Basis Vectors in Persona Space: Stable Low-Rank Geometry with Limited Causal Steering Gains

Ari Holtzman and Idea-Explorer

Abstract

Persona steering in large language models is often driven by manually crafted vectors, but it remains unclear whether persona representations share a compact, reusable basis in residual space. We test this question on a real pretrained model (QWEN2.5-0.5B) using three datasets: PERSONAHUB for vector-bank construction, PERSONA-CHAT for behavioral evaluation, and MYPERSONALITY for external trait alignment. Our pipeline extracts persona-conditioned residual vectors, fits PCA, and evaluates geometry, stability, and causal interventions.

We find strong low-rank structure: the top 10 principal components (PCs) explain 57.14% of variance and the top 20 explain 70.06%. Subspace stability is high across bootstrap splits (mean overlap 0.905 ± 0.016), and variance capture substantially exceeds random 10D bases. However, causal steering gains are limited. In directional Wilcoxon tests, PC- versus base is significant ($p = 0.0228$, $d = -0.309$), while PC+ does not significantly outperform either base or RANDOM+ ($p = 0.1587$ for both comparisons). External validity is weak: no PC–Big-5 correlations survive Benjamini-Hochberg correction, and the maximum absolute raw correlation is $|r| = 0.0918$.

These results support a clear geometric claim: persona vectors occupy a stable low-dimensional subspace. At the same time, they caution against treating top PCs as universally reliable control knobs. We conclude that PCA is promising for persona analysis and compression, but robust persona control needs stronger objectives and richer evaluation metrics.

1 Introduction

Persona control is central to personalized and safe LLM deployment, yet most steering practice still relies on brittle, handcrafted directions. Our main question is simple: can persona behavior be organized into a reusable basis in residual space, rather than a large set of one-off vectors?

Why this matters. If persona representations share a low-rank structure, we can reduce control complexity, improve interpretability, and transfer directions across prompts more reliably. This directly affects controllable generation in dialogue systems, assistant personalization, and mechanistic interpretability.

What is missing in prior work? Recent studies show that residual streams encode useful geometry and that activation steering can change behavior [Shai et al., 2024, Cao et al., 2024, Højer et al., 2025, Venhoff et al., 2025]. However, prior work has not directly tested whether a broad persona vector bank decomposes into stable principal components that remain causally useful under intervention. At the same time, reliability studies show steering can degrade across prompt and domain conditions [Braun, 2026, You et al., 2026]. Persona-focused evaluations also report mismatches between intended and produced traits [Liu et al., 2024, Shin et al., 2025].

Our approach. We build an end-to-end pipeline: collect persona vectors from PERSONAHUB, fit PCA in the residual stream of QWEN2.5-0.5B, test subspace stability under bootstrap splits, perform directional interventions on PERSONA-CHAT, and evaluate external alignment with Big-5 labels from MYPERSONALITY. We report both effect sizes and significance tests.

What we find. Geometry is strong and stable: top-10 and top-20 PCs explain 57.14% and 70.06% variance, with bootstrap overlap 0.905 ± 0.016. Causal effects are asymmetric: PC- significantly shifts lexical steering score relative to base, but PC+ does not significantly beat base or random steering. External trait transfer is weak after multiple-testing correction.

Contributions.

- We propose a reproducible pipeline for persona-basis discovery in residual space using PCA and causal interventions.
- We conduct a multi-part evaluation across geometry, subspace stability, steering outcomes, and external trait alignment.
- We show a clear separation between descriptive geometry and controllable behavior: low-rank structure is strong, while steering gains are modest.
- We provide empirical evidence that persona components are layer-sensitive, with low cross-layer overlap between middle and late layers.

Paper organization. Section 2 situates our study in activation steering and persona modeling. Section 3 describes datasets, extraction, PCA, and evaluation. Section 4 reports quantitative findings. Section 5 interprets implications and limitations, and section 6 concludes.

2 Related Work

Residual-stream geometry and linear structure. A growing line of work argues that transformer residual states encode meaningful geometry, including belief-state organization and task-relevant linear directions [Shai et al., 2024]. Representation-engineering methods further show that simple activation directions can improve downstream behavior in some settings [Højer et al., 2025, Venhoff et al., 2025]. Building on this evidence, we test whether persona-specific vectors also admit a compact linear basis.

Persona steering and personalization. Personalized control methods increasingly use inference-time interventions rather than full fine-tuning. Bi-directional preference optimization learns steering vectors for personalization [Cao et al., 2024], and persona-conditioned dialogue traces back to PERSONA-CHAT [Zhang et al., 2018]. Unlike these methods, we focus on decomposition: whether many persona vectors can be summarized by a few shared PCs that remain stable and useful.

Reliability limits of steering vectors. Recent work highlights unstable steering behavior across prompts and domains [Braun, 2026]. Spherical steering proposes norm-preserving rotations as an alternative to additive interventions [You et al., 2026]. Our results are consistent with these concerns: we observe strong geometry but only partial causal gains with additive PC steering.

Evaluation of persona faithfulness. Persona-driven generation can exhibit bias, inconsistency, and out-of-character behavior [Liu et al., 2024, Shin et al., 2025]. These findings motivate our use of held-out dialogue prompts and external Big-5 alignment checks, rather than relying only on in-distribution geometric summaries.

Cross-layer decomposition and interpretability. Multi-layer sparse autoencoder analyses show that representation structure can vary by layer and that cross-layer sharing is nontrivial [Lawson et al., 2024]. Our cross-layer overlap result (0.280 between layers 12 and 23 for top-10 subspaces) similarly suggests layer-local persona bases.

3 Methodology

Research question. We ask whether persona representations in a transformer residual stream decompose into shared basis directions that are stable and causally meaningful.

3.1 Data and Preprocessing

We use three datasets. PERSONAHUB provides diverse persona descriptions for vector-bank construction. PERSONA-CHAT provides held-out dialogue prompts for steering evaluation. MYPERSONALITY provides text with binary Big-5 labels for external alignment.

Preprocessing follows a fixed pipeline: whitespace normalization, uniform sampling without replacement ($n = 800$ for persona vectors and trait analysis), persona prompt construction in the template "System: You are defined by this persona: ...", and Big-5 label parsing from JSON answers. Data-quality checks report 0% missing values in audited subsets, 0 duplicates in checked PERSONAHUB slices, and 0/500 parse failures for MYPERSONALITY labels.

3.2 Model, Representation Extraction, and PCA

We use QWEN2.5-0.5B (24 layers, hidden size 896), extracting residual activations primarily at layer 23 (with a cross-layer comparison at layer 12). For each persona prompt, we collect hidden-state vectors with max length 128 and batch size 64. We center vectors and fit PCA with 20 components.

Geometry metrics. We report component-wise and cumulative explained variance ratio (EVR), including top-10 and top-20 EVR.

Stability metrics. We estimate subspace overlap between top-k PCs across bootstrap splits using principal-angle overlap. We also compare to random orthonormal bases and compute a cross-layer top-10 overlap score.

3.3 Causal Steering Evaluation

We intervene during generation by adding or subtracting principal directions at a target residual layer with intervention strength $\alpha = 8.0$. Conditions are: base (no intervention), PC+, PC-, and RANDOM+ (random direction with matched norm). We generate up to 24 new tokens per prompt on held-out PERSONA-CHAT validation histories (48 prompts, 45 usable).

Steering score. We use a lexical score defined as pos_hits $-$ neg_hits. This metric is intentionally simple and fast but may miss semantic changes not expressed through lexicon tokens.

Statistical testing. We run directional Wilcoxon signed-rank tests for pairwise comparisons, report 95% confidence intervals for mean deltas, and include standardized effect sizes (d).

3.4 External Trait Alignment

For external validity, we project MYPERSONALITY texts onto PC1–PC5 and compute Pearson correlations with Big-5 binary labels. We apply Benjamini-Hochberg FDR correction across tested correlations.

3.5 Implementation and Reproducibility

We use Python 3.12.8, torch 2.5.1+cu124, transformers 4.51.3, datasets 4.6.1, scikit-learn 1.8.0, scipy 1.17.1, and statsmodels 0.14.6. All random sources are seeded at 42 (Python, NumPy, PyTorch, CUDA). Experiments run on 2× NVIDIA RTX 3090 GPUs (24GB each) with mixed precision enabled. Runtime is approximately 22 seconds per full run.

4 Results

Overview. We report geometry, stability, causal steering, and external trait alignment. Across analyses, we see a consistent pattern: strong low-rank structure but only partial behavioral leverage from top PCs.

Table 1: Geometry and stability metrics for persona vectors in layer 23 unless noted. Higher is better for EVR and overlap metrics.

Metric	Value
EVR top-10	**0.5714**
EVR top-20	**0.7006**
PC1 EVR	0.1515
PC2 EVR	0.0872
Split subspace overlap (mean ± sd)	**0.9050** ± **0.0163**
Random 10D basis variance ratio (mean)	0.0113
Cross-layer overlap (L12 vs L23, top-10)	0.2802

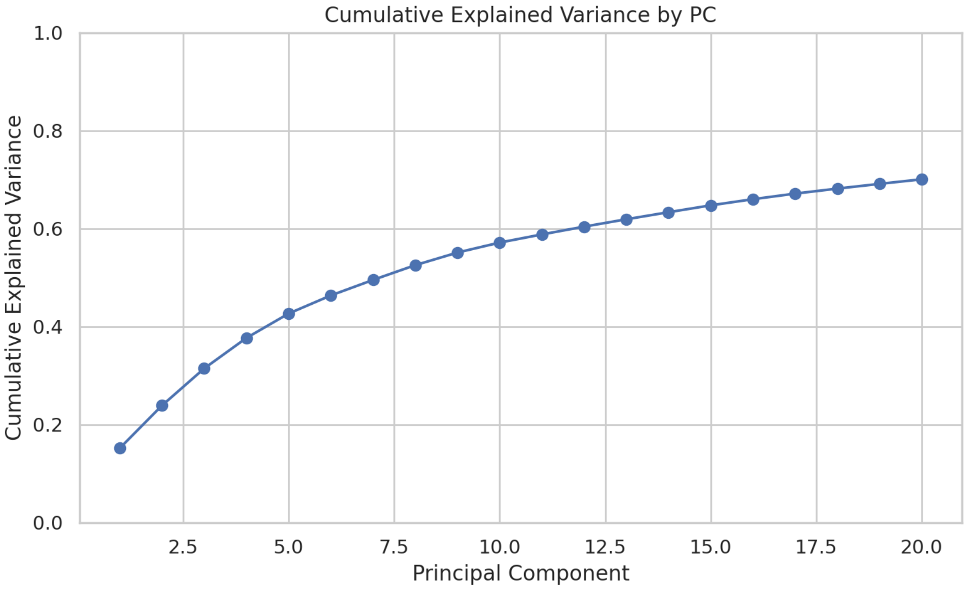

Figure 1: PCA scree plot for persona vectors at layer 23. The first components capture most variance, indicating strong low-rank structure.

4.1 Low-Rank Geometry and Stability

Figure 1 shows a steep early eigenvalue spectrum, and table 1 quantifies the effect: top-10 PCs explain 57.14% variance and top-20 explain 70.06%. This is far above the random 10D basis variance ratio (0.0113), supporting H1 (low-dimensional structure). Subspace overlap across bootstrap splits is high (0.9050 ± 0.0163), supporting H4 (within-layer stability).

Cross-layer transfer is weaker: top-10 overlap between layers 12 and 23 is 0.2802. This suggests persona bases are at least partly layer-specific.

4.2 Causal Steering Effects

Condition means show directional separation in lexical steering score: Base $= 0.0417$, PC+ $=$ 0.1250, PC- $= -0.0417$, and RANDOM+ $= 0.1042$. Figure 2 visualizes this spread.

Directional Wilcoxon tests (abreftab:steering) show one significant result: PC- vs base ($p = 0.0228$, $d = -0.309$). By contrast, PC+ is not significant against base or RANDOM+ ($p = 0.1587$ in both tests). This partially supports H2 (causal usefulness) but fails to show robust advantage over random positive steering.

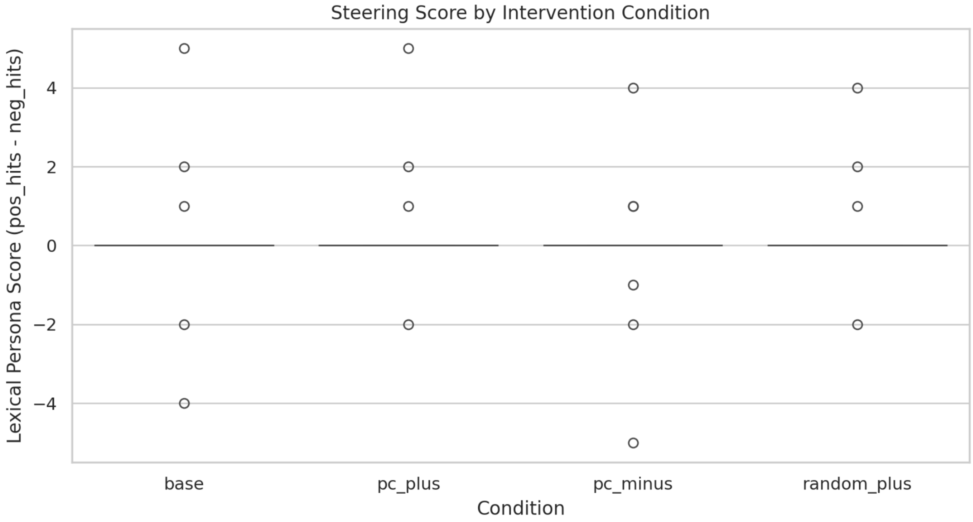

Figure 2: Distribution of lexical steering scores by intervention condition. PC- shifts downward relative to base; PC+ does not clearly separate from random steering.

Table 2: Steering outcomes on held-out PERSONA-CHAT prompts. We report condition means and directional Wilcoxon tests with 95% confidence intervals. Best directional significance is in **bold**.

Comparison	Mean Δ	95% CI	p-value	Effect size (d)
PC+ vs Base	0.0889	[0.0000, 0.2667]	0.1587	0.149
PC+ vs RANDOM+	0.0222	[0.0000, 0.0667]	0.1587	0.149
PC- vs Base	-0.0889	[-0.1778, -0.0222]	**0.0228**	-0.309

4.3 External Big-5 Alignment

Trait alignment is weak. No PC–trait correlation survives BH-FDR correction, and the maximum absolute raw correlation is $|r| = 0.0918$. Therefore H3 is not supported in this setup.

4.4 Failure Patterns

Manual inspection of generation rows identifies three recurring issues: (1) generic fallback completions reduce lexical sensitivity, (2) prompt echoing limits persona expression, and (3) semantic persona shifts can occur without lexicon hits, which lowers measured score gains.

5 Discussion

Interpreting the geometry-behavior gap. The central empirical pattern is a mismatch between descriptive structure and controllability. PCA reveals a stable low-rank persona manifold, but top directions do not consistently deliver strong positive steering gains. This supports the view that not every high-variance component is a strong causal control direction [Braun, 2026].

What is likely happening? A plausible explanation is that leading PCs mix multiple factors, including format and style artifacts, while lexical scoring captures only a narrow slice of persona behavior. The significant PC- effect suggests some components do encode behaviorally relevant polarity, but additive interventions along PC1 alone are insufficient for robust positive control.

Implications for persona steering. PCA remains useful for compression, diagnostics, and subspace characterization. For deployment-grade persona control, however, stronger objectives are needed:

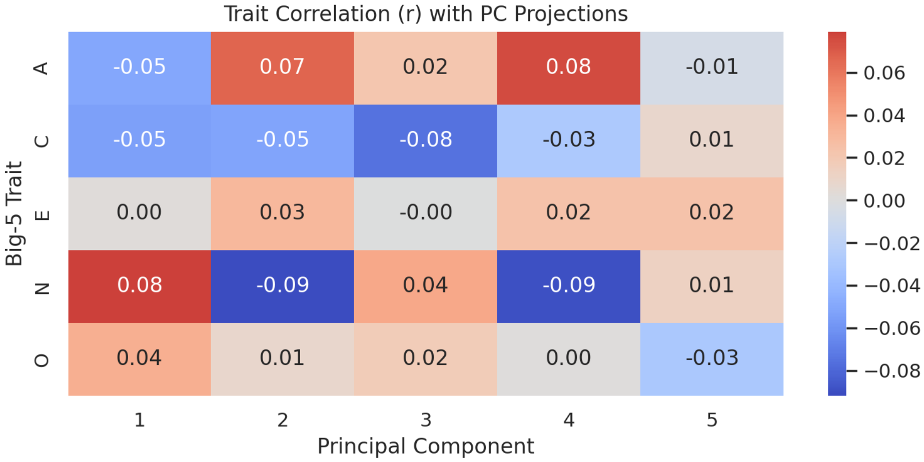

Figure 3: Correlation heatmap between PC1–PC5 projections and Big-5 labels on MYPERSONALITY. Correlations are uniformly small.

classifier-based fidelity targets, out-of-character detectors [Shin et al., 2025], and interventions that better respect geometry (e.g., rotational methods [You et al., 2026]).

Limitations. Our study uses one model (QWEN2.5-0.5B) and modest sample sizes, with a coarse lexical metric and no human fidelity judgments. We do not include spherical steering or SAE-based decomposition baselines in this run, and cross-model generalization is not yet tested.

Broader impact. Better persona control can improve user experience and consistency, but it can also increase risks of manipulation and identity overfitting. Mechanistic transparency and robust evaluation should be treated as safety requirements, not optional diagnostics.

6 Conclusion

We studied whether persona representations in transformer residual streams decompose into shared basis vectors. Using QWEN2.5-0.5B and three persona-related datasets, we found strong, reproducible low-rank geometry: top-10 PCs explain 57.14% of variance, top-20 explain 70.06%, and bootstrap overlap reaches 0.905 ± 0.016.

Causal results were mixed. Negative-direction steering was significant relative to baseline ($p = 0.0228$, $d = -0.309$), while positive-direction steering did not significantly outperform baseline or random directions. External Big-5 alignment was weak after multiple-testing correction.

The key takeaway is that PCA basis discovery is effective for persona representation analysis, but geometric prominence alone does not guarantee robust controllability. Immediate next steps are to replace lexical scoring with classifier-based fidelity and out-of-character metrics, add layer/strength sweeps, and replicate across larger model families.

References

Joschka Braun. Understanding unreliability of steering vectors in language models. *arXiv preprint arXiv:2602.17881*, 2026.

Yuanpu Cao et al. Personalized steering of llms via bi-directional preference optimization. *arXiv preprint arXiv:2406.00045*, 2024.

Bertram Højer et al. Improving reasoning performance via representation engineering. *arXiv preprint arXiv:2504.19483*, 2025.

Tim Lawson et al. Residual stream analysis with multi-layer saes. *arXiv preprint arXiv:2409.04185*, 2024.

Andy Liu, Mona Diab, and Daniel Fried. Evaluating llm biases in persona-steered generation. *arXiv preprint arXiv:2405.20253*, 2024.

Adam S. Shai et al. Transformers represent belief state geometry in their residual stream. *arXiv preprint arXiv:2405.15943*, 2024.

Jisu Shin et al. Spotting out-of-character behavior. *arXiv preprint arXiv:2506.19352*, 2025.

Constantin Venhoff et al. Understanding reasoning in thinking lms via steering vectors. *arXiv preprint arXiv:2506.18167*, 2025.

Zejia You et al. Spherical steering. *arXiv preprint arXiv:2602.08169*, 2026.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.