Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

How Far Can a Single Vector Carry a Language? Mechanistic Limits of Inference-Time Steering for Low-Resource Devanagari Languages

Sumit Yadav
Pulchowk Campus https://orcid.org/0009-0009-9984-2956
Santosh Giri
Pulchowk Campus https://orcid.org/0009-0003-0577-7345
Ganesh Gautam
ganesh@pcampus.edu.np
Pulchowk Campus https://orcid.org/0000-0002-1926-1492

Research Article

Keywords: representation steering , mechanistic interpretability, multilingual language models, low resource languages, activation engineering, cross-lingual transfer

Posted Date: September 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10978527/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

How Far Can a Single Vector Carry a Language? Mechanistic Limits of Inference-Time Steering for Low-Resource Devanagari Languages

Sumit Yadav Santosh Giri Ganesh Gautam*

Department of Electronics and Computer Engineering,
IOE, Pulchowk Campus, Tribhuvan University, Lalitpur, Nepal
076bct088.sumit@pcampus.edu.np ganesh@pcampus.edu.np

Abstract

Large multilingual decoder language models publish closed lists of supported languages: Llama-3.1 and Aya-23 both claim Hindi but not its lower-resource Devanagari sisters, Maithili, Nepali, and Bhojpuri. We ask a mechanistic question: *where*, and *how far*, can a single inference-time direction move such a model toward an unclaimed sister language, using only its Hindi representations and no fine-tuning, no new data, and no weight updates? Because the four languages share the Devanagari script and most subword tokens, any shift must reflect representation geometry rather than vocabulary. Across two architecturally distinct eight-billion-parameter decoders we report four findings. First, both models exhibit a shared *hourglass* geometry (languages distinct at the surface, merged mid-network, and re-separated near the output), yet language identity is linearly decodable at every layer (about 99.5% accuracy). Second, adding the difference of per-language activation centroids at the re-divergence onset shifts generation toward the target (peak language adherence 4.03 of 5 for Nepali on Llama, scored by a three-judge cross-family panel) but only at the surface: fluency collapses into repetition, and no combination of model, target, and coefficient is simultaneously target-adherent and fluent. Third, few-shot prompting does reach fluent sisters, so the ceiling is a limit of single-direction steering, not model competence; steerability tracks linear separability, not proximity. Fourth, Hindi outperforms English as the steering source despite English's pretraining dominance. Steering thus serves as a causal probe of where language identity is manipulable. We release a four-way parallel Devanagari corpus and an evaluation harness.

Keywords: representation steering | mechanistic interpretability | multilingual language models | low-resource languages | activation engineering | cross-lingual transfer

1 Introduction

Instruction-tuned large language models (LLMs) publish closed, named lists of the languages they support. Llama-3.1 names eight languages, with Hindi the only Indo-Aryan member; Aya-23 names twenty-three, again with Hindi the sole Indo-Aryan inclusion. Neither claims Maithili, Nepali, or Bhojpuri, Devanagari-script languages spoken by hundreds of millions of people, yet under-served by current systems. This yields a clean contrastive setting: a claimed, high-resource anchor (Hindi) alongside closely related, *unclaimed* targets that share its script and much of its vocabulary.

We ask a *mechanistic* question. *Where* in the network, and *how far*, can a single inference-time direction move such a model toward an unclaimed sister language, using only its Hindi representations, with no fine-tuning, no new training data, and no weight updates? Representation steering adds a fixed direction to the residual stream during the forward pass; here that direction is the difference of per-language

*Corresponding author: Ganesh Gautam, ganesh@pcampus.edu.np

activation centroids. Because Hindi, Maithili, Nepali, and Bhojpuri share the Devanagari script and a large fraction of subword tokens, tokenizer sharing is held fixed as a controlled constant: any shift in the output must come from the geometry of the model's internal representations, not from vocabulary overlap. We therefore treat steering not as a deployment recipe (few-shot prompting elicits these languages far more fluently) but as a *causal probe* of where language identity becomes a manipulable direction.

This framing matters because it separates two questions that interpretability work often conflates. A linear probe can certify that language identity is *decodable* from the hidden states; it cannot tell us *where* that identity is a causally usable control direction. Steering supplies exactly that test and, as we show, returns a counter-intuitive answer: the layers where a single direction steers most effectively are *not* the layers where the languages are most geometrically separated.

Summary of findings. Our study spans two architecturally distinct decoders and yields a single coherent account. (i) The four languages occupy a shared *hourglass* geometry: distinct at the surface, merged in a common core, re-separated for output, yet linearly decodable throughout. (ii) A single direction steers across an early-to-middle band of layers but is inert once the languages re-separate near the output, and even within the band it carries generation only to *surface* language identity, not to fluent text. (iii) The effect is direction-specific (a matched-magnitude random direction is inert), is not improved by multi-layer composites, and is bounded by the target's linear *separability* rather than its raw proximity to Hindi, while few-shot in-context learning reaches fluent sisters, proving the ceiling is steering-specific, not a competence limit. (iv) Hindi beats English as the steering source despite English's pretraining dominance, so transfer follows linguistic proximity, not resource level.

Contributions.

1. A shared hourglass geometry documented across two models, with a probe-versus-geometry dissociation that explains *where* single-direction steering can work (Section 5.1).
2. Single-layer steering at the re-divergence onset as a *mechanistic probe* of this geometry, with a fully mapped adherence–fluency frontier and a degeneracy-aware reading of the headline result (Section 5.3.1).
3. A control suite (matched-magnitude random direction, multi-layer composite, a per-layer single-layer sweep, and a few-shot in-context baseline) that localizes both the cause and the limit of the effect, plus independent (non-LLM) metric validation (Section 5.3.2).
4. A source-language comparison isolating linguistic proximity, not resource level, as the driver of transfer (Section 6.1); and a discussion of what a *causal* steering probe reveals that a decodability classifier cannot (Section 6). We release the four-way parallel corpus and the evaluation harness.

2 Related Work

2.1 Directions and vectors in representation space

That meaningful semantic relations correspond to linear directions dates to word-embedding arithmetic (Mikolov et al., 2013) and re-emerges inside transformers: in-context tasks collapse to *task* and *function* vectors (Hendel et al., 2023; Todd et al., 2024), and behaviours from truthfulness to refusal are mediated by single directions (Li et al., 2023; Arditi et al., 2024). We treat *language* as one such direction and ask where in the network it is a usable control axis.

2.2 Representation steering

Adding task- or concept-aligned directions to the residual stream steers generation without weight updates (Turner et al., 2023; Rimsky et al., 2024; Liu et al., 2023; Zou et al., 2023; Stolfo et al., 2024);

a complementary line edits a small set of parameters or features instead (Wu et al., 2024). SafeConstellations (Maskey et al., 2025), the direct methodological parent of this work, uses per-task trajectory steering to mitigate over-refusal; we reuse its memory-bank construction and single-layer hook, replacing the task-pair contrast (refuse/comply) with a language-pair contrast (source/target).

2.3 Shared multilingual representations

Multilingual decoders are argued to process non-English input through a shared, English-leaning mid-layer latent space (Wendler et al., 2024; Zhao et al., 2024), where language identity concentrates in language-specific neurons (Tang et al., 2024; Gurgurov et al., 2025) and English-centric models exhibit language confusion (Nie et al., 2025); see Qin et al. (2025); Resck et al. (2025) for surveys. This continues a longer line on the partial language-neutrality of multilingual encoders: multilingual BERT transfers across scripts and typologies (Pires et al., 2019), and its representations decompose into a large language-neutral component plus a language-specific one (Libovický et al., 2020). Prior cross-lingual transfer largely aids languages a model already knows (Tanwar et al., 2023; Shi et al., 2022); we instead target languages a model never claimed. Our hourglass refines this view with a probe-versus-geometry dissociation: identity is linearly decodable (Belinkov, 2022) throughout, but only geometrically prominent at the surface and the deep layers, a distinction interpretability work (Rai et al., 2024; Elhage et al., 2022) rarely separates.

2.4 Language steering, and how we differ

Closest to our recipe is language steering by source–target activation differences, used to improve multilingual in-context learning (Kirtane and Huang, 2026) and for cross-lingual control (Pokharel et al., 2026). We differ on three axes. First, we use the vector to *induce low-resource target output that a model never claimed*, rather than to aid in-context learning in languages it already supports. Second, we give a *mechanistic* account of *which* layers are steerable and why: an early-to-middle band, not a single grid-searched "best" layer. Third, we show the source language must be linguistically proximate to the target (Hindi $\gg$ English), not merely high-resource.

2.5 Low-resource Devanagari languages

maiBERT (Yadav et al., 2026) documents the difficulty of building Maithili capability; we target Maithili, Nepali, and Bhojpuri jointly and measure transfer from a model that never claimed them. To our knowledge this is the first study to treat these three Devanagari sisters as a controlled family for testing inference-time steering.

3 Method

Figure 1 summarizes the three stages: offline extraction of per-language representations, computation of a centroid-difference steering vector, and inference-time injection of that vector through a forward hook.

3.1 Memory banks

For each model M and language $L \in \{\text{hin, mai, npi, bho, en}\}$ we run the $N{=}250$ parallel passages through M with hidden-state output enabled and mean-pool the non-padding residual-stream states at every layer. This yields, per passage, a tensor of shape $(N, L_{\text{layers}}{+}1, H)$ and per-layer centroids $\mu_{M,L,\ell} \in \mathbb{R}^H$, where H is the hidden width. Index 0 is the token-embedding output; indices $1..L_{\text{layers}}$ are the residual stream after each decoder block. Mean-pooling over content tokens makes each centroid a passage-averaged, position-agnostic summary of how the model represents that language at that depth.

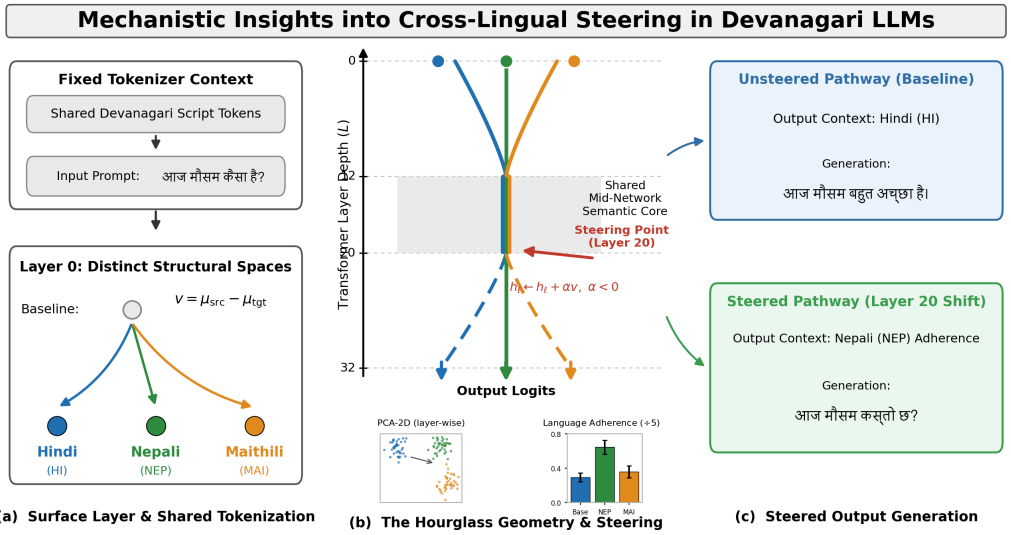


Figure 1: Overview. **(a)** The languages share the Devanagari tokenizer yet at the surface occupy *distinct* representation spaces; we form $v = \mu_{\text{src}} - \mu_{\text{tgt}}$ from per-layer centroids. **(b)** With depth the languages follow an *hourglass* (distinct $\rightarrow$ merged core $\rightarrow$ re-separated); adding v at the layer-20 onset of re-divergence steers a Hindi-prompted pass toward the target (insets: layer-wise 2-D projections; measured adherence). **(c)** Unsteered output is Hindi; steered output is the target (illustrative)

3.2 Steering vector

For a source (anchor) language and a target, $v_\ell = \mu_{\text{src},\ell} - \mu_{\text{tgt},\ell}$. With a source-language prompt, adding $\alpha\, v_\ell$ with $\alpha < 0$ moves the residual stream from the source centroid toward the target centroid, pushing generation away from the source and toward the target. The magnitude $|v_\ell|$ is itself informative (Section 5.1): it grows with depth and orders the targets by linguistic distance.

3.3 Intervention

A forward hook on decoder block ℓ sets $h_\ell \leftarrow h_\ell + \alpha\, v_\ell$ at *all* token positions on every forward pass (both prompt prefill and each generated step), so the edit propagates to all later layers and the language-model head. We use a single layer unless stated otherwise; the multi-layer variant (Section 5.3.2) hooks several blocks, with each v_ℓ norm-equalized to unit length so every hooked layer contributes equal magnitude. Decoding is greedy with 90 new tokens. We sweep $\alpha \in [-3, 0]$ per target (Llama in steps of 0.5; Aya on a finer 0.1–0.2 grid around its narrower effective window). Algorithm 1 states the full procedure; injecting at all positions, rather than only generated positions, is a deliberate choice we revisit in Section 5.3.2.

4 Experimental Setup

4.1 Models

We study Llama-3.1-8B-Instruct (Meta) (Grattafiori et al., 2024) and Aya-23-8B (Cohere Labs) (Aryabumi et al., 2024). They sit at opposite ends of the multilingual-training spectrum (English-centric versus multilingual-by-design), and both name Hindi as their only Indo-Aryan language, making them a deliberate test of whether any finding is architecture-specific. We load each model in 4-bit NormalFloat (NF4) quantization with bfloat16 (bf16) compute on a single 16 GB NVIDIA T4 GPU. Section 5.3.2 re-extracts the geometry and re-runs the headline steering in full bf16 to confirm that neither is a quantization artifact.

Algorithm 1: Centroid steering vector and intervention

Input: model M; parallel corpus $\{x_i\}_{i=1}^{N}$; source s, target t; layer ℓ; coefficient $\alpha < 0$.

// build memory banks (offline, once)

for $L \in \{s, t\}$:

  for $i = 1 \dots N$: $H \leftarrow M(x_i^L)$ (hidden states, all layers);

    $b_i^L \leftarrow \text{MeanPool}_{\neg\text{pad}}(H) \quad \in \mathbb{R}^{(L_{\text{layers}}+1)\times H}$

  $\mu_L \leftarrow \frac{1}{N}\sum_i b_i^L$ (per-layer centroid)

// steering vector at layer ℓ

$v_\ell \leftarrow \mu_s[\ell] - \mu_t[\ell]$

// steered generation (inference)

prompt p in source language s;

register hook on block ℓ: $h_\ell \leftarrow h_\ell + \alpha\, v_\ell$ (all positions);

return M.generate(p) ($\alpha < 0$ moves $s \to t$)

4.2 Corpus

The memory-bank corpus is 250 long Hindi passages sampled from XLSum, translated to Maithili, Nepali, and Bhojpuri by a three-frontier-LLM consensus and to English by a single frontier model. Content is parallel across all five languages, so per-language centroids isolate the language signal rather than topical differences. We release the corpus as a paper artifact. We note a validity caveat that we return to in Section 7: corpus construction and the in-context exemplars draw on frontier LLMs. The evaluation judges are a cross-family panel that shares no family with the generators (mitigating self-preference), and an independent non-LLM language-identification check (Section 5.3.2) further guards language identity; native-speaker assessment of translation adequacy remains future work.

4.3 Evaluation

The evaluation set is a held-out collection of instruction prompts, disjoint from the memory-bank corpus; we report n=30 prompts per (model, target, α) cell. For each (prompt, α) we generate greedily (90 new tokens) and score the output with a *panel of three LLM judges drawn from disjoint model families* (`gemini-3.1-pro-preview`, `gpt-5.5`, and `gemini-3.5-flash`); crucially, none shares a model family with the Llama (Meta) or Aya (Cohere) generators, so judge self-preference bias is structurally avoided (Verga et al., 2024; Panickssery et al., 2024). Each judge scores independently on a strict 1–5 rubric covering *language adherence* (5 = fully target language, 3 = even mix, 1 = wrong language), *fluency*, *faithfulness*, *coherence*, and *overall* quality; we report the panel *mean* (with the median and the per-judge spread as robustness). Language adherence is the diagnostic metric throughout. Inter-judge agreement is high on the two reported dimensions (Krippendorff ordinal α=0.87 for adherence and 0.84 for fluency; ICC$(2, k)$=0.96 and 0.95; within-one-point agreement 99–97%), and the headline is invariant to panel composition under leave-one-judge-out and family-balanced re-weighting (Section 5.3.2). We additionally pair adherence with a non-LLM language identifier and a degeneracy measure (Section 5.3.2). The judge prompt and rubric are given in Appendix A.

5 Analysis

We organise the analysis in three parts: the depth-wise *geometry* of the four languages (Section 5.1), where language identity is *decodable* versus geometrically prominent (Section 5.2), and a *causal* analysis that uses inference-time steering as a probe of where that identity is a manipulable control direction (Section 5.3).

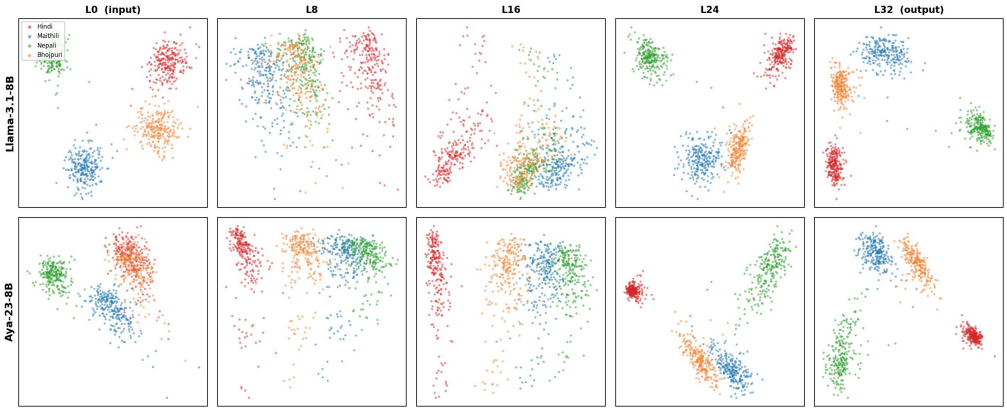


Figure 2: The hourglass replicates across architectures. Per-sample 2-D projections (250 samples $\times$ 4 languages) at fixed depths; Llama (top), Aya (bottom). The languages are distinct at the input, merge mid-network, and re-separate at the output

5.1 Geometry of language

Projecting the per-language centroid trajectories through a shared UMAP embedding (Figure 3) and the per-sample activations (Figure 2) reveals the same shape on both models: the four languages are distinct at the surface, collapse into a shared region in the middle layers, and re-separate for output. The Hindi–Nepali centroid cosine bottoms at about 0.69 on Llama (layers 26–27) and 0.627 on Aya (layer 26), with closeness tracking linguistic distance (Bhojpuri $>$ Maithili $>$ Nepali in proximity to Hindi).

Finding 1 (hourglass geometry). On both models the four languages are geometrically distinct in the early layers, merge into a shared region in the middle layers, and separate again near the output.

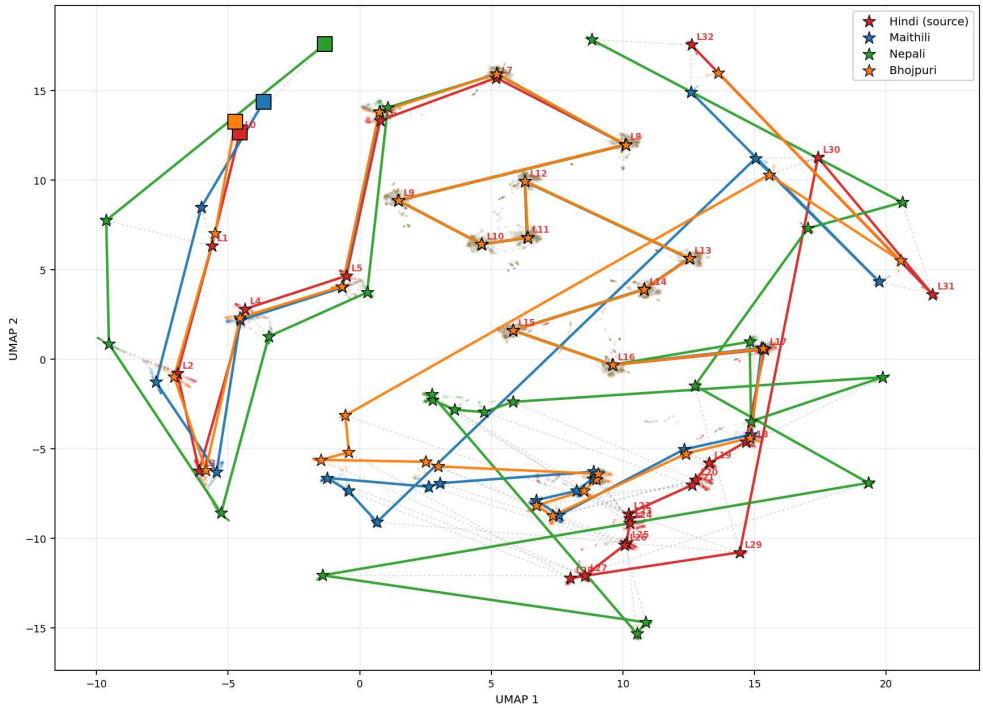


Figure 3: Centroid *constellation* (Llama, layers 0–32): each language's per-layer centroid traced as a path (square = embedding, star = layer) in one shared UMAP embedding; dashed lines are the per-layer Hindi–sister gap. Paths bunch in the merged middle and fan out at the ends

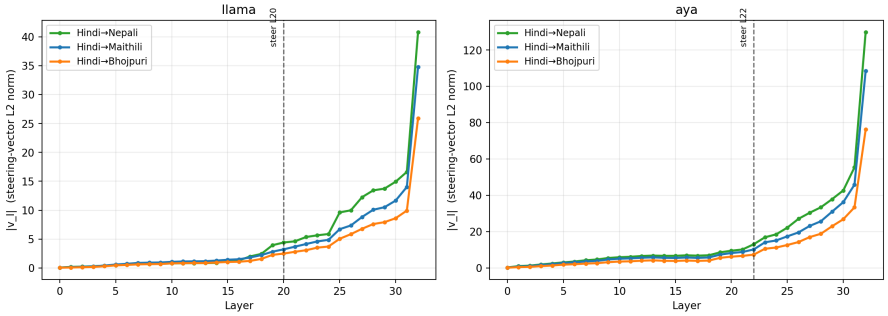


Figure 4: Steering-vector magnitude $|v_\ell|$ by layer for both models. The magnitude grows with depth and orders the targets by linguistic distance (Nepali > Maithili > Bhojpuri)

Vector magnitude and linguistic distance. The steering-vector norm $|v_\ell|$ grows with depth and, at any given layer, orders the targets by linguistic distance (Nepali > Maithili > Bhojpuri; Figure 4). This metric structure, with larger difference vectors for more distant languages, reappears in the steering dial below: per-target coefficients mostly compensate for $|v|$ to reach a common effective steering window.

5.2 Where is language identity decodable?

A four-way linear probe (logistic regression, 5-fold cross-validation on the top-50 principal components) scores about 0.995–1.0 accuracy at *every* layer, including the merged middle, while the silhouette score (a measure of cluster separation) dips from about 0.26 at the surface to 0.06 at Llama layer 14 and recovers to 0.36 at layer 31 (Figure 5). Language identity is therefore always *decodable*; what changes across depth is its geometric prominence.

Finding 2 (decodable everywhere, prominent only at the ends). Language identity is linearly decodable at every layer, including the merged middle, yet geometrically prominent only at the surface and the re-separated deep layers.

Decodability does not, however, predict *steerability*. As the per-layer sweep shows (Section 5.3.2), a single direction steers across an early-to-middle band (including the most *merged* layers) and fails only once the languages re-separate near the output. Geometric separation and steerability thus run in *opposite* directions with depth, a dissociation a probing classifier alone could never reveal and that we make the central interpretability point of the paper (Section 6).

5.3 Causation analysis

Inference-time steering supplies a *causal* test that decodability cannot: does writing to a language direction at a given depth actually change what the model generates? We treat steering as a probe, reporting a single-layer intervention (Section 5.3.1), a battery of controls, and a validation of the metrics that does not rest on any single judge.

5.3.1 Single-layer steering

The adherence-fluency frontier. At the onset of re-divergence (Llama layer 20, Aya layer 22), $\alpha < 0$ shifts generation toward the target: adherence rises and then collapses, while fluency declines monotonically (Tables 1 and 2). **Nepali on Llama reaches** $4.03/5$ **adherence at** $\alpha = -2.0$ (95% bootstrap CI [3.82, 4.22]), majority target language, with Nepali grammatical markers (the copula *ho*, the relativizer *jun*, the locative *-maa*, and the verb *bhanincha*; native script in Table 3). This is far above the sister targets (Maithili 1.82, Bhojpuri 1.72) and *above* Aya's single-layer ceiling of about 1.34 (Table 2), so the ceiling is a property of a given model, not a universal limit. The peak's confidence interval [3.82, 4.22] lies well above both the sister peaks and the Hindi control, so the effect is not $n{=}30$ noise; it is also in-

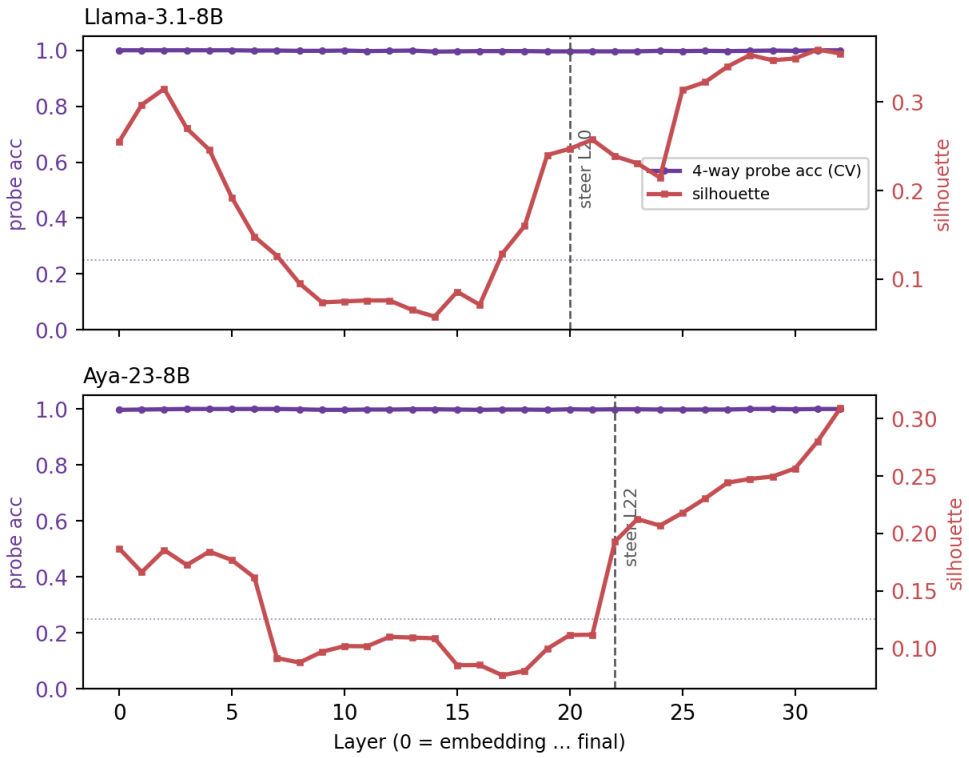


Figure 5: Probe versus geometry, *both* models (Llama top, Aya bottom). The four-way linear probe stays near-perfect at every layer while the silhouette score traces the hourglass: identity is always decodable but only geometrically prominent at the surface and the re-separated deep layers, the dissociation replicates across architectures

variant to judge composition, holding under every leave-one-judge-out and family-balanced re-weighting (Section 5.3.2).

Surface shift, not fluency. Language adherence rewards target-language tokens without penalizing degeneracy, so it measures surface language identity rather than usable generation. At the Nepali peak the distinct-bigram ratio falls from 0.96 to 0.79 (i.e. heavy repetition). The trade-off is unavoidable here: *no* (model, target, α) cell has mean adherence *and* fluency both above 3, and only 1.3% of Llama generations (6/450; 0% for Aya) are individually both adherent and fluent (each ≥ 3), peaking at Nepali $\alpha = -1.5$ (13%). We therefore read steering as inducing target-language *identity*, not fluent *competence*. The effect is robust to precision: the Nepali peak that clears the "mostly target" line under NF4 (4.03) is reproduced in full bf16 (4.02; Section 5.3.2).

Example generations. Table 3 makes the trade-off concrete. At $\alpha = -1.0$ the Nepali generation is readable yet already carries target morphology (the locative भारतमा where Hindi uses भारत में), scoring adherence 2 at fluency 3; by the peak ($\alpha = -2.0$) it is unmistakably Nepali (वास्तु हो, जुनको) but collapses into a repetition loop (भुटान, भुटान, ...), adherence 5 at fluency 1. Maithili and Bhojpuri shift partway (छ, करे ता, जे के ले) but never become fluent. The failure-mode breakdown for Llama-Nepali (Appendix B) confirms a code-switch-dominant regime peaking at $\alpha = -1.5$, with repetition only at $\alpha = -2.0$ and fluent output rare throughout: steering induces target-language *identity* well before *fluency*, so the usable operating point is an intermediate α, not the adherence maximum.

Effective magnitude across models. Within a model the *effective* perturbation at the sweet spot, $|\alpha| \cdot |v_\ell|$, is roughly constant (about 9 residual-space units for Llama; about 10–13 for Aya), while $|v_\ell|$ varies

Table 1: Llama single-layer (layer 20) sweep, n=30. Control is α=0 (Hindi). "peak adh" is the best language adherence over α

Target	ctrl adh	peak adh (α)	Δ	flu@peak
Nepali	1.00	**4.03** (-2.0)	$+3.03$	1.68
Maithili	1.00	1.82 (-2.5)	$+0.82$	1.60
Bhojpuri	1.00	1.72 (-3.0)	$+0.72$	1.97

Table 2: Aya single-layer (layer 22) sweep, n=30. Adherence caps near 1.34

Target	ctrl adh	peak adh (α)	Δ	flu@peak
Nepali	1.00	1.34 (-0.8)	$+0.34$	1.31
Maithili	1.00	1.24 (-0.7)	$+0.24$	2.27
Bhojpuri	1.00	1.26 (-1.4)	$+0.26$	1.40

Table 3: Steered generations (Llama, single-layer L20), first 14 tokens of each output. Each block is one prompt traced across the steering coefficient α (α=0 is the unsteered Hindi control). Adh/Flu are the LLM-judge language-adherence and fluency scores (1–5). The trade-off is explicit: at the adherence peak ($\alpha = -2.0$, Nepali) the text is target-language but degenerates into a repetition loop (fluency 1); the readable shift sits at smaller $|\alpha|$.

α	Adh	Flu	Generation (first 14 tokens, Devanagari)
Nepali target. Prompt: पोंगल त्योहार दक्षिण भारत में क्यों और कैसे मनाया जाता है? इसकी मुख्य विशेषताएँ क्या हैं? *(Why and how is Pongal celebrated in South India?)*			
$+0.0$ (ctrl)	2	4	पोंगल त्योहार दक्षिण भारत में एक महत्वपूर्ण त्योहार है, जो मकर संक्रांति के अवसर
-1.0	2	3	पोंगल त्योहार दक्षिण भारतमा एक महत्वपूर्ण त्योहार है, जो तमिल नृवंशी लोग द्वारा मनाया
-2.0	5	1	पोंगल त्योहार दक्षिण भारतमा एक महत्वपूर्ण वास्तु हो, जुनको नेपाल, भुटान, भुटान, भुटान, भुटान,
Maithili target. Prompt: भारतीय शास्त्रीय संगीत के प्रमुख रूपों का संक्षिप्त परिचय दें। *(Briefly introduce the major forms of Indian classical music.)*			
$+0.0$ (ctrl)	1	5	भारतीय शास्त्रीय संगीत एक समृद्ध और विविध परंपरा है, जिसमें विभिन्न प्रकार के संगीत
-2.5	3	1	भारतीय शास्त्रीय संगीतके प्रमुख रूप छ दु: 1. भक्ति गीत: भक्ति गीत, जे भक्ति
Bhojpuri target. Prompt: भांगड़ा नृत्य पंजाब की संस्कृति का प्रतिनिधित्व कैसे करता है? इसकी विशेषताएँ बताएँ। *(How does Bhangra dance represent Punjabi culture?)*			
$+0.0$ (ctrl)	1	5	भांगड़ा नृत्य पंजाब की संस्कृति का एक महत्वपूर्ण हिस्सा है, जो प्रेम और उत्साह
-2.5	3	2	भांगड़ा नृत्य पंजाब की संस्कृति का प्रतिनिधित्व करे ता है, जे के ले पंजाबी

by target (Figure 4). The per-target α ladder is thus mostly compensating for $|v|$ to land in the same absolute steering window. The mechanism and the effective-magnitude window transfer across models; the raw α dial does not.

Finding 3 (surface identity, not fluency). A single centroid-difference direction can push generation into the target language (Nepali adherence 4.03/5 on Llama), but fluency collapses in step, and no (model, target, coefficient) setting is simultaneously adherent and fluent.

5.3.2 Controls

Direction specificity. To rule out generic perturbation, we replace v with ten random unit directions scaled to the *same* magnitude ($|v_{20}^{\text{npi}}|$=4.60) and sweep the headline coefficients $\alpha \in \{-1.5, -2.0\}$. Random directions are inert (Figure 6): mean adherence 1.00 (0 of 300 generations reach adherence ≥ 3), with fluency preserved near the control level. The inertness holds across the reference-free metrics too: at the peak ($\alpha = -2.0$) random directions stay at the control level on meaning preservation (LaBSE 0.86 vs control 1.00), native-LM perplexity (3.96 vs 3.49), and diversity (distinct-2 0.96 vs 0.97), whereas the real vector moves all of them (semantic 0.69, perplexity 18.6, distinct-2 0.80). The real vector (4.03) thus

Table 4: Multi-layer (layers 18+20+22, norm-equalized) versus single-layer (layer 20) peak adherence, Llama

Target	single-layer L20	multi-layer L18+20+22
Nepali	**4.03**	4.03
Maithili	1.82	1.59
Bhojpuri	1.72	2.03

does direction-specific work ($\Delta \approx +3.0$ on adherence); the model is otherwise robust to perturbations of this magnitude.

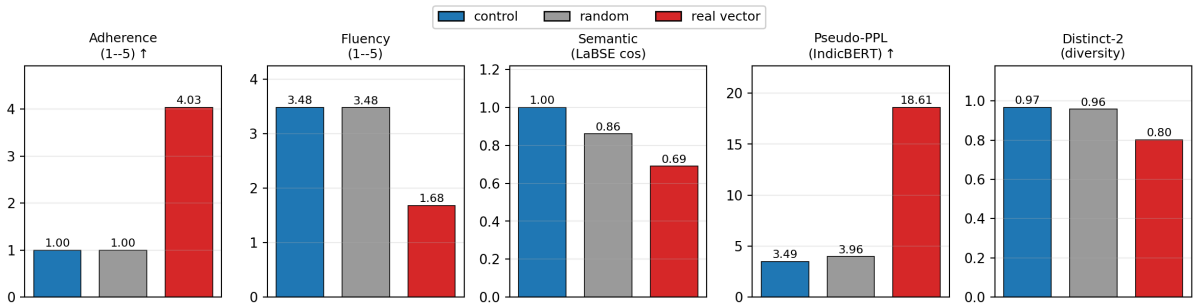


Figure 6: Direction specificity across the full metric suite (Llama, Nepali, at the adherence peak $\alpha = -2.0$): control (α=0), a matched-magnitude *random* direction (mean over 10 seeds), and the real $\mu_{\text{hin}} - \mu_{\text{npi}}$ vector, on judge adherence and fluency, LaBSE semantic preservation, IndicBERT pseudo-perplexity, and distinct-2. The real vector shifts *every* metric (adherence up; fluency, meaning preservation, and diversity down; perplexity up), while random directions of the same magnitude sit at the control level on all of them, so the effect is direction-specific, not generic perturbation

Multi-layer steering. Hooking layers 18, 20, and 22 simultaneously (norm-equalized) does *not* break the wall (Table 4): Maithili and Bhojpuri stay near their single-layer values (about 1.6–2.0, within n=30 noise) and Nepali is *no better* than single-layer. This is not an under-injection artifact: a *raw* (non-normalized) composite, which at $\alpha \leq -1$ injects more total magnitude than the single-layer peak, peaks *lower* (Nepali 3.46, Maithili 1.54, Bhojpuri 1.92) and collapses toward 1.0 as $|\alpha|$ grows: more steering only degrades output. The residual ceiling is therefore a limit of *single-direction steering*, not of model competence, since few-shot in-context learning elicits fluent sister-language text (below).

Depth-trajectory steering. A stronger version of the multi-layer test asks whether injecting the direction along the model's whole depth *trajectory*, rather than at one or three layers, breaks the ceiling. We inject each layer's own v_ℓ (norm-equalized) at *every* layer of the steerable band (L8–20, 13 layers) and sweep α.[1] It does not help. On the one target that steers well, Nepali peaks at 3.47 (at $\alpha = -0.5$), *below* the matched single-layer peak on the same judges (4.50); Maithili and Bhojpuri stay at 1.78 and 2.13. Worse, accumulating the direction across all 13 layers *over-steers*: adherence collapses to the 1.0 floor for $|\alpha| \geq 1$ as generation degenerates. Neither multi-layer composites nor the full depth trajectory exceed single-layer steering, so the ceiling is a property of the single centroid-difference direction, not of how many layers it is applied at.

Depth of the steerable band. A magnitude-controlled single-layer sweep (v norm-equalized) shows the effect is not pinned to the onset layer (Llama-Nepali peak adherence: layer 8 3.72, layer 14 3.41,

[1]This late control was scored by the two Gemini panel judges (the OpenAI seat was temporarily unavailable); the qualitative conclusion is unaffected and is consistent with the leave-one-judge-out and family-balanced re-weightings of Section 5.3.3.

Table 5: Few-shot in-context baseline (Llama, no steering) versus single-layer steering. k exemplars of (Hindi $\rightarrow$ gold target); adherence/fluency (1–5), n=30

	ICL k=3		steering (L20) peak	
Target	adh	flu	adh	flu
Nepali	3.01	**3.59**	**4.03**	1.68
Maithili	**2.39**	**3.41**	1.82	1.60
Bhojpuri	**2.30**	**3.43**	1.72	1.97

layer 20 3.91, versus layer 26 1.00, layer 31 1.00). An early-to-middle band (layers 8–20), including the most *merged* layer 14, steers, while *late*, re-separated layers are wholly inert (adherence at the 1.0 floor, fluency preserved). The required magnitude grows with depth (layer 8 peaks at $\alpha = -4$, layer 20 at $\alpha = -8$), which is why a fixed raw vector makes the mid-network look special. A single direction redirects language while the representation is still merging, not once it has re-separated; layers 20 and 22 are a representative point in this band, not a unique optimum. This is the central evidence that steerability and geometric separation run opposite (Section 6).

In-context learning baseline. Does the sister-language ceiling reflect the model's competence or the reach of a single linear direction? We compare against few-shot in-context learning, with no steering: each test prompt is preceded by k exemplars of (Hindi prompt $\rightarrow$ gold target-language response), drawn from held-out rows, plus a target-language instruction (Table 5). With k=3, all three targets are generated at *usable fluency* (Nepali 3.59, Maithili 3.41, Bhojpuri 3.43), far above the fluency single-layer steering ever reaches. Adherence under ICL is moderate (3.01/2.39/2.30): for Maithili and Bhojpuri, ICL exceeds steering on adherence *and* fluency; for Nepali, steering can push raw adherence *higher* (4.03 versus 3.01) but only by destroying fluency, whereas ICL stays fluent. The decisive comparison is on the axis steering fails: the models plainly *can* produce fluent sister-language text, so the ceiling is a property of the single centroid-difference direction, not a model-competence limit. This positions steering as a mechanistic probe of where language information is linearly accessible, not as the practical recipe for eliciting these languages.

Finding 4 (a band-limited, steering-specific ceiling). A single direction steers only across an early-to-middle band and is inert once the languages re-separate, so steerability runs opposite to geometric separation with depth; because few-shot prompting reaches fluent sister-language text, the ceiling is a limit of single-direction steering, not of model competence.

5.3.3 Metric validation and robustness

The evaluation does not rest on any single judge. *First*, the three judges agree strongly on the diagnostic dimension and moderately on fluency: Krippendorff ordinal α=0.87 for adherence and 0.84 for fluency, ICC$(2, k)$=0.96 and 0.95, with exact agreement 88%/60% and within-one-point agreement 99–97% (adherence/fluency), and Gwet's AC1 of 0.87 for adherence but only 0.50 for fluency (over n=3880 scored generations); fluency is thus the less reliable of the two dimensions, and two of the three judges share the Google family, so agreement should be read as convergent rather than fully independent. *Second*, the headline is robust to panel composition in *direction* though not exact in *magnitude*: recomputing the Nepali peak under every leave-one-judge-out panel and under a family-balanced weighting (which offsets the two Google seats against the single OpenAI seat) keeps it within $[3.72, 4.50]$, so the reported bootstrap intervals (prompt sampling at a fixed panel) understate the total uncertainty, and the qualitative claims, Nepali the most steerable Llama target and Llama-Nepali above Aya's best, holds and holds under every variant. *Third*, a non-LLM check confirms the language genuinely moves: GlotLID (Kargaran et al., 2023), a fastText identifier covering all four languages in native Devanagari, independently labels 83% of Nepali-peak generations Nepali (versus 0% at control, and only 3% still Hindi), while Maithili

Table 6: Reference-free automatic metrics, control (α=0, Hindi) $\rightarrow$ adherence peak (Llama, layer 20). LaBSE = sentence-embedding cosine to the control answer (meaning retained, 1=identical); pseudo-PPL = IndicBERT-v2 masked-LM perplexity (median; fluency floor, lower is better); CMI = code-mixing index; rep = adjacent-repetition rate

Target	LaBSE cos (peak)	pseudo-PPL (ctrl$\rightarrow$peak)	CMI (peak)	rep (peak)
Nepali	0.69	3.49$\rightarrow$18.61	27.62	0.14
Maithili	0.77	3.49$\rightarrow$10.79	4.50	0.04
Bhojpuri	0.82	3.49$\rightarrow$7.51	3.39	0.04

and Bhojpuri stay majority-Hindi (at most 10% target), corroborating both the transfer and the sister ceiling with no LLM in the loop. The shift is lexical, not cosmetic: word-level Jaccard overlap of each steered Nepali output against its own Hindi control falls from 0.56 at $\alpha = -0.5$ to 0.08 at the peak.

Finally, neither the geometry nor the effect is a 4-bit quantization artifact: re-extracting the banks in full bf16 reproduces the same hourglass (per-layer silhouette curves correlate at r=0.998 with the NF4 ones; the probe is still 1.00 at the steering layer), and bf16 steering reproduces the Nepali shift: under the three-judge panel it reaches 4.02 at $\alpha = -2.0$, matching the NF4 panel peak (4.03) and likewise clearing the 3.0 "mostly target" line. (Under a single judge the bf16 peak had scored below NF4; the panel scores them equivalent, so that earlier gap was a judge-scoring artifact, not a precision effect.)

Reference-free automatic metrics. Fluency alone cannot certify that steering worked: the base model is already fluent in Hindi, so the trivial way to maximise fluency is to *ignore* the intervention and keep answering in Hindi: maximum fluency is the *failed* steer. We therefore treat fluency as a necessary floor and read language adherence together with *meaning preservation* as the objective, using three families of reference-free automatic metrics (no LLM judge, no gold reference) that triangulate the panel (Table 6). *(i) Meaning preservation.* The cosine of a multilingual sentence embedding (LaBSE; Feng et al., 2022) between each steered output and its own α=0 Hindi answer measures whether the content survives the language shift. At the Nepali peak it is 0.69 on Llama, well above the gibberish floor, so the steered text still carries the source meaning; it is higher for the less-moved sisters (Maithili 0.77, Bhojpuri 0.82). *(ii) Fluency floor.* Pseudo-perplexity under a native Indic masked LM (IndicBERT v2; Doddapaneni et al., 2023), which covers all three targets, rises from 3.49 (Hindi control) to 18.61 at the Nepali peak (median; lower is more native-like), an independent, non-judge confirmation that steering breaches the fluency floor. *(iii) Code-mixing.* A per-token Code-Mixing Index (CMI; Gambäck and Das, 2016) rises from 1.21 to 27.62 for Nepali (M-index 0.59, I-index 0.36), quantifying the Hindi$\leftrightarrow$target mixing we describe qualitatively; the sisters mix far less (CMI 4.50, 3.39). Degeneracy rises in step (distinct-2 0.97$\rightarrow$0.80, repetition 0.00$\rightarrow$0.14). Together the automatic metrics tell the panel's story without any judge: the Nepali peak attains target-language *identity* (high adherence, substantial meaning retention, heavy code-mixing) while *violating the fluency floor*: surface identity, not usable competence. These automatic metrics moreover *agree with the panel per generation*, not merely in aggregate: across all 900 scored outputs, panel fluency tracks IndicBERT pseudo-perplexity (Spearman $\rho = -0.62$), panel adherence tracks the code-mixing index (ρ=0.42 overall, 0.56 on Llama), and higher adherence tracks *lower* semantic similarity to the Hindi control ($\rho = -0.29$), the adherence-fluency-meaning trade-off confirmed with no judge in the loop. (Per-word language identification on closely related Devanagari sisters is noisy; we report CMI as an order-of-magnitude mixing signal, not a precise rate.)

6 Discussion

6.1 The role of the source language

Is the transfer about Hindi specifically, or about anchoring on any high-resource language? We rebuild the pipeline with English as the source ($v = \mu_{\text{en}} - \mu_{\text{tgt}}$, English prompts) and compare to Hindi-source

Table 7: Peak language adherence by steering source (Llama, layer 20)

Target	Hindi-source	English-source	advantage
Nepali	**4.03**	1.14	+2.89
Maithili	1.82	1.09	+0.73
Bhojpuri	1.72	1.08	+0.64

at matched α (Table 7, Figure 7). **Hindi beats English on all three targets**, most sharply for Nepali (4.03 versus 1.14). English never reaches the target: adherence caps near 1.1 and degrades to fragments at strong α. This is not a coefficient artifact, the English vector is about $1.7\times$ larger at layer 20 (7.93 versus 4.60) and its adherence peaks at an *interior* $\alpha = -1.0$ (effective magnitude about 7.9, near Hindi's about 9.2), so the English optimum was found and still falls short.

Why is geometric (Figure 8): at the steering layer the Hindi centroid is far closer to each sister than the English centroid (cosine: Nepali 0.81 versus 0.55, Maithili 0.89 versus 0.62, Bhojpuri 0.93 versus 0.65), and in per-sample projections English forms a cluster disjoint from the Devanagari group at every depth. Because English dominates pretraining yet fails, the driver is Hindi's proximity, not resource level. The pattern replicates in direction on Aya (which barely steers from either source): Hindi-source matches or exceeds English on Nepali (1.34 versus 1.12) and Maithili (1.24 versus 1.14), with Bhojpuri a tie at the noise floor (1.26 versus 1.29).

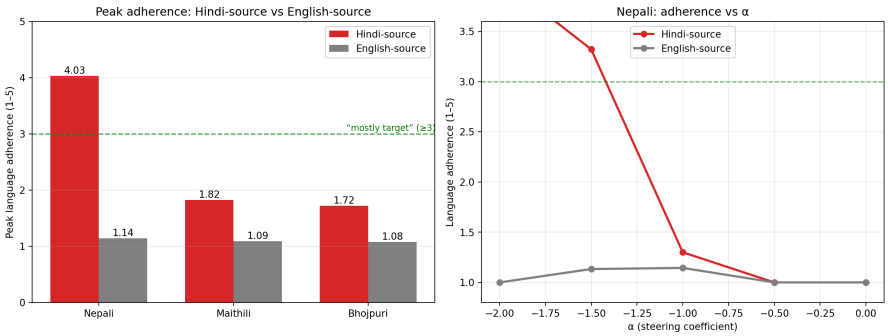


Figure 7: Hindi versus English as the steering source: English never reaches the sisters

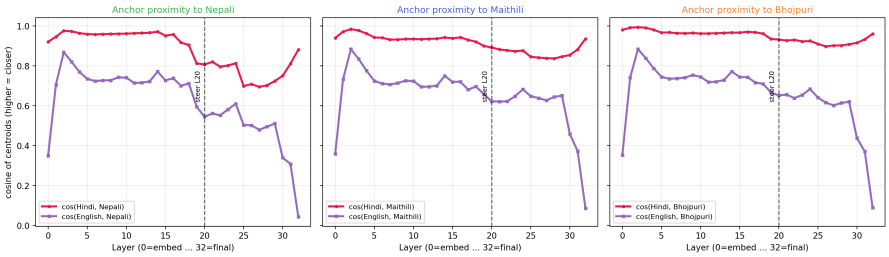


Figure 8: Anchor proximity by layer: Hindi is closer to every sister than English, especially at the steering layer, the geometric mechanism behind Hindi $\gg$ English

6.2 Steering versus probing

A linear probe certifies that language identity is *decodable* at every layer (about 99.5%), but not *where* it is a causally usable control direction. Steering supplies that test, with a counter-intuitive answer. Steerability dissociates from *both* decodability and geometric *separation*: a single direction steers across an early-to-middle band, most effectively in the *merged*, low-silhouette layers, and is inert once the languages re-separate near the output. The boundary tracks how much network remains to propagate the

edit (a depth-of-intervention effect), not where the language axis is geometrically most prominent. This is the scientific payload of the paper: it is precisely the kind of *causal* claim a decodability classifier cannot make. A probe answers "is the information present here?"; steering answers "does writing to this direction here change the model's behaviour?", and the two answers come apart with depth. For a model that is the right object of study, an intervention, not a read-out, this dissociation is the finding.

6.3 Predictors of steerability

Which *targets* steer is set by separability, not raw proximity. By cosine, the *closest* sister to Hindi at layer 20 is Bhojpuri (0.93) > Nepali (0.81), the *reverse* of the steerability ranking; whereas the Hindi–target centroid distance over pooled within-language spread is 1.96 (Nepali) > 1.41 (Maithili) > 1.07 (Bhojpuri), matching adherence (4.03 > 1.82 > 1.72). Bhojpuri is entangled with Hindi and data-poor, yet the same models produce fluent Bhojpuri under few-shot prompting, so the bound on steering is separability, not competence. This distinguishes our account from best-layer grid search: we explain *which* layers steer, *which* targets, and *why* Hindi (Figure 9).

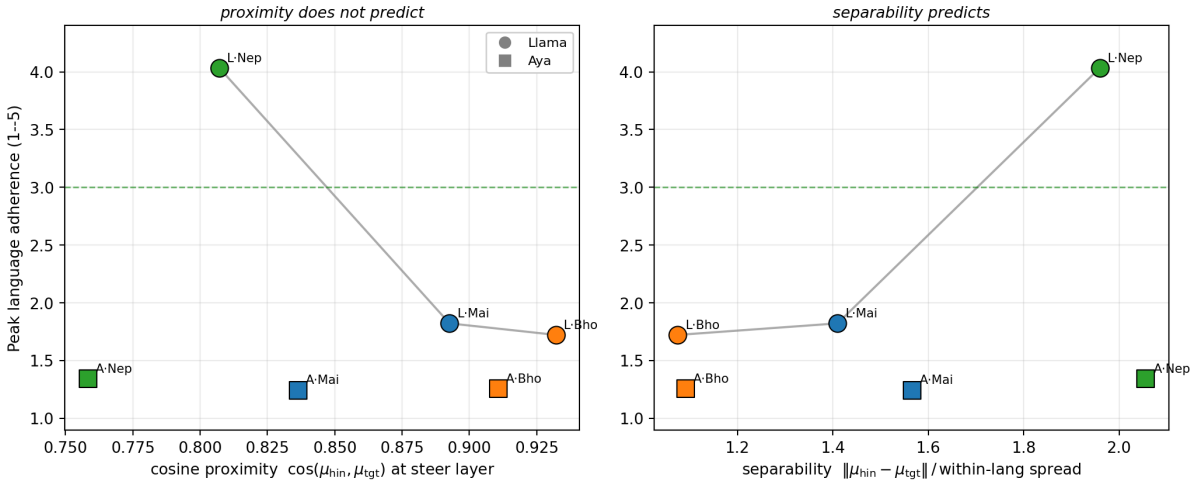


Figure 9: Proximity does not predict steerability, but separability does (Llama circles, Aya squares; the grey line is the Llama within-model trend). *Left*: peak adherence versus Hindi–target cosine proximity at the steering layer; the Llama trend is *reversed*, its closest sister (Bhojpuri) being its least steerable. *Right*: peak adherence versus linear separability ($\|\mu_{\text{hin}} - \mu_{\text{tgt}}\|$ over the within-language spread), which rises monotonically with adherence across Llama's targets. Aya stays at its model-level steering floor on both axes regardless of separability, the cross-model ceiling discussed in Section 6

6.4 Cross-model differences in steerability

Llama-Nepali breaks Aya's single-layer ceiling (about 1.34), and the gap is not explained by the geometry, which replicates across both architectures. The most likely explanation is training composition: Aya is multilingual-by-design and may already allocate distinct, tightly packed subspaces to many languages, leaving less "slack" along any single Hindi–sister axis for one direction to exploit, whereas Llama's English-centric training leaves the Devanagari sisters more entangled with Hindi and hence more movable by the Hindi–target vector. We cannot adjudicate this from two models; explaining the cross-model gap is the clearest target for the broader-model study we flag below.

7 Limitations

The sweeps use n=30 culture-domain prompts; broader-domain results are future work and likely smaller. Corpus construction and the in-context exemplars reuse frontier LLMs; the evaluation judges are a three-

family panel with high agreement (α=0.87), and GlotLID gives a non-LLM cross-check, but the automatic metrics gauge language identity rather than translation adequacy, so native-speaker ratings remain important future work. The corpus is modest (N=250), though the steering direction is data-efficient: a vector rebuilt from 50 passages has cosine at least 0.998 with the full vector (Appendix C). Most importantly, our model-dependence claim rests on two 8B decoders: the *geometry* (hourglass, decodable-everywhere probe, separability ordering) replicates across both, but whether the *steerability* spectrum (Llama near one pole, Aya near the other) generalizes is an n=2 observation. Placing more architectures on that spectrum (e.g. Gemma-2-9B, Qwen2.5-7B, and Indic-specialized decoders) is the clearest next step and would test whether separability predicts steerability out of sample; why Aya barely steers (about 1.34) while Llama does is, for now, only hypothesized (Section 6). Finally, steering of this kind could be misused (for example, low-quality dialect impersonation), and the corpus may carry regional bias; we discuss this under Section 8.

8 Conclusion

We asked how far a single inference-time direction, built only from a model's Hindi representations, can carry a large language model toward Devanagari sister languages it does not officially support. Across two architecturally distinct eight-billion-parameter decoders the answer is consistent. The four languages share an hourglass geometry in which their identity is linearly decodable at every layer yet geometrically prominent only at the ends. Used as a causal probe, steering is effective only across an early-to-middle depth band and runs opposite to geometric separation, so where a direction is writable is not where the language axis is most prominent. A single direction reaches target-language identity but not fluent generation, with no operating point both adherent and fluent; which targets steer is governed by linear separability rather than surface proximity; and the source language must be linguistically proximate, Hindi far outperforming English despite English's pretraining dominance. A three-judge cross-family panel, four families of reference-free automatic metrics, and per-generation convergent-validity checks agree on this account. Steering thus serves less as a deployment recipe than as a map of where, and how far, language identity is a manipulable direction in a multilingual model, and of the limits a single linear direction faces.

Statements and Declarations

Funding. This work was supported by the University Grants Commission (UGC), Nepal.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data availability. The four-way parallel Devanagari corpus (Hindi, Maithili, Nepali, Bhojpuri, and English) generated for this study is machine-translated and released, clearly labelled as synthetic (LLM-generated and not native-speaker-validated), at `https://huggingface.co/datasets/rockerritesh/devanagari-parallel-250` (made public upon acceptance). It should not be treated as gold parallel data for these languages.

Code availability. The steering and multi-judge evaluation harness, and the scripts that reproduce every figure and statistic, are released at `https://github.com/rockerritesh/devanagari-steering`.

Ethics and responsible use. This work studies open-weight models on publicly sourced text and involves no human subjects. Inference-time steering toward low-resource languages could be misused for

low-quality dialect impersonation, and the machine-translated corpus may encode regional bias; we release the corpus and harness to enable scrutiny and recommend native-speaker validation before any downstream use.

Use of large language models. LLMs are used as part of the *methodology* and are documented in Section 4: a three-frontier-LLM consensus produced the sister-language translations, and a three-judge cross-family panel (`gemini-3.1-pro-preview`, `gpt-5.5`, `gemini-3.5-flash`) served as the evaluation panel. No LLM is an author, and no LLM performed generative authorship of this manuscript.

Author contributions. **Sumit Yadav**: conceptualization, methodology, software development, data curation, investigation, experimentation, formal analysis, visualization, preparation of figures and tables, and writing of the original manuscript. **Santosh Giri**: funding acquisition, project administration, and critical review of the manuscript. **Ganesh Gautam**: supervision, conceptual guidance, validation of the research approach, interpretation of results, and critical review, revision, and editing of the manuscript. All authors reviewed the manuscript critically for important intellectual content and read and approved the final version.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 136037–136083, 2024.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. Aya 23: Open weight releases to further multilingual progress, 2024.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12402–12426, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. In *Transformer Circuits Thread*, 2022.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 878–891, 2022.

Björn Gambäck and Amitava Das. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1850–1855, 2016.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models, 2024.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Bäumel, Josef van Genabith, and Simon Ostermann. Language arithmetics: Towards systematic language neuron identification and manipulation, 2025. Accepted to IJCNLP-AACL 2025.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, 2023.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218. Association for Computational Linguistics, 2023.

Neeraja Kirtane and Kuan-Hao Huang. Language steering for multilingual in-context learning, 2026. Language vectors from source–target activation differences for multilingual ICL.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 41451–41530, 2023.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.150.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering, 2023.

Utsav Maskey, Sumit Yadav, Mark Dras, and Usman Naseem. SafeConstellations: Mitigating over-refusals in LLMs through task-aware representation steering, 2025. To appear, ACL 2026.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 746–751, 2013.

Ercong Nie, Helmut Schmid, and Hinrich Schütze. Mechanistic understanding and mitigation of language confusion in English-Centric large language models, 2025.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493.

Rhitabrat Pokharel, Ameeta Agrawal, and Tanay Nagar. Cross-lingual activation steering for multilingual language models, 2026.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. A survey of multilingual large language models. *Patterns*, 6(1), 2025.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models, 2024.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 20454–20486, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, 2024.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners, 2022.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering, 2024.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, 2024.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, 2023.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.15213.

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

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. In *Conference on Language Modeling (COLM)*, 2024.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sumit Yadav, Raju Kumar Yadav, Utsav Maskey, Gautam Siddharth Kashyap, Ganesh Gautam, and Usman Naseem. MaiBERT: A pre-training corpus and language model for low-resourced Maithili language. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 444–452, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-377-7. doi: 10.18653/v1/2026.loreslm-1.38. URL `https://aclanthology.org/2026.loreslm-1.38/`.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 15296–15319, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency, 2023.

A LLM-as-judge prompt and rubric

Each of the three panel judges (`gemini-3.1-pro-preview`, `gpt-5.5`, `gemini-3.5-flash`) independently receives the source prompt, the generated continuation, and the same fixed rubric, and returns

integer 1–5 scores for language adherence, fluency, faithfulness, coherence, and overall quality, with adherence anchored as 5 = fully target language, 3 = even mix of target and source, 1 = wrong language; we report the panel mean. Inter-judge agreement and the composition-robustness test are reported in Section 5.3.2. The full prompt template and the held-out instruction set are included in the released harness (Section 8).

B Error-mode taxonomy

Beyond the headline scores, we classify each Llama-Nepali generation as fluent target, code-switch (mixed target/source morphology), repetition loop, or fragment. The regime is code-switch-dominant and peaks at $\alpha = -1.5$; repetition loops appear at $\alpha = -2.0$ (the adherence peak), and fluent target output is rare throughout. This is why the usable operating point is an intermediate α rather than the adherence maximum.

C Data efficiency of the steering vector

Rebuilding the centroid-difference vector from a 50-passage subsample (rather than the full 250) yields a direction with cosine 0.999 (Nepali), 0.998 (Maithili), and 0.998 (Bhojpuri) to the full-corpus vector, indicating the language direction is recoverable from far less parallel data than the full corpus provides.

D Supplementary geometry

Figure 10 adds English to the per-layer constellation, showing that English sits in a cluster disjoint from the Devanagari group at every depth, the per-sample counterpart of the centroid-proximity result in Section 6.1.

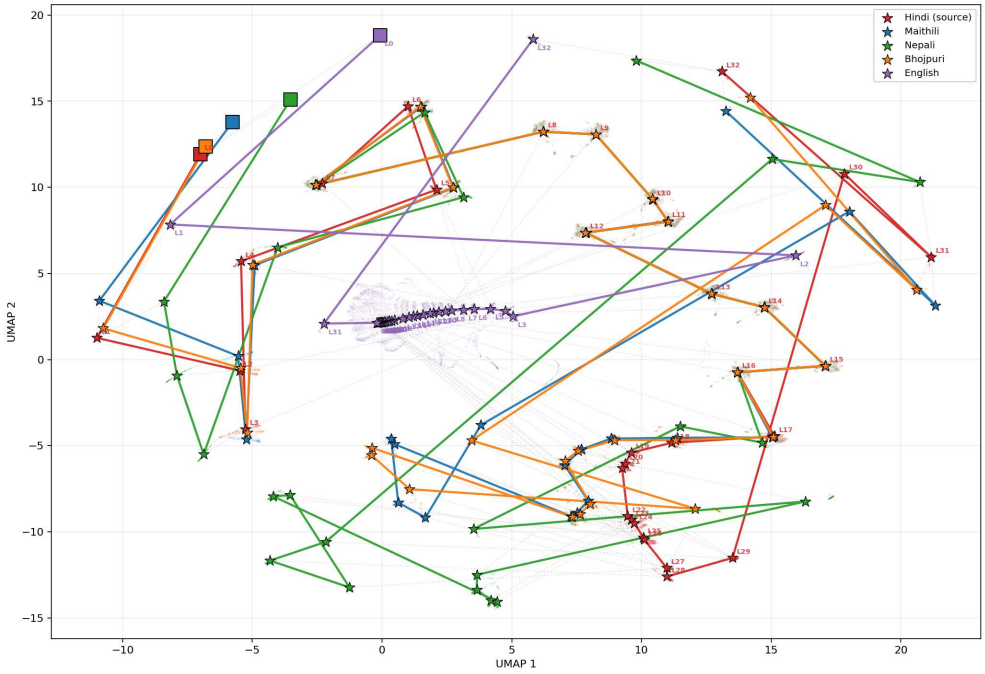


Figure 10: English-inclusive constellation (Llama, layers 0–32). English (added to the four Devanagari languages) traces a path that stays separate from the Devanagari group throughout, unlike Hindi, which sits inside it

Figure 11 illustrates what *depth-trajectory* steering (Section 5.3.2) does geometrically: applying each layer's own vector v_ℓ along the whole steerable band bends the Hindi representation path toward the Nepali one. As the sweep in Section 5.3.2 shows, this bending does not raise the adherence ceiling and, pushed harder, collapses generation, so the figure is an illustration of the intervention rather than a route to fluent output.

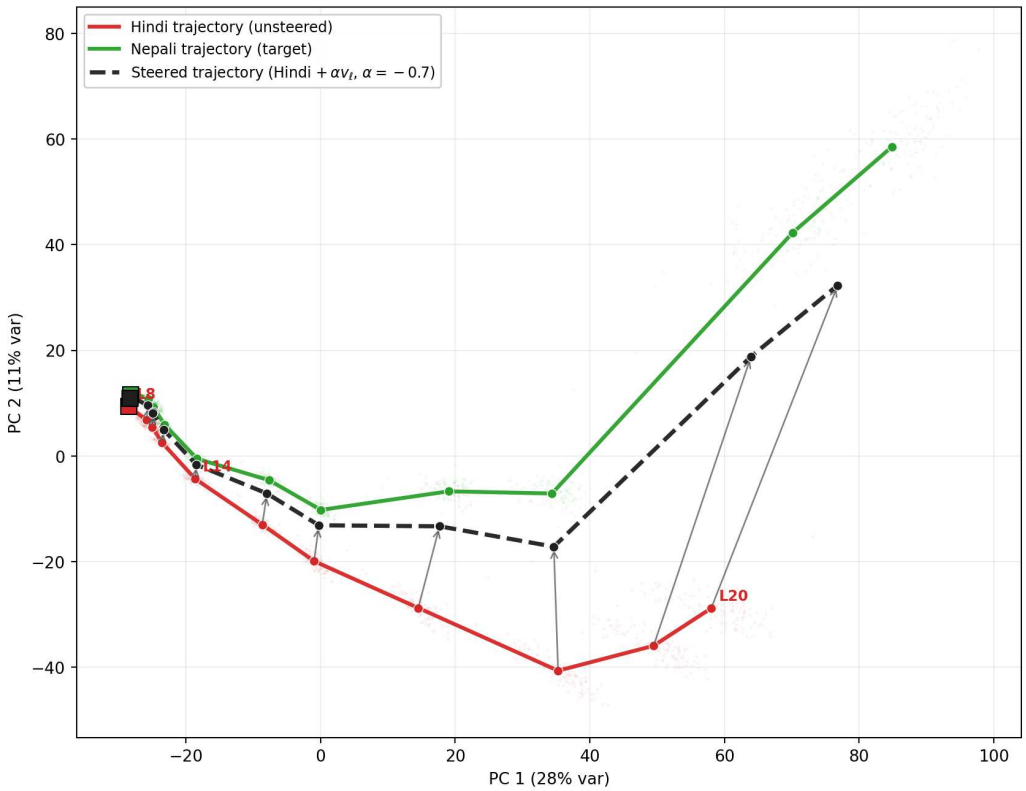


Figure 11: Depth-trajectory steering, illustrated (Llama, PCA-2D of the steerable band). Per-layer centroid paths across L8–L20 for Hindi (source) and Nepali (target), and the steered path obtained by adding each layer's steering vector αv_ℓ to the Hindi centroid ($\alpha = -0.7$, illustrative; arrows show the per-layer push). Injecting the direction along the entire trajectory shifts the Hindi path toward Nepali, yet it does not exceed single-layer steering and over-steering degenerates (Section 5.3.2). A linear (PCA) projection is used because it renders the additive edit $\mu_{\text{hin},\ell} + \alpha v_\ell$ as a visible translation; a nonlinear embedding does not