Directional Personalization: Preference Geometry and Multi-Team Decision Making

Author One
Affiliation
email

Author Two
Affiliation
email

Abstract

Adapting large language models to team-specific decision preferences typically requires fine-tuning millions of parameters per team. For structured decision tasks (binary and multiclass classification with explicit candidate labels), we show this conflates two separable subproblems: *extracting* a compact preference signal and *coupling* it to decisions. We find that the geometry of preference encoding makes coupling surprisingly simple. We characterize the depth-dependence of activation steering through a two-phenomenon decomposition: *encoding quality* (peaking at 78–96% depth) is distinct from *routing sensitivity* (peaking at 46–48% depth at 7B scale, shifting later at 14B). Extracting at the encoding peak yields $3.2\times$ larger logit swing than mid-layer extraction, motivating layer-asymmetric extraction-application protocols.

For structured decisions, a single scalar γ along an extracted direction suffices for per-team coupling. We propose a Hypothesis Token (HYP) architecture that learns extraction once via a shared context compressor and reduces per-team adaptation to one scalar. On PrefEval Zhao et al. (2025c), a standard personalization benchmark, directional steering achieves 100.0% on Mistral-7B (+9.89pp) and 73.7% on Qwen-2.5-7B (+5.22pp); the direction-flip control inverts Mistral accuracy to 9.2%, confirming saturation reflects real directional signal rather than a ceiling artifact. On CL-Bench, HYP+scalar coupling achieves $90.5\% \pm 1.5\%$ while outperforming a 1.97M-parameter MLP (73.3%). Free-form generation, where the preference signal is distributed throughout a long output rather than localized at decision tokens, is out of scope.

1 Introduction

When organizations deploy large language models for decision-making, different teams need different behaviors from the same model. A risk-management team requires conservative escalation; an operations team prioritizes throughput; a compliance team enforces regulatory review. These *team-specific reasoning preferences* go beyond simple instruction-following: the model must internalize a preference, maintain it throughout multi-step reasoning, and act on it consistently even when subsequent context creates pressure toward a different decision. Our method is designed for tasks with explicit candidate labels (binary and multiclass classification); free-form generation is beyond the scope of this work.

Current approaches (fine-tuning, adapters, RLHF, DPO) treat personalization as a capacity problem, which is vastly overparameterized for structured decisions. We show personalization conflates two separable subproblems: *extracting* a compact preference state from context, and *coupling* that state to decisions. The structure of preference encoding is geometric (preferences are *directions*), and this geometry makes coupling trivial: once the shared axis is identified, per-team adaptation reduces to a

Preprint.

single scalar gain. We learn extraction once via a hypothesis-token compressor and reduce per-team coupling to one scalar γ.

Geometric structure. Preferences encode as directions: 180° flip reverses decisions by 23.38 logits while halving the norm has zero effect (Appendix O). Preferences in a family share a direction ($\cos = 0.84$–0.90 within template, 0.64 across paraphrases), consistent with block-diagonal structure Park et al. (2024). Depth-dependence decomposes: *routing* peaks at 46–48% (7B), *encoding* at 78–96%; encoding-peak extraction yields $3.2\times$ larger logit swing.

Contributions.

1. Two-phenomenon depth decomposition: encoding (78–96%) and routing (46–48% at 7B) peak at distinct depths, motivating extraction at the encoding peak with $3.2\times$ larger logit swing than mid-layer extraction (§4, §5.4).
2. Decoupling principle: shared extraction + per-team scalar γ, with characterized boundaries on both sides (§5, §8).
3. Template-conditional block-diagonal structure: within-family cosine 0.84–0.90 (vs. CAA's 0.58) (§4.1).
4. Standard benchmark validation: directional steering achieves 100% on PrefEval Mistral-7B ($+9.89$pp, flip 9.22%), $+5.22$pp on Qwen-7B. On CL-Bench (implicit preference), HYP+logit-coupling achieves $90.5\% \pm 1.5\%$ where basic hidden-state steering fails (48.8%); per-team adaptation reduces to one scalar (shared compressor amortized), outperforming 1.97M per-team MLP adapters (73.3%) (§7).

2 Related Work

Activation steering and the linear representation hypothesis. Representation engineering Zou et al. (2023), activation addition Turner et al. (2023), and contrastive activation addition Panickssery et al. (2024) control LLM behavior at inference time via direction vectors in hidden states. The Linear Representation Hypothesis Park et al. (2024) provides theoretical grounding for the block-diagonal direction structure we extend to preference families. Recent variations include rotation-based interventions Vu and Nguyen (2025); Pham and Nguyen (2024), prototype-based steering Kayan and Zhang (2025), SAE-derived steering Wu et al. (2025), and depth-wise or hypernetwork-parameterized variants. We contribute the first encoding/routing depth decomposition, identifying that extraction and application target distinct depths—a structural insight orthogonal to coupling-parameterization choices, refining the practitioner heuristic of 60–75% depth Mitra (2026). Recent geometry-first work highlights that linear direction-based interventions can go off-manifold in curved activation spaces Park et al. (2024); geometric coupling is effective when preferences correspond to natural activation directions.

Reading-writing asymmetry. Zou et al. (2023) notes that effective reading does not imply effective control; Wu et al. (2025) shows difference-of-means is best for detection while prompting is best for steering. We quantify this asymmetry and show it is substantially layer-dependent (§6).

Personalization via steering and preference learning. CHAMELEON Tan et al. (2024) and SteerX Zhao et al. (2025a) apply steering to personalization; BiPO Cao et al. (2024) learns steering vectors via bi-directional optimization. A complementary literature studies preference elicitation outside the activation-space framing Shao et al. (2023); Oh et al. (2024); Lu et al. (2023); Huang et al. (2023); Park et al. (2024b); Wolfert (2026); Cao et al. (2026); Drago et al. (2025); Dong et al. (2025); Xie et al. (2025). Our contribution is complementary: we identify geometric structure that allows per-team coupling to reduce to a single scalar at zero per-team training cost.

Low-rank adaptation. LoRA Hu et al. (2022) and TinyLoRA Morris et al. (2026) demonstrate low-dimensional structure in weight space; our findings are complementary in activation space.

Positioning relative to per-team training methods. CoPL Choi et al. (2025) and BiPO Cao et al. (2024) require per-user training data; our setting is zero-shot per-team adaptation via a single scalar with no per-user training. SteerX Zhao et al. (2025a) and CHAMELEON Tan et al. (2024) use

different mechanisms (disentangled steering, personalized PEFT); compute-matched comparison left to future work.

3 Problem Formulation

Consider a model f with L layers and hidden dimension d. Given context x with personalization signals P (team criteria, risk preferences), we study how P encodes in hidden states $\mathbf{h}^{(\ell)} \in \mathbb{R}^d$ and causally determines output. The *preference direction* $\mathbf{d}_P^{(\ell)} \in \mathbb{S}^{d-1}$ is the unit vector maximizing separation between preference-aligned and baseline hidden states at layer ℓ; the *preference projection* $\pi^{(\ell)} = \mathbf{h}^{(\ell)} \cdot \mathbf{d}_P^{(\ell)}$ tracks alignment with the active preference.

We test three hypotheses: (1) preference encoding is *directional*: causal effect depends on direction $\mathbf{d}_P$, not magnitude; (2) within narrow semantic families, different preference types share a *block-diagonal* direction structure Park et al. (2024); (3) encoding quality and routing sensitivity peak at *different depths*, decomposing the commonly observed depth-dependence of steering effectiveness.

4 Two-Phenomenon Decomposition of Steering Depth

Prior work on activation steering observes that interventions are most effective at middle-to-late layers Turner et al. (2023); Zou et al. (2023), with practitioner heuristics placing the optimal range at 60–75% of network depth Mitra (2026). We show that this conflates two distinct phenomena, *encoding quality* and *routing sensitivity*, that peak at different depths and have different mechanistic origins.

We validate this decomposition on Qwen-2.5-7B-Instruct and Mistral-7B-Instruct-v0.3 using a controlled team-decision preference task: given a scenario and a preference (e.g., "escalate to senior team"), the model produces a binary decision aligned with the preference. Directions are extracted via constraint-injection contrast pairs (§5.1); we vary preference type and prompt surface form to isolate the geometric properties of preference encoding.

4.1 Two Distinct Phenomena

We measure two layer-wise quantities for each preference:

- **Encoding quality**: For each layer ℓ, extract a preference direction $\mathbf{d}_P^{(\ell)}$ from layer ℓ's hidden states via mean-difference, then apply $\mathbf{d}_P^{(\ell)}$ as steering at the same layer ℓ. Measures: how discriminative is the preference representation at layer ℓ? (We validate that this same-layer protocol does not conflate representation quality with write efficacy below.)
- **Routing sensitivity**: Extract a single reference direction $\mathbf{d}_P^{\text{ref}}$ once (at 64% depth), then apply it as steering at varying layers ℓ. Measures: how effectively does an intervention at layer ℓ propagate to the output decision?

These measure different things (one varies extraction, one varies application) and have different curves (Figure 1).

Results. Routing peaks at 48% depth in Qwen-7B and 47% in Mistral-7B, within 1 percentage point despite architectural differences, with width 3–9 layers. Encoding peaks much later: 96% in Qwen and 78% in Mistral, tracking cumulative norm growth ($r = 0.89$ between direction norm and steering effect in Qwen). The two peaks are separated by 31–50 percentage points, confirming that late layers store cleaner representations while mid layers are required for interventions to propagate. This gap motivates §5.4: extract at the encoding peak, apply at the routing peak.

This refines the 60–75% practitioner heuristic Mitra (2026): routing peaks earlier (46–47%), extraction targets later layers than application (Appendix M). Direction dominates magnitude Vu and Nguyen (2025); Pham and Nguyen (2024); quantified via 180° flip (Appendix O).

Methodology validation. A purer encoding sweep (apply all directions at fixed routing peak $L = 15$) peaks at the same layer (25, 78%) as the same-layer curve, isolating representation quality from

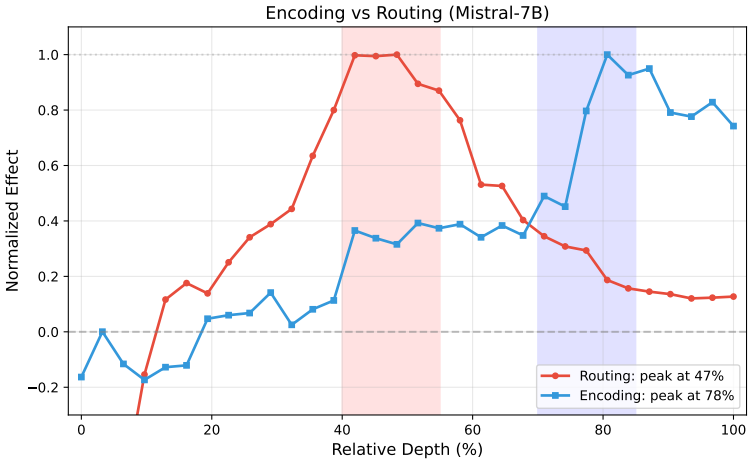

Figure 1: Encoding vs. routing in Mistral-7B. Routing (red) peaks at 47% depth; encoding (blue) peaks late at 78%. The two phenomena are distinct.

write efficacy. The routing peak is direction-independent across mid-layer references $L \in [10, 20]$. Full validation in Appendix U.

Scaling beyond 7B. The routing peak shifts later at 14B scale (Qwen-2.5-14B L29, 62% depth vs. 47–48% at 7B); relative-depth invariance holds across 7B architectures but not cleanly across scales (Appendix U).

4.2 Block-Diagonal Structure of Preference Directions

Park et al. (2024) establishes that semantically related concepts share directions in activation space (block-diagonal structure), while unrelated concepts are approximately orthogonal. We verify and quantify this for preference families.

We extract preference directions for 5 families using constraint-injection format (§5.1): team decisions, movie genres, political positions, writing styles, and AI risk personas. Within-family pairwise cosine similarities are 0.84–0.90 for 4 of 5 families (team decisions 0.90, movie genres 0.90, political positions 0.89, AI risk personas 0.84); writing styles is the exception (0.55). Cross-family similarities average 0.19, confirming near-orthogonality across semantic domains.

Extraction format matters. The same team decision preferences yield mean within-family cosine 0.58 when extracted via standard CAA multi-choice format Panickssery et al. (2024) but 0.90 when extracted via constraint-injection (holding scenario constant, varying only the preference constraint). Constraint-injection isolates the preference dimension from prompt-level variation, revealing block-diagonal structure that standard formats obscure.

What the universal axis claim does and does not entail. The high within-family cosine (0.84–0.90) reported above measures direction alignment across varied preference *content* under a fixed extraction template. To test whether this universality survives template variation, we extracted directions using five paraphrased surface forms of the same preference. Within-paraphrase pairwise cosine is 0.64 (range 0.45–0.76), demonstrating that constraint-injection extraction is partially template-dependent. The universal preference axis claim is thus best understood as *template-conditional*: within a fixed extraction template, preferences in a family share a direction; across templates, the extracted directions differ but remain causally effective on natural test inputs.

Finer-grained preferences (e.g., multiclass tag selection) span sub-clusters rather than sharing a single axis (Appendix N).

Preference and computation axes are nearly orthogonal ($\cos = -0.012$), suggesting steering does not interfere with reasoning (Appendix P).

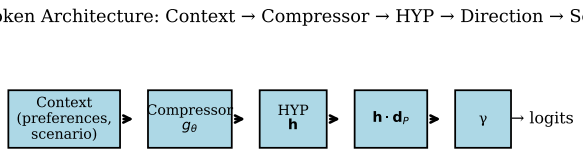

Figure 2: Context compressed to embedding, projected onto $\mathbf{d}_P$, scaled by per-team γ.

5 Method: Hypothesis Token with Decoupled Coupling

The decomposition of §4 motivates a personalization architecture that exploits both the block-diagonal structure of preference directions and the encoding-routing gap. The architecture has three stages: a one-time direction discovery (§5.1), a learned context compressor shared across teams (§5.2), and per-team coupling reduced to a single scalar (§5.3; Figure 2). Encoding-peak extraction (§5.4) is an additional ingredient.

Two coupling mechanisms. **Logit-level coupling** (Eq. 2, used in §7) adds a signed term to decision-token logits for labels $\{y_+, y_-\}$, with $s(y) = \pm 1$ for aligned/alternative; natural for binary classification. **Hidden-state coupling** adds $\sum_k z_k \cdot \mathbf{d}_k$ at a chosen layer, used for multi-axis decisions (§7.5) where signal distributes across leading tokens. For A/B decisions, hidden-state steering shifts both candidates' logits equally (§7.2), leaving margin unchanged; asymmetric $s(y) = \pm 1$ is necessary for differential scoring.

5.1 Stage 0: Preference Direction Discovery (One-Time)

The preference direction $\mathbf{d}_P$ is extracted once via mean-difference between preference-present and preference-absent hidden states on calibration examples (30 for CL-Bench, 100 for PrefEval; last 35% of tokens), then normalized. For PrefEval experiments (§7.1) we use layer $\ell = 20$; for CL-Bench (§7.2) we use $\ell = 22$ ($\sim 80\%$ depth, near the encoding peak for Qwen-7B). For layer-asymmetric application (§5.4) we use the encoding peak $\ell^{\text{enc}} = 25$. The protocol is deterministic and hyperparameter-free. The direction is a model property shared across preference types within a semantic family (§4.1).

5.2 Stage 1: Hypothesis Token Compressor (Shared)

A learned compressor g_θ maps the personalization context to a single token embedding $\mathbf{h}_{\text{HYP}} = g_\theta(x_{\text{context}}) \in \mathbb{R}^d$. The compressor is trained with a pairwise contrastive loss that discriminates preference-aligned from baseline projections onto $\mathbf{d}_P$:

$$\mathcal{L}_{\text{compress}} = -\log \frac{\exp(\pi_{\text{aligned}}/\tau)}{\exp(\pi_{\text{aligned}}/\tau) + \exp(\pi_{\text{baseline}}/\tau)}, \tag{1}$$

where $\pi = \mathbf{h}_{\text{HYP}} \cdot \mathbf{d}_P$. Cross-entropy training creates a shortcut (Appendix B); contrastive loss requires HYP projection to be the only information path. Like Stage 0, this is a one-time shared investment. The compressor is a Qwen-2.5-1.5B instance (same family as the 7B base, frozen); HYP is prepended at position 0 with its output embedding replacing the HYP slot. Architectural details in Appendix A.

5.3 Stage 2: Per-Team Scalar Coupling

With Stages 0 and 1 fixed, adapting to a new team reduces to a single scalar γ:

$$\text{logit}_{\text{steered}}(y) = \text{logit}(y) + s(y) \cdot \gamma \cdot (\mathbf{h}_{\text{HYP}} \cdot \mathbf{d}_P), \quad s(y) = \begin{cases} +1 & y = y_+ \\ -1 & y = y_- \end{cases} \tag{2}$$

The binding score is added to the preference-aligned label and subtracted from the alternative. The scalar is fitted by grid search over $\gamma \in [0.01, 1.0]$; the learned value $\gamma = 0.17$. With fixed injection strength β instead of learned γ, accuracy is invariant across $\beta \in [0.1, 1.0]$, confirming directional sufficiency: the scalar's role is *coupling*, not *discovery*.

Table 1: Layer-asymmetric steering on Mistral-7B. Encoding-peak extraction yields 3.2–$3.3\times$ larger logit swing than mid-layer baseline; layer-asymmetric application achieves the largest swing.

Configuration	Extract	Apply	Logit Swing
Routing baseline (mid-mid)	15	15	0.892
Encoding baseline (late-late)	25	25	2.873
Extract-late, apply-mid	**25**	**15**	**2.912**
Extract-mid, apply-late	15	25	0.365

5.4 Layer-Asymmetric Application

The encoding-routing decomposition predicts that extraction and application should target different layers. We test this on Mistral-7B-Instruct-v0.3 with the encoding peak ($\ell^{\text{enc}} = 25$) and routing peak ($\ell^{\text{rout}} = 15$), measuring logit swing on the correct decision token under $\gamma = 2.0$ steering.
Extract-late, apply-mid achieves the largest swing (2.912, $3.3\times$ over the mid-mid routing baseline). The opposite combination (mid-extract, late-apply) is dramatically worse (0.365), confirming that extraction-layer and application-layer roles are not interchangeable.

Swing vs. accuracy. The encoding/routing decomposition is mechanistically diagnostic: larger logit swings confirm distinct extraction and application phenomena, but accuracy gain depends on whether the task is swing-limited. CL-Bench accuracy is already strong via mid-layer coupling; PrefEval saturates at 100%. Layer-asymmetric application benefits tasks where decision margin is the bottleneck (e.g., tail outputs in free-form generation), as predicted by the reading–writing asymmetry analysis (§6).

6 From Single-Axis to Multi-Axis Steering

The single-axis method (§5) does not address multi-dimensional judgment profiles or the reading-writing asymmetry Zou et al. (2023). Naive multi-axis steering with two near-orthogonal axes ($\cos = -0.147$) fails on conservative profiles: the bottleneck is *logit leverage*, not directional interference (Appendix T).

Reading-writing asymmetry. Late-layer extraction (L25) yields $140\times$ larger writing swing than L20, indicating substantial layer-choice artifact (Appendix V).

6.1 End-to-End Learned Steering

We resolve the multi-axis failure through end-to-end learned steering, extending HYP with K learnable directions $D \in \mathbb{R}^{K \times d}$ and context-dependent coefficients $z = W_z \cdot g_\theta(x)$:

$$\text{steering} = \sum_{k=1}^{K} z_k \cdot \mathbf{d}_k. \tag{3}$$

The system is trained end-to-end with behavioral loss; no axis labels or manual γ are required. PCA on 17 preference types fails to recover risk tolerance (Appendix Q), motivating supervised learning.

7 Experiments

7.1 Standard Personalization Benchmark: PrefEval

We validate on PrefEval Zhao et al. (2025c) (ICLR 2025), a standard benchmark for LLM preference-following. We evaluate on the implicit_choice subset, which provides native multiple-choice structure where each example contains preference-aligned and preference-violating candidate responses. Scoring uses sequence log-probability over the candidate options, which is the benchmark's native evaluation protocol.

Table 2: PrefEval implicit_choice ($N = 1000$). Directional steering improves both architectures with validated causal structure; HYP matches basic steering on Qwen-7B, confirming HYP is non-disruptive when preference is explicit.

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
Architectural extension:		
HYP + scalar coupling	—	73.33

Setup. We evaluate Mistral-7B-Instruct-v0.3 and Qwen-2.5-7B-Instruct on the full $N = 1000$ test set. Preference directions are extracted via mean-difference between hidden states at the last token of preference-aligned option text versus non-aligned option text (§5.1) on 100 held-out calibration examples disjoint from the test set. Layer 20 and $\gamma = 3.0$ are selected once via grid search on a separate 15-example validation split (also disjoint from test) and held fixed across both models; no benchmark-specific test information is used for hyperparameter selection. This mean-difference extraction with additive steering corresponds to the standard activation addition protocol Zou et al. (2023); Panickssery et al. (2024); "basic steering" in Table 2 is thus a direct implementation of prior baselines applied to personalization.

Results. On Mistral-7B, basic steering saturates at 100% ($+9.89$pp); the direction flip inverts accuracy to 9.22% (80.9pp causal effect), and random direction stays near baseline (92.33%), confirming saturation reflects directional signal rather than ceiling artifact. Qwen-2.5-7B is non-saturating ($68.44 \rightarrow 73.67\%$, $+5.22$pp) with consistent causal structure (flip 61.67%, random 69.00%).

Cross-template robustness. 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 $\cos(\mathbf{d}_P^{(A)}, \mathbf{d}_P^{(B)}) = 0.91$, confirming that constraint-injection extraction recovers a shared preference axis under moderate template variation, with scalar coupling effective under template mismatch between extraction and inference.

HYP value emerges on implicit-preference tasks. PrefEval shows directional steering already achieves the benchmark's accessible ceiling: HYP compressor matches basic steering (73.33% vs. 73.67%). When preference is explicit, basic extraction captures the signal directly. HYP's architectural value emerges on implicit-preference tasks (CL-Bench), where basic hidden-state steering fails (48.8%, see §7.2).

7.2 Diagnostic Benchmark: CL-Bench

We validate on CL-Bench Dou et al. (2026) (1,899 examples, 4 categories), reformulated as A/B binding discrimination (correct answer vs. binding-degraded alternative) because open-ended CL-Bench exceeds 7B capabilities (6% task-solving rate). Anti-shortcut measures include random position swap, neutral framing, and joint log-probability scoring (Appendix J).

Key findings. The compressor is necessary (removing it $\rightarrow 0\%$); direction is specific (random direction 73.3% vs. true direction 86.7%); and per-team capacity beyond a scalar is unnecessary (1.97M-parameter per-team MLP achieves only 73.3%, LoRA 40%; Appendix I). The scalar succeeds because geometric priors (correct direction) dominate raw capacity in this low-data regime.

Why basic hidden-state steering fails. Basic activation steering ($+\mathbf{d}_P$ at layer 22) achieves $48.8\% \pm 2.5\%$ ($N = 378$, 3 seeds), indistinguishable from baseline (51.9%, flip and random also $\sim 51.9\%$). Hidden-state injection shifts both candidates' logits by approximately equal amounts,

Table 3: CL-Bench (reformulated): HYP scalar-γ (1 param) matches injection, beats 1.97M MLP.

Method	Params	Accuracy	Δ vs. Baseline
Baseline (no HYP)	—	46.7%	—
Basic steering ($+\mathbf{d}_P$)‡	—	48.8%	+2.1
HYP + no injection	—	60.0%	+13.3
HYP + injection ($\beta = 0.1$)	—	**86.7%**	+40.0
MLP adapter (2-layer)	918K	73.3%	+26.7
MLP adapter (3-layer)	1.97M	73.3%	+26.7
Scalar-γ adapter	**1**	**86.7%**	+40.0
Ablations:			
No compressor†	—	0.0%	−46.7
Random direction ($\beta = 0.1$)	—	73.3%	+26.7
Random direction ($\beta = 0.5$)	—	80.0%	+33.3

†HYP token replaced with fixed vector. ‡Basic steering on $N = 378$.

leaving the A–B margin unchanged. The asymmetric $s(y) = \pm 1$ coupling (Eq. 2) is what enables HYP's $+41.7$pp gain; HYP's contribution is the coupling mechanism, not direction quality.

Scaling and robustness. The adapter scales from 82.2% ($N = 45$) to $90.5\% \pm 1.5\%$ at full scale (3 seeds; Appendix C). Performance is consistent across CL-Bench categories (macro-avg 91.1%; Appendix D); 30-example calibration is sufficient (low cross-seed variance ± 1.5pp). Split details in Appendix E.

Injection robustness. Accuracy is invariant across $\beta \in \{0.1, 0.2, 0.5, 1.0\}$ (all 86.7%), confirming directional sufficiency.

Isolating geometric coupling. HYP-token-only (60%, $+13.3$pp) reflects implicit attention-pathway signal; explicit coupling adds 26.7pp ($60 \rightarrow 86.7\%$), isolating projection onto $\mathbf{d}_P$. Three controls rule out shortcuts: constant-HYP (0%), flip ($90.5\% \rightarrow 0\%$), context-swap (0.1%).

7.3 Causal Validation: Both Components Are Necessary

To validate the causal chain underlying the decoupled architecture, we break individual components at evaluation time while holding the rest fixed. Both components are causally necessary, with effects on the full stratified CL-Bench (3 seeds): random direction $82.5\% \pm 3.2\%$ (-8pp), constant HYP 0.0%, direction-flip inverts predictions ($90.5\% \rightarrow 0.0\%$), and context-swap yields 0.1%.

Why random direction still works partially. Random-direction 82.5%: HYP compressor alone provides structural advantage; 8pp gap isolates learned direction. Flip ($-\mathbf{d}_P \rightarrow 0\%$) deterministically inverts predictions (Appendix F).

7.4 Cross-Category Generalization

4-fold LOCO on CL-Bench achieves $91.0\% \pm 2.8\%$ on unseen categories, comparable to in-distribution (Appendix G). Cross-architecture generalization: Mistral/Qwen (§7.1); Gemma (Appendix K).

7.5 Multi-Axis Steering Results

Multi-axis experiments evaluate short generated responses classified by decision verbs (approve/deny, recommend/decline), isolating the multi-axis coupling mechanism in a structured decision setting where the output signal is localized to the leading tokens.

Results. Single-axis steering achieves 100% ($20/20$) across three team profiles on loan decisions, confirming per-team scalar coupling suffices for single-dimension preferences. Multi-axis E2E

steering achieves 91.2% ($146/160$, $N = 40$ per team) with two learned axes ($\cos = 0.117$), and transfers zero-shot to hiring decisions at 100% (Appendix R).

7.6 Geometric Structure of Preference Encoding

Three properties motivate the directional framing: 100% linear separability of preference-encoding hidden states (PCA, $L \geq 22$), $2.2\times$ trajectory divergence after commitment (0.77 vs. 0.35, $p < 0.001$), and $2.3\times$ amplified logit gap at commitment tokens. Full measurements in Appendix S.

8 Analysis

Why 1 scalar beats 1.97M parameters. The scalar-γ adapter succeeds via the decoupling principle: extraction and coupling are separated, with coupling receiving the correct inductive bias (the preference direction). MLPs must discover both direction and scaling from 45 examples; geometric priors dominate raw capacity in this low-data regime.

When learned transformations hurt. When direct extraction already yields a near-optimal direction (as observed when preference is explicitly stated in prompt, e.g., PrefEval), adding learned parametric transformations (HYP compressor) plateaus at comparable performance, providing no marginal value. The decoupling principle thus has two regimes: HYP compression adds value when preference is implicit in dynamic context (CL-Bench); when preference is explicit and directly extractable, basic steering suffices.

9 Discussion and Conclusion

Personalization in LLMs decomposes more cleanly than current methods assume: extraction is a one-time shared investment, and per-team coupling on binary tasks reduces to a single scalar. The depth-dependence of activation steering decomposes into two distinct phenomena (encoding at 78–96%, routing at 46–48% at 7B, later at 14B), and extracting at the encoding peak yields $3.2\times$ larger logit swing than mid-layer extraction. The reading-writing asymmetry observed by Zou et al. (2023) is at least partially a layer-choice artifact: late-layer extraction yields $140\times$ larger writing effect than at intermediate layers.

Practical implications and open problems. For structured decisions, deploying personalized models for hundreds of teams requires only a base model paired with per-team scalars, each interpretable as preference intensity. The picture is more nuanced for fine-grained multiclass personalization: when the preference space spans multiple sub-clusters within a domain, a single scalar per team may not suffice, motivating the mixture-of-directions extension noted in future work.

Limitations. Depth decomposition is validated on two 7B architectures; the 14B routing peak shifts to 62%, indicating relative-depth invariance does not extend uniformly; 70B+ untested. **Scope: structured decisions.** Our coupling is designed for tasks with explicit candidate labels (binary and multiclass classification), where the geometric decoupling principle directly applies. Free-form generation requires distributed coupling mechanisms beyond the scope of this work. Constraint-injection shows template variation (0.64 across paraphrases vs. 0.84–0.90 within), though scalar coupling remains effective under template mismatch (§7.1, 73.9% vs. 73.7%). Direction stability is supported by low CL-Bench variance (± 1.5pp) and LOCO generalization ($91.0\% \pm 2.8\%$). Steering could entrench biases; per-team scalars provide audit interpretability. Compute-matched comparison to CoPL Choi et al. (2025) and BiPO Cao et al. (2024) is left to future work.

Future work. The encoding-routing decomposition motivates mechanistic study of other interpretable phenomena (refusal, factuality, sentiment). Fine-grained multiclass personalization suggests three directions: *retrieval-based selection* (precomputed direction library indexed by user-history), *hybrid context-direction selection* (combining HYP context with retrieval in a learned selector), and *mixture-of-directions coupling* (soft mixtures over candidate directions).

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.

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.

John X. Morris, Niloofar Mireshghallah, Mark Ibrahim, and Saeed Mahloujifar. Learning to reason in 13 parameters. *arXiv preprint arXiv:2602.04118*, 2026.

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.

Subhadip Mitra. Activation Steering in 2026: A Practitioner's Field Guide. `https://subhadipmitra.com/blog/2026/activation-steering-field-guide/`, 2026.

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.

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

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.

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.

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.

Minhyeon Oh, Seungjoon Lee, and Jungseul Ok. Comparison-based Active Preference Learning for Multi-dimensional Personalization. *arXiv preprint arXiv:2411.00524*, 2024.

Yijiang River Dong, Tiancheng Hu, Yinhong Liu, Ahmet Üstün, and Nigel Collier. When Personalization Meets Reality: A Multi-Faceted Analysis of Personalized Preference Learning. *arXiv preprint arXiv:2502.19158*, 2025.

Junlin Lu, Patrick Mannion, and Karl Mason. Inferring Preferences from Demonstrations in Multi-objective Reinforcement Learning: A Dynamic Weight-based Approach. *arXiv preprint arXiv:2304.14115*, 2023.

Simone Drago, Marco Mussi, and Alberto Maria Metelli. Towards Theoretical Understanding of Sequential Decision Making with Preference Feedback. In *ICML*, 2025.

Tian Huang, Shengbo Wang, and Ke Li. Direct Preference-Based Evolutionary Multi-Objective Optimization with Dueling Bandit. *arXiv preprint arXiv:2311.14003*, 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.

Xianwei Cao, Dou Quan, Zhenliang Zhang, and Shuang Wang. Learning What Matters Now: Dynamic Preference Inference under Contextual Shifts. *arXiv preprint arXiv:2603.22813*, 2026.

A. R. M. Wolfert. Preference-Based Optimisation in 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.

A Experimental Details

Models. Discovery experiments (§4) use Qwen-2.5-7B-Instruct. Hypothesis Token experiments (§7) use Qwen-2.5-1.5B-Instruct (compressor) with Qwen-2.5-7B-Instruct (base model). Cross-model experiments use GPT-4o, GPT-4.1, GPT-5.2, DeepSeek-V3, Kimi-K2, o3, and DeepSeek-R1 via API. DPO experiments fine-tune Qwen-2.5-7B-Instruct with LoRA (rank 16, $\alpha = 32$).

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. We use 30 calibration examples (not included in train or test).

Hypothesis Token training. Stage 1: 45 training examples, 5 epochs, learning rate 5×10^{-5}, contrastive loss with temperature $\tau = 0.07$. Stage 2: scalar-γ fitted on 15 validation examples via grid search over $\gamma \in [0.01, 1.0]$.

E2E multi-axis training. $K = 2$ directions, trained on 32 examples (8 scenarios $\times$ 4 teams) for 10 epochs with Adam (lr $= 5 \times 10^{-4}$), behavioral loss (cross-entropy on A/B decision token). No regularization or scheduling. Convergence: loss $1.28 \rightarrow 0.00$ in $\sim$10 epochs.

HYP architecture details. Compressor model: Qwen-2.5-1.5B-Instruct (1.5B parameters). **Base model**: Qwen-2.5-7B-Instruct (7B parameters, frozen). **Token position**: HYP token is prepended to the prompt at position 0; the compressor's output embedding replaces the HYP position in the base model's input embeddings. **Embedding dimension**: $d = 3584$ (matching Qwen-7B hidden size). **Training**: Adam optimizer, learning rate 5×10^{-5}, batch size 1, 5 epochs on 45 training examples. **Contrastive loss**: temperature $\tau = 0.07$, pairwise (one preference-aligned, one baseline per example). **Gradient isolation**: base model weights frozen via `torch.no_grad()` on the base forward pass; only compressor parameters receive gradients. **Injection site**: final logits at the decision token (last generated position before the A/B choice). This architectural separation prevents the model from bypassing the compressor by extracting preference information directly from prompt tokens.

B Additional Ablations

CE vs. contrastive training. Cross-entropy trained compressors achieve 73.3% clean accuracy (matching MLP adapters) but with zero information flow through HYP: the compressor output is constant across inputs (cosine similarity > 0.99 between all output pairs). This confirms that CE enables prompt-token shortcuts that bypass genuine preference compression.

C Scaling Behavior

Table 4: Scaling behavior across CL-Bench (3 seeds each). $N = 45$ is a pilot subset; $N = 1500$ (stratified) is the main result.

	$N = 45$	$N = 200$	$N = 1500$ (strat.)
Scalar-γ adapter	82.2 ± 6.3	85.3 ± 6.2	$\mathbf{90.5 \pm 1.5}$
Injection	77.8 ± 12.6	85.3 ± 4.1	90.5 ± 1.5
Random dir. injection	73.3	54.0 ± 5.9	82.5 ± 3.2

D Per-Category Breakdown

Table 5: Per-category BDA on the stratified full CL-Bench (3 seeds, 399 test samples).

Category	N_{test}	BDA (%)	$\pm$ 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

E 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.

F Causal Validation Details

Table 6: Causal ablation on CL-Bench (3 seeds). Both direction and compressor are necessary; flip/swap controls rule out shortcuts.

Condition	BDA (%)	Δ	Interpretation
Full system (main)	$\mathbf{90.5 \pm 1.5}$	—	—
Random direction	82.5 ± 3.2	-8.0	Direction matters
Constant HYP	0.0 ± 0.0	-90.5	Compressor essential
Direction flip ($-\mathbf{d}_P$)	0.0 ± 0.0	—	Semantic, not positional
Context swap	0.1 ± 0.2	—	Anti-cheat validated

Compressor is essential. Replacing the compressor output with a fixed random vector produces 0% across all seeds; identical margins for all inputs, confirming that context-specific compression is necessary.

Direction is semantic. Random-direction injection achieves 82.5%, above chance but significantly below the main result ($\Delta = -8.0\%$). The partial effectiveness is expected: the HYP bottleneck alone provides structural advantage, but the *learned* direction provides the additional geometric signal. The direction-flip control ($-\mathbf{d}_P \rightarrow 0.0\%$) provides stronger causal evidence: negating the true direction deterministically inverts all predictions, confirming semantic content.

Anti-cheat: direction flip and context swap. Negating the preference direction inverts all predictions ($90.5\% \rightarrow 0.0\%$), ruling out position bias. Swapping contexts produces the same inversion (0.1%), confirming the model uses actual context content.

G LOCO Analysis Details

Table 7: Leave-one-category-out cross-validation on CL-Bench. The compressor generalizes to entirely unseen reasoning categories.

Held-Out Category	N_{test}	Adapter	Random Dir	Const. HYP
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

The scalar-γ adapter achieves $91.0\% \pm 2.8\%$ on completely unseen categories (comparable to in-distribution 90.5%), with low variance across folds. Because the compressor learns category-agnostic binding, the mechanism transfers to entirely unseen reasoning regimes. Ablation controls replicate out-of-distribution: random direction degrades to 51.7%; constant HYP produces 0%. This validates the decoupling principle: extraction captures generic binding structure, not category-specific patterns.

H Distillation: Preference-Filtered Training

We apply preference-filtered distillation: the student is trained on teacher reasoning trajectories that successfully maintained preference identity, with explicit commitment structure.

Table 8: Preference-filtered distillation with domain-appropriate guidance.

Model	CommitAcc	95% CI	WUPV	DIFR
Teacher (structured)	95.4%	[93.7, 97.1]	0.061	0.089
Student (distilled)	**96.7%**	[94.8, 98.3]	**0.044**	**0.061**

The student's WUPV drops by 28% (0.044 vs. 0.061), and DIFR drops by 31% (0.061 vs. 0.089). Domain-appropriate guidance is critical: abstract commitment instructions drop CommitAcc from 96.7% to 39.2%.

I Why Explicit Coupling Is Necessary

A natural question is whether trajectory shaping alone (without the HYP architecture's explicit geometric coupling) can improve CL-Bench accuracy. We report a series of trajectory-shaped DPO experiments on Qwen-2.5-7B-Instruct that demonstrate a consistent finding: *representation separation improves, but decisions do not flip*.

Setup. 60 CL-Bench examples (41 Rule System, 12 Procedural, 7 Domain Knowledge); 45 train, 15 test. LoRA at layers [24,25,26,27] (180K parameters). DPO + trajectory loss ($w = 1.0$), 5 epochs. Scoring: mean log-probability (corrected from sum to address length bias).

Table 9: CL-Bench trajectory DPO debugging progression. Margin improves steadily, but accuracy never budges from 40% until the Hypothesis Token + logit injection approach (§7) directly couples the geometric signal to decisions.

Exp	Method	Accuracy	Margin
2	Per-instance self-direction	$40\% \rightarrow 40\%$	$+0.27 \rightarrow +1.02$
3	+ EMA z-score normalization	Fundamentally flawed*	
4	+ Frozen μ_0/σ_0 z-score	Math bug discovered**	
5	Global binding direction (fixed)	$40\% \rightarrow 40\%$	Δ_z positive
6	+ α calibration	(in progress)	

Experiment progression. *EMA z-score "eats" improvement: as the model improves, the running mean adapts, treating gains as the new normal. $\Delta_{\text{traj},z}$ collapses from $0.398 \rightarrow 0.020$ while raw Δ_{traj} increases from $313 \rightarrow 335$.

**Per-instance direction $\mathbf{d}_x = (\mathbf{h}_c - \mathbf{h}_r)/\|\mathbf{h}_c - \mathbf{h}_r\|$ yields projection = L2 norm (always positive), not a meaningful alignment score. Switching to a global binding direction (pre-computed from training data mean difference) resolves this.

Interpretation. These experiments provide an ablation of the HYP architecture's design: trajectory shaping improves the model's internal geometry (margin $3.8\times$) but this improved representation does not automatically translate to correct decisions (accuracy flat at 40%). The HYP architecture's explicit geometric coupling (projecting the compressed binding state onto the preference direction and injecting into logits) bridges exactly this gap, achieving 86.7% on the same CL-Bench pilot tasks ($N = 45$). This stark contrast (LoRA with 180K parameters stuck at 40% versus a single scalar parameter reaching 86.7%) demonstrates that *knowing the right direction* matters far more than model capacity for preference steering. This validates a core design principle: when the relevant structure is known to be directional, explicit coupling along that direction outperforms implicit learning from data.

J 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.

Table 10: Cross-architecture direction validation. All models show consistent directional structure ($\cos > 0.5$).

Model	Layers	Dir. 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
Gemma-4-E4B-it	42	0.52	15pp

K Cross-Architecture Direction Validation

Direction consistency ($\cos > 0.5$) confirms that conservative/aggressive preferences encode as separable directions regardless of architecture.

L Gemma-2-2B Depth Analysis

Gemma-2-2B-it shows weak steering effects overall, with maximum routing effect 0.009 compared to 0.111 (Qwen-7B) and 2.10 (Mistral-7B). The encoding-routing decomposition pattern is present but attenuated: routing peaks at 77% depth (layer 20 of 26) and encoding peaks at 96% (layer 25). The weak absolute effects prevent meaningful comparison of layer-asymmetric application on this architecture. We attribute this to smaller model capacity rather than architectural differences in how preferences are encoded.

M Relation to Prior Work

Practitioner observations of steering effectiveness at 60–75% depth Mitra (2026) reflect the routing sensitivity phenomenon: the mid-to-late layer window where interventions propagate. Our decomposition refines this in two ways. First, we identify the routing peak more precisely (46–47% for 7B models), which is earlier than common practitioner choices (60–75%). Second, we show that direction *extraction* should target later layers (78–96%) than direction *application*, inverting the common assumption that extraction and application should occur at the same layer.

The Linear Representation Hypothesis Park et al. (2024) provides theoretical grounding for why preferences encode as directions, and our cross-architecture results extend Park et al. (2024)'s block-diagonal characterization to preference families with quantitative depth dependence.

N 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. Block-diagonal structure thus operates at multiple granularity levels: narrow semantic clusters share an axis, but a domain spanning multiple sub-clusters does not.

O Direction Dominates Magnitude

To verify the directional assumption underlying our method, we decompose preference interventions into direction and magnitude components.

The 180° flip produces a 23.38-logit swing while $3\times$ magnitude scaling produces only 2.34. Halving the norm produces no measurable effect ($\Delta = 0.00$). This confirms the directional assumption: interventions act on the direction component, not the magnitude.

Table 11: Directional vs. radial interventions. Consistent with prior work establishing the primacy of direction 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
	Orthogonal rotation	1.44
	90° rotation	0.75
	180° flip	**23.38**

P Computation Orthogonality

A natural concern is whether preference steering interferes with the model's reasoning capacity. We extract a "computation axis" from math-reasoning tasks and measure its similarity to the preference axis: $\cos(\mathbf{d}_{\text{pref}}, \mathbf{d}_{\text{comp}}) = -0.012$ (Qwen-2.5-7B-Instruct). The two axes are effectively orthogonal: personalization and computation occupy independent subspaces. We confirm this empirically through the absence of measurable reasoning quality degradation under preference steering at recommended layers, though this property is verified on a single model and may not generalize to all architectures.

Q PCA Analysis

Unsupervised decomposition via PCA on 17 diverse preference types finds that PC1 aligns with explanation depth ($\cos = 0.830$ with manual axis) but risk tolerance does not emerge ($\cos = -0.436$). Top-3 PCs explain only 57.4% of variance. PCA recovers high-variance surface features (verbosity, transparency) but misses low-variance judgment axes (risk tolerance), precisely the axes that matter for team differentiation. This motivates supervised E2E learning for discovering behaviorally relevant steering directions.

R Cross-Domain Transfer

Table 12: Cross-domain transfer: E2E compressor trained on loan decisions, tested zero-shot on hiring decisions ($N = 20$ pilot). Larger-scale evaluation ($N = 40$) achieves 91.2%.

Team	Loan (train domain)	Hiring (zero-shot)
A (Conservative + Brief)	20/20 = 100%	10/10 = 100%
B (Conservative + Detailed)	20/20 = 100%	10/10 = 100%
C (Aggressive + Brief)	20/20 = 100%	10/10 = 100%
D (Aggressive + Detailed)	20/20 = 100%	10/10 = 100%

The E2E compressor, trained exclusively on loan decisions, achieves 100% accuracy when tested zero-shot on hiring decisions. This demonstrates that learned steering captures team judgment style rather than domain-specific patterns.

S Geometric Structure: Detailed Measurements

Linear separability and semantic encoding. Hidden states at the preference-encoding position are 100% linearly separable (5-fold CV) via PCA at layer 22+ on 1.5B models, with separability persisting across surface form variations, confirming that the model encodes preference *semantics*, not surface patterns.

Dynamic trajectory divergence. After the commitment point in dynamic reasoning, correct decisions show significantly higher trajectory divergence from baseline (0.77 vs. 0.35 at position 6, $p < 0.001$), indicating that successful preference binding corresponds to a measurable geometric departure in activation space.

Attention geometry and decision confidence. Under structured preference binding, attention entanglement decreases by 33.6% and effective rank increases by 24%, indicating more structured attention patterns. At the commitment token, the logit gap is 2.3× larger for preference-aligned conditions (0.77 vs. 0.33), revealing that preference binding amplifies decision confidence.

T Naive Multi-Axis Composition: Detailed Analysis

We define two near-orthogonal judgment axes ($\cos = -0.147$): risk tolerance and explanation depth, yielding four team profiles (A–D). Naïve multi-axis steering injects $\gamma_{\text{risk}} \cdot \mathbf{d}_{\text{risk}} + \gamma_{\text{exp}} \cdot \mathbf{d}_{\text{exp}}$ at layer 20.

Team A fails (0%) even after orthogonalization ($\cos \to 0.00002$): the bottleneck is not directional interference but *logit leverage*: the model has a ∼17-point logit gap favoring aggressive actions that hidden-state steering at layer 20 cannot overcome.

U Routing Curve Validation and Scaling Details

Methodology validation. The same-layer encoding protocol could conflate representation quality with write efficacy (ease of intervention at that layer). To isolate representation quality, we run a *purer* encoding sweep: extract at varying layers ℓ but apply all directions at the *fixed* routing peak (layer 15 for Mistral). This measures only how discriminative the extracted direction is, independent of where it is applied. The purer curve peaks at the same layer (25, 78% depth) as the same-layer curve, confirming that the late-layer encoding peak reflects true representation quality, not write efficacy.

Robustness to reference direction. The routing 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 routing 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, confirming the routing peak is direction-independent for mid-layer reference extraction. 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 routing site (layer 15), indicating that the mid-layer routing peak ($\sim 47\%$) remains the direction-independent component of the routing curve while late-layer directions additionally engage layer-local mechanisms.

Scaling beyond 7B. The routing peak measured at 46–48% depth is consistent across two 7B architectures (Mistral-7B at 47%, Qwen-7B at 48%). We test whether this relative-depth invariance extends to larger scale by running the same routing sweep on Qwen-2.5-14B-Instruct (48 layers) with identical protocol. The Qwen-14B routing peak occurs at layer 29 (62% depth), shifted later than the Qwen-7B peak (48% depth) and the Mistral-7B peak (47% depth). The relative-depth invariance therefore holds across architectures at 7B scale but does not extend cleanly to 14B within the same architecture family. Two interpretations are consistent with this pattern: the routing peak may track absolute layer count rather than relative depth (Qwen-14B's L29 is approximately twice Qwen-7B's L13, matching the model size ratio), or larger models may engage richer late-layer processing that extends the effective routing region. Distinguishing these hypotheses requires intermediate scales we did not test. For the layer-asymmetric application, we use the empirically measured routing peak for the specific model under study rather than relying on a fixed relative-depth heuristic.

V Reading-Writing Asymmetry: Detailed Quantification

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 show high reading accuracy ($> 70\%$) combined with near-zero writing swing.

Layer dependence. Comparing the same extraction mechanism at different layers reveals that writing swing is strongly layer-dependent: extraction at the encoding 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.