READ–WRITE FACTORIZATION: PREFERENCE GEOMETRY FOR PERSONALIZED DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalizing large language models requires both *reading* a preference from context and *writing* it into a decision, yet activation steering typically ties these operations. We show that this tying fails along three geometric degrees of freedom: *where* to read and write, *how many* coordinates are needed, and *which directions* support readout versus causal control. Across two 7B architectures, extraction-by-application intervention surfaces are approximately separable, with rank 1 explaining 81–91% of variance in homogeneous-preference conditions. Heterogeneous team preferences are low-rank but not scalar (effective rank 3.14), with the dominant coordinate tracking preference severity ($\rho = 0.952$). Most strikingly, high-decoding read directions are nearly orthogonal to a gradient-defined write direction ($|\cos| \leq 0.026$), and their span captures only 0.07% of its squared norm: decodability does not determine causal control. A sign-symmetric decomposition further separates directional control from sign-invariant perturbation. Together, these results establish *read–write factorization*: preference representation and causal control need not share a single steering geometry. Behaviorally, the dissociation transfers to PrefEval's original implicit-preference classification task: write but not read steering improves accuracy ($34\% \rightarrow 51\%$). On a CL-Bench contextual-binding task, separately parameterizing representation and coupling raises binding accuracy from 46.7% to $90.5\%\pm1.5\%$, with ablations attributing most of the gain to coupling the context-dependent representation rather than to the learned direction alone.

1 INTRODUCTION

When the same language model serves different users or teams, the desired judgment can change with context. Personalization must therefore both *read* the relevant preference and *write* it into a decision. Activation-space methods commonly tie these operations by extracting and applying one direction, often at one layer and along a scalar preference axis. We test this tying assumption along three degrees of freedom—depth, rank, and causal direction—and find that it fails along each. We call this separation *read–write factorization*.

Reading and writing favor different depths. Across all primary conditions, extraction peaks deeper than application; crossed intervention confirms this functionally: late-to-mid retains the late–late effect (2.91 vs. 2.87), whereas mid-to-late falls to 0.37.

Preference geometry is low-rank, not scalar. Heterogeneous team directions have effective rank 3.14; their first three PCs explain 88.7% of variance, while PC1 tracks preference severity ($\rho = 0.952$, $p < 10^{-4}$). Preference geometry is therefore compact but not scalar.

Reading is not writing. Four read estimators perfectly decode a controlled contrast (5-fold CV AUC $= 1.0$), yet mean-difference, logistic, and LDA directions are nearly orthogonal to the local write direction ($|\cos| \leq 0.026$). Their four-dimensional span captures only 0.07% of its squared norm, comparable to a random subspace. Thus decodability does not determine causal control.

Directional versus perturbational effects. Raw intervention gains can mix sign-sensitive control with sign-invariant perturbation. We introduce a sign-symmetric decomposition, Δ_{odd} and Δ_{even}

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(§6.2), that separates the two. Some directions show large one-sided gains but negligible odd effects, so standard evaluations can overstate directional control. Together, a single steering axis conflates preference depth, rank, and causal direction.

Contributions.

1. **Depth factorization:** We show that steering depth splits into distinct extraction and application coordinates. Cross-layer intervention matrices are approximately separable across two 7B architectures (rank-1: 81–91% in homogeneous conditions), crossed interventions reveal functional asymmetry, which replicates significantly at 14B and 72B, with a same-sign unresolved estimate at 32B (§4).
2. **Low-rank, non-scalar preference geometry:** We show that heterogeneous team preferences occupy a compact but genuinely multidimensional space (effective rank 3.14), whose dominant coordinate tracks preference severity ($\rho = 0.952$; §5).
3. **Reading is not writing:** We show that high-decoding read directions can contain essentially none of the local causal write direction. Read directions are nearly orthogonal to a gradient-defined write direction ($|\cos| \leq 0.026$), and their span captures only 0.07% of its squared norm despite both geometries being individually stable. A sign-symmetric decomposition further distinguishes directional control from sign-invariant intervention effects (§6). This directional dissociation persists in targeted replications from 7B to 72B and across model families.
4. **Operational consequence:** On PrefEval's original implicit-preference classification task, read steering leaves accuracy unchanged at 34%, whereas write steering reaches 51%. On a contextual preference-binding task, separately parameterizing context representation and decision coupling raises BDA from 46.7% to $90.5\% \pm 1.5\%$, with ablations attributing most of the gain to coupling the context-dependent representation rather than to the learned direction alone (§8).

2 Problem Formulation

Consider a model f with L layers and hidden dimension d, and let $\mathbf{h}^{(\ell)} \in \mathbb{R}^d$ denote its hidden state at layer ℓ. We distinguish two operations on this representation: *reading* preference information from $\mathbf{h}^{(\ell)}$ and *writing* a perturbation that causally changes the downstream decision. We further separate the layer at which a direction is extracted from the layer at which it is applied.

Read direction. A read direction $\mathbf{d}^{(\ell)}_{\text{read}}$ captures variation in $\mathbf{h}^{(\ell)}$ that distinguishes preference conditions. Its quality is measured by held-out decodability (e.g., AUC).

Write direction. A write direction $\mathbf{d}^{(\ell)}_{\text{write}}$ is an intervention direction whose perturbation changes a downstream decision. Its causal efficacy is measured by the resulting change in a task-aligned decision score. In general, $\mathbf{d}^{(\ell)}_{\text{write}} \neq \mathbf{d}^{(\ell)}_{\text{read}}$.

Local first-order write direction. For a shared unit-norm intervention direction $\mathbf{d}$ and small strength γ, let G_n denote the task-aligned downstream decision score. A first-order expansion gives

$$\frac{1}{N}\sum_n [G_n(\mathbf{h}_n + \gamma\mathbf{d}) - G_n(\mathbf{h}_n)] = \gamma\,\bar{\mathbf{g}}^\top\mathbf{d} + O(\gamma^2), \qquad \bar{\mathbf{g}} = \frac{1}{N}\sum_n \nabla_{\mathbf{h}_n} G_n(\mathbf{h}_n).$$

Under $\|\mathbf{d}\|_2 = 1$, Cauchy–Schwarz gives the calibration-optimal first-order direction

$$\mathbf{d}^*_{\text{write}} = \frac{\bar{\mathbf{g}}}{\|\bar{\mathbf{g}}\|_2}.$$

We call this the *calibration write direction*: a label-dependent local oracle for causal leverage, not a deployment method. We evaluate its leverage using sign-symmetric slopes in a verified small-intervention regime; larger interventions are reported separately and need not preserve the local ordering.

Application sensitivity. Even for a fixed intervention direction, causal efficacy can vary with application depth. For extraction layer i and application layer j, let M_{ij} denote the signed behavioral

effect of extracting a preference direction at i and applying it at j. We test whether

$$M_{ij} \approx \mu + E_i A_j, \tag{1}$$

where E_i and A_j capture extraction- and application-side sensitivity, respectively. For heterogeneous teams, we ask whether preference directions $\{\mathbf{d}_t\}_{t=1}^{T}$ share a single axis, $\mathbf{d}_t = \gamma_t \mathbf{d}$, or instead occupy a low-dimensional subspace

$$\mathbf{d}_t \approx D\mathbf{z}_t, \qquad D \in \mathbb{R}^{d\times K}, \quad K \ll d. \tag{2}$$

Together these distinctions define the read–write parameterization

$$z = R^\top h_c, \qquad \Delta h = W\,\mathrm{diag}(a)z, \tag{3}$$

where $R, W \in \mathbb{R}^{d\times K}$ are separate read and write operators, h_c is a context representation, and a modulates how its coordinates affect the downstream state. Conventional single-direction steering implicitly imposes three tying assumptions:

$$\ell_R = \ell_W, \qquad K = 1, \qquad W = R.$$

Sections 4–6 test each independently.

3 Related Work

Activation steering and intervention geometry. Representation engineering Zou et al. (2023), activation addition Turner et al. (2023), and contrastive activation addition Rimsky et al. (2024) control LLM behavior through hidden-state directions, motivated by approximately linear representation structure Park et al. (2024). Extensions include rotation-based steering Vu and Nguyen (2025); Pham and Nguyen (2024), SAE-derived directions Wu et al. (2025), and geometry-aware interventions You et al. (2026). Layer-selection methods identify effective intervention sites Sun et al. (2025); Dang and Ngo (2026), whereas we separately vary the depth at which a preference direction is extracted and the depth at which it is causally applied.

Adaptive steering and causal control. Adaptive and gradient-informed methods relax fixed global steering. CLAS Hsu et al. (2026) learns context-dependent strengths for fixed directions, SADI Wang et al. (2025) makes interventions input-dependent, SVF Li et al. (2026a) derives local directions from gradients of a learned concept field, and GrAInS Nguyen et al. (2026) uses Integrated-Gradients attribution for output-aware intervention. Recent work also studies the mechanisms and limits of activation control Xu et al. (2026); Liang et al. (2026), including settings where decodable information is not behaviorally actionable. We characterize this gap geometrically: read and write directions can be individually stable yet nearly orthogonal, separating preference decodability from local causal sensitivity.

Personalization and preference geometry. Activation-space personalization represents user preferences, styles, or personas through steerable internal features Zhang et al. (2025); Zhao et al. (2025a); Cao et al. (2024); Choi et al. (2025); Tan et al. (2024); Bo et al. (2025); Zhang et al. (2025b). Complementary work studies preference elicitation and personalized alignment Shao et al. (2023); Park et al. (2024b); Wolfert (2026); Xie et al. (2025). We instead characterize the geometry of preference families themselves: heterogeneous preferences occupy a compact low-dimensional but non-scalar activation subspace, complementary to low-rank adaptation in weight space such as LoRA Hu et al. (2022).

4 Factorizing Steering Depth into Extraction and Application

Activation-steering interventions are often localized to particular network depths Turner et al. (2023); Zou et al. (2023). We ask whether a single "best layer" conflates two distinct choices: where a preference direction is extracted and where it is applied. We therefore measure the full extraction-layer $\times$ application-layer intervention surface. We study Qwen-2.5-7B-Instruct and Mistral-7B-Instruct-v0.3 on a controlled team-decision task in which the model makes a binary decision under a specified preference (e.g., "escalate to senior team"). Preference directions are extracted from constraint-injection contrast pairs (§e); we vary preference type and prompt surface form to test the stability of the resulting geometry.

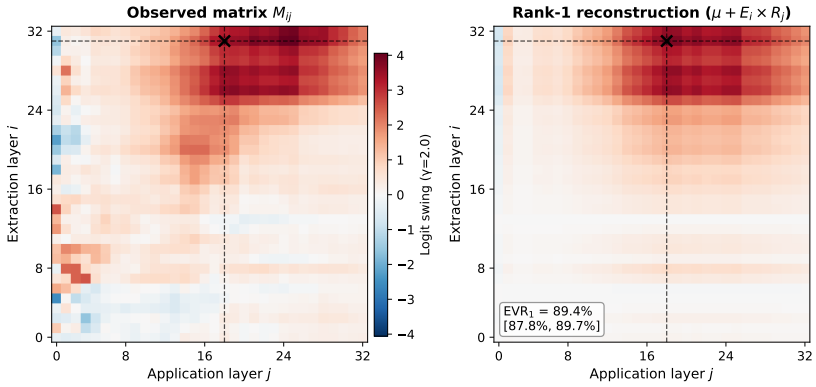


Figure 1: Extraction-by-application intervention matrix (Mistral-7B, 32×32; Qwen-7B sampled every second layer in Appendix a). **Left:** observed M_{ij}. **Right:** rank-1 reconstruction $\mu + E_i A_j$. Dashed lines mark the extraction peak (layer 30, 97%) and application peak (layer 17, 55%). Rank 1 explains 89.4% of variance on Mistral and 82.7% on Qwen.

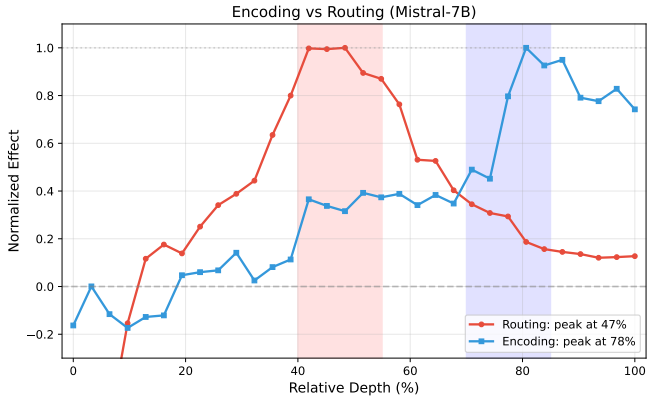


Figure 2: Extraction and application profiles on Mistral-7B. Application peaks at 47% of network depth, while extraction peaks later at 78%.

4.1 Full Extraction-by-Application Matrix

To test depth separability directly, we measure the extraction-layer $\times$ application-layer matrix M_{ij} (all 32 layers on Mistral-7B; every second layer on Qwen-7B). Rank 1 explains 89.4% of variance on Mistral (95% CI $[87.8, 89.7]$) and 82.7% on Qwen ($[81.0, 84.0]$), while rank 3 explains 98.1% and 95.9%, respectively (Figure 1). Thus the intervention surface approximately factorizes into distinct extraction- and application-side depth profiles. Separability weakens with preference heterogeneity. Rank-1 EVR is 90.5% on Mistral PrefEval and 68.7% on heterogeneous Qwen PrefEval, rising to 81.3% when Qwen is restricted to a single domain. Across all primary contrastive-direction conditions, extraction peaks deeper than application (Appendix a).

4.2 Extraction and Application Profiles

We next isolate the two depth factors using complementary profiles: the *extraction profile* varies extraction depth at a fixed application site, while the *application profile* varies application depth for a fixed reference direction (Figure 2).

Results. Application peaks at 47–48% depth in both 7B models, whereas extraction peaks later: 78% on Mistral and 96% on Qwen for the team task, and 61–81% on PrefEval. Across all primary contrastive-direction conditions, extraction peaks deeper than application. The gap ranges from 7–

Table 1: Layer-asymmetric steering on Mistral-7B (bootstrap 95% CIs). Crossing extraction and application depths produces strongly asymmetric intervention effects.

Configuration	Extract	Apply	Logit Swing (95% CI)
Mid–mid baseline	15	15	$0.89\ [0.78, 1.01]$
Late–late baseline	25	25	$2.87\ [2.83, 2.92]$
Extract-late, apply-mid	25	15	$2.91\ [2.82, 3.01]$
Extract-mid, apply-late	15	25	$0.37\ [0.34, 0.39]$

50 percentage points; on Qwen PrefEval it is only one grid cell and therefore unresolved at the sampled granularity.

Validation and scaling. Fixing application at $L = 15$ preserves Mistral's extraction peak at $L = 25$, and a coarse Qwen-2.5-14B sweep preserves the extraction-later-than-application ordering (98% vs. 69% depth; Appendix s).

4.3 Depth-Asymmetric Intervention

If extraction and application are distinct depth factors, crossing their preferred regions should produce asymmetric effects. We test this with the profile-derived late extraction reference ($L = 25$) and mid-depth application reference ($L = 15$) on Mistral-7B-Instruct-v0.3. These differ from the full-matrix SVD peaks ($L = 30$ and $L = 17$); we measure correct-token logit swing at $\gamma = 2.0$.

The crossed interventions are strongly asymmetric. A late-extracted direction retains its full effect when applied at the mid-depth site: $25 \to 15$ reaches 2.91, statistically indistinguishable from $25 \to 25$ at 2.87. In contrast, moving a mid-extracted direction late collapses the effect to 0.37. Thus late extraction transfers across application sites, whereas late application does not rescue a mid-extracted direction; changing extraction depth also changes the direction itself.

The crossed-depth asymmetry also replicates beyond 7B: paired late-to-mid versus mid-to-late contrasts exclude zero at 14B (D_{cross}=0.053, 95% CI $[0.032, 0.072]$) and 72B (0.051, $[0.034, 0.068]$); the 32B estimate has the same sign but is statistically unresolved (0.015, $[-0.008, 0.036]$). Full results appear in Appendix n.

5 Low-Rank Geometry of Team Preferences

Having separated *where* preferences are extracted and applied, we next ask *how many* coordinates represent heterogeneous preferences. Related directions are highly aligned within families (0.84–0.90 vs. 0.19 cross-family; Appendix o).

Team preferences are low-rank but not scalar. Team-specific directions exhibit substantial shared structure: mean pairwise cosine similarity is 0.671. PCA gives $\text{EVR}_1 = 68.2\%$, $\text{EVR}_{1:2} = 83.1\%$, and $\text{EVR}_{1:3} = 88.7\%$, with effective rank 3.14. The dominant coordinate is interpretable: PC1 strongly tracks preference severity (Spearman $\rho = 0.952$, $p < 10^{-4}$). Thus preference severity has a dominant scalar coordinate, while team variation spans a compact multidimensional subspace.

Behavioral validation. A two-axis representation of risk tolerance and explanation depth achieves 100% in-domain across three seeds (80/80 per seed). Trained only on loan decisions, it transfers zero-shot to hiring at $73.3\% \pm 4.7\%$, with errors concentrated in aggressive profiles (Appendix r). This partial transfer shows that multiple preference coordinates can carry behaviorally relevant structure beyond the training domain.

6 Read–Write Directional Asymmetry

Directions that best expose preferences need not be those that causally control decisions. For gradient-based comparisons, gradients and interventions are matched at the final-token state; conventional steering experiments retain their all-position protocol (Appendix v).

6.1 Read and Write Directions Occupy Distinct Geometry

Across 15 preference directions, read and write scores correlate only weakly ($r = 0.31$; Appendix t). Directly, four read estimators perfectly decode a controlled contrast (5-fold CV AUC $= 1.0$), yet their directions are nearly orthogonal to the calibration write direction ($|\cos| \leq 0.037$). Under matched comparison, mean-difference, logistic, and LDA remain nearly orthogonal (all $|\cos| \leq 0.026$). Further, the span of four high-decoding read estimators captures only 0.07% of the write direction's squared norm, consistent with random four-dimensional subspace overlap. This is not an instability artifact: read and write directions remain individually stable under calibration resampling (0.999 and 0.983 cosine) while remaining nearly orthogonal. Thus decodability does not determine causal control: the measured read subspace contains essentially none of the local write direction.

Read–write separation persists across scale and model family. Across Qwen models from 7B to 72B and targeted Llama-3.1-8B and Gemma-2-9B replications, the LDA read direction remains weakly aligned with the local write direction ($|\cos| \leq 0.042$); detailed subspace and stability results are reported in Appendix n.

6.2 Separating Directional Control from Perturbation

One-sided steering gains can mix sign-sensitive directional control with sign-invariant perturbation. We therefore decompose

$$\Delta_{\text{odd}}(\mathbf{d}) = \frac{G(+\mathbf{d}) - G(-\mathbf{d})}{2}, \qquad \Delta_{\text{even}}(\mathbf{d}) = \frac{G(+\mathbf{d}) + G(-\mathbf{d})}{2} - G(0).$$

The odd component changes sign under $\mathbf{d} \mapsto -\mathbf{d}$, whereas the even component captures sign-invariant perturbation. For example, PC8 shows a large one-sided improvement but $\Delta_{\text{odd}} \approx 0.002$ and $\Delta_{\text{even}} \approx 0.131$, indicating that almost all of its apparent benefit is sign-invariant. We therefore use odd effects when assessing directional causal control.

6.3 Causal Consequences and Robustness

In the primary setting, geometric separation corresponds to a causal difference: the calibration write direction has substantially greater local signed leverage than the mean-difference read direction (Appendix v). We estimate this effect using sign-symmetric slopes in a verified small-γ regime, consistent with the first-order construction in Section 2; larger interventions enter a nonlinear regime.

The separation is not explained by downstream supervision alone. A margin-regression direction trained on the same continuous target remains nearly orthogonal to the calibration write direction ($\cos = 0.026$) and has substantially smaller finite-intervention effect (Appendix d). The result also replicates on PrefEval: read-direction odd-effect confidence intervals include zero, whereas the gradient direction yields $0.042\,[0.022, 0.060]$ and remains nearly orthogonal to the read directions ($|\cos| < 0.04$).

6.4 Behavioral Validation on PrefEval (Original Implicit Regime)

We test whether directional dissociation predicts behavior on an external benchmark. We use PrefEval's original 4-way classification task under its implicit choice-based preference condition, where the relevant preference must be inferred from conversation history. We score the original candidate options by sequence log-probability and retain the benchmark's classification-accuracy metric. This evaluation is distinct from Section 8.1, where preference-relevant information is already available in the decision input rather than inferred from conversation history.

The unsteered model achieves 34.0% accuracy (N=200). Steering along the read direction selects γ=0 on validation and remains at 34.0%, whereas the write direction estimated from 50 labeled calibration examples reaches 51.0% ($+17$ points). The two directions are nearly orthogonal ($|\cos(R, W)|$=0.007). Thus read–write dissociation translates into a behavioral difference on an established external preference-following benchmark under its original classification task format.

The magnitude of the causal advantage is nevertheless distribution-sensitive. Across calibration resamples, the mean paired write-minus-read slope is $+0.27$, but its confidence interval includes

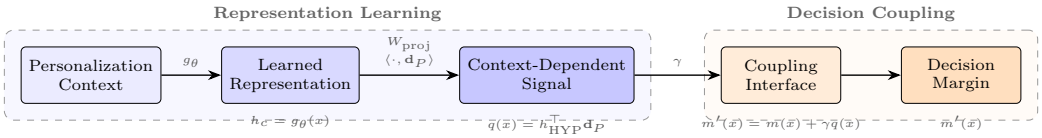


Figure 3: Operational realization of factorized context-to-decision coupling. A learned context representation h_c produces a context-dependent preference signal $q(x)$ that is coupled separately to the frozen task model's decision margin. Representation learning and downstream coupling are trained separately; implementation details appear in Appendix j.

zero (Appendix v). We therefore treat stronger write leverage as a primary-setting consequence of read–write separation, not as a universal superiority claim.

6.5 Propagation of Write Perturbations

Through late layers, gradient-defined perturbations develop substantially more decision-readout alignment than read-derived perturbations (Appendix u).

Extension to autoregressive generation. Structured decisions admit a localized write direction, but free-form generation unfolds over a trajectory of intermediate states. A targeted diagnostic examines local write directions W_t at positions $t \in \{1, 4, 8, 16\}$ along generation trajectories. The measured cross-position alignments are near zero ($|\cos| \leq 0.048$), indicating that autoregressive generation does not expose a single stationary write vector (Appendix x).

7 From Geometric Factorization to Personalized Decisions

The geometric results motivate a corresponding operational question: when does separating preference representation from downstream coupling matter behaviorally? If the relevant preference is directly available, standard steering may be sufficient; when it must first be recovered from context, representation and expression become distinct problems. We test this regime contrast by separately parameterizing context representation and downstream decision coupling.

For structured decisions, we instantiate this separation as

$$m'(x) = \mathcal{C}\big(m(x), q(x); \theta_c\big), \tag{4}$$

where $m(x)$ is the frozen task model's context-conditioned decision margin, $q(x)$ is a learned context-dependent preference signal, and $\mathcal{C}$ is a separately parameterized coupling interface. In our implementation, $q(x) = h_{\text{HYP}}^\top \mathbf{d}_P$ and $m'(x) = m(x) + \gamma q(x)$. The compressed representation h_{HYP} is prepended to the task model at position 0, defining the context-conditioned margin $m(x)$, and is separately projected onto $\mathbf{d}_P$ to produce $q(x)$. Eq. 4 operationalizes the broader separation between representation and downstream coupling; it is not a direct implementation or causal test of the linear R, W parameterization in Eq. 3.

We train the representation mechanism separately from the coupling interface and evaluate the resulting factorization on implicit context-to-decision binding. The representation learner is trained once and frozen before fitting the low-dimensional downstream coupling; our comparison therefore tests factorization rather than parameter efficiency.

8 Experiments

8.1 Preference Information Available at Decision Time: PrefEval

We evaluate on the implicit choice subset of PrefEval Zhao et al. (2025c), where each example provides preference-aligned and preference-violating candidate responses. Following the benchmark protocol, we score options by sequence log-probability.

Table 2: PrefEval implicit choice accuracy ($N = 1000$). Standard steering improves both models, while factorized coupling matches standard steering on Qwen.

Method	Mistral-7B	Qwen-7B
Baseline (no steering)	90.11	68.44
Basic steering ($+\mathbf{d}_P$)	**100.00**	**73.67**
Causal controls:		
Flip ($-\mathbf{d}_P$)	9.22	61.67
Random direction	92.33	69.00
Read–write factorization:		
Factorized context coupling	—	73.33

Setup. We evaluate Mistral-7B and Qwen-7B on $N = 1000$ test examples. Preference directions use 100 held-out calibration examples; layer and γ are selected on a separate 15-example validation split and then fixed. We compare standard activation addition with sign-flip and random-direction controls, and on Qwen with the factorized coupling of Section 7.

Results. Basic steering improves Mistral from 90.11% to 100% and Qwen from 68.44% to 73.67%. Sign reversal sharply reduces performance, while random directions remain near baseline, supporting a direction-sensitive rather than generic perturbational effect. On Qwen, a direction extracted from a paraphrased template performs similarly to the in-template direction (73.89% vs. 73.67%; cosine 0.91; Appendix e).

Factorized context coupling reaches 73.33% on Qwen, essentially matching basic steering (73.67%). Thus factorization provides no measurable advantage in this setting, where preference information is already directly available to the model. We next test the harder case in which the relevant preference must first be recovered from context.

8.2 Contextual Preference Binding

We use the CL-Bench setup Dou et al. (2026) to construct a controlled preference-binding task in which the model must bind a preference expressed in context to a downstream decision. Unlike PrefEval, the relevant preference must be recovered from accumulated context rather than supplied directly. We construct an A/B binding-discrimination task over $1{,}899$ examples from four CL-Bench categories. We measure Binding Discrimination Accuracy (BDA),

$$\mathrm{BDA} = \frac{1}{N} \sum_{n} \mathbf{1}[m_n^{\mathrm{full}} > m_n^{\mathrm{trunc}}],$$

where m is the correct-choice logit margin. BDA tests whether accumulated context is preferentially coupled to the downstream decision; it is a diagnostic rather than a measure of open-ended CL-Bench performance. Randomized labels and neutral framing reduce shortcut cues (Appendix k).

Results. Factorized coupling reaches $90.5\% \pm 1.5\%$ BDA, versus 46.7% at $\gamma = 0$. Standard additive steering (48.8%), Contrastive Activation Addition (CAA; 51.1%), and Semantics-Adaptive Directional Intervention (SADI; 50.8%) all remain near baseline. This comparison does not attribute the full gain specifically to directional untying, since the factorized system also includes an explicit coupling interface (Appendix h). The representation learner never sees decision labels; it is trained only to separate full- from truncated-context projections, an objective closely aligned with BDA itself. The steering baselines receive no such optimization, so the gap reflects coupling a representation optimized for binding, rather than label information or a like-for-like comparison among steering methods.

Causal controls. The effect depends on both the context representation and its coupling orientation: reversing the learned direction reduces BDA to 0%, swapping context to 0.1%, and replacing the context representation with a constant produces only ties (0% strict BDA). A random direction with the inherited coupling reaches $82.5\% \pm 3.2\%$, showing that much of the improvement is carried

Table 3: CL-Bench binding discrimination on the held-out test set ($N = 399$). Factorized coupling substantially improves BDA; causal ablations test dependence on context signal and its orientation.

Method	BDA (%)
Baseline ($\gamma = 0$)	46.7
Factorized coupling	$\mathbf{90.5 \pm 1.5}$
Steering baselines:	
Standard additive steering	48.8
CAA Rimsky et al. (2024)[†]	51.1
SADI Wang et al. (2025)	50.8
Causal ablations:	
Random direction (inherited coupling)	82.5 ± 3.2
Direction flip	0.0
Context swap	0.1
Constant context representation	0.0

[†]CAA selects γ=0 on validation; nonzero strengths yield 44–46%.

by the context-dependent representation pathway and explicit coupling interface. The learned direction adds BDA in this intervention (90.5% vs. 82.5%) and, more clearly, cross-category stability: leave-one-category-out performance is $91.0\% \pm 2.8\%$ (range 88.2–94.5%), versus 14.5–80.9% for random inherited directions. Thus this experiment demonstrates the behavioral relevance of separating representation from coupling, but does not identify the full gain with $W \neq R$ alone. Additional ablations appear in Appendix h.

9 Discussion and Conclusion

Our results identify a systematic failure of the tying assumption underlying standard activation-space personalization. The assumption fails along all three geometric degrees of freedom we test: extraction and application separate across depth, heterogeneous preferences are low-rank but not scalar, and, most strikingly, high-decoding read directions span essentially none of the local causal write direction. Read–write factorization is therefore an empirical property of the measured geometry: *decodability does not determine causal control*. Sign-symmetric evaluation further shows why one-sided steering success is insufficient evidence of directional control.

Scope and limitations. Full cross-layer surfaces are measured on two 7B architectures, while targeted replications test the central geometric findings across Qwen models from 14B to 72B and on Llama-3.1-8B and Gemma-2-9B (Appendix n). Read–write directional separation persists across all tested models; crossed-depth asymmetry is significant at 14B and 72B and has the same sign but is unresolved at 32B. Preference geometry also shows some template sensitivity, although cross-template extraction preserves PrefEval performance.

Our causal claims are local. The calibration write direction is first-order optimal only for its calibration objective, and larger interventions can enter nonlinear regimes. Although the read and write directions are geometrically stable across calibration resamples, the magnitude of the write-over-read leverage gap varies. We therefore treat stronger write leverage as a primary-setting consequence of read–write separation, not a universal property. Likewise, geometric dissociation does not imply that untied parameterizations universally outperform tied ones. When both saturate a downstream task, their performance can coincide (Appendix w).

Our trajectory diagnostic suggests that extending from discrete decisions to free-form generation may require trajectory-dependent rather than stationary write operators (Appendix x).

More broadly, personalization need not reduce to finding a better steering vector. The representation that exposes a preference, the coordinates that organize preference variation, and the intervention that expresses those preferences can obey different geometries. Read–write factorization turns these from a single tied design choice into separate measurable objects. Our results therefore motivate treating where to read, how many coordinates to retain, and how to write them into behavior as separate empirical questions, rather than assuming that one direction can serve all three roles.

ETHICS STATEMENT

The loan and hiring scenarios used in this work are fully synthetic and contain no real applicant data or protected-attribute proxies. Team profiles represent organizational policy parameters rather than demographic groups or user identities. This work studies the mechanics of activation steering and does not advocate its deployment in consequential decision making; such applications would require additional fairness, safety, and robustness evaluation.

REPRODUCIBILITY STATEMENT

Code for reproducing and verifying the main experimental analyses is provided in the supplementary material. The appendix documents the models, data splits, calibration and intervention protocols, statistical procedures, and implementation details needed to reproduce the paper's main experiments.

AI USE STATEMENT

Generative AI tools were used to provide feedback on research methodology and experimental design, generate synthetic decision scenarios and team profile descriptions for Section 5, assist with sanity-checking the interpretation of experimental results, check the first-order derivation in Section 2 and the projection identity in Appendix v for correctness, refine presentation and prose, generate simulated peer reviews used to identify weaknesses before submission, organize manuscript text, assist with code implementation, draft and edit portions of the appendix based on experiment logs. We did not use generative AI tools to develop the conceptual framework, generate original claims, propose hypotheses, translate, or clean datasets. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

REFERENCES

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

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

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition. In *ACL*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, et al. CL-Bench: A benchmark for context learning. *arXiv preprint arXiv:2602.03587*, 2026.

Xiaoyan Zhao, Ming Yan, Yilun Qiu, Haoting Ni, Yang Zhang, Fuli Feng, Hong Cheng, and Tat-Seng Chua. SteerX: Disentangled Steering for LLM Personalization. *arXiv preprint arXiv:2510.22256*, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *ICML*, 2024.

Hieu M. Vu and Tan Minh Nguyen. Angular Steering: Behavior Control via Rotation in Activation Space. In *NeurIPS*, 2025.

Van-Cuong Pham and Thien Huu Nguyen. Householder Pseudo-Rotation: A Novel Approach to Activation Editing in LLMs with Direction-Magnitude Perspective. In *EMNLP*, 2024.

Ceyhun Efe Kayan and Li Zhang. Prototype-Based Dynamic Steering for Large Language Models. *arXiv preprint arXiv:2510.05498*, 2025.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. In *ICML*, 2025.

Yijing Zhang, Dyah Adila, Changho Shin, and Frederic Sala. Personalize Your LLM: Fake it then Align it. In *Findings of NAACL*, 2025.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing Large Language Models via Personalized Parameter-Efficient Fine-tuning. In *EMNLP*, 2024.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization. In *NeurIPS*, 2024.

Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering Vector Fields for Context-Aware Inference-Time Control in Large Language Models. *arXiv preprint arXiv:2602.01654*, 2026.

Brandon Hsu, Daniel Beaglehole, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Contextual Linear Activation Steering of Language Models. *arXiv preprint arXiv:2604.24693*, 2026.

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical Steering: Geometry-Aware Activation Rotation for Language Models. In *ICML*, 2026.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs. In *ICLR*, 2025.

Youngbin Choi, Seunghyuk Cho, Minjong Lee, MoonJeong Park, Yesong Ko, Jungseul Ok, and Dongwoo Kim. CoPL: Collaborative Preference Learning for Personalizing LLMs. In *EMNLP*, 2025.

Han Shao, Lee Cohen, Avrim Blum, Yishay Mansour, Aadirupa Saha, and Matthew R. Walter. Eliciting User Preferences for Personalized Multi-Objective Decision Making through Comparative Feedback. *arXiv preprint arXiv:2302.03805*, 2023.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. RLHF from Heterogeneous Feedback via Personalization and Preference Aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

A. R. M. Wolfert. Preference-Based Optimisation for Integrated Design and Group Decision-Making. *arXiv preprint arXiv:2603.19050*, 2026.

Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, et al. A Survey on Personalized and Pluralistic Preference Alignment in Large Language Models. *arXiv preprint arXiv:2504.07070*, 2025.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-Adaptive Activation Intervention for LLMs via Dynamic Steering Vectors. In *ICLR*, 2025.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs. In *ACL*, 2026.

Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. LayerNavigator: Finding Promising Intervention Layers for Efficient Activation Steering in Large Language Models. In *NeurIPS*, 2025.

Quy-Anh Dang and Chris Ngo. Selective Steering: Norm-Preserving Control Through Discriminative Layer Selection. In *Findings of ACL*, 2026.

Ziwen Xu, Chenyan Wu, Hengyu Sun, Haiwen Hong, Mengru Wang, Yunzhi Yao, Longtao Huang, Hui Xue, Shumin Deng, Zhixuan Chu, Huajun Chen, and Ningyu Zhang. Why Steering Works: Toward a Unified View of Language Model Parameter Dynamics. In *ACL*, 2026.

Man Liang, Xinzhao Cheng, and Faizan Wajid. Encoded but Not Actionable: Auditing the Decode-Generate-Steer Gap in Frozen LLMs for Geometric Constraints. *arXiv preprint arXiv:2608.17843*, 2026.

Jessica Y. Bo, Tianyu Xu, Ishan Chatterjee, Katrina Passarella-Ward, Achin Kulshrestha, and D. Shin. Steerable Chatbots: Personalizing LLMs with Preference-Based Activation Steering. *arXiv preprint arXiv:2505.04260*, 2025.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. Personalized Text Generation with Contrastive Activation Steering. In *ACL*, 2025.

A Full Cross-Layer Factorization

Protocol. We measure the matrix on Mistral-7B-Instruct-v0.3 (protocol below) and replicate on Qwen-2.5-7B-Instruct. For Mistral, we evaluate all 32 layers; for Qwen, we sample 15 layers (every second layer plus the final). We use $N = 100$ evaluation examples and intervention strength $\gamma = 2.0$. For team-task cells, the matrix is constructed on a fixed controlled team-decision preference family (binary escalate/review); calibration and evaluation examples vary business scenarios while preserving the preference semantics. PrefEval cells use PrefEval calibration pairs (see Cross-task analysis below). For each extraction layer i and application layer j, we extract a mean-difference preference direction from calibration contrasts at layer i, normalize it, and intervene at layer j. Team-task cells use 30 calibration pairs; PrefEval cells use 100 (single-domain: 50, drawn from the 221-example subset). Each entry M_{ij} is the mean signed change in the task-aligned decision margin (logit difference for the correct vs. incorrect choice) relative to the no-intervention baseline. Within each task, all cells use the same calibration and evaluation examples so that variation across M_{ij} reflects only the extraction/application layer pair.

Low-rank structure. Let $\widetilde{M} = M - \bar{M}$ denote the mean-centered intervention matrix. Its singular-value decomposition gives cumulative explained-variance ratios:

$$\mathrm{EVR}_1 = 89.4\%, \qquad \mathrm{EVR}_{1:2} = 94.5\%, \qquad \mathrm{EVR}_{1:3} = 98.1\%.$$

The dominant left and right singular factors peak at extraction layer 30 (97% depth) and application layer 17 (55%), respectively. Thus the empirical matrix is well approximated by

$$M_{ij} \approx \mu + E_i A_j,$$

where μ is the global mean, E captures extraction strength, and A captures application sensitivity. The rank-1 approximation explains nearly 90% of variance, supporting an approximately separable extraction-by-application surface with distinct depth profiles.

Bootstrap confidence interval. We bootstrap evaluation examples with replacement for 1,000 replicates, recompute every matrix entry, center the resulting matrix, and recompute the singular spectrum for each replicate. The observed matrix has $\mathrm{EVR}_1 = 89.4\%$; the bootstrap distribution has mean 88.9% and 95% percentile interval $[87.8\%, 89.7\%]$. Extraction peak layer 30 appears in 100% of bootstrap samples; the modal application peak is layer 17, occurring in 98% of samples, confirming stability to evaluation resampling of the rank-1 EVR and the extraction–application peak separation.

Qwen-2.5-7B replication. We replicate the layer matrix on Qwen-2.5-7B-Instruct with 15×15 sampled layers (every second layer from 0 to 26, plus 27), $N = 100$ evaluation examples, and $\gamma = 2.0$. The low-rank structure replicates:

$$\mathrm{EVR}_1 = 82.7\%, \qquad \mathrm{EVR}_{1:2} = 92.8\%, \qquad \mathrm{EVR}_{1:3} = 95.9\%.$$

Because Qwen uses a coarser layer grid, absolute EVR values should not be interpreted as a pure architectural comparison. The important replication is qualitative: the matrix remains approximately low-rank and extraction peaks later than application. The extraction factor peaks at layer 27 (100% depth) and the application factor at layer 22 (81%). Bootstrap 95% CI for EVR_1: $[81.0\%, 84.0\%]$; extraction peak stability: 100%; application peak stability: 100%. Figure 4 shows four additional conditions; Mistral single-domain appears only in Table 4. The qualitative structure persists across conditions.

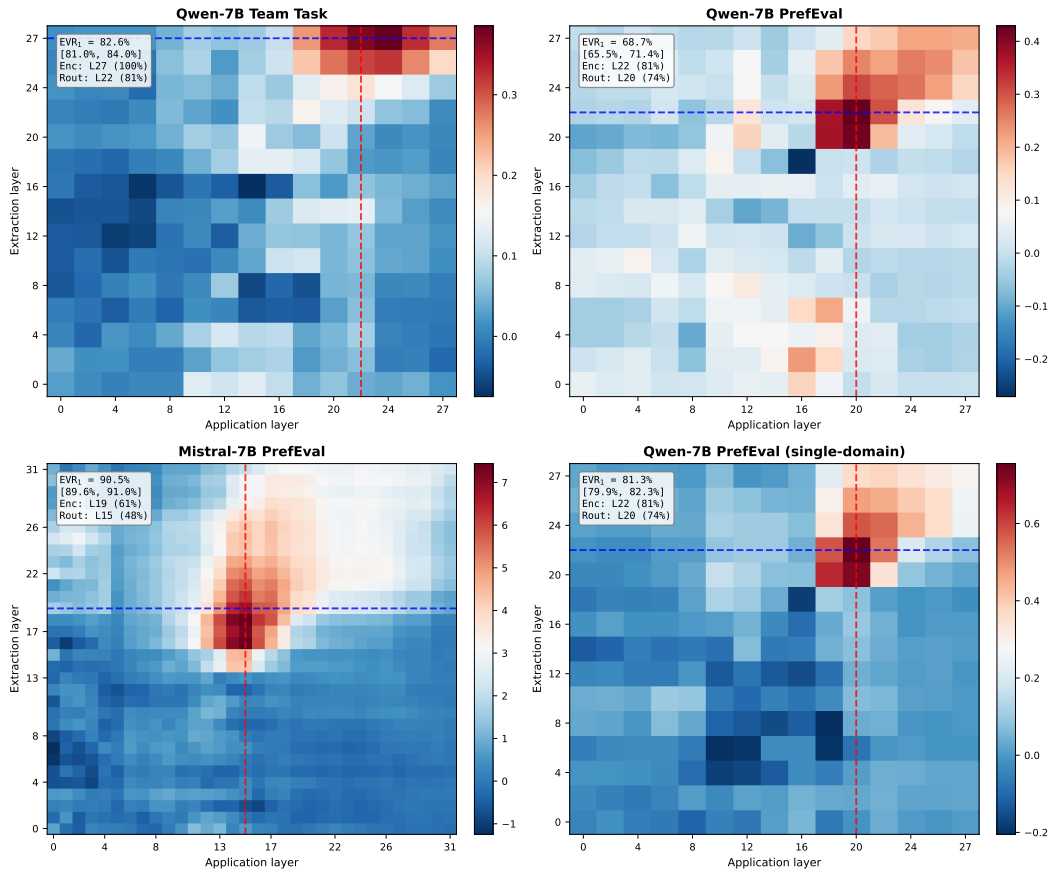


Figure 4: Extraction-by-application matrices across four conditions. Each panel shows the intervention matrix M_{ij} with EVR_1, bootstrap CI, and peak locations. **Top left:** Qwen-7B team task (15×15). **Top right:** Qwen-7B PrefEval (15×15); weaker separability (68.7%) due to preference heterogeneity. **Bottom left:** Mistral-7B PrefEval (32×32); robust separability (90.5%). **Bottom right:** Qwen-7B PrefEval single-domain (15×15); separability recovers (81.3%) with homogeneity. Dashed lines mark extraction (blue, horizontal) and application (red, vertical) peaks. The ordering $\ell_{\text{ext}} > \ell_{\text{app}}$ holds in all panels. Color scales differ per panel; note Qwen's $\sim 10\times$ smaller effect magnitudes. Mistral single-domain (88.4%) appears in Table 4 only.

Extraction-depth-dependent application. The matrix reveals that the application peak shifts with extraction depth: mid-layer extracted directions apply mid, while late-layer directions apply later. On Qwen, row 18 (extraction at 67% depth) peaks at layer 14 (52%), while rows 24–27 (89–100%) peak at layers 22–24 (81–89%). On Mistral, late-layer references ($L \geq 25$) apply later than mid-layer references ($L \in [10, 20]$); see Appendix s. This extraction-depth–application interaction is captured by the second singular component, explaining why Qwen's rank-1 EVR is lower (82.7% vs. 89.4%): the interaction is stronger on Qwen. The consistent finding is the ordering $\ell_{\text{ext}} > \ell_{\text{app}}$ on both architectures.

Cross-task analysis: PrefEval. We replicate the layer matrix on PrefEval Zhao et al. (2025c). Table 4 summarizes all six cells.

Mistral-7B (32×32): The factorization is robust: $\text{EVR}_1 = 90.5\%$ (CI $[89.6\%, 91.0\%]$), comparable to the team task (89.4%). Extraction peak: layer 19 (61% depth); application peak: layer 15 (48% depth), both with 100% bootstrap stability.

Qwen-2.5-7B (15×15): The factorization weakens: $\text{EVR}_1 = 68.7\%$ (CI $[65.5\%, 71.4\%]$), $\text{EVR}_{1:3} = 93.3\%$. Extraction peak: layer 22 (81%); application peak: layer 20 (74%).

Table 4: Summary of layer matrix experiments across six cells (model × task × domain). All use $\gamma = 2.0$. "ent" = entertainment single-domain subset ($n = 221$). Extract./Apply. = extraction/application peak as % of network depth. Extraction peaks have 100% bootstrap stability; application peaks are also highly stable (98–100%).

Model	Task	Grid	N	EVR_1 [CI]	$\text{EVR}_{1:3}$	Extract.	Apply.
Mistral	Team	32^2	100	89.4 [87.8, 89.7]	98.1	97%	55%
Mistral	PrefEval	32^2	100	90.5 [89.6, 91.0]	98.8	61%	48%
Mistral	PrefEval-ent	32^2	100	88.4 [85.8, 89.9]	98.0	68%	48%
Qwen	Team	15^2	100	82.7 [81.0, 84.0]	95.9	100%	81%
Qwen	PrefEval	15^2	100	68.7 [65.5, 71.4]	93.3	81%	74%
Qwen	PrefEval-ent	15^2	100	81.3 [79.9, 82.3]	95.3	81%	74%

Single-domain analysis: Restricting Qwen to a single preference domain (entertainment, $n = 221$ examples in the subset, from which 50 calibration and 100 evaluation examples are drawn; same 15×15 grid, $N = 100$, $\gamma = 2.0$ as the Qwen team-task run) recovers separability: $\text{EVR}_1 = 81.3\%$ (CI $[79.9\%, 82.3\%]$), $+12.6$pp from the full PrefEval heterogeneous set. Mistral on the same single-domain subset (32×32) yields $\text{EVR}_1 = 88.4\%$ (CI $[85.8\%, 89.9\%]$), within 2pp of its full-PrefEval value (CIs overlap), consistent with a stronger heterogeneity effect on Qwen in this setting. The ordering $\ell_{\text{ext}} > \ell_{\text{app}}$ persists across all six cells.

B Multi-Team Low-Rank Geometry

Scalar probe attempt. We attempted a behavioral rank-1 probe: fitting one signed scalar $\gamma_t \in [-2, 2]$ per team along a fixed shared direction (escalate/review mean-difference from 30 calibration pairs, layer 22). Per-team accuracy remained at chance (46–48%) across the entire grid, so fitted values collapsed to the boundary and carried no information about severity. This probe is therefore uninformative about the rank-1 hypothesis in either direction; the geometric analysis below is the relevant evidence.

Direction geometry. We independently extract one normalized direction for each of the $T = 20$ teams and stack the directions into a $T \times d$ matrix. Mean pairwise cosine is 0.671. SVD of the centered direction matrix gives:

$$\text{EVR}_1 = 68.2\%, \qquad \text{EVR}_{1:2} = 83.1\%, \qquad \text{EVR}_{1:3} = 88.7\%,$$

with effective rank $r_{\text{eff}} = 3.14$, computed as the entropy-based effective rank:

$$r_{\text{eff}} = \exp\left(-\sum_k p_k \log p_k\right), \qquad p_k = \frac{\sigma_k^2}{\sum_j \sigma_j^2}.$$

The first principal coordinate is strongly monotonic with the predefined severity factor (Spearman $\rho = 0.952$, $p < 10^{-4}$).

Severity annotation protocol. Severity is not post-hoc annotated. It is the predefined ordinal conservatism parameter used to construct the team profiles (1=conservative, 2=cautious, 3=neutral, 4=permissive, 5=aggressive). The Spearman correlation therefore measures how well the unsupervised PC1 recovers the known ordinal generative factor. Because each severity level occurs in four domain contexts, severity ranks contain ties; we use the standard tie-corrected Spearman correlation. Statistical significance is computed via permutation test with 10,000 shuffles of the severity labels; $\rho = 0.952$ exceeds all null values, giving $p < 10^{-4}$.

Neighboring semantic profiles are also geometrically close: conservative/cautious have cosine 0.930, while permissive/aggressive have cosine 0.961.

C Sign-Symmetric Read–Write Diagnostics

Protocol. All sign-symmetric diagnostics use the controlled team-decision preference task of §4 (binary escalate/review). Both gradient computation and intervention are restricted to the final-token

state in these read–write diagnostics. Measurements are on Qwen-2.5-7B-Instruct at intervention layer $\ell = 15$ (47% depth, near the application peak). We use $N = 50$ evaluation examples disjoint from the 30-example calibration set used for direction extraction. Intervention strength is $\gamma = 0.1$. The task-aligned score G is the logit margin for the preference-aligned choice minus the alternative. Supervised read directions (mean-diff, logistic, LDA, Mahalanobis) are extracted from calibration contrasts; the 15 random directions are sampled uniformly on the unit sphere and evaluated under matched positive/negative interventions. The gradient direction is computed via backpropagation of the task-aligned score G (logit margin for the preference-aligned choice) with respect to the final-token hidden state at baseline (no intervention). Per-example gradients are computed on the same 30-example calibration set used for read-direction extraction, then averaged and normalized to unit length. This construction uses preference labels to define the target logit margin. Statistical test: two-sided t-test against zero for the random-direction odd effects ($N_d = 15$).

For each direction $\mathbf{d}$, we evaluate matched interventions at $+\mathbf{d}$ and $-\mathbf{d}$ and decompose the task-aligned score as:

$$\Delta_{\text{odd}}(\mathbf{d}) = \frac{G(+\mathbf{d}) - G(-\mathbf{d})}{2}, \qquad \Delta_{\text{even}}(\mathbf{d}) = \frac{G(+\mathbf{d}) + G(-\mathbf{d})}{2} - G(0).$$

The odd term changes sign under $\mathbf{d} \to -\mathbf{d}$ and measures orientation-sensitive intervention; the even term is sign-invariant and measures the sign-invariant finite-intervention response, which may include generic perturbational or nonlinear effects.

Table 5: Sign-symmetric read–write diagnostics at $\gamma = 0.1$ with bootstrap 95% CIs (matched protocol: same N and γ for all rows). Read directions perfectly decode the calibration contrast but have uniformly weak causal effects; the gradient direction has a $\sim 14\times$ larger finite-γ Δ_{odd} than the strongest read direction while remaining nearly orthogonal to all read directions ($|\cos| \leq 0.037$). Local first-order slopes are analyzed separately in Appendix v.

Direction	Read AUC*	Δ_{odd}	95% CI
Mean-diff	1.0	$+0.0045$	$[+0.0009, +0.0082]$
Logistic	1.0	$+0.0110$	$[+0.0085, +0.0137]$
LDA	1.0	$+0.0001$	$[-0.0029, +0.0032]$
Mahalanobis	1.0	$+0.0097$	$[+0.0061, +0.0134]$
Gradient (write)	—	$+0.159$	$[+0.135, +0.183]$
Random (null)	—	$+0.0006$	$[-0.0023, +0.0036]$

*5-fold cross-validated AUC on calibration pairs, grouped by pair.

Semantic orientation sanity check. The sign of every supervised read direction is fixed independently using the calibration preference labels: positive-condition examples must have higher mean projection than negative-condition examples. This ensures a consistent semantic convention across estimators.

Random null. For $N_d = 15$ random directions under the same N and γ, $\mathbb{E}[\Delta_{\text{odd}}] = +0.0006 \pm 0.0058$ ($p = 0.69$ vs. zero); the random-direction mean is not significantly different from zero. The key observation is not sign uniformity but the lack of correspondence between read quality and causal efficacy: estimators with identical AUC can differ by two orders of magnitude in write effect.

PrefEval replication. We replicate the sign-symmetric protocol on PrefEval (Qwen-2.5-7B, layer 19 at 71% depth—the layer validated for Qwen PrefEval steering in §8.1—with $\gamma = 0.1$; Δ_{odd} is approximately proportional over the tested range $\gamma \in \{0.05, 0.1, 0.2\}$; $N_{\text{eval}} = 100$, $N_{\text{calib}} = 100$ pairs, 15 random directions, 1,000 bootstrap replicates). Mean-difference steering also improves PrefEval accuracy by $+5.23$pp at its separately validated intervention strength (§8.1).

The dissociation holds even when read directions achieve AUC as high as 0.90: all read-direction CIs include zero while $\mathbf{d}_{\text{grad}}$ excludes zero. Thus successful large-γ steering on PrefEval does not imply strong per-unit causal leverage of the read direction.

Table 6: Sign-symmetric diagnostics on PrefEval (Qwen-2.5-7B, layer 19, $\gamma = 0.1$). Read directions have non-saturated AUC (0.78–0.90) yet all Δ_{odd} CIs include zero; $\mathbf{d}_{\text{grad}}$ is orthogonal to all read directions ($|\cos| < 0.04$) and its CI excludes zero.

Direction	CV AUC	Δ_{odd}	95% CI	$\cos(\mathbf{d}_{\text{grad}})$
Mean-diff	0.78	$+0.0021$	$[-0.005, +0.009]$	$+0.026$
Logistic	0.90	$+0.0004$	$[-0.002, +0.003]$	$+0.031$
LDA	0.82	$+0.0005$	$[-0.002, +0.003]$	$+0.011$
Mahalanobis	0.90	-0.0002	$[-0.002, +0.002]$	$+0.019$
Gradient	—	$+0.042$	$[+0.022, +0.060]$	—
Random (15)	—	-0.0001	$[-0.0009, +0.0007]$	—

Random: $p = 0.81$ vs. zero ($N_d = 15$).

D Perturbation and Gradient Diagnostics

Note: The PCA-spectrum diagnostic and supervised read-direction diagnostic (§c) use different evaluation protocols and normalization. Absolute effect magnitudes should not be directly compared across the two analyses; the qualitative pattern (odd vs. even dominance) is the relevant finding.

Principal-component sign decomposition. We evaluate the first 20 activation PCs under matched positive and negative interventions. Some PCs exhibit substantial odd effects, whereas others are dominated by the even component. For example, PC8 produces $\Delta_{\text{odd}} \approx 0.002$ and $\Delta_{\text{even}} \approx 0.131$, indicating that its large one-sided gain is almost entirely invariant to intervention sign. In contrast, PC1, PC4, and PC7 exhibit sizeable odd components, demonstrating that orientation-sensitive and sign-invariant finite-intervention effects coexist.

Gradient-based write direction. At intervention layer ℓ, we define the local decision-gradient direction:

$$\mathbf{d}_{\text{grad}} = \frac{\mathbb{E}_n\left[\nabla_{\mathbf{h}_n^{(\ell)}}[\text{logit}(y_+) - \text{logit}(y_-)]\right]}{\left\|\mathbb{E}_n\left[\nabla_{\mathbf{h}_n^{(\ell)}}[\text{logit}(y_+) - \text{logit}(y_-)]\right]\right\|}.$$

Protocol details: The gradient is computed at layer $\ell = 15$ (47% depth) on the baseline (no-intervention) forward pass, averaged over $N = 30$ calibration examples disjoint from the $N = 50$ evaluation set. The gradient direction is nearly orthogonal to all four read directions ($|\cos| \leq 0.037$; random baseline ≈ 0.017 in $d = 3584$). Crucially, $\mathbf{d}_{\text{grad}}$ transfers to held-out examples with $\Delta_{\text{odd}} = +0.159$ (CI $[+0.135, +0.183]$), roughly $14\times$ the best read direction (Table 5). The effect is positive for both preference orientations (escalate-preferred, $n = 23$: $\Delta_{\text{odd}} = +0.243$; review-preferred, $n = 27$: $+0.088$), so $\mathbf{d}_{\text{grad}}$ increases the preference-aligned margin regardless of which preference is stated; the asymmetry reflects the calibration set's 17:13 escalate/review composition. The gradient thus defines a distinct write direction that achieves large directional effects while remaining geometrically orthogonal to the read directions. We report this finite effect separately from the first-order comparison: a dedicated small-γ sweep identifies the regime in which both read and write effects retain their expected orientation, yielding a $\sim 6.5\times$ slope gap (Appendix v). Because $\mathbf{d}_{\text{grad}}$ is estimated on the $N = 30$ calibration examples and evaluated on the disjoint $N = 50$ set, this is an out-of-sample comparison between directions estimated from the same calibration examples but with different objectives and supervision.

Interpretation caveat: $\mathbf{d}_{\text{grad}}$ requires per-example downstream decision labels (y_+, y_-) at calibration time—information that read directions do not use. It therefore serves as a local first-order write oracle (§2)—a diagnostic of causal leverage rather than a deployable write operator. The key finding is not that $\mathbf{d}_{\text{grad}}$ is practically useful, but that the causal leverage it reveals lies orthogonal to every read-derived direction, showing that decodability and local causal leverage select substantially different directions in this setting.

Matched-supervision control. One might hypothesize that the read–write gap arises because $\mathbf{d}_{\text{grad}}$ uses continuous margin supervision while read directions use binary labels. To control for

this, we fit a *margin-regression* direction that regresses hidden states onto the same continuous margin used by $\mathbf{d}_{\text{grad}}$ ($R^2 = 0.84$ on calibration). Even when given the same continuous margin supervision, the margin-regression direction is nearly orthogonal to $\mathbf{d}_{\text{grad}}$ ($\cos = +0.026$, random baseline ≈ 0.017) and achieves $\Delta_{\text{odd}} = +0.014$. Its finite-intervention odd effect is also substantially smaller than that of the gradient direction under the same diagnostic protocol. Thus matching the downstream supervision target is not sufficient to recover the gradient-defined write geometry or its finite-intervention effect.

Motivating sign-symmetric evaluation. Initial one-sided random-direction evaluation produced the counterintuitive observation that random directions slightly outperformed the mean-difference read direction (baseline 47.6%, true direction 46.9%, random $50.7 \pm 5.5\%$). This motivated paired sign-symmetric evaluation. Under sign-symmetric measurement, random directions have substantial mean even effect while their mean odd effect is not significantly different from zero ($p = 0.69$; Table 5). Thus the apparent random-direction advantage was due to even (sign-invariant) perturbation effects rather than directional control.

E Experimental Details

Models. Unless otherwise noted, experiments use the following models. Discovery experiments (§4) use Qwen-2.5-7B-Instruct and Mistral-7B-Instruct-v0.3; the full extraction-by-application matrix (§4.1) is measured on Mistral-7B and Qwen-7B. Factorized coupling experiments (§8) use a 1.5B-parameter context compressor trained to map interaction context to a compact representation, with a frozen Qwen-2.5-7B-Instruct task model. A learned projection W_{proj} maps the compressor output h_c into the task model's representation space, yielding $h_{\text{HYP}} = W_{\text{proj}} h_c$. This projected representation is prepended to the task model at position 0 and separately projected onto the preference direction to obtain $q(x) = h_{\text{HYP}}^{\top} \mathbf{d}_P$. The compressor is frozen after pretraining, and only the downstream coupling parameters are fit for the BDA experiments (Appendix j).

Preference direction extraction. The preference direction $\mathbf{d}_P$ is extracted by collecting hidden states at the preference-encoding span for personalized and baseline inputs, computing the mean difference, and normalizing to unit length. Calibration set sizes: 30 examples for CL-Bench, 100 for PrefEval. CL-Bench direction-calibration and γ-validation examples are drawn from the development partition.

PrefEval cross-template extraction. To test whether direction extraction is robust to surface form variation, we extract a direction $\mathbf{d}_P^{(B)}$ using a paraphrased template (varying preference framing, question prefix, and option labels: `A/B/C/D` $\rightarrow$ `1/2/3/4`), then evaluate on test prompts in the original template format. The cross-template direction yields 73.89%, essentially matching the in-template baseline (73.67%, $\Delta = +0.22$pp). The two directions have $\cos(\mathbf{d}_P^{(A)}, \mathbf{d}_P^{(B)}) = 0.91$, supporting stability of mean-difference direction extraction under this moderate template variation and showing that steering remains effective under extraction–inference template mismatch.

Factorized coupling implementation. The factorized coupling experiments instantiate Eq. 4 with three logical components: a context representation map $g_\theta : x \mapsto h_c$, a preference read operator $Q_\psi : h_c \mapsto q(x)$, and a downstream coupling operator $\mathcal{C}_{\theta_c}$. g_θ and Q_ψ are learned from the training partition and frozen before downstream coupling is selected on validation data; all reported test examples remain held out. The task model remains frozen throughout.

The formulation does not require a particular representation architecture: g_θ may be implemented by any context encoder producing a fixed-dimensional state, while Q_ψ may produce either a scalar preference coordinate or a low-dimensional vector. Likewise, $\mathcal{C}$ may act in hidden-state or decision space depending on the task structure. In our implementation, Q_ψ produces a scalar preference coordinate; $\mathcal{C}$ is the scalar decision-space coupling $m'(x) = m(x) + \gamma q(x)$, with γ selected on the validation partition as described in Appendix j. The representation learner is substantially larger than the downstream coupling interface. Accordingly, our comparison evaluates the effect of factorizing representation and downstream coupling rather than parameter efficiency.

Notation: γ across experiments. We use γ for activation-intervention strength; its value is protocol-specific and should not be compared across PrefEval steering and mechanistic diagnostics.

F Per-Category Breakdown

Table 7: Per-category BDA on the stratified full CL-Bench (3 seeds, 399 test samples).

Category	N_{test}	BDA (%)	± Std
Procedural Task Execution	96	93.4	2.6
Empirical Discovery	50	92.0	5.3
Rule System Application	117	89.9	0.4
Domain Knowledge Reasoning	136	88.8	2.2
Macro-average	—	**91.1**	1.8
Micro-average	399	**90.5**	1.5

G Split Methodology

Train/test split is stratified by category and shuffled per seed. Earlier non-stratified splits produced inflated accuracy ($\approx 95\%$) due to unbalanced category representation in the test set; the stratified $90.5\% \pm 1.5\%$ reported is the definitive main result. Split is by CL-Bench example ID (no scenario or template deduplication); we verify that no exact context string appears in both splits.

H Causal Validation Details

Table 8: CL-Bench ablations of the context coupling pathway (3 seeds). Random-direction performance is high under inherited coupling; the representation ablation and random-direction ensemble below isolate the learned direction's role.

Condition	BDA (%)	Δ	Interpretation
Full system (main)	$\mathbf{90.5 \pm 1.5}$	—	—
Random direction	82.5 ± 3.2	-8.0	Learned config. improves
Constant representation	0.0 ± 0.0	-90.5	Context representation essential
Direction flip ($-\mathbf{d}_P$)	0.0 ± 0.0	—	Sign of coupling matters
Context swap	0.1 ± 0.2	—	Context dependence validated

Context-dependent representation is essential. Replacing the context representation model output with a fixed random vector produces 0% across all seeds (100% ties); identical margins for all inputs, confirming that context-dependent representation variation is necessary for nontrivial BDA under this architecture.

Learned versus random coupling. Random-direction injection achieves 82.5%, above chance but lower than the main result ($\Delta = -8.0\%$). The partial effectiveness shows that substantial performance arises from the context-dependent representation together with the explicit coupling pathway. The learned direction contributes additional BDA and, more clearly, cross-category stability. Random coupling is highly unstable across held-out categories (LOCO range: 14.5–80.9%), whereas learned coupling remains both stronger and substantially more stable (88.2–94.5%; Table 9). The direction-flip control ($-\mathbf{d}_P \rightarrow 0.0\%$) confirms sign-sensitivity: negating the learned direction deterministically reverses the coupling effect.

Ablating the task-relevant representation component. Ablating the representation component identified by the preference-direction diagnostic and refitting the downstream coupling parameter reduces BDA from 88.6% to 37.0% on the corresponding $N = 378$ diagnostic split. Residual margins

are concentrated near zero; among non-tied examples the split is exactly balanced. The below-chance strict BDA therefore reflects ties and signal destruction rather than a systematic inversion.

Random-direction ensemble. The 82.5% figure above inherits the coupling parameter fitted for the learned direction, isolating the direction while holding scale fixed. To test whether arbitrary directions can match the learned direction when given separately refit coupling parameters, we independently refit the downstream coupling parameter for each of 50 random unit directions. The ensemble achieves $49.5\% \pm 29.4\%$ BDA (range 2.3–89.5%), revealing high variability across arbitrary coupling directions. The learned direction reaches $90.5\% \pm 1.5\%$. The two protocols answer different questions—whether the direction contributes with the inherited coupling fixed, and whether an arbitrary direction can match the learned direction after separately refitting its coupling parameter.

Anti-cheat: direction flip and context swap. Negating the preference direction inverts all predictions ($90.5\% \rightarrow 0.0\%$), confirming that the downstream effect is sign-sensitive under the learned coupling pathway. Position bias is addressed separately by randomized label assignment. Swapping contexts produces the same inversion (0.1%), confirming the model uses actual context content.

I LOCO Analysis Details

Table 9: Leave-one-category-out cross-validation on CL-Bench. Factorized context coupling generalizes to held-out CL-Bench categories. Random-direction column inherits the coupling parameter fitted for the learned direction (Appendix h).

Held-Out Category	N_{test}	Factorized	Random Dir	Const. Repr
Domain Knowledge	663	88.2%	56.1%	0.0%
Empirical Discovery	199	89.4%	80.9%	0.0%
Rule System	399	91.7%	14.5%	0.0%
Procedural	399	94.5%	55.4%	0.0%
Mean ± Std	1660	$\mathbf{91.0 \pm 2.8}$	51.7	0.0

For LOCO evaluation, Domain Knowledge (663) and Empirical Discovery (199) are evaluated in full, while the Rule System and Procedural folds use 399 randomly subsampled examples from their respective 566- and 471-example categories, yielding $1{,}660$ evaluated examples across the four folds.

The factorized coupling achieves $91.0\% \pm 2.8\%$ on held-out categories (comparable to in-distribution 90.5%), with low variance across folds. The strong LOCO performance is consistent with category-general context-to-decision binding rather than purely category-specific patterns. Ablation controls replicate out-of-distribution: random inherited directions average 51.7% but vary widely across held-out categories (14.5–80.9%); constant representation produces 0%. This provides evidence that the factorized context-to-decision pathway generalizes across the four held-out CL-Bench categories tested here.

J Context Compressor Details

The factorized coupling experiments use Qwen-2.5-1.5B trained to compress full interaction context into a fixed-dimension representation h_c.

- **Base model:** Qwen-2.5-1.5B
- **Representation dimension:** $h_c \in \mathbb{R}^{1536}$
- **Projection:** $W_{\text{proj}} \in \mathbb{R}^{3584 \times 1536}$ maps h_c into the 7B task model's embedding space
- **Training objective:** contrastive alignment loss $\mathcal{L} = -\log \frac{\exp(\pi^+/\tau)}{\exp(\pi^+/\tau)+\exp(\pi^-/\tau)}$, where $\pi = h_{\text{HYP}} \cdot \mathbf{d}_P$; π^+ uses full context, π^- uses truncated context
- **Training data:** the 1,500-example development partition (test set of 399 held out); γ is selected on a 15-example subset after freezing the representation

- **Training hyperparameters:** Adam, lr 5×10^{-5}, batch size 1, 5 epochs, temperature $\tau = 0.07$
- **Injection:** h_{HYP} is prepended to the task-model input at position 0; the task prompt contains the decision query only (no raw context), so context enters solely through h_{HYP}. Coupling acts at the decision-token logits
- **Stage 2 (downstream coupling):** compressor and W_{proj} are frozen; γ selected by grid search over $[0.01, 1.0]$

The preference signal $q(x) = h_{\mathrm{HYP}}^{\top} \mathbf{d}_P$, where $h_{\mathrm{HYP}} = W_{\mathrm{proj}} h_c$ (Figure 3), is computed using the same extracted direction $\mathbf{d}_P$ tested in the causal ablations (Table 8): flipping the direction ($-\mathbf{d}_P$) inverts the signal, confirming sign-sensitivity. Architectural innovation is not claimed; the operational point is that separating context representation from decision coupling enables the factorized intervention tested in §8.2.

K MCQ Evaluation Protocol

Token collision in label scoring. A subtle failure mode arises when MCQ labels share the same first token under the model's tokenizer. For example, "Option1" and "Option2" both tokenize to the same initial token (ID 5341 = "Option"), making single-token logit comparison ill-posed: the prior difference is exactly 0, and the model cannot distinguish between labels regardless of context.

Joint log-probability scoring. We fix this by computing the joint log-probability of the entire label string: $\log p(\text{label} \mid \text{prompt}) = \sum_t \log p(\text{token}_t \mid \text{prompt}, \text{token}_{<t})$. This restores distinguishability: for the Option1/Option2 pair, single-token scoring yields prior difference $= 0.000$ (degenerate), while joint scoring yields 1.206 (non-degenerate). Main task accuracy on the $N = 15$ sanity subset is unaffected (93.3%) because the A/B labels have distinct first tokens, but joint scoring is used throughout as a robustness measure.

Label-position bias and swap calibration. Raw MCQ scoring exhibits inherent label-position priors (e.g., the model assigns higher probability to position 1 regardless of content). We found global null calibration insufficient for small sample sizes; instead, we use per-instance swap calibration: each sample is scored twice with labels in both orderings, and the final score is the average: $\text{score}(a) = \frac{1}{2}[\log p(a \mid a@\text{pos1}) + \log p(a \mid a@\text{pos2})]$. This reduces the position gap from 75% to 50% and improves MCQ accuracy from 50% to 66.7% in sanity tests. For the main A/B binding discrimination task reported in the paper, we use randomized label assignment (anti-shortcut measure (i)) rather than swap calibration, as the A/B task is scored via direct logit comparison at the decision token, not MCQ-style label selection.

Inference-time information flow. During evaluation, full-context and truncated-context prompts are forwarded independently through the model and assigned scores solely from their content. The orientation sign used to construct the preference direction is applied only during direction extraction and is not recomputed or supplied during inference. The evaluator subsequently checks whether $m^{\mathrm{full}} > m^{\mathrm{trunc}}$; this use of ground truth is metric computation, not model input. The 0% constant-representation result is not evidence of oracle application: the intervention produces exactly equal full/truncated margins on every example, and strict BDA counts all ties as incorrect.

L Within-Model Direction Consistency Across Architectures

Across the tested architectures, matched conservative/aggressive contrasts exhibit nontrivial within-model directional consistency, although sensitivity to prompt variation and steering magnitude vary substantially by model.

M Gemma-2-2B Depth Analysis

Gemma-2-2B-it shows weak steering effects overall, with maximum application effect 0.009 compared to 0.111 (Qwen-7B) and 2.10 (Mistral-7B). The extraction-later-than-application ordering is present but strongly attenuated: application peaks at 77% depth (layer 20 of 26) and extraction peaks

Table 10: Within-model directional consistency across matched preference contrasts on multiple architectures. Consistency is the cosine between directions extracted from conservative vs. aggressive preference pairs using different calibration prompts; lower values indicate higher sensitivity to prompt variation.

Model	Layers	Within-model consistency (cos)	Steering Swing
Qwen-2.5-7B (reference)	28	0.92	45pp
Gemma-2-2B-it	26	0.62	35pp
Mistral-7B-v0.3	32	0.63	15pp

at 96% (layer 25). The weak absolute effects prevent meaningful comparison of layer-asymmetric application on this architecture. The attenuation may reflect model scale, architecture, or instruction-tuning differences; the present experiment does not distinguish among these explanations.

N Cross-Scale and Cross-Family Replication

We performed a targeted scaling replication on Qwen-2.5-14B-Instruct (48 layers) using the team-decision task (N_{calib}=30, N_{eval}=50), preserving the protocols of the corresponding 7B analyses rather than retuning them at 14B.

Read–write geometry. At layer 33 (68.8% depth), read directions remained strongly separated from the local gradient-defined write direction: $\cos(R_{\text{LDA}}, W) = 0.011$, $\cos(R_{\text{mean}}, W) = 0.085$, and $\cos(R_{\text{logistic}}, W) = 0.076$. The span of these three read directions captured only 0.85% of the write direction's squared norm. This separation was not explained by unstable estimation: calibration-bootstrap stability was 0.993 for the read direction and 0.992 for the write direction. Thus the geometric read–write dissociation observed at 7B persists under this targeted 14B replication, although its exact magnitude is model-dependent.

Crossed-depth intervention. Using the independently selected mid-depth application site (layer 33) and late extraction site (layer 47), with the same $\gamma = 2.0$ intervention protocol, the four crossed effects were 0.062 ($33 \to 33$), 0.032 ($47 \to 47$), 0.052 ($47 \to 33$), and -0.001 ($33 \to 47$). The point estimates reproduce the qualitative asymmetry observed at 7B: late-extracted preference information transfers to the mid-depth application site, whereas the reverse mid-to-late transfer is near zero. Paired bootstrap resampling ($n = 10{,}000$) yields a 95% CI for the contrast $D_{\text{cross}} = (47 \to 33) - (33 \to 47)$ of $[0.032, 0.072]$, excluding zero and confirming the asymmetry is statistically reliable at 14B. (The same-depth effects are not monotonic with depth; the replicated claim concerns crossed transfer asymmetry.)

32B replication. A further replication on Qwen-2.5-32B-Instruct (64 layers) at layer 44 (68.8% depth) yielded $\cos(R_{\text{LDA}}, W) = 0.042$ and read-span coverage 0.44%, with bootstrap stability 0.987 (read) and 0.978 (write). The crossed-depth four-cell at layers 44/63 showed $+0.017$ ($63 \to 44$) versus $+0.002$ ($44 \to 63$), preserving the qualitative asymmetry. Table 11 summarizes the scaling pattern.

72B replication. On Qwen-2.5-72B-Instruct (80 layers) at layer 55 (68.8% depth), $\cos(R_{\text{LDA}}, W) = 0.033$ with read-span coverage 0.53% and bootstrap stability 0.988 (read) and 0.985 (write). The crossed-depth four-cell at layers 55/78 showed the same qualitative asymmetry: $D_{\text{cross}} = 0.051$ with 95% CI $[0.034, 0.068]$, excluding zero.

Cross-family replication. On Llama-3.1-8B-Instruct at layer 22 (68.8% depth), $\cos(R_{\text{LDA}}, W) = 0.038$ with read-span coverage 9.75% and bootstrap stability 0.892 (read) and 0.997 (write). The elevated coverage reflects architectural differences but the directional separation persists. On Gemma-2-9B-it at layer 28 (66.7% depth), $\cos(R_{\text{LDA}}, W) = 0.018$ with read-span coverage 0.04% and bootstrap stability 0.996 (write); read stability is undefined because all calibration margins share the same sign. Table 11 summarizes the full scaling and cross-family pattern.

Table 11: Scale and cross-family summary: geometric dissociation persists across model scale and architecture.

Model	Scale	$\lvert\cos(R_{\text{LDA}}, W)\rvert$	Coverage	R stab	W stab
Qwen-2.5	7B	≤ 0.026	0.07%	0.999	0.983
Qwen-2.5	14B	0.011	0.85%	0.993	0.992
Qwen-2.5	32B	0.042	0.44%	0.987	0.978
Qwen-2.5	72B	0.033	0.53%	0.988	0.985
Llama-3.1	8B	0.038	9.75%	0.892	0.997
Gemma-2	9B	0.018	0.04%	—	0.996
Mistral	7B	0.087	5.6%	0.920	0.991

O Preference Granularity

Within a single domain, finer-grained preferences span sub-clusters rather than sharing a single axis. Multiclass tag preferences (e.g., genre labels like action, comedy, dark comedy) yield lower mean within-domain cosine (~ 0.36), with hierarchical clustering revealing sub-clusters that roughly correspond to broad genres. Family-structured geometry therefore appears at multiple granularity levels: narrow semantic clusters share an axis, but a domain spanning multiple sub-clusters does not.

P Directional Sensitivity

To characterize the relative causal weight of orientation and magnitude, we decompose preference interventions into direction and magnitude components and compare their causal effects.

Table 12: Directional vs. radial interventions. Consistent with prior work emphasizing directional structure for steering Vu and Nguyen (2025); Pham and Nguyen (2024); Kayan and Zhang (2025), the $180°$ flip produces a $10\times$ larger logit swing than $3\times$ magnitude scaling.

Type	Intervention	$\lvert\Delta\rvert$
Radial (magnitude)	$0.5\times$ scale	0.00
	$2.0\times$ scale	1.38
	$3.0\times$ scale	2.34
Directional	Random direction	0.03
	Random orthogonal	1.44
	$90°$ in-plane	0.75
	$180°$ flip	**23.38**

The $180°$ flip produces a 23.38-logit swing while $3\times$ magnitude scaling produces only 2.34 ($10\times$ smaller). Halving the norm produces no measurable effect ($\Delta = 0.00$). Behavior is substantially more sensitive to orientation than to radial scaling over the tested range. These results provide complementary evidence that intervention orientation carries substantial causal information over the tested range.

Q PCA Analysis

Unsupervised decomposition via PCA on 17 diverse preference types finds that PC1 aligns strongly with explanation depth ($\lvert\cos\rvert = 0.830$ with the manual axis), whereas alignment with risk tolerance is weaker ($\lvert\cos\rvert = 0.436$). Top-3 PCs explain only 57.4% of variance. PCA preferentially recovers the higher-variance explanation-depth factor but does not cleanly isolate the lower-variance risk-tolerance axis—a lower-variance axis used in our team-differentiation experiments. This motivates supervised identification of behaviorally relevant steering directions.

R MULTI-AXIS EVALUATION AND CROSS-DOMAIN TRANSFER

Hidden-state coupling. Unlike logit-level coupling which directly modifies decision margins, hidden-state coupling adds $\sum_k z_k \cdot \mathbf{d}_k$ at a chosen layer, distributing signal across leading tokens. This mechanism is used for multi-axis decisions where teams differ along multiple preference coordinates simultaneously.

Multi-axis in-domain evaluation. The main text reports 100% ($80/80$, three seeds, $N = 20$ per team). This evaluation uses the loan-decision domain with 4 teams (Conservative+Brief, Conservative+Detailed, Aggressive+Brief, Aggressive+Detailed), each evaluated on 20 held-out scenarios. Total: $4 \times 20 = 80$ in-domain test examples.

Table 13: Cross-domain transfer: context representation model trained on loan decisions ($N = 48$, 4 teams $\times$ 12), tested zero-shot on hiring decisions ($N = 40$, 4 teams $\times$ 10). Three seeds; single-seed pilot reported 100% and is superseded.

Team	Loan (in-domain)	Hiring (zero-shot)
A (Conservative + Brief)	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
B (Conservative + Detailed)	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
C (Aggressive + Brief)	$100.0\% \pm 0.0\%$	$40.0\% \pm 0.0\%$
D (Aggressive + Detailed)	$100.0\% \pm 0.0\%$	$53.3\% \pm 18.9\%$
Average	$100.0\% \pm 0.0\%$	$73.3\% \pm 4.7\%$

Cross-domain zero-shot transfer. The context representation model, trained exclusively on loan decisions, transfers to hiring decisions with conservative profiles (A, B) at 100% and aggressive profiles (C, D) at 40–53% ($N = 10$ per team). The source of this profile-dependent transfer gap remains open.

S APPLICATION CURVE VALIDATION AND SCALING DETAILS

Methodology validation. The same-layer extraction protocol could conflate representation quality with write efficacy (ease of intervention at that layer). To isolate extraction-side quality, we run a *fixed-application* extraction sweep: extract at varying layers ℓ but apply all directions at the *fixed* application peak (layer 15 for Mistral). Holding the application site fixed removes variation due to application depth, so differences across the sweep more directly reflect properties of the extracted direction. The fixed-application curve peaks at the same layer (25, 78% depth) as the same-layer curve, supporting the interpretation that the late extraction peak is not an artifact of layer-specific write efficacy.

Robustness to reference direction. The application curve definition uses a single reference direction extracted at one layer, which could bias the measured peak location. We test sensitivity by re-running the application sweep on Mistral-7B with five reference directions extracted at layers $\{10, 15, 20, 25, 30\}$, spanning 31–94% depth. References extracted at layers 10–20 all peak in the 45–48% depth range, showing that the application peak is stable across the tested mid-layer reference directions. Late-layer references (layers 25 and 30) show peak locations at 58% and 77% depth respectively, but retain 79% and 67% of peak effect at the mid-layer application site (layer 15), indicating a stable mid-layer application component, while late-extracted directions additionally favor later application sites.

Scaling beyond 7B. The application peak measured at 47–48% depth is consistent across two 7B architectures (Mistral-7B at 47%, Qwen-7B at 48%). We test whether the core structural ordering $\ell_{\text{ext}} > \ell_{\text{app}}$ extends to larger scale with a coarse 8×8 sampled-layer check on Qwen-2.5-14B-Instruct (48 layers), using the same baseline-subtracted margin definition as the 7B matrices with $\gamma = 2.0$. In this check, extraction again peaks later than application (98% vs. 69% depth), preserving the ordering $\ell_{\text{ext}} > \ell_{\text{app}}$. Because this check uses a different task and sampled-grid protocol from the main 7B full matrices, we do not interpret its rank-1 EVR as a direct scale comparison; the

relevant finding is that the same depth ordering is also observed in this 14B check. For the layer-asymmetric application, we use the empirically measured application peak for the specific model under study rather than relying on a fixed relative-depth heuristic.

T DECODABILITY AND FINITE-INTERVENTION EFFICACY

Setup. For each of 15 constraint-injection directions, we measure *reading accuracy* (linear probe, 5-fold CV on natural prompts) and *writing swing* (logit shift under $\gamma = 2.0$ steering).

Results. Across 15 directions extracted at layer 20 (Qwen-2.5-7B-Instruct), mean reading accuracy is 68.4% (range 13–100%, mean AUC 0.71), while mean writing swing is only 0.021 logits (range -0.026 to $+0.078$). Reading and writing are weakly correlated ($r = 0.31$); 8 of 15 directions exceed 70% reading accuracy while falling in the lower half of the observed writing-swing distribution.

Layer dependence. Comparing the same extraction mechanism at different layers reveals that writing swing is strongly layer-dependent: extraction at the extraction peak (layer 25) yields $140\times$ larger writing swing than extraction at layer 20. On multiclass benchmarks (e.g., CL-Bench), larger writing swings do not improve accuracy, indicating the reading-writing phenomenon has both a layer-dependent component (swing magnitude) and a selection-dependent component (choosing which direction to apply) that layer choice alone does not address.

U DECISION-DIRECTION PROPAGATION

We investigate why $\mathbf{d}_{\text{grad}}$ achieves substantially larger behavioral effects than the read directions. The key difference emerges in how the induced perturbations propagate through the remaining layers.

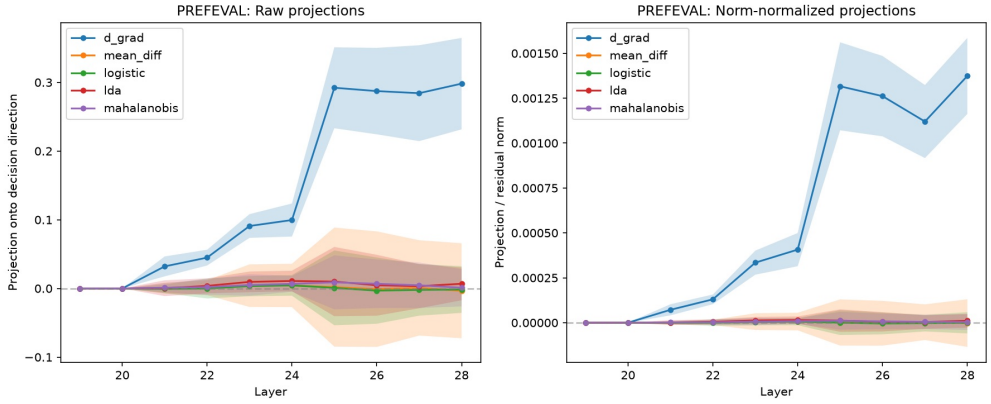
Protocol. For each intervention direction $\mathbf{d}$, we define the propagated perturbation at layer ℓ as $\boldsymbol{\delta}_n^{(\ell)}(\mathbf{d}) = \mathbf{h}_{n,\text{steered}}^{(\ell)} - \mathbf{h}_{n,\text{base}}^{(\ell)}$. We then measure its projection onto an output-readout contrast. We use Qwen-2.5-7B-Instruct with $N = 20$ test examples per task (PrefEval and team); the smaller sample reflects the cost of computing per-layer projections for five directions across ten layers per example. For each example, y_+ and y_- are assigned per the ground-truth label: if the correct answer is A, then $y_+ = \text{A}$ and $y_- = \text{B}$; otherwise reversed. Directions are unit-normalized at injection with $\gamma = 0.1$. The intervention layer is 19 for PrefEval (matching the sign-symmetric protocol). For the team task, we measure from layer 22; the sign-symmetric Δ_{odd} results use layer 15, so the propagation analysis starts deeper than the layer used for the primary team-task causal diagnostic. We measure projections through layer 28. Bootstrap CIs use $1{,}000$ replicates over 20 examples.

Projection onto the decision direction. For each example, we define the decision direction as $\mathbf{u}_{\text{dec}} = \mathbf{u}(y_+) - \mathbf{u}(y_-)$, where $\mathbf{u}(\cdot)$ denotes the normalized unembedding vector. We measure each perturbation's projection $\boldsymbol{\delta}^{(\ell)\top}\mathbf{u}_{\text{dec}}$ at the intervention layer and at the final layer before the language modeling head.

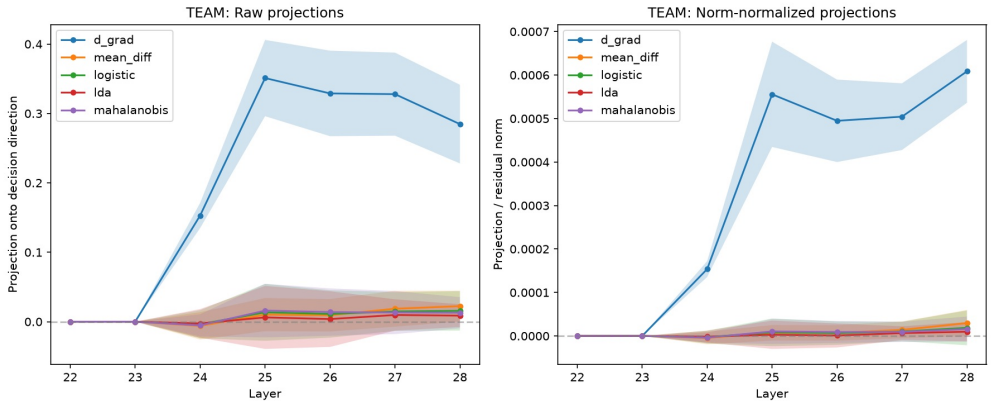
At the intervention layer, perturbations induced by both $\mathbf{d}_{\text{grad}}$ and the read directions have comparably small projections onto the decision direction ($|\cdot| \approx 0.01$–0.04). Yet read directions achieve large class separation (cross-validated AUC 0.78–0.90 on PrefEval), indicating their discriminability lies in a subspace with little direct decision alignment.

The difference emerges in propagation. Figure 5 shows the projection magnitude as a function of layer depth. On PrefEval, $\mathbf{d}_{\text{grad}}$'s projection grows from 0.03 at layer 19 to 0.30 at layer 28—but the growth is not gradual. The projection is flat through layer 23 ($\approx$0.04–0.10), then jumps sharply between layers 24 and 25 ($0.10 \rightarrow 0.29$), and plateaus thereafter. This jump occurs just after the Qwen-PrefEval extraction peak (layer 22, Table 4), suggesting a possible connection between the extraction peak and subsequent write-relevant amplification, which we do not establish causally here. Read directions' projections remain near zero throughout on PrefEval.

On the team task, $\mathbf{d}_{\text{grad}}$ grows from 0.15 to 0.28, while read directions show smaller but nonzero growth.



(a) PrefEval (left: raw, right: norm-normalized)



(b) Team task (left: raw, right: norm-normalized)

Figure 5: Perturbation projection onto the decision direction vs. layer depth. Each subfigure shows raw projections (left panel) and norm-normalized projections (right panel). The perturbation induced by $\mathbf{d}_{\text{grad}}$ (blue) grows sharply through late layers while perturbations from read directions (other colors) remain flat on PrefEval and grow far less on the team task. Shaded regions indicate bootstrap 95% CIs over $N = 20$ examples. The growth pattern survives norm normalization, so simple residual-stream norm growth does not account for the effect.

Norm growth. A second component of the write advantage is perturbation survival. All directions are unit-normalized at injection ($\gamma = 0.1$), but at the final layer, $\mathbf{d}_{\text{grad}}$'s perturbation has norm $\|\boldsymbol{\delta}\| = 1.30$ while read directions' perturbations have $\|\boldsymbol{\delta}\| = 0.42$–$0.46$. The perturbation along $\mathbf{d}_{\text{grad}}$ grows $\sim 3\times$ more than read directions through propagation.

Together, these observations identify two correlates of the write advantage: the gradient-defined perturbation both *survives* more strongly (larger $\|\boldsymbol{\delta}\|$) and *becomes substantially more aligned with the decision direction* (larger projection onto $\mathbf{u}_{\text{dec}}$). On PrefEval, read directions' perturbations neither grow nor align; on the team task they grow and align far less than $\mathbf{d}_{\text{grad}}$.

Norm control. The projection growth could be a norm artifact: the residual stream grows in norm through depth, so any fixed direction's projection can grow mechanically. We control for this by dividing projections by the layer-wise residual-stream norm ($\|\mathbf{h}\| \approx 290$ at the final layer). The normalized projections show the same pattern. At the final layer on PrefEval:

- $\mathbf{d}_{\text{grad}}$: 1.57×10^{-3}

- mean-diff: 4.08×10^{-5}
- LDA: 1.15×10^{-5}

The ratio $\mathbf{d}_{\text{grad}}$/mean-diff ≈ 38 is unchanged by normalization, confirming the growth is not a norm artifact.

Sub-finding: mean-difference on team task. One exception to the flat-propagation pattern is mean-difference on the team task: 80% of examples have positive final-layer projections, with mean $+0.022$ (95% CI $[+0.013, +0.032]$). This provides convergent evidence that at least some read-derived perturbations can acquire downstream decision alignment.

Interpretation. Read directions and $\mathbf{d}_{\text{grad}}$ start with comparably small decision alignment at the intervention layer. The observed write advantage is therefore consistent with downstream transformation of the perturbation, rather than being explained solely by its initial alignment with the decision direction. On PrefEval, $\mathbf{d}_{\text{grad}}$'s projection grows $\sim 10\times$ through the remaining layers while read directions remain nearly flat; on the team task, measured from layer 22, it grows $\sim 2\times$ and read directions grow less. The team-task propagation analysis starts deeper than the layer used for the primary team-task causal diagnostic, so we interpret it as mechanistic support rather than a complete account of that effect. Behaviorally, this aligns with Table 1: on Mistral, a mid-extracted read direction applied late collapses ($15 \rightarrow 25$: 0.37), a related failure mode.

This provides support for the read–write factorization: the gradient-defined perturbation acquires substantially greater output-readout alignment than the read-derived perturbations.

V Robustness of Read–Write Dissociation

Token-position protocol. For gradient-based read–write comparisons, both gradients and interventions are defined at the final-token state: interventions modify only the same token position with respect to which the gradient is computed. Conventional activation-steering and depth-surface experiments retain their standard all-position intervention protocol; we do not treat these intervention operators as interchangeable.

Layer and intervention robustness. Across layers 10–25, read–write alignment varies with depth ($|\cos| = 0.003$–0.20) but remains far from collinearity. Larger interventions exhibit substantial nonlinearity, including orientation changes, so we estimate first-order leverage only from the verified local regime $\gamma \in \{0.0025, 0.005, 0.01\}$. In the primary setting at layer 15, the write oracle has a $\sim 6.5\times$ larger signed odd-effect slope than the mean-difference read direction ($+0.451$ vs. $+0.069$). Thus geometric dissociation persists across depth, while the causal-leverage comparison is confined to the regime in which its first-order interpretation is valid.

Alternative read estimators. The geometric dissociation is not specific to mean-difference readout. Logistic and LDA read directions also remain nearly orthogonal to the calibration write direction (all $|\cos| \leq 0.026$):

Read estimator	$\cos(d_{\text{read}}, d_W)$	Signed slope
Mean-diff	0.005	-0.258
Logistic	0.026	$+0.208$
LDA	0.024	$+0.585$

In this separate estimator-comparison realization, the calibration write direction has $\cos = 1.0$ by construction and signed slope $+0.064$ on held-out examples. The write direction is first-order optimal only for the calibration objective used to construct it; held-out slopes measure transfer and therefore need not preserve this ordering. Indeed, LDA attains a larger held-out slope in this realization despite remaining nearly orthogonal to the calibration write direction.

Cosine null. In $d = 3584$, independent random unit vectors have $\mathbb{E}|\cos| \approx 0.013$ (sd ≈ 0.017; 97.5th percentile ≈ 0.033). The observed read–write cosines (0.005–0.026 across estimators)

lie within this null distribution: read–write cosines fall within the random-vector null. Meanwhile, the directions themselves are stable across calibration resamples ($\cos(R_b, R_{\text{full}}) = 0.999$, $\cos(W_b, W_{\text{full}}) = 0.983$; see below). The dissociation therefore reflects two well-estimated directions with no detectable alignment, not estimation noise.

Leverage lies outside the read subspace. Let Q span the four read directions and $P = QQ^\top$. Only 0.07% (95% CI $[0.02\%, 0.21\%]$) of $\|\mathbf{d}_{\text{write}}\|^2$ lies in $\text{span}(D_{\text{read}})$, against $k/d \approx 0.11\%$ for a random 4-dimensional subspace. Since first-order leverage is $\bar{\mathbf{g}}^\top \mathbf{d}$ (§2) and $\mathbf{d}_{\text{write}} \propto \bar{\mathbf{g}}$, the leverage of the in-subspace component is $\bar{\mathbf{g}}^\top P \mathbf{d}_{\text{write}} = \|\bar{\mathbf{g}}\| \, \|P\mathbf{d}_{\text{write}}\|^2$, so this fraction is also the share of $\mathbf{d}_{\text{write}}$'s first-order leverage recoverable from the read subspace. Empirical odd-effect slopes of the projected components are not resolvable at $\gamma \leq 0.01$ (all CIs span zero); the leverage decomposition therefore rests on the first-order identity above rather than on a direct finite-γ measurement.

Calibration sensitivity. The magnitude of the held-out causal advantage is sensitive to calibration composition. Across 50 calibration resamples:

- Mean paired difference $D = S_W - S_R$: $+0.27$
- Median D: $+0.31$
- 95% CI: $[-0.88, +1.27]$
- $P(D > 0)$: 70% (35/50 resamples)

The positive mean and median indicate an average tendency toward greater write leverage, but the wide interval reflects substantial variation across calibration resamples. We therefore interpret the write direction as a local, distribution-conditioned diagnostic rather than a universally superior intervention.

Calibration stability analysis. To diagnose whether the 70% finding reflects direction estimation instability or effect heterogeneity, we conduct three additional analyses.

Direction stability: Resampled directions remain highly aligned with full-calibration directions ($\cos(W_b, W_{\text{full}}) = 0.983$, $\cos(R_b, R_{\text{full}}) = 0.999$). The correlation between D and direction alignment is weak ($\rho = 0.26$, $p = 0.07$), indicating that direction instability is not the primary driver of D variance.

Calibration-size curve: We vary calibration size from $0.25N$ to N (30 resamples each). $P(D > 0)$ plateaus at $\sim$73% from $0.5N$ onwards—within the tested range, increasing calibration size beyond $0.5N$ does not further increase the observed proportion.

Variance decomposition: Bootstrap decomposition yields $\text{Var}_{\text{cal}}(D)/\text{Var}_{\text{test}}(D) = 1.6$. Both calibration-composition and evaluation-sampling variability contribute substantially; neither alone accounts for the observed variation.

These diagnostics indicate that the variability is not primarily explained by unstable estimation of the read or write direction. Instead, the magnitude of the held-out causal advantage is sensitive to calibration and evaluation composition: the directions themselves remain highly stable, while the write-minus-read effect varies across resamples. We therefore treat stronger causal leverage as a distribution-sensitive consequence observed in the primary setting, rather than a universal property of the write direction.

Score dependence. The gradient-defined oracle also depends on the downstream score G. Token-margin and sequence-probability objectives yield substantially different directions ($\cos(d_{\text{tok}}, d_{\text{seq}}) \approx -0.29$), and both remain weakly aligned with the read direction ($|\cos| < 0.16$). The token-margin oracle has the expected positive local slope under token-margin evaluation ($+0.48$). We therefore interpret the write direction as score-conditioned, $d^*_{\text{write}}(G, \ell)$; the present experiment establishes objective sensitivity of the geometry, not universal finite-intervention optimality across objectives.

W TIED–UNTIED RANK SWEEP (NULL RESULT)

Untied read–write formulation. The findings of §4–§6 motivate a factorized design space with independent read and write degrees of freedom. A general form is

$$z = R^\top \mathbf{h}_c, \qquad \Delta \mathbf{h}_t = W \,\mathrm{diag}(a_t) z, \tag{5}$$

where $R \in \mathbb{R}^{d\times K}$ reads K preference coordinates from the context representation $\mathbf{h}_c$, $W \in \mathbb{R}^{d\times K}$ writes them back, and $a_t \in \mathbb{R}^K$ modulates coupling at token t. Equation 5 makes explicit two architectural degrees of freedom suggested by our experiments: representational rank K and the separation of read and write operators. Heterogeneous team geometry shows that preference families can occupy a multidimensional low-rank subspace (§5), while causal intervention experiments show that directions optimized for reading need not coincide with directions associated with local decision sensitivity (§6).

To test whether the directional read–write dissociation necessarily implies a behavioral advantage for untied low-rank operators, we compare tied ($W = R$) and untied ($W \neq R$) parameterizations on a controlled 5-way team-classification task.

Task and setup. We train low-rank personalization models on a 5-way team classification task derived from hidden-state features. Teams correspond to the five severity levels (conservative, cautious, neutral, permissive, aggressive) used in the multi-team experiments. Each example is an LLM hidden state extracted at layer 15 from Qwen-2.5-7B-Instruct with a policy-prefixed decision prompt.

Data. $N_{\text{train}} = 200$ (40 per team), $N_{\text{test}} = 100$ (20 per team). Hidden states are extracted once and stored; training operates on pre-computed features.

Models. For each rank $K \in \{1, 2, 3, 4, 8\}$, we compare two parameterizations:

- **Tied** ($W = R$): Read $z = R^\top h$, write $\tilde{h} = h + Rz$, classify via $C\tilde{h}$. Effective actuator: $A = RR^\top$ (symmetric, rank-K). Parameters: $dK + d \cdot 5 + 5$.
- **Untied** ($W \neq R$): Read $z = R^\top h$, write $\tilde{h} = h + Wz$, classify via $C\tilde{h}$. Effective actuator: $A = WR^\top$ (general rank-K). Parameters: $2dK + d \cdot 5 + 5$.

Critically, both R and W participate in the forward computation that drives gradient updates.

Training. Adam optimizer, learning rate 0.1, 200 epochs, no regularization. We report mean and standard deviation over $n = 3$ random seeds. All runs achieve 100% training accuracy.

Results.

K	Tied ($W = R$)	Untied ($W \neq R$)	Δ
1	100.0 ± 0.0	100.0 ± 0.0	0.0
2	100.0 ± 0.0	100.0 ± 0.0	0.0
3	100.0 ± 0.0	100.0 ± 0.0	0.0
4	100.0 ± 0.0	100.0 ± 0.0	0.0
8	100.0 ± 0.0	100.0 ± 0.0	0.0

Both parameterizations achieve 100% held-out accuracy for every tested rank $K \in \{1, 2, 3, 4, 8\}$. This *null result* indicates that the classification task is not sufficiently constraining to distinguish tied from untied operators: when even the most restrictive rank-1 symmetric projection saturates the task, there is no room for asymmetry to help. The null result does not contradict the causal read–write dissociation reported in the main text; it shows only that this particular downstream task cannot serve as a behavioral probe for the distinction.

X TRAJECTORY WRITE ROTATION

Structured decisions admit a localized write direction W defined by the gradient of a decision score. Autoregressive generation, however, unfolds over a trajectory of intermediate states. We examine whether local write directions remain stable along the generation trajectory.

Protocol. For each of $n = 20$ PrefEval implicit-choice examples, we generate a 30-token completion and compute local write directions W_t at positions $t \in \{1, 4, 8, 16\}$ using gradients of a next-token preference score (log-probability of the preference-aligned continuation minus the preference-violating one) with respect to the hidden state.

Results. Local write directions show near-zero alignment across the measured trajectory positions:

Pair	$\cos(W_t, W_{t'})$
W_1, W_4	0.002
W_1, W_8	0.005
W_1, W_{16}	0.048
W_4, W_8	-0.020
W_8, W_{16}	-0.018

The alignment with the generation-matched read direction R_{gen} is similarly near-zero across all positions ($|\cos(R_{\text{gen}}, W_t)| < 0.03$).

Interpretation. Unlike structured decisions, autoregressive generation does not expose a single stationary write vector. Local causal write directions change substantially—indeed, to near-orthogonality—along the generation trajectory. This suggests that extending decision-local steering to free-form generation may require trajectory-conditioned write operators rather than a single stationary direction.