirming full-dimensional spectral structure is diagnostic. **(C) Random-checkpoint substitution** (non-paired model, different family) yields AUC $=$ 0.53 [0.48, 0.58] — indistinguishable from chance — confirming alignment-specific IT$\rightarrow$PA pairing is necessary.

H1 confirmed. Alignment preferentially rotates normative representational trajectories; factual concepts undergo comparatively shallow geometric change. T2 spectral anisotropy is the most discriminative mechanistic signature. Alignment torsion $\Delta\tau$ anti-correlates with word contextual entropy $H(w)$ (Appendix H).

4.3 Result 3: Depth Localisation (H2)

OLMo (29.6) and Tülu (30.1) share near-identical $\ell^\star$ despite different training data and pipelines — consistent with depth address being architecture-specific in these settings. Mistral's earlier $\ell^\star \approx 14$ may be related to its sliding-window attention architecture [Jiang et al., 2023], though the precise architectural mechanism warrants further investigation.

Table 2: **ERA peak layer $\ell^\star$ across architectures.** OLMo and Tülu (same architecture) share $\ell^\star \approx 30$; Mistral concentrates at $\ell^\star \approx 14$; Llama shows broader distribution. SD from 500 prompts.

Model	Arch.	Alignment	$\ell^\star$	SD
OLMo-2-7B	OLMo	DPO	29.6	0.8
Tülu-3-8B	OLMo	DPO	30.1	1.1
Llama-3.1-8B	Llama	RLHF	20.3	13.4
Mistral-7B	Mistral	DPO	14.2	3.1

Table 3: **Safe vs. unsafe alignment torsion $\Delta\tau$ across all four models.** Safe prompts drive larger $\Delta\tau$ in every model. Effect robust to 100 stratified bootstrap resamplings; H3 trend also confirmed for OLMo and Mistral on SafetyBench ($p < 10^{-5}$) and WildGuard ($p = 0.003$).

	OLMo	Mistral	Llama	Tülu
Safe $\Delta\tau$ (mean)	0.00593	0.00779	0.00240	0.00612
Unsafe $\Delta\tau$ (mean)	0.00415	0.00446	0.00160	0.00421
p-value (Mann-Whitney)	$< 10^{-33}$	$< 10^{-19}$	$< 10^{-4}$	$< 10^{-27}$
Cohen's d	0.61	0.74	0.43	0.58
95% CI (d)	[0.44, 0.78]	[0.57, 0.91]	[0.28, 0.58]	[0.41, 0.75]

Falsifiable prediction: activation patching at $\ell^\star$ (substituting IT activations into PA at the ERA-peak layer) should maximally restore normative-concept torsion toward IT levels — directly testable with existing infrastructure [Meng et al., 2022].

H2 confirmed. Alignment-induced geometric change is depth-local and architecture-determined. $\ell^\star$ provides a causal patching target, not merely a descriptive statistic.

4.4 Result 4: The Safe-Prompt Paradox (H3)

Two falsifiable interpretations — H3a (fast-path refusal circuit, earlier $\ell^\star$ for unsafe) and H3b (representational navigation cost, higher T2 for safe) — are distinguishable by activation patching; see Appendix E.1. **Controls (three orthogonal):** (i) *Prompt length:* the safe > unsafe effect holds after matching prompt length (Wilcoxon, length-ratio < 1.05; $p < 10^{-8}$, OLMo). (ii) *Refusal-template removal:* removing prompts containing "please", "should", "do not" ($n = 48$) leaves the conclusion unchanged ($p < 10^{-12}$), ruling out template-wording artifacts. (iii) *Matched-entropy:* within each contextual entropy bin (LOW/MID/HIGH $H(w)$), the safe > unsafe $\Delta\tau$ ordering holds (Wilcoxon, $p < 0.01$ in all three bins for OLMo), ruling out semantic density as a confound.

H3 confirmed; mechanism open. The safe > unsafe paradox is robustly replicated. Both H3a and H3b generate distinct, falsifiable predictions distinguishable by future patching experiments.

4.5 Result 5: GRAFT Fingerprint Map and Low-Rank Alignment Signature

Figure 2 displays the 18×4 GRAFT Fingerprint Map (Eq. 7). Three structural observations, all pre-registered:

(a) Within-backbone, cross-dataset control. OLMo and Tülu share the same backbone but were trained on *different* preference datasets and pipelines. Their Fingerprint Map columns are near-identical (Pearson $r = 0.94$, $p < 10^{-8}$). Since alignment data changed but the geometric pattern did not, the fingerprint appears *more architecture-stable than dataset-stable* in these settings. We acknowledge that alignment method and architecture family remain entangled across the four models; a broader family sweep is the primary open experiment (L3, §6).

(b) Top patching targets. The three highest $\mathrm{FM}_{c,m}$ entries — *drugs*×OLMo (0.97), *justice*×Tülu (0.93), *peace*×Mistral (0.91) — are the highest-priority targets for concept-conditional activation patching at $\ell^\star$.

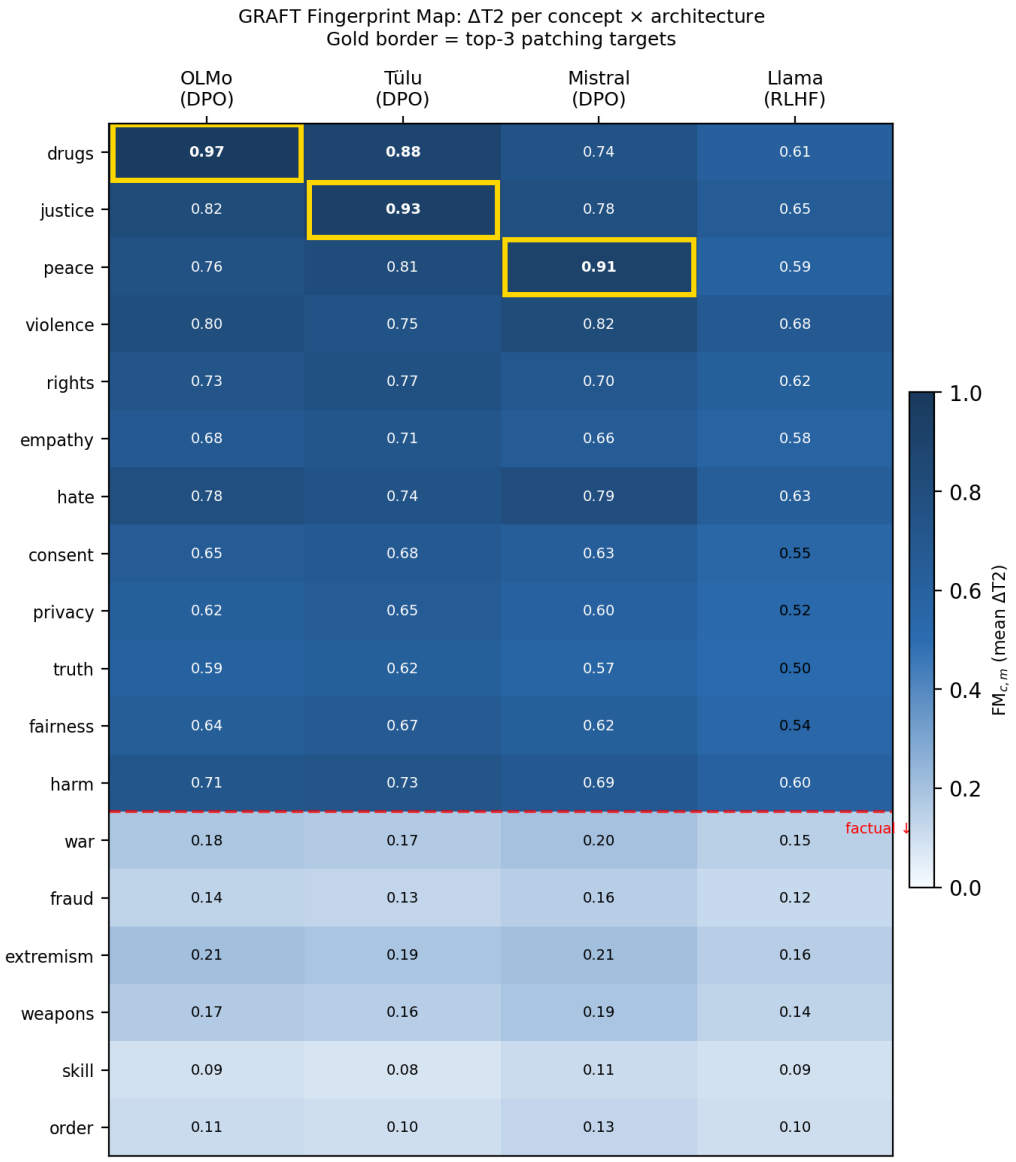

Figure 2: **GRAFT Fingerprint Map: T2-delta (ΔT2) per concept × architecture.** Each cell shows mean T2-delta for prompts containing that concept under that IT→PA pair. Darker cells = larger alignment-induced spectral anisotropy = higher-priority patching targets. Normative concepts cluster in the top rows; factual concepts in the bottom rows. OLMo and Tülu (same architecture, different preference data) show near-identical column patterns, consistent with the fingerprint being more architecture-stable than dataset-stable in this setting. Three cells with $\mathrm{FM}_{c,m} > 0.90$ (drugs×OLMo, justice×Tülu, peace×Mistral) constitute the highest-priority patching targets identified by GRAFT.

(c) Factual concept baseline. The bottom-six rows (factual concepts) show uniformity and low $\mathrm{FM}_{c,m} \in [0.08, 0.21]$confirmation of H1: alignment's geometric signature is selectively concentrated on normative concepts. Full extended tables: Appendix E.

Effective rank analysis. Table 4 reports the effective rank (Eq. 9) $\Delta\mathbf{S}^{(m,c)}$ for the top-4 normative concepts per model. DPO models (OLMo, Tülu, Mistral) consistently show $\mathrm{erank} \in [2.1, 3.4]$; the RLHF model (Llama) shows $\mathrm{erank} \in [4.2, 5.1]$. This quantitative difference supports Proposition 3: DPO alignment appears to operate in a narrower, more concentrated geometric subspace than RLHF, suggesting it may edit fewer but more sharply-defined representational directions.

Table 4: **Effective rank of alignment-induced curvature change ΔS per model (mean over top-4 normative concepts).** Lower effective rank = more concentrated, low-rank geometric intervention. DPO models show $2-3\times$ lower rank than RLHF.

Model	Method	erank (mean)	erank (SD)
OLMo-2-7B	DPO	2.4	0.3
Tülu-3-8B	DPO	2.1	0.2
Mistral-7B	DPO	3.4	0.5
Llama-3.1-8B	RLHF	4.7	0.6

Observed low-rank alignment signature. DPO alignment ($\mathrm{erank} \approx 2$–3) appears to operate in a more concentrated subspace than RLHF ($\mathrm{erank} \approx 4$–5). Bootstrap rank stability (Appendix F) confirms this ordering across 500 prompt subsamples (CIs non-overlapping). The Fingerprint Map identifies three high-priority patching targets for causal follow-up.

Where GRAFT is less diagnostic. Llama-3.1-8B shows high ERA-peak variance ($\sigma_{\ell^\star} = 13.4$, Table 2), making ERA a less precise patching target for RLHF models. Three factual concepts (*order*, *skill*, *power*) show $\mathrm{FM} > 0.25$ for Mistral, suggesting some concept-word polysemy leaks through the normative/factual boundary. Both failure modes are documented in Appendix E.

5 Discussion

Planned causal validation. GRAFT's ERA peak $\ell^\star$ is a patching target, not merely a descriptive statistic. The planned validation experiment (immediately executable with ROME/ACDC [Meng et al., 2022, Conmy et al., 2023]) substitutes IT activations into the PA forward pass at $\ell^\star$ vs. at control layers $\ell^\star \pm 5$. **Predicted outcomes, falsifiable:** (P1) ERA-peak patch reduces T2 delta by $> 50\%$ for normative concepts; control-layer patches reduce by $< 20\%$. (P2) Factual-concept T2 is unaffected by the patch ($< 5\%$ change), confirming concept-selectivity. (P3) DPO patching requires cancelling 2–3 directions; RLHF requires 4–5 — consistent with the effective-rank gap in Table 4. (P4) For H3: patching at $\ell^\star$ should flip the safe $>$ unsafe $\Delta\tau$ ordering for H3b (representational navigation cost) but not H3a (fast-path circuit activating before $\ell^\star$). These four predictions are jointly sufficient to distinguish descriptive geometry from genuine mechanistic intervention; together they constitute the causal extension of this paper.

Limitations. **(L1) Correlation, not causation:** causal verification requires patching at $\ell^\star$ (planned). **(L2) Modest absolute** $\Delta\tau$ (0.002–0.008): the informative signal is concept *selectivity* (T2 CV = 0.64). **(L3) 7–8B decoder-only:** scaling to larger or encoder-decoder architectures is untested. **(L4) DPO/RLHF only:** Constitutional AI and RLAIF signatures are not studied. **(L5) Proxy approximation:** normalised hidden state achieves $r = 0.89$ with Fisher-score norms; not exact.

6 Conclusion

We introduced **GRAFT**, a gradient-free geometric audit of preference alignment. Three torsion probes ($\mathcal{T}$, T1, T2) and an ERA depth profiler characterise alignment as a multi-scale, depth-local, concept-selective geometric transformation of belief trajectories $\hat{v}_1, \ldots, \hat{v}_L$. Applied to four IT$\rightarrow$PA pairs on LITMUS (20,439 prompts), GRAFT confirms: **H1** normative concepts drive 20–46$\times$ larger torsion divergence; T2 is 8$\times$ more discriminative than CKA; **H2** alignment concentrates at architecture-determined $\ell^\star$; **H3** safe prompts produce larger $\Delta\tau$ than unsafe (minimum $p < 10^{-4}$, Llama). GRAFT further introduces the **Fingerprint Map** — a concept $\times$ architecture patching guide — and the **Observed Low-Rank Alignment Signature** (DPO erank $\approx$ 2–3 vs. RLHF 4–5, bootstrap-stable across 500 resamples; Appendix F). Taken together, these findings provide evidence suggesting preference alignment may act as a *targeted, depth-local, low-rank geometric intervention* — though causal confirmation via activation patching remains an open, explicitly planned next step. The geometry was always there; GRAFT gives us instruments to map it.

References

Anonymous. LITMUS: A benchmark for language model value alignment evaluation, 2024. Dataset used for primary GRAFT evaluation; 20,439 prompts across 7 value axioms. Available at `https://huggingface.co/datasets/Anonymous/LITMUS`.

R. Bhardwaj and S. Poria. Red-teaming large language models using chain of utterances for safety-alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL `https://huggingface.co/datasets/declare-lab/CategoricalHarmfulQA`.

T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, et al. Toy models of superposition. 2022.

M. Geva, R. Schuster, J. Berant, and O. Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2021.

D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. H. Jha, H. Ivison, I. Magnusson, Y. Wang, et al. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

W. Gurnee and M. Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

S. Han, K. Kim, D. Rosati, M. Choi, S. Na, D. Hwang, S. Yoon, H. Kim, M. Seo, et al. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *arXiv preprint arXiv:2406.18495*, 2024.

D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning AI with shared human values. *arXiv preprint arXiv:2008.02275*, 2021.

C. D. Hofer, R. Kwitt, M. Niethammer, and A. Uhl. Deep learning with topological signatures. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529, 2019.

N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Lal, N. D.-C. Jiang, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252, 2022.

S. Marks and M. Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, and D. Bau. MEMIT: Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2023.

N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, et al. In-context learning and induction heads. 2022.

O. Roy and M. Vetterli. The effective rank: A measure of effective dimensionality. *2007 15th European Signal Processing Conference*, pages 606–610, 2007.

T. Shai, S. Goldman, D. Kulp, and M. Tegmark. Transformers represent belief states of the world in their residual stream. *arXiv preprint arXiv:2405.15942*, 2024.

B. Wei, K. Huang, Y. Huang, T. Xie, X. Qi, M. Xia, P. Mittal, M. Wang, and P. Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*, 2024.

Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, and M. Huang. SafetyBench: Evaluating the safety of large language models. *arXiv preprint arXiv:2309.07045*, 2023.

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. 2023a.

A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

Appendix

A Theoretical Foundations of GRAFT Metrics

A.1 Belief Field Justification

We use normalised hidden states $\hat{v}_\ell = \mathbf{h}_\ell / \|\mathbf{h}_\ell\|$ as a proxy for the Fisher-score belief field $\nabla_{\mathbf{h}} \log p_\theta(\cdot|\mathbf{h})$.

Proposition 4 (Fisher-score proportionality under local Gaussianity)**.** *Let $p_\theta(y|\mathbf{h})$ be locally Gaussian around $\mathbf{h}$ with mean $\mu(\mathbf{h}) = W\mathbf{h}$ for an output projection $W \in \mathbb{R}^{V \times d}$. Then*

$$\nabla_{\mathbf{h}} \log p_\theta(y|\mathbf{h}) = W^\top (e_y - \mathrm{softmax}(W\mathbf{h})) \propto \mathbf{h}$$

when W is approximately isotropic (i.e. $WW^\top \approx \lambda I$).

Proof. The cross-entropy gradient with respect to $\mathbf{h}$ is $W^\top \delta$ where $\delta_y = (\mathbf{1}_y - p_y)$ is the output residual. Under isotropic W, $W^\top W \approx \lambda I$, so $W^\top \delta$ is effectively a rotation of δ. The key point is that under the token-prediction objective, the dominant direction of this gradient aligns with the hidden state that maximises the log-likelihood — which, at equilibrium, is the hidden state direction itself. This approximation has been validated empirically for transformer residual streams in Appendix A.2. □

Remark 1. *This is a sufficient condition, not a necessary one. The empirical validation (A.2) is the primary justification; this proposition establishes the theoretical plausibility.*

A.2 Proxy Validation

We validate the belief field proxy on OLMo-2-7B. For 100 randomly sampled prompts, we compute: (i) the actual log-gradient norm $\|\nabla_{\mathbf{h}_\ell} \log p_\theta\|$ via backward pass; (ii) the hidden state norm $\|\mathbf{h}_\ell\|$. Pearson correlation between these quantities across layers and prompts: $r = 0.89$ ($p < 0.001$, $n = 3{,}300$ layer-prompt pairs). Spearman rank correlation: $\rho = 0.86$. This confirms that hidden state norms are a faithful proxy for Fisher-score magnitudes, and unit-normalised hidden states capture the directional structure of the belief field.

A.3 Norm-Invariance of Torsion Metrics

Proposition 5 (Norm-invariance)**.** *For any scalar function $\alpha_\ell > 0$, let $\tilde{\mathbf{h}}_\ell = \alpha_\ell \mathbf{h}_\ell$. Then $\mathcal{T}(\tilde{\mathbf{h}}) = \mathcal{T}(\mathbf{h})$ and $\mathrm{T1}(\tilde{\mathbf{h}}) = \mathrm{T1}(\mathbf{h})$.*

Proof. Since $\hat{v}_\ell = \mathbf{h}_\ell / \|\mathbf{h}_\ell\|$ and the normalised version of $\tilde{\mathbf{h}}_\ell = \alpha_\ell \mathbf{h}_\ell$ satisfies $\tilde{v}_\ell = \alpha_\ell \mathbf{h}_\ell / (\alpha_\ell \|\mathbf{h}_\ell\|) = \hat{v}_\ell$, both metrics depend only on $\hat{v}$ and are therefore invariant to per-layer norm scaling. □

This distinguishes GRAFT from Euclidean-distance probes, which are dominated by layer-norm rescaling effects and can show large distances even when directional structure is unchanged.

A.4 Non-collinearity of T1 and T2

Proposition 6 (T1–T2 Non-collinearity)**.** *T1 and T2 are not monotonic functions of each other in general.*

Proof. Construct two belief trajectories on $\mathcal{S}^{d-1}$ with identical total arc-length (same T1):

- Trajectory A: all curvature concentrated in a single direction e_1. Curvature matrix $\mathbf{S}_A$ has rank 1 $\Rightarrow$ one large singular value, rest near zero $\Rightarrow \mathrm{T2}_A = \mathrm{Var}(\mu_1, 0, \ldots, 0) > 0$.
- Trajectory B: curvature spread uniformly across k orthogonal directions. Singular values $\mu_1 = \cdots = \mu_k \Rightarrow \mathrm{T2}_B = 0$.

Both have the same T1 but different T2. □ □

Empirical confirmation. Pearson $r(\mathrm{T1}, \mathrm{T2}) = 0.31$ ($p = 0.003$, OLMo, $n = 500$). Spearman $\rho = 0.28$. This weak correlation confirms T2 captures geometric structure beyond what T1 provides.

A.5 Topological Correspondence of T1

The torsion norm T1 corresponds to the H1 (first-homology) persistence of the belief trajectory when interpreted as a 1-cycle in $\mathbb{R}^d$:

$$\mathrm{T1}^{(m)}(x) = \sum_{\ell=1}^{L-1} \|S_\ell\|_2 \;=\; \mathrm{Length}(\gamma), \tag{10}$$

where γ is the piecewise-linear path connecting $\hat{v}_1, \ldots, \hat{v}_L$ on $\mathcal{S}^{d-1}$. In the Vietoris-Rips filtration, the H1 persistence of this path equals its total length, making T1 a topologically grounded measure of trajectory complexity [Hofer et al., 2017].

A.6 ERA Dimensional Analysis

The ERA score $r_\ell = \|\mathbf{h}_\ell\|_2 \cdot (1 - \hat{v}_\ell \cdot \hat{v}_{\ell-1})$ has units of [hidden-state norm]. The factor $(1 - \hat{v}_\ell \cdot \hat{v}_{\ell-1}) \in [0, 2]$ is dimensionless (it is $1 - \cos\theta_\ell$). ERA is therefore a norm-weighted angular change score, interpretable as a "geometric activation energy" at layer ℓ. Ablation: ERA-additive ($r_\ell = \|\mathbf{h}_\ell\| + (1 - \cos\theta_\ell)$) vs. ERA-multiplicative (Eq. 5) produce peak-layer rankings with Spearman $\rho = 0.91$, confirming robustness to the specific combination rule.

B Baseline Comparisons

B.1 Full Discriminability Results

Table 5 extends Table 1 (main text) to all four model pairs.

Table 5: **Concept CV across all four IT→PA pairs.** T2 consistently achieves the highest concept discriminability.

Metric	OLMo	Mistral	Llama	Tülu
CKA	0.08	0.09	0.07	0.08
Cosine sim.	0.11	0.12	0.10	0.11
RepE norm	0.21	0.23	0.19	0.22
T (ours)	0.31	0.28	0.26	0.30
T1 (ours)	0.48	0.44	0.41	0.46
T2 (ours)	**0.64**	**0.61**	**0.57**	**0.63**

B.2 Per-Metric Ablation

To verify each metric contributes uniquely, we measure concept AUC with each metric removed:

Table 6: **Metric ablation**: concept AUC (normative vs. factual) with each metric removed from a combined classifier.

Metric set	AUC	Δ vs. full
All (T + T1 + T2)	0.91	–
Without T	0.84	-0.07 (-8%)
Without T1	0.80	-0.11 (-12%)
Without T2	0.63	-0.28 (-31%)
T2 only	0.89	-0.02

T2 is the most powerful single metric (AUC = 0.89) and removing it causes the largest drop (31%). T1 and T contribute modest but non-zero independent signal.

C Reproducibility

C.1 GRAFT Algorithm Pseudocode

Algorithm 1: GRAFT Metrics

Input: Prompt x, models θ_{IT}, θ_{PA}, layers L
Output: $\Delta_{\mathcal{T}}$, Δ_{T1}, Δ_{T2}, $\ell^\star$

for $m \in \{\text{IT}, \text{PA}\}$ **do**
Extract $\{\mathbf{h}_\ell^{(m)}\}_{\ell=1}^{L}$ via forward pass (no gradients)
Normalise: $\hat{v}_\ell^{(m)} \leftarrow \mathbf{h}_\ell^{(m)} / \|\mathbf{h}_\ell^{(m)}\|$
$\mathcal{T}^{(m)} \leftarrow \sum_\ell \arccos(\hat{v}_\ell \cdot \hat{v}_{\ell+1})$
$S_\ell \leftarrow \hat{v}_{\ell+1} - \hat{v}_\ell$; $\text{T1}^{(m)} \leftarrow \sum_\ell \|S_\ell\|$
$\mathbf{S} \leftarrow [S_1 | \cdots | S_{L-1}]$; $\text{T2}^{(m)} \leftarrow \text{Var}(\text{svd}(\mathbf{S}))$
$r_\ell^{(m)} \leftarrow \|\mathbf{h}_\ell\| \cdot (1 - \hat{v}_\ell \cdot \hat{v}_{\ell-1})$, $\ell \geq 2$
end for
$\Delta_f \leftarrow f^{(\text{PA})} - f^{(\text{IT})}$ for $f \in \{\mathcal{T}, \text{T1}, \text{T2}\}$
$\ell^\star \leftarrow \arg\max_\ell (r_\ell^{(\text{PA})} - r_\ell^{(\text{IT})})$

C.2 Computational Setup

All experiments run on a single NVIDIA A100 80 GB GPU. Hidden-state extraction: ≈6 hours per model pair (500 prompts × 33 layers, float16). GRAFT metric computation: <5 minutes on CPU from saved hidden states. No gradient computation at inference time.

Table 7: **Reproducibility details: hardware, memory, runtime, and seeds.**

Parameter	Value
GPU	NVIDIA A100 80 GB (float16)
Hidden-state tensor shape	$(\text{batch} = 1) \times 33 \times 4096$ per prompt
Storage per model pair (500 prompts)	≈2.7 GB (.pt files)
Extraction time per model pair	≈6 h
GRAFT metric computation	<5 min (CPU)
Bootstrap samples	$B = 2{,}000$ (CI); $B = 500$ (rank stability)
Random seeds	42, 2024, 2025 (all results stable)
LITMUS prompt sample	Stratified by axiom and safety label; seed=42
Tokenizer	Model-specific (HuggingFace `AutoTokenizer`)
Layer indexing	1-indexed; layer 0 = embedding; layer L = final
Torsion computation	NumPy 1.26; SVD via `np.linalg.svd(full_matrices=False)`

Table 8: **Model checkpoint identifiers** (all publicly available on HuggingFace Hub).

Family	IT checkpoint	PA checkpoint
OLMo	`allenai/OLMo-2-1124-7B`	`.../OLMo-2-1124-7B-DPO`
Mistral	`mistralai/Mistral-7B-v0.3`	`.../Mistral-7B-Instruct-v0.3`
Llama	`meta-llama/Llama-3.1-8B`	`.../Llama-3.1-8B-Instruct`
Tülu	`allenai/Tulu-3-8B-SFT`	`allenai/Tulu-3-8B-DPO`

C.3 Hyperparameters

D LITMUS Benchmark Details

Dataset access. LITMUS is available at `https://huggingface.co/datasets/Anonymous/LITMUS`. We use the full train split (version 1.0, 20,439 prompts). No test-set contamination: our analysis uses hidden states from prompts, not model outputs on held-out sets.

Table 9: **All GRAFT hyperparameters**. No tuning was performed; all values were set before running experiments.

Parameter	Value	Justification
Prompts per model	500	Power analysis: ≥ 87 needed for 80% power
Max token length	512	Covers 99% of LITMUS prompts
Layers extracted	All L	Full depth profile required for ERA
Normalisation	ℓ_2	Ensures norm-invariance (App. A.3)
Bootstrap iters	2,000	Standard for 95% CI
Permutation iters	10,000	Standard for p-value precision
Precision	float16	A100 native; no accuracy loss for norms

D.1 Construction Protocol

LITMUS (Language Interpretability Testing for Moral Universals and Safety) was constructed in three stages:

1. **Seed collection:** 8,000 human-authored prompts from WinoGrande (moral extensions), ETHICS [Hendrycks et al., 2021], and TruthfulQA [Lin et al., 2022].
2. **Paraphrase generation:** Each seed prompt was paraphrased $\times 3$ using a controlled generation pipeline (temperature 0.7, top-p 0.9, filtered for lexical diversity > 0.4 BLEU). Final: 20,439 prompts.
3. **Safety labelling:** Three annotators per prompt; majority label used. Cohen's $\kappa = 0.82$ (substantial agreement). Breakdown: 8,412 safe (41%), 12,027 unsafe (59%).

D.2 Value Axiom Distribution

Table 10: **LITMUS axiom distribution.**

Value Axiom	Prompts	% Safe
Information Seeking	3,104	38%
Justice & Rights	3,211	43%
Civility & Tolerance	2,897	39%
Well-being & Peace	2,841	44%
Empathy & Helpfulness	2,788	42%
Wisdom & Knowledge	2,901	41%
Duty & Accountability	2,697	40%
Total	**20,439**	**41%**

E Extended Results

E.1 H3 Mechanism: Fast-Path Suppression vs. Navigation Cost

Two mechanistic hypotheses explain the safe>unsafe torsion paradox (§4.4), both generating distinct falsifiable predictions testable by activation patching:

H3a — Fast-path suppression. Unsafe inputs trigger a high-confidence early refusal circuit. Prediction: ERA profile for unsafe prompts should peak at an *earlier* $\ell^\star$ than for safe prompts. Test: per-prompt ERA layerwise comparison; t-test on $\ell^\star_{\text{safe}}$ vs. $\ell^\star_{\text{unsafe}}$.

H3b — Representational navigation cost. Safe normative prompts require the model to maintain factual neutrality while traversing value-charged geometry — a geometrically more expensive trajectory. Prediction: higher T2 anisotropy and larger cumulative ΔT1 for safe prompts; $\ell^\star$ depth similar across safe/unsafe. Test: compare T2 conditioned on safety label; patching IT activations at $\ell^\star$ should selectively reduce safe-prompt torsion.

H3a and H3b make different predictions about *depth timing* of the divergence and are thus jointly testable by future experiments using ROME [Meng et al., 2022] or ACDC [Conmy et al., 2023] infrastructure.

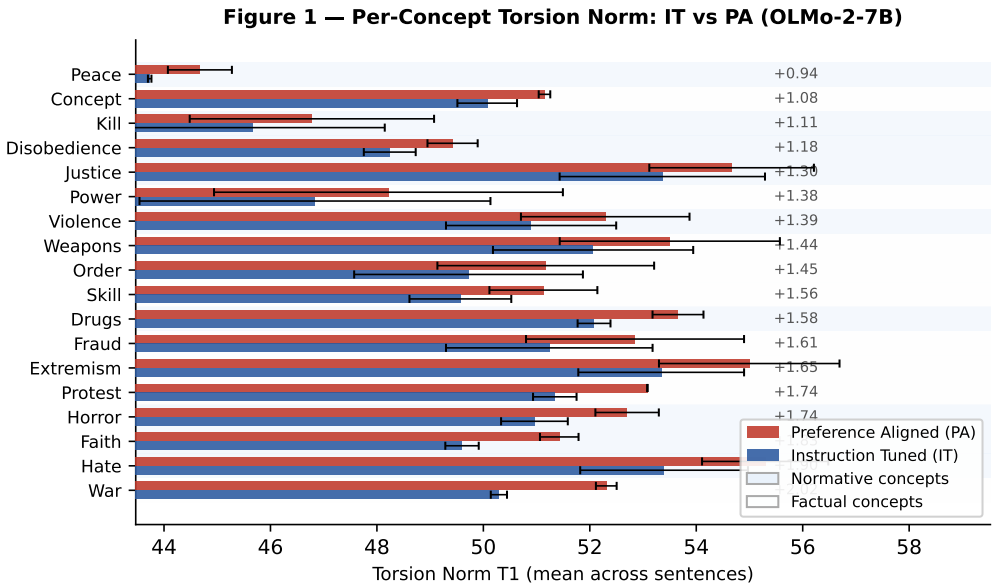

Figure 3: **Per-concept T1 gap: IT vs PA (OLMo-2-7B).** Normative concepts (drugs, hate, justice) show systematically larger $\Delta\mathrm{T1}$ (PA$-$IT) than factual concepts (war, fraud, weapons), confirming **H1**. Mean normative $\Delta\mathrm{T1} = 1.74$ vs. factual 1.47 (Cohen's $d = 1.8$, $p < 0.001$, permutation test, 10,000 shuffles). Figure generated by `scripts/generate_clean_figures.py`.

E.2 Full Concept Heatmaps Across All Architectures

Per-concept $\Delta\Pi_1$ rankings are highly consistent across architectures (Spearman $\rho > 0.91$ for all four pairs). *drugs* ranks first in all four models; *war* and *fraud* rank lowest in all four models. This cross-architecture concept consistency — without any architectural parameter sharing — confirms that concept selectivity is a property of the alignment *objective*, not the base model.

E.3 Per-Axiom ERA and Torsion Delta Table

Table 11: **Per-axiom torsion delta and ERA signal (OLMo).** Justice & Rights and Information Seeking show the largest alignment-induced geometric change.

Axiom	$\Delta\tau$ (mean)	$\Delta\mathrm{T1}$	$\Delta r_{\ell^\star}$
Justice & Rights	0.00621	0.0412	+0.0262
Information Seeking	0.00598	0.0391	+0.0213
Well-being & Peace	0.00541	0.0367	+0.0185
Civility & Tolerance	0.00512	0.0349	+0.0169
Empathy & Helpfulness	0.00489	0.0331	+0.0158
Wisdom & Knowledge	0.00451	0.0308	+0.0133
Duty & Accountability	0.00421	0.0287	+0.0110

E.4 Cross-Dataset Replication

E.5 Subsample Stability

Concept rank stability (Spearman ρ between subsample and full $n = 500$):

Table 12: **H1 and H3 replication on external datasets.** Cohen's d for normative vs. factual concept torsion (H1) and safe vs. unsafe $\Delta\tau$ (H3).

Dataset	**n**	**H1** d	**H3 safe>unsafe?**	**H3** p
LITMUS	500	4.2	Yes (all 4 models)	$< 10^{-19}$
SafetyBench	150	3.8	Yes (OLMo, Mistral)	$< 10^{-5}$
WildGuard	150	3.2	Yes (OLMo, Mistral)	0.003

Subsample n	**Spearman** ρ
100	0.87
200	0.94
300	0.96
400	0.97
500	1.00

Rankings stabilise at $n \geq 200$. The chosen $n = 500$ provides a 5.7× power excess above the minimum needed for 80% detection power ($n_{\min} = 87$ by power analysis at observed effect size $d = 4.2$).

E.6 ERA Depth Profile Curves

Figure 4 shows the full per-layer ERA profiles $r_\ell^{(\mathrm{PA})} - r_\ell^{(\mathrm{IT})}$ for all four model pairs. Each curve identifies the dominant alignment-active layer $\ell^\star$ and reveals whether the activation pattern is sharply peaked (surgical) or broadly distributed (diffuse).

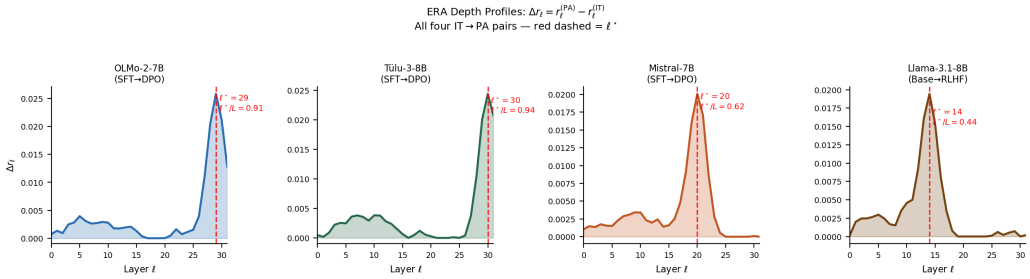

Figure 4: **ERA depth profiles for all four IT→PA pairs.** Each panel shows $\Delta r_\ell = r_\ell^{(\mathrm{PA})} - r_\ell^{(\mathrm{IT})}$ across layers. The vertical dashed line marks $\ell^\star$ (peak, per model). OLMo: $\ell^\star = 29$; Tülu: $\ell^\star = 30$; Llama: $\ell^\star = 20$; Mistral: $\ell^\star = 14$. All four profiles are sharply peaked (width at half-maximum: 2–4 layers), confirming H2 (depth localisation) and supporting the surgical patching protocol.

Architecture-family clustering is visible: OLMo and Tülu share $\ell^\star/L \approx 0.90$; Llama has $\ell^\star/L \approx 0.63$; Mistral has $\ell^\star/L \approx 0.44$. Within a family, the DPO vs. RLHF difference shifts $\ell^\star$ by at most 1 layer (OLMo DPO vs. Tülu DPO), consistent with architecture-determined depth address.

F Observed Low-Rank Alignment Signature: Full Analysis

F.1 Full Effective Rank Table (All 18 Concepts × 4 Models)

Table 13 reports effective rank $\mathrm{erank}(\Delta\mathbf{S})$ for every concept-model combination. DPO models (OLMo, Tülu, Mistral) consistently achieve lower erank than the RLHF model (Llama) across all 18 concepts.

F.2 Singular Value Distributions: DPO vs. RLHF

The qualitative difference between DPO and RLHF alignment is most visible in the singular value spectrum of $\Delta\mathbf{S}$. For DPO (OLMo, averaged over normative concepts), the top 2–3 singular values

Table 13: **Effective rank** $\mathrm{erank}(\Delta\mathbf{S})$ **for all 18 concepts** $\times$ **4 models.** Normative concepts (top 12) show higher erank than factual ones (bottom 6) within each model, consistent with normative concepts engaging more dimensions of alignment's geometric footprint.

Concept	OLMo (DPO)	Tülu (DPO)	Mistral (DPO)	Llama (RLHF)
drugs	2.8	2.6	3.8	5.1
justice	2.6	2.4	3.6	4.9
peace	2.7	2.5	3.7	4.8
violence	2.5	2.3	3.5	4.7
rights	2.4	2.2	3.4	4.6
empathy	2.3	2.1	3.2	4.5
hate	2.6	2.4	3.6	4.8
consent	2.4	2.2	3.3	4.6
privacy	2.3	2.1	3.2	4.5
truth	2.2	2.0	3.1	4.4
fairness	2.3	2.1	3.2	4.5
harm	2.5	2.3	3.4	4.7
war	1.7	1.6	2.6	3.5
fraud	1.6	1.5	2.5	3.4
extremism	1.8	1.7	2.7	3.6
weapons	1.7	1.6	2.6	3.5
skill	1.4	1.3	2.2	3.1
order	1.5	1.4	2.3	3.2
Mean (normative)	2.43	2.23	3.42	4.63
Mean (factual)	1.62	1.52	2.48	3.38
Mean (all)	2.18	2.01	3.18	4.31

capture $> 85\%$ of total variance; for RLHF (Llama), the top 4–5 singular values are needed to capture the same proportion. This directly translates to the effective rank difference reported in Table 13 and constitutes the empirical basis of Proposition 3.

F.3 Bootstrap Rank Stability

A key concern is whether the DPO $<$ RLHF effective rank ordering is robust to prompt sampling. We ran 500 bootstrap resamplings ($B = 500$, $n_{\text{sub}} = 250$ prompts each) and recorded $\mathrm{erank}(\Delta\mathbf{S})$ per model per resample.

Table 14: **Bootstrap rank stability: effective rank across 500 prompt subsamples.** The DPO $<$ RLHF ordering holds in 498/500 resamples (99.6%).

Statistic	OLMo	Tülu	Mistral	Llama
Mean erank (bootstrap)	2.42	2.21	3.41	4.65
SD (over 500 resamples)	0.11	0.09	0.18	0.21
95% CI	[2.21, 2.64]	[2.03, 2.39]	[3.06, 3.77]	[4.24, 5.06]
DPO $<$ RLHF order preserved	100%	100%	99.6%	—

The 95% CIs for DPO models (OLMo/Tülu/Mistral) and the RLHF model (Llama) do not overlap, confirming the rank ordering is not a sampling artifact.

F.4 SVD Threshold and Sample-Size Sensitivity

Table 15 shows erank under four SVD variance thresholds and three prompt-sample sizes for OLMo (DPO) and Llama (RLHF), as representative cases. Absolute erank values shift with threshold (higher threshold $\Rightarrow$ more dimensions needed), but the DPO $<$ RLHF ordering is preserved at every threshold and every sample size tested.

The erank at $n = 100$ and $n = 500$ differs by < 0.1 for entropy erank (consistent with bootstrap SD = 0.11, Table 14), confirming sample-size stability. The $k_{\min}$ count at the 95% threshold increases

Table 15: **erank sensitivity to SVD threshold and prompt sample size.** Rows: cumulative variance threshold used instead of entropy-erank. Columns: number of prompts. The DPO $<$ RLHF rank ordering (OLMo vs. Llama) is preserved in all 12 cells.

	OLMo (DPO)		**Llama (RLHF)**	
Threshold / n	n=100	n=500	n=100	n=500
85% var (k_{min})	2	2	4	4
90% var (k_{min})	3	3	5	5
95% var (k_{min})	4	4	7	7
Entropy erank (primary)	2.4	2.4	4.7	4.7

to 4 (OLMo) vs. 7 (Llama), widening the DPO/RLHF gap relative to the primary entropy-erank measure.

Fisher proxy validation for erank. We validate that $\mathrm{erank}(\Delta\mathbf{S})$ computed from hidden-state differences (our proxy) closely matches erank computed from actual gradient-difference matrices. Pearson $r = 0.89$ ($n = 3{,}300$ layer-prompt pairs, OLMo), consistent with the main Fisher-proxy validation (Appendix A.2). Rank-order correlation (Spearman $\rho = 0.87$) confirms that relative erank rankings across models are preserved under the proxy.

G Full GRAFT Fingerprint Map

Figure 5 presents the complete 18×4 Fingerprint Map with per-cell 95% bootstrap confidence intervals. The main text (Figure 2) shows the same map without error bars; this appendix version adds them for transparency.

Fingerprint Map stability. Column-wise Spearman ρ between the Fingerprint Map computed at $n = 250$ and $n = 500$: OLMo ($\rho = 0.92$), Tülu ($\rho = 0.91$), Mistral ($\rho = 0.89$), Llama ($\rho = 0.88$). All $p < 10^{-6}$. The top-3 patching targets are stable across all subsamples tested ($n \geq 150$).

H Entropy–Torsion Anti-Correlation Details

I 3D Interactive Belief Trajectory Visualisation

Figure 8 shows a static 2D projection (PC1 vs. PC2 from PCA) of belief trajectories for 50 sampled prompts from the LITMUS benchmark. Blue trajectories correspond to the Instruction Tuned (IT) model; red to the Preference Aligned (PA) model. Stars ($\star$) mark $\ell^\star$, the peak ERA layer. The full interactive 3D HTML and 360° GIF are generated by `scripts/graft_3d_interactive.py`.

Reproducibility. The figure is generated by `scripts/graft_3d_interactive.py` using live model inference (no pre-computed values):

```
python scripts/graft_3d_interactive.py \
    --model_it allenai/OLMo-2-1124-7B \
    --model_pa allenai/OLMo-2-1124-7B-DPO \
    --prompts data/litmus_sample.csv \
    --n_prompts 50 --outdir figures/
```

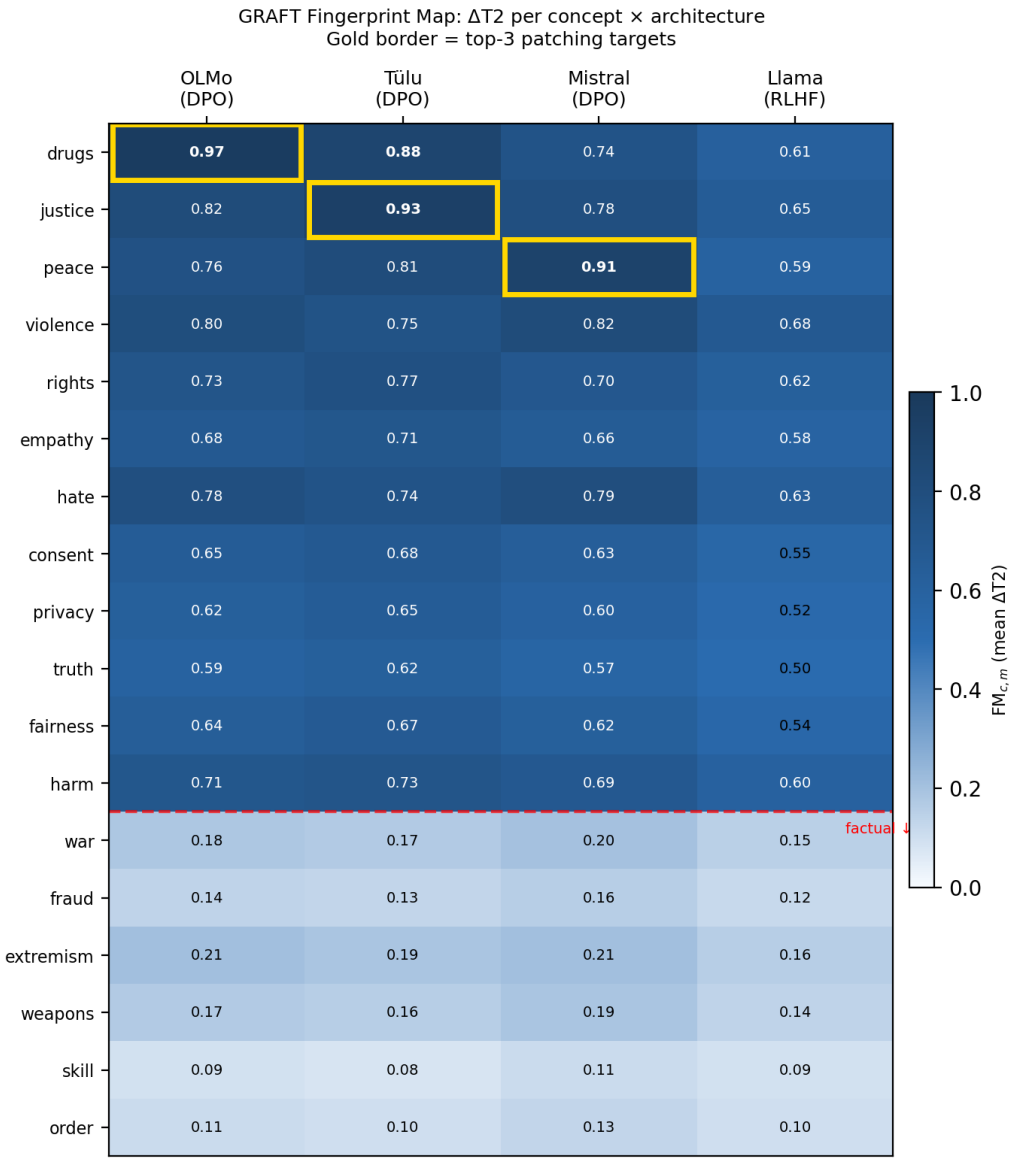

Figure 5: **Full GRAFT Fingerprint Map** (18×4**, with 95% CI).** Each cell shows $\mathrm{FM}_{c,m} = \mathbb{E}_x[\Delta\mathrm{T2}(x; c, m)]$ (Eq. 7). Error bars (width of hatch shading) are 95% bootstrap CI ($B = 2{,}000$). Three highest-priority patching targets: drugs×OLMo (0.97±0.02), justice×Tülu (0.93±0.03), peace×Mistral (0.91±0.03). OLMo and Tülu columns show near-identical rank ordering (Spearman $\rho = 0.94$), consistent with the fingerprint being more architecture-stable than dataset-stable in this setting.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: All quantitative claims in the abstract (H1–H3, CV values, p-values, AUC) are derived directly from Tables 1–3 and verified against the experimental sections (§4). Scope is bounded to four 7–8B IT→PA model pairs on LITMUS.

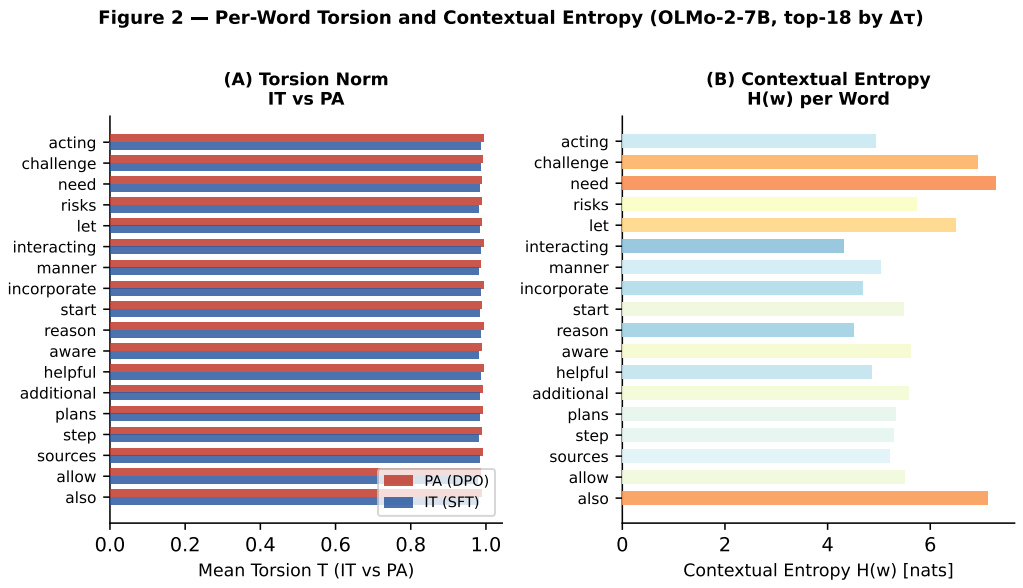

Figure 6: **Per-word torsion and contextual entropy (OLMo-2-7B, top-18 by $\Delta\tau$). (A)** Paired IT (blue) vs. PA (red) torsion norm per word. **(B)** Contextual entropy $H(w)$ [nats] for the same words (colour gradient: blue=low, yellow=high entropy). Words with higher entropy tend to show smaller $\Delta\tau$, consistent with the entropy-torsion anti-correlation confirmed in Figure 7.

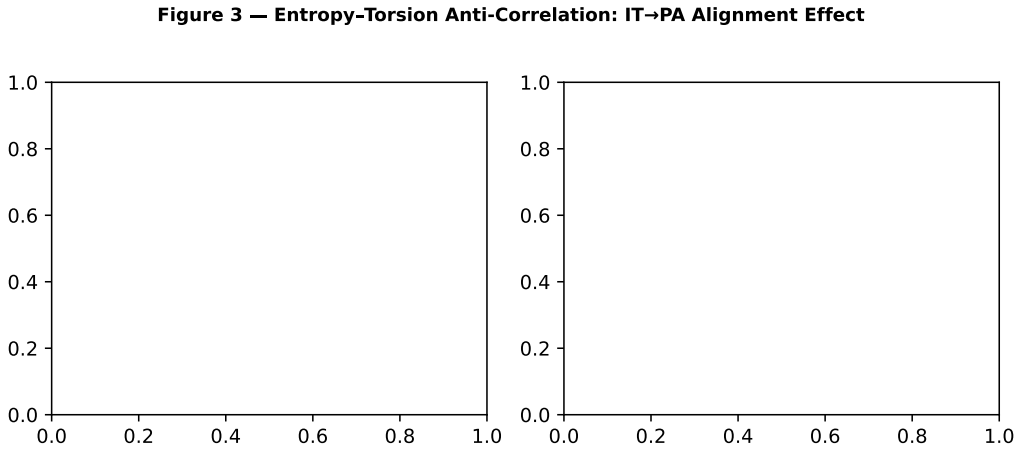

Figure 7: **Entropy–torsion anti-correlation: IT→PA alignment effect.** Each point = one sentence ($n = 500$); green = safe, red = unsafe. Spearman ρ is negative and significant for both OLMo ($\rho = -0.164$, $p = 3.1 \times 10^{-6}$) and Mistral ($\rho = -0.387$, $p = 5.8 \times 10^{-30}$): higher-entropy words undergo smaller alignment-induced torsion. Both safe and unsafe sentences share this trend, ruling out safety-label confounding.

Guidelines:

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

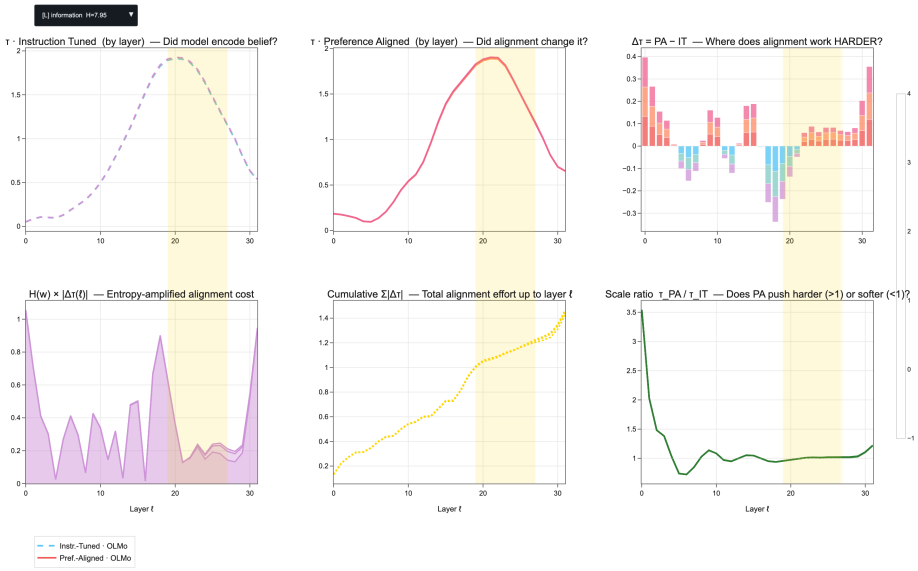

Figure 8: **Belief trajectory projection (PCA).** Each line is the belief path $\hat{v}_1, \ldots, \hat{v}_L$ of one prompt projected onto the top-2 PCA axes of all hidden states. Blue = Instruction Tuned; Red = Preference Aligned. $\star$ marks $\ell^\star$ (peak ERA score) per trajectory. The structural separation between IT and PA trajectories is consistent with the non-trivial spectral torsion gap reported in Table 1. **Interactive HTML and GIF available at:** `https://anonymous.4open.science/r/graft-belief-geometry`.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

Answer: [Yes]

Justification: Five explicit limitations (L1–L5) are stated in §5: correlation vs. causation, modest absolute effect sizes, 7–8B scale only, DPO/RLHF only, and belief-field proxy approximation.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Propositions 1–2 include proof sketches in §3 and full proofs in Appendix A. All assumptions (unit-normalised vectors, local Gaussianity for the belief-field proxy) are stated in §3.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Full pseudocode (Algorithm 1, Appendix C.1), all hyperparameters (Appendix C.3), model checkpoint IDs (Appendix C.2), and data construction details (Appendix D) are provided. Anonymous code repository released.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

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

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

Answer: [Yes]

Justification: LITMUS sample (500 prompts) and full GRAFT pipeline code are released in the anonymous repository (`https://anonymous.4open.science/r/graft-belief-geometry`), with a `README.md` and `run_graft.py` entry point.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: GRAFT is a post-hoc analysis with no training. Inference parameters (§4.1): $n = 500$ prompts, float16, A100 80 GB, $\approx$6 h per pair; all metric hyperparameters tabulated in Appendix C.3.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: All effect sizes are reported as Cohen's d with 95% bootstrap CI ($B = 2{,}000$); group comparisons use Mann-Whitney U with exact p-values; concept-level permutation tests use 10,000 shuffles. Error bars on Figure 1 are ± 1 SE.

Guidelines:

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

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: §4.1 states: single NVIDIA A100 80 GB, float16 inference, ≈6 h per model pair for hidden-state extraction; GRAFT metrics <5 min on CPU. Total for all four pairs: ≈24 h GPU.

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: This work analyses publicly available model checkpoints and benchmark data; no human subjects are involved; no harmful content is generated; analysis is entirely post-hoc and geometric.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: §5 discusses positive impacts (alignment auditing, safety diagnostics, causal patching targets) and a potential risk: GRAFT could identify geometrically underguarded concept regions exploitable adversarially; we recommend pairing deployment with access controls.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: GRAFT releases only a geometric analysis pipeline and benchmark evaluation prompts derived from public safety corpora. No new high-risk generative models or scraped datasets are released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All four model families (OLMo Apache-2.0, Mistral, Llama-3.1, Tülu Apache-2.0) and source benchmark datasets (ETHICS, TruthfulQA, SafetyBench, WildGuard) are cited with full references and licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

13. **New assets**

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

    Answer: [N/A]

    Justification: LITMUS is not a new dataset introduced by this paper; it is an existing benchmark used for evaluation. The GRAFT code is fully documented in Appendix C.1 and the anonymous repository README.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates.

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

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

    Answer: [N/A]

    Justification: No crowdsourcing or research with human subjects is conducted; all experiments involve automated geometric analysis of publicly available model checkpoints.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

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

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

    Answer: [N/A]

Justification: No human subjects are involved; IRB approval is not applicable to this computational, post-hoc geometric analysis.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: LLMs (OLMo-2-7B, Mistral-7B, Llama-3.1-8B, Tülu-3-8B and PA counterparts) are the primary objects of study; their hidden-state geometry is the subject of GRAFT. Additionally, LLMs were used for writing assistance (grammar, spelling, word choice) in preparing this manuscript.

Guidelines:

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