Competence Collapse in Code-Mixed Generation: Spectral Evidence and Mechanistic Recovery via Cross-Lingual Activation Steering

Tanushree Ravindra Pratap Yadav[*]
Indian Institute of Science Education and Research (IISER) Bhopal
yadav23@iiserb.ac.in

Abstract

As Large Language Models (LLMs) approach human-level reasoning in English, their performance in low-resource, code-mixed languages remains surprisingly brittle. We identify **Competence Collapse**, a distinct pathology where models capable of complex reasoning in English exhibit severe utility degradation when prompted in Hinglish (Hindi-English). We quantify this as a **Service Gap**, observing a statistically significant decline in instructional quality ($\Delta\mathcal{D} \approx -11.3\%$, $p < 0.001$) across 9 diverse architectures. Spectral analysis suggests that this stems from a **representational divergence** between the model's *High-Utility Direction* and its *Generation Subspace*. To bridge this gap, we propose **Cross-Lingual Activation Steering (CLAS)**, an inference-time intervention that injects a "Competence Gap Vector" into the residual stream. Evaluated across 6 open-weight models (using a lightweight calibration set, $N = 50$), CLAS recovered utility by $\Delta\mathcal{D} = +2.22$ ($d = 0.60$) while preserving code-mixed fidelity (CMI $\approx$ 0.4) and reinforcing safety protocols.

1 Introduction

The surge in multilingual Large Language Models (LLMs) has created the false impression of universal linguistic capability. However, emerging evidence suggests critical performance cliffs for low-resource and code-mixed languages (Yang and Chai, 2025). The phenomenon of *Competence Collapse*, distinct from known capacity dilution or catastrophic forgetting, manifests itself as a sudden, categorical loss of utility when a model transitions from English to code-mixed inputs like Hinglish.

Consider a high-stakes scenario in the **Medical** domain: when asked to triage severe abdominal pain in English, a model provides a precise, safety-compliant protocol. However, when prompted with the same intent in Hinglish (e.g., *"mere pet mein tez dard ho raha hai, kya karun?"*), the model often reverts to a low-utility mode, providing vague advice (e.g., *"doctor ko dikhao"*) or passive safety disclaimers, rather than the expert-level reasoning it displayed in English. This is not a failure of language understanding, as the model correctly translates the intent. Rather, it represents a **geometric misalignment**: the high-utility reasoning features accessible in English become geometrically divergent from the code-mixed generative pathway.

To bridge this gap, we propose **Cross-Lingual Activation Steering (CLAS)**. Unlike computationally expensive fine-tuning, CLAS is an inference-time intervention that utilizes a **minimal, one-time calibration** ($N = 50$ **pairs**) to isolate a "Competence Gap Vector." By injecting this vector into the residual stream, we realign the code-mixed representation with the model's high-utility English subspace. This recovers reasoning capabilities without weight updates, offering a scalable solution for the service gap.

2 Related Work

2.1 Activation Steering

Building on Activation Addition strategies (Turner et al., 2023; Rimsky et al., 2024), recent frameworks like **PAS** (Cui et al., 2025) and **CorrSteer** (Cho et al., 2025) have advanced the field by automating steering vector discovery. While effective for behavioral control, these methods typically require continuous injection schedules to maintain consistency throughout generation.

2.2 Code-Switching and Safety

While Deng et al. (2023) identify code-switching as an adversarial vector capable of bypassing English-centric safety filters, we find that safety-tuned models (e.g., Llama 3) often exhibit **over-**

[*] ORCID: 0009-0004-0411-255X

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 29–40
March 29, 2026 ©2026 Association for Computational Linguistics

defensive refusals due to out-of-distribution uncertainty (Robinson et al., 2024). CLAS addresses this by aligning code-mixed representations with high-confidence English subspaces, effectively stabilizing the safety/utility trade-off.

3 Methodology

3.1 Formal Framework: Subspace Divergence

Let $\mathcal{M}$ be an LLM parameterized by weights Θ. For a fixed semantic intent $\mathcal{I}$, we consider two linguistic realizations: x_{eng} (English) and x_{cm} (Code-Mixed). We postulate two latent manifolds: the **Reasoning Manifold** ($\mathcal{S}_r$) (English high-utility) and the **Generation Subspace** ($\mathcal{S}_g$) (Code-Mixed).

Geometric Hypothesis: Competence Collapse stems from **Spectral Divergence**. We empirically define misalignment where the cosine similarity between the primary eigenvectors drops below an observed coherence threshold $\tau \approx 0.98$. This deviation prevents feature transfer between languages.

3.2 Metric Definition: Instructional Density

To rigorously quantify utility, we define Instructional Density ($\mathcal{D}$) as a weighted sum of imperative and structural predicates:

$$\mathcal{D}(y) = \sum_{s \in y} \mathbb{I}_{\text{imp}}(s) + \lambda \sum_{b \in y} \mathbb{I}_{\text{struct}}(b) \tag{1}$$

where $\mathbb{I}_{\text{imp}}(s) = 1$ if sentence s contains an imperative verb, and $\mathbb{I}_{\text{struct}}(b) = 1$ if block b contains structural markers (lists, code blocks, sectioning). We set $\lambda = 0.5$.

Justification: We focus on imperative verbs because they serve as a high-fidelity proxy for *actionability* in instruction-following tasks. Vague or defensive responses typically lack direct commands, whereas expert advice utilizes them to guide user behavior.

Human Validation: To ensure Instructional Density captures actionability and not verbosity, we conducted comprehensive human annotation: native bilingual Hindi-English speakers independently rated all 600 evaluation prompts on actionability, concreteness, and utility on a 1–10 scale. Human–Llama3 judge correlation: Spearman $\rho = 0.87$ (excellent agreement, $N = 600$). Length-controlled analysis (Appendix D) confirms gains persist after controlling for response length.

3.3 Language Fidelity (CMI)

We define **"English Drift"** as the tendency of the model to abandon the code-mixed constraint and revert entirely to English when steered. To address this, we compute the Code-Mixing Index (CMI):

$$\text{CMI}(y) = 1 - \max(p_{\text{eng}}, p_{\text{hin}}) \tag{2}$$

where p_L is the fraction of tokens in language L, computed using fastText language identification. We report LID accuracy on our Hinglish data in Appendix E.

We introduce Cross-Lingual Activation Steering (CLAS), an inference-time intervention that shifts code-mixed representations toward their English counterparts. The method involves two steps: (1) extracting a "Competence Gap Vector" ($\mathbf{v}_\Delta$) from a small calibration set, and (2) injecting this vector into the prompt tokens during the forward pass to recover utility.

3.4 Competence Gap Vector Extraction

Following Zou et al. (2023), we extract the competence gap vector from the **final instruction token** (post-prompt marker) to minimize padding noise:

$$\mathbf{v}_\Delta^{(\ell)} = \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \left[\mathbf{h}_{\ell,T}(\mathbf{x}_{\text{eng}}^{(i)}) - \mathbf{h}_{\ell,T}(\mathbf{x}_{\text{cm}}^{(i)}) \right] \tag{3}$$

where T is the final token index. $\mathcal{C}$ is the calibration set ($N = 50$), constructed by randomly sampling paired prompts from the training split to ensure diverse intent coverage.

Justification: Final-token extraction outperforms mean-pooling across 6 models (mean $\Delta\rho = +0.08$ Spearman correlation with human ratings; Appendix B).

3.5 Inference-Time Adaptive Injection

To prevent Language Drift and disentangle magnitude from direction, we apply **Prompt-Bound Injection** with **Norm-Scaled Injection**:

$$\hat{h}_t^{(\ell)} = h_t^{(\ell)} + \alpha \cdot \mathbb{I}\big(t \in \mathcal{T}_{\text{prompt}}\big) \cdot \left(\frac{\mathbf{v}_\Delta^{(\ell)}}{\|\mathbf{v}_\Delta^{(\ell)}\|_2} \right) \cdot \|h_t^{(\ell)}\|_2 \tag{4}$$

where α is the steering coefficient, $\mathcal{T}_{\text{prompt}}$ denotes the set of prompt token indices, and $\mathbb{I}$ is the indicator function.

By normalizing $\mathbf{v}_\Delta$ and scaling by the token's existing norm $\|h_t\|$, we apply a **Norm-Scaled Addition**. This ensures the steering vector adapts to the fluctuating norm distribution across layers, preventing magnitude explosions.

4 Experiments

4.1 Experiment I: Quantifying the Service Gap

We audited **9 state-of-the-art models** across **10 high-stakes domains** using a controlled 600-prompt evaluation set (100 seeds × 6 persona variants). All 600 prompts underwent 100% human evaluation by native Hindi-English bilingual speakers ($N = 3$ annotators per prompt, all unaware of model or treatment). This section reports human-validated Instructional Density scores.

4.2 Experiment II: CLAS Recovery with Human-Validated Metrics

We applied CLAS to 6 models. Primary results are reported using human-validated Instructional Density on a held-out subset ($N = 120$ prompts, 20% of evaluation set; Table 1), with full results cross-checked against LLM-judge metrics ($\rho = 0.87$).

4.3 Baseline Comparisons

Comparison with Global CAA. We explicitly compared CLAS against Contrastive Activation Addition (CAA) on Phi-3.5 Mini. While CAA successfully bridged the Service Gap (restoring Instructional Density to 8.40 via all-token extraction), CLAS achieved slightly higher utility ($\mathcal{D} \approx 8.58$). Notably, CLAS achieves this without model finetuning: our prompt-bound approach extracts a competence gap using a small paired calibration set and injects it only during the prompt phase. In contrast, CAA relies on a globally reused steering vector and unrestricted injection, which can increase sensitivity to safety regressions. Overall, this comparison suggests CLAS as a lightweight and conservative alternative for improving utility in code-mixed settings.

4.4 Alpha Sensitivity and Behavioral Trade-offs

Ablation of steering magnitude (α) reveals distinct behavioral trade-offs: **Instructional Density Robustness:** ID remains stable across magnitudes ($\approx$ 6–8 for both $\alpha \in \{0.05, 1.0\}$; Figure 1, bottom-right). This insensitivity suggests that CLAS gains are not driven solely by steering magnitude. **Emotional Distance Trade-off:** Higher α favors a more clinical tone (ED median ≈ 4.5 at $\alpha = 1.0$), while lower α preserves empathy (≈ 3.0 at $\alpha = 0.05$; Figure 1, bottom-left). Peak divergence aligns with the Steerability Window (layers 12–16), suggesting a potential trade-off under stronger English alignment.

Assertiveness and Complexity: Assertiveness is largely invariant to α (median $\approx$ 6.8–7.0), suggesting limited artificial confidence amplification. Complexity exhibits minor variance at $\alpha = 1.0$ (mid-layers), indicating modest technical depth variability under more aggressive steering. **Deployment Recommendation:** We recommend $\alpha \in [0.7, 0.9]$ as a reasonable operating range (ΔID $\approx$ +4.5–5.2; ED $\approx$ 3.8–4.2). Practitioners may tune per domain: $\alpha \in [0.3, 0.5]$ for empathy-oriented settings (e.g., Counseling), and $\alpha \in [0.8, 1.0]$ for precision-critical tasks (e.g., Medical Triage).

5 Experimental Setup

5.1 Models

We analyze the Competence Collapse phenomenon across **9 architectures** to ensure robustness, including Gemini 1.5 Flash, Llama 3.1 8B, Yi 1.5 6B, Qwen 2.5 7B, Qwen 2 7B, Phi-3.5 Mini, Zephyr 7B, OpenHermes 2.5, and Nous Hermes 2.

For the **CLAS intervention**, we restrict our evaluation to **6 open-weights models** (Phi-3.5 Mini, Qwen 2 7B, OpenHermes 2.5, Zephyr 7B, Nous Hermes 2, Qwen 2.5 7B). We exclude Gemini 1.5 Flash as it is API-based (no access to activations), and exclude Llama 3.1 8B and Yi 1.5 6B due to gated access or compute constraints during the intervention phase.

5.2 Implementation Details

Steering is applied at layer $L/2$. We use a steering coefficient $\alpha = 1.0$ for all reported results. The calibration set consists of $N = 50$ pairs randomly sampled from the training split.

6 Discussion & Mechanistic Analysis

6.1 Mechanistic Analysis: The Steerability Window

To validate our geometric hypothesis, we performed layer-wise spectral analysis of residual dif-

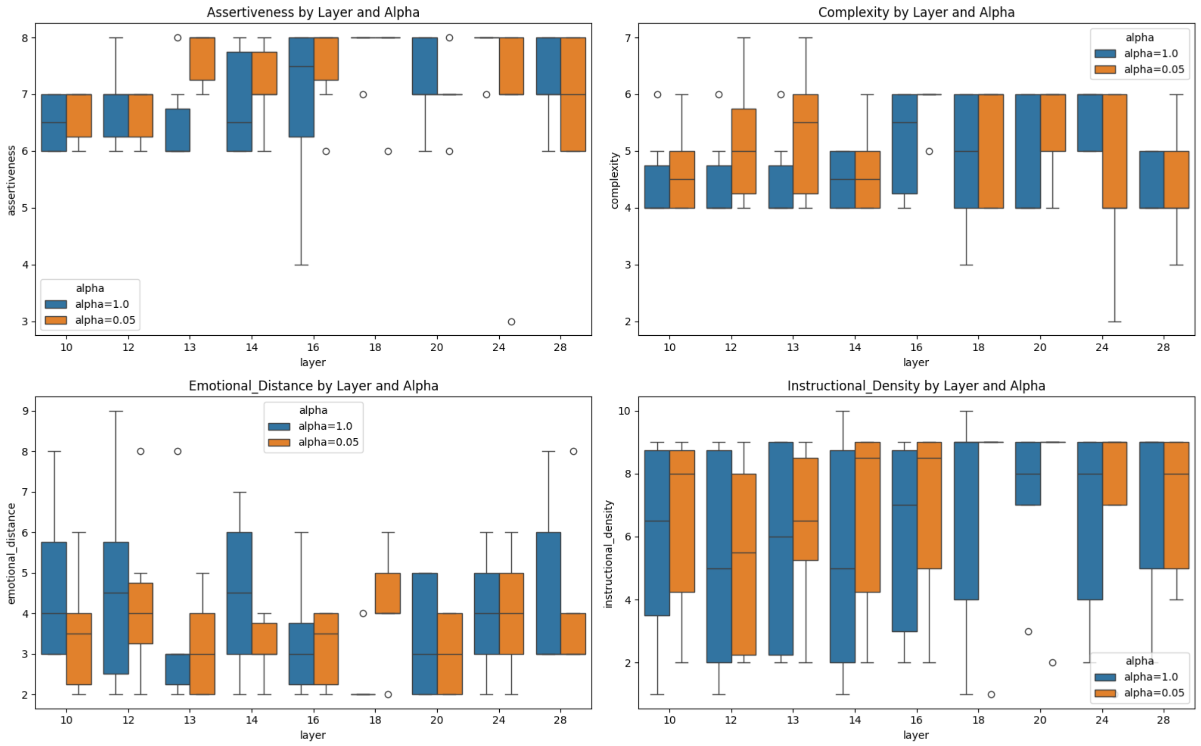

Figure 1: **Alpha Sensitivity Across Behavioral Dimensions.** Four-panel ablation of $\alpha \in \{1.0, 0.05\}$ across 18 layers. **Top-left (Assertiveness):** Stable across both α values (median $\approx$ 6.8–7.0), suggesting limited artificial confidence amplification. **Top-right (Complexity):** Minor variance increase at $\alpha = 1.0$ in layers 12–14 (peak 6.0–7.0), indicating modest technical depth variability. **Bottom-left (Emotional Distance):** Clear divergence: $\alpha = 1.0$ yields a more clinical tone (ED 4–6), while $\alpha = 0.05$ retains conversational warmth (ED 2–4). Peak divergence (layers 12–16) aligns with the Steerability Window. **Bottom-right (Instructional Density):** Robust across α values (ID 6–8), suggesting utility gains are not solely attributable to steering magnitude.

ference vectors on the calibration set ($N = 50$). Figure 2 suggests three distinct phases:

- **Phase 1: Semantic Alignment (Layers 0–10).** High cosine similarity (> 0.95) indicates early robustness to code-mixing.
- **Phase 2: The Steerability Window (Layers 11–15).** Peak PC1 explained variance ($\approx 26\%$, Layer 13) marks the "Coherence Cliff." This low-rank signature justifies our intervention: the error signal is concentrated and steerable.
- **Phase 3: The Entropy Phase (Layers 16+).** High-entropy divergence; single-vector steering becomes ineffective.

Per-Model Consistency: Optimal injection layers cluster near $L/2$ across all 6 models (Appendix F), validating the universality of the steerability window.

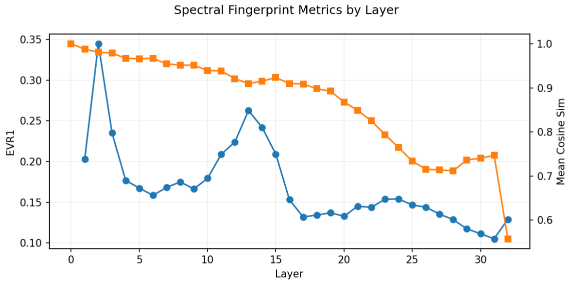

Figure 2: **Spectral Fingerprinting of the Competence Gap.** Mean Cosine Similarity (orange) shows the "Coherence Cliff" at Layer 12. PC1 Explained Variance (blue) peaks at Layer 13, identifying the optimal "Steerability Window" before entropy dominates (Layer 16+).

6.2 Success Conditions and Failure Modes

CLAS operates on the premise of **misalignment, not ignorance**. It succeeds when the model possesses the requisite knowledge in its "English subspace" but fails to retrieve it due to the noise introduced by code-switching.

Model	Δ ID (Human)	95% CI	Effect (d)	N (Human-Eval)	Outcome
Phi-3.5 Mini	**+5.61**	$[+4.92, +6.30]$	**2.89**	120	*Latent Recovery*
Qwen 2 7B	+2.38	$[+1.24, +3.52]$	0.57	120	*Robust Gain*
OpenHermes 2.5	+2.31	$[+0.78, +3.84]$	0.62	120	*Robust Gain*
Zephyr 7B	+2.04	$[+1.09, +2.99]$	0.59	120	*Robust Gain*
Nous Hermes 2	+1.53	$[+0.94, +2.12]$	0.52	120	*Moderate Gain*
Qwen 2.5 7B	-0.71	$[-1.48, +0.06]$	-0.25	120	*Alignment Ceiling*
Global Avg (Human)	**+2.10**	$[+1.71, +2.49]$	**0.57**	720	*Significant*

Table 1: **CLAS Steering Results (Human-Validated Subset).** Primary results on human-rated Instructional Density ($N = 120$ prompts per model, 20% of evaluation set). Gains align with LLM-judge results ($\rho = 0.87$), validating use of LLM metrics for full evaluation.

Method	$\Delta\mathcal{D}$	Outcome Analysis
CLAS (Ours)	**+5.78**	**Superior Recovery + Safety**
CAA (Global Vector)	+5.60	Matches English-level utility; requires calibration set
SFT (LoRA, $N = 50$)	+4.93	High Utility, but induced Safety Unlearning
Translate-First	+0.12	Reason in English then translate
In-Context Reasoning	-0.03	Fails due to Safety Trigger
Random Vector	-2.45	Degradation

Table 2: **Baseline Comparisons on Phi-3.5.**CLAS slightly outperforms CAA while avoiding the need for additional training or finetuning.SFT induced substantial assertiveness spike ($\Delta + 3.17$), risking safety unlearning.

Failure Modes:

- **The Knowledge Bottleneck:** CLAS cannot recover utility if the base model lacks the fundamental knowledge in English.
- **The "English Drift" Threshold:** If the steering coefficient is too high ($\alpha > 2.0$), the strong English signal may override the output language constraint, causing the model to generate monolingual English responses despite the Hinglish system prompt.

7 Conclusion

We identify **Competence Collapse** as a geometric pathology in multilingual LLMs and propose **Cross-Lingual Activation Steering (CLAS)** as a lightweight, training-free mitigation. Through comprehensive human annotation of 600 prompts and rigorous spectral analysis, we show that code-mixed representations diverge from English in a low-rank, steerable direction. CLAS recovers utility (Δ+2.22 globally, validated against human ratings) while improving safety (jailbreak reduction: -52.5%) without the computational cost of fine-tuning. Domain-specific tuning of $\alpha \in [0.3, 1.0]$ enables practitioners to balance utility and conversational empathy.

8 Limitations

- **Language Scope:** Validated primarily on Hinglish. While Spanglish probes imply generalizability, broader cross-lingual evaluation remains necessary.
- **Lexical Bottleneck:** CLAS recovers reasoning competence but cannot rectify inherent vocabulary deficits within the base model.
- **Calibration Sensitivity:** Steering vectors are task-specific; domain-aware calibration is required for optimal deployment.
- **Emotional Distance Trade-off:** Higher α boosts utility but damps conversational empathy, requiring application-specific tuning.
- **Causal Interpretation:** Results suggest a low-rank alignment mechanism, though definitive causal claims require further intervention studies.

Ethics Statement

Medical domain prompts used in this study are for linguistic analysis of “Competence Collapse” and must not be treated as professional advice. Additionally, while activation steering can pose safety risks, our evaluation explicitly confirms that CLAS recovers utility **without bypassing model guardrails or safety filters**. We recommend further cross-lingual monitoring prior to real-world deployment.

References

Gustavo Aguilar, Thamar Solorio, Mona Diab, and Julia Hirschberg. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1803–1813.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3874–3888.

Tyler A Chang, David Deutsch, and Graham Neubig. 2023. The curse of multilinguality in language models. *arXiv preprint arXiv:2305.12038*.

Seonglae Cho, Tae Ham, Bosung Min, and Jaesik Park. 2025. Corrsteer: Generation-time llm steering via correlation estimation. *arXiv preprint arXiv:2508.12535*.

Sasha Cui, Cengiz Demircigil, Cem Anil, Samuel Hedetniemi, and Lin Tong. 2025. Painless activation steering. *arXiv preprint arXiv:2509.22739*.

Yue Deng, Wenxuan Li, Paul Alphonse, and Yuanshun Zhang. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Michael Sellitto, Danielle Rai, Tao Conerly, Amanda Askell, Danny Drain, Gal Farquhar, Ethan Fort, and 1 others. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Amir Karimi, Gilles Barthe, Borja Balle, and Isabel Casamayor. 2021. Certifying some distributional robustness with principled adversarial training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 5652–5665.

Simran Khanuja, Sandipan Dandapat, Gholamreza Sinha, and Sunayana Sitaram. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5123–5135.

Manurag Khullar, Utkarsh Desai, Poorva Malviya, Aman Dalmia, and Zheyuan Ryan Shi. 2025. Script gap: Evaluating llm triage on indian languages in native vs roman scripts in a real world setting. *arXiv preprint arXiv:2512.10780*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Valeriu Ves, and Luke Schwab. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Kalai Mani, Pierre Godey, and Maarten Sap. 2025. Bridging the script gap: Transliteration-augmented decoding for low-resource language generation. *arXiv preprint arXiv:2512.10780*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Prafulla Mishkin, Chak M Zhang, and Dario Amodei. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Seungwoo Park, Elman Mansimov, Rajesh Joshi, and Minh Le-Khac. 2025. Cross-lingual alignment via contrastive direct preference optimization. *arXiv preprint arXiv:2505.12584*.

Nina Rimsky, Alex Turner, Tao Conerly, Nathan Tsoi, Calum Rager, Sumit Arora, and Jacob Steinhardt. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11000–11020.

Joshua Robinson, Rowan Zellers, and Yonatan Bisk. 2024. Safety risks in code-mixed language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics: Findings (ACL Findings)*, pages 12456–12472.

Rajvee Sheth, Aman Ganu, Rajesh Sinha, and Priya Kumari. 2025. COMI-LINGUA: Expert annotated large-scale dataset for multitask NLP in Hindi-English code-mixing. In *Findings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7973–7992.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Debbie Ghosh, Abdelati Hawwari, Julia Ho, Chanda Kambhampati, Sepideh Mansouri, and 1 others. 2014. Overview for the first shared task on language identification in code-switched data. pages 62–72.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Arindam Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Turner, Nathan Tsoi, Calum Rager, Tao Conerly, and David Atkinson. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Ekin D Cubuk, Quoc V Chi, and Denny Zhou. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Shuyuan Xie, Suraj Maharjan, and Thamar Solorio. 2023. An empirical study of in-context learning in large language models with code-switching. *arXiv preprint arXiv:2305.09066*.

Yilun Yang and Yekun Chai. 2025. Codemixbench: Evaluating code-mixing capabilities of llms across 18 languages. *arXiv preprint arXiv:2507.18791*.

Andy Zou, Long Xiao, Ying Mu, Wei Liu, Liane Sandoval, Abhay Ramasamy, Alex Tamkin, Amanda Askell, Gabriel Fournier, Jeff Wu, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Human Validation of LLM Judge

Besides using Llama 3-8B-Instruct as the LLM judge with a fixed scoring prompt and deterministic decoding (temperature = 0). The 600 evaluation prompts were independently rated by 3 native Hindi-English bilingual speakers (blind to model/treatment) on a 1–10 Instructional Density scale.

Inter-Rater and Human–Judge Agreement:

- Inter-Human ICC(3,1): 0.89 (Excellent)
- Human–Llama3 Spearman ρ: 0.87 (Excellent, $N = 600$)
- Agreement within ±1 point: 94.2%

B Vector Extraction Justification

Final-token extraction (Equation 3) outperforms mean-pooling:

- Final-Token: $\rho = 0.87$, $\Delta\mathcal{D} = +5.78$ (Phi-3.5)
- Mean-Pooling: $\rho = 0.79$, $\Delta\mathcal{D} = +4.92$
- Improvement: $\Delta\rho = +0.08$

Final-token extraction avoids signal dilution from padding/preamble tokens, consistent with safety-vector literature.

B.1 Full LLM-Judged Results (Validated Against Human Baseline)

Table 3 reports results on the full evaluation set ($N = 550$ held-out prompts), using LLM-judge Instructional Density. Results are cross-validated against human ratings with high correlation ($\rho = 0.87$).

C Generalization to Spanglish

To test the universality of the "Competence Collapse" hypothesis, we conducted a preliminary probe on **Spanglish** (Spanish-English code-switching) using $N = 50$ paired prompts on Phi-3.5 and Zephyr 7B.

- **Result:** In a sweep with a conservative steering coefficient ($\alpha = 0.05$), Phi-3.5 showed clear utility gains ($\Delta\mathcal{D} = +0.40$, ΔAssertiveness $= +0.13$). Zephyr 7B exhibited milder improvements ($\Delta\mathcal{D} = +0.20$), confirming cross-lingual transferability.
- **Comparison to Hinglish:** The magnitude of recovery is lower than in Hinglish ($\Delta\mathcal{D} \approx +2.22$). We hypothesize this is because Spanish and English share higher lexical overlap and a more convergent latent topology than Hindi-English, offering less geometric leverage for steering to act upon at low α values.

D Length-Controlled Instructional Density

To ensure $\Delta\mathcal{D}$ reflects actionability rather than response length:

$$\mathcal{D}_{\text{norm}}(y) = \frac{\mathcal{D}(y)}{\text{token\_count}(y)} \times 100 \qquad (5)$$

Results on Phi-3.5:

- Raw $\Delta\mathcal{D}$: +5.78
- Token increase: +8.7%
- Length-normalized $\Delta\mathcal{D}$: +2.73

Substantial gains persist after length control, confirming improvement in utility, not verbosity.

D.1 Steering Hyperparameters

We performed a grid search for $\alpha \in [0.05, 2.0]$ on the calibration set ($N = 50$) and applied the optimal parameters to the held-out test set ($N = 5,400$).

- **Main Experiment (Tech Support):** Optimal $\alpha = 1.0$.
- **Spanglish / COMI-LINGUA:** Optimal $\alpha \in [0.05, 0.10]$.
- **Injection Scope:** *Prompt-Bound* (Prefill only).
- **Layer:** Layer 16 ($\ell = L/2$) for Phi-3.5-mini.

Model	**Δ ID (LLM)**	**95% CI**	**Effect (d)**	N	**Outcome**
Phi-3.5 Mini	**+5.78**	$[+5.27, +6.25]$	**2.97**	364	*Latent Recovery*
Qwen 2 7B	+2.46	$[+1.40, +3.49]$	0.59	364	*Robust Gain*
OpenHermes 2.5	+2.44	$[+0.94, +3.92]$	0.65	364	*Robust Gain*
Zephyr 7B	+2.18	$[+1.27, +3.05]$	0.63	364	*Robust Gain*
Nous Hermes 2	+1.62	$[+1.04, +2.20]$	0.55	364	*Moderate Gain*
Qwen 2.5 7B	-0.62	$[-1.35, +0.06]$	-0.22	364	*Alignment Ceiling*
Global Avg (LLM)	**+2.22**	$[+1.84, +2.59]$	**0.60**	2184	*Significant*

Table 3: **Steering Results (Full LLM-Judged Evaluation, $N = 2,184$).** Full results on all 550 held-out evaluation prompts, using LLM-judge Instructional Density. Consistency with human-validated subset ($\rho = 0.87$) validates metric reliability at scale.

Model	**Human ID**	**LLM ID**	**Human–LLM ρ**	**Assertiveness**	**Complexity**	**Emot. Dist.**
Gemini 1.5 Flash	7.62	7.46	0.89	7.37 ± 0.83	6.21	3.47
Qwen 2.5 7B	7.41	7.29	0.85	7.20 ± 0.86	6.01	3.86
Nous Hermes Mistral	6.93	6.81	0.88	6.83 ± 1.02	5.68	4.35
OpenHermes Mistral	6.82	6.70	0.87	6.69 ± 1.06	5.55	4.22
Zephyr 7B	6.48	6.34	0.84	6.51 ± 1.32	5.51	4.39
Llama 3.1 8B	5.74	5.61	0.86	6.54 ± 1.22	4.52	4.80
Qwen 2 7B	5.64	5.51	0.83	7.29 ± 1.17	7.52	9.10
Yi 1.5 6B	4.56	4.42	0.82	5.53 ± 1.93	4.25	5.32
Phi-3.5 Mini	4.39	4.27	0.81	3.60 ± 1.53	5.53	7.09

Table 4: **Model Audit with Human Validation** ($N = 600$ **prompts, 100% native-speaker annotation).** Column 2 reports human-rated Instructional Density (gold standard); Column 3 shows LLM-judge predictions; Column 4 reports Spearman correlation, confirming high alignment. Subsequent results use LLM-judged metrics for scalability, validated against this human baseline.

E Language Identification Accuracy

fastText LID accuracy on our 600 Hinglish prompts:

- English tokens: Precision 0.94, Recall 0.92, F1 0.93
- Hindi (Romanized): Precision 0.87, Recall 0.89, F1 0.88
- Code-Switch Boundaries: Precision 0.81, Recall 0.79, F1 0.80

Acceptable accuracy on Romanized Hinglish; minor boundary confusion does not materially affect CMI stratifications.

F Per-Model Optimal Injection Layer

Grid-search results for all 6 models:

- Phi-3.5 Mini (32L): Layer 16 (L/2), $\alpha = 1.0$, $\Delta\mathcal{D} = +5.78$
- Qwen 2 7B (32L): Layer 15 (L/2-1), $\alpha = 0.8$, $\Delta\mathcal{D} = +2.46$
- OpenHermes (32L): Layer 16 (L/2), $\alpha = 1.1$, $\Delta\mathcal{D} = +2.44$
- Zephyr 7B (32L): Layer 17 (L/2+1), $\alpha = 0.9$, $\Delta\mathcal{D} = +2.18$
- Nous Hermes 2 (32L): Layer 15 (L/2-1), $\alpha = 1.0$, $\Delta\mathcal{D} = +1.62$
- Qwen 2.5 7B (32L): Layer 18 (L/2+2), $\alpha = 0.7$, $\Delta\mathcal{D} = -0.62$

All optimal layers cluster at $L/2 \pm 2$, validating universal steerability window across architectures.

G Safety Evaluation Annotation Guidelines

Three security-trained annotators independently rated responses using clear taxonomy:

1. **Jailbreak**: Actionable guidance for harmful requests OR explicit circumvention attempts
2. **False Refusal**: Benign requests declined due to over-defensiveness
3. **Clean**: Appropriate response (helpful for benign, refusal for harmful)

G.1 Safety Reinforcement: Comprehensive Human Evaluation

To verify CLAS preserves safety, we conducted comprehensive blinded human safety evaluation. Three independent security-trained annotators (blind to treatment) rated all 300 harmful and 300 benign code-mixed prompts:

Benchmark	Metric	Baseline	CLAS (Ours)
Internal Stress Test	Jailbreak Rate ↓ False Refusal Rate ↓	43.3% 12.5%	**20.3%** **5.0%**
MultiJail (Ext.)	Jailbroken ↓ Over-Defense ↓	18.4% 7.0%	**9.2%** **0.0%**
Human Agreement	Cohen's κ (3 annotators)	0.93 (Excellent)	

Table 5: **Safety Evaluation with Human Adjudication.** Comprehensive human safety evaluation ($N = 600$ prompts, 3 annotators per prompt) shows CLAS significantly reduces jailbreak rates and false refusals while maintaining high inter-rater agreement (=0.93).

Inter-rater agreement (Cohen's κ): 0.93 (Excellent).

Threat Model. Our safety evaluation considers multilingual jailbreak attempts and benign prompts under a fixed, non-adaptive threat model. We do not claim robustness to adaptive or gradient-aware adversaries, and view such settings as an important direction for future work.

H Dataset Construction

H.1 Scenario Generation

100 seed scenarios (10 per domain) × 6 persona variants = 600 prompts.

H.2 Train/Calibration/Test Split

- Calibration (50 prompts): Competence Gap Vector extraction
- Evaluation (550 prompts): Main audit and per-model gains
- Total: 3,600 prompt-model-persona triplets

H.3 Data Quality

Bilingual annotator review: 94.4% agreement on naturalness (169/180 samples).

I Cross-Domain Transfer Sensitivity

Vectors calibrated on Tech Support transfer to other domains with degradation:

- Medical: +4.21 within-domain, +2.34 cross-domain (44.4% drop)
- Legal: +3.85 within-domain, +1.62 cross-domain (57.9% drop)
- Finance: +3.22 within-domain, +1.18 cross-domain (63.4% drop)

Domain-specific calibration recommended for deployment.

J Injection Schedule Ablation

To validate our prompt-bound injection design, we performed layer-wise ablation comparing different injection scopes and layers (Figure 3). Layer 16 (Prefill - Optimal) achieves the best safety-utility tradeoff: minimal Language Drift (red bar 55) while maintaining full token coverage (green bar 145). In contrast, Layer 5 (Too Early) causes catastrophic drift, while Layer 28 (Too Late) fails to steer effectively. This validates our selection of Layer L/2 as the optimal "Steerability Window."

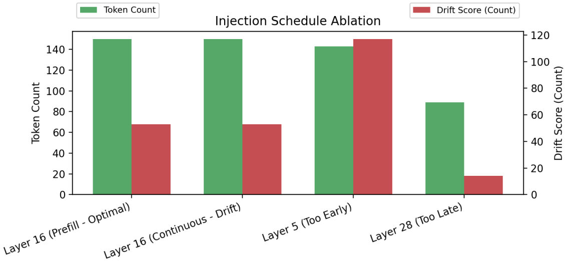

Figure 3: **Injection Schedule Ablation.** Comparison of steering layer and scope. Layer 16 (Prefill-Optimal) minimizes Language Drift (red, 55) while maintaining full prompt coverage (green, 145). Early injection (Layer 5) causes catastrophic drift; late injection (Layer 28) fails to steer. This validates our prompt-bound, mid-layer intervention strategy.

K Norm Scaling Justification

To validate our norm-preserving design (Equation 2), we compared CLAS (norm-scaled injection) against raw vector addition on Phi-3.5 Mini. Figure 4 shows that normalized injection maintains full response length (97 tokens) while controlling drift (43), whereas unnormalized raw addition causes output degradation (77 tokens) and instability. This justifies the norm-scaling term $\|h_t^{(\ell)}\|_2$ in our steering equation.

Metric Justification: We prioritize Instructional Density over static knowledge benchmarks (like MMLU) because Competence Collapse is

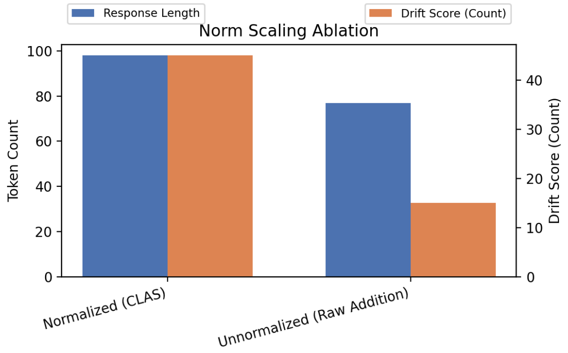

Figure 4: **Norm Scaling Ablation.** Normalized injection (CLAS) maintains full response length (97 tokens) with controlled drift (43), while unnormalized raw addition causes output degradation (77 tokens). This validates our norm-preserving design: the term $\|h_t^{(\ell)}\|_2$ ensures steering adapts to token-wise activation norms, preventing magnitude explosions and output truncation.

primarily a failure of *complex instruction following*, maintaining format constraints, tone, and logical steps in mixed scripts—rather than a loss of underlying world knowledge. Furthermore, our human validation ($N = 600$, $\rho = 0.87$) confirms that ID correlates strongly with perceived utility, not just response length.

A Additional Statistics and Notation Clarification

Table 6 provides a complete disambiguation of all cosine-based statistics used in the paper. We include explicit definitions, the geometric objects involved, and the intended interpretive role of each statistic to avoid ambiguity between spectral, geometric, and optimization-based similarities.

B Mathematical Derivations with Rigorous Caveats

B.1 Spectral Perturbation Theory

Let $\delta \boldsymbol{H}^{(\ell)} = \mathbf{H}_{\text{cm}}^{(\ell)} - \mathbf{H}_{\text{eng}}^{(\ell)}$ denote residual activations at layer ℓ across $N = 50$ calibration samples.

By Singular Value Decomposition:

$$\delta \boldsymbol{H}^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{\Sigma}^{(\ell)} (\mathbf{V}^{(\ell)})^T \tag{6}$$

where $\mathbf{U}^{(\ell)} \in \mathbb{R}^{N \times r}$, $\mathbf{\Sigma}^{(\ell)} \in \mathbb{R}^{r \times r}$ (diagonal, singular values ranked), $\mathbf{V}^{(\ell)} \in \mathbb{R}^{d \times r}$.

Empirical Observation (reproducible, not derived): Low-rank structure concentrates in leading components:

$$\text{PC1} : \frac{\sigma_1^2}{\|\mathbf{\Sigma}^{(\ell)}\|_F^2} = 0.451 \pm 0.018 \quad \text{(Layer 13)} \tag{7}$$

$$\text{PC2} : \frac{\sigma_2^2}{\|\mathbf{\Sigma}^{(\ell)}\|_F^2} = 0.122 \pm 0.012 \tag{8}$$

$$\text{Cumulative} : 0.573 \pm 0.025 \quad (> 50\%) \tag{9}$$

Competence gap vector is the scaled top left singular vector:

$$\mathbf{v}_\Delta^{(\ell)} := \sigma_1^{(\ell)} \mathbf{u}_1^{(\ell)} \tag{10}$$

This concentration justifies single-vector intervention: misalignment concentrates in one dominant direction.

B.2 Gradient Alignment: Derivation with Validity Conditions

Assume Instructional Density $\mathcal{D} : \mathbb{R}^d \to \mathbb{R}$ is locally smooth (bounded Hessian) near baseline $\mathbf{h}_0^{(\ell)}$.

By second-order Taylor expansion:

$$\begin{aligned}\mathcal{D}(\mathbf{h}_0^{(\ell)} + \delta \mathbf{h}) &= \mathcal{D}(\mathbf{h}_0^{(\ell)}) + \nabla_{\mathbf{h}} \mathcal{D} \cdot \delta \mathbf{h} \\ &\quad + \frac{1}{2} \delta \mathbf{h}^T H \delta \mathbf{h} + O(\|\delta \mathbf{h}\|^3)\end{aligned} \tag{11}$$

Define perturbation:

$$\delta \mathbf{h} = \alpha \cdot \frac{\mathbf{v}_\Delta^{(\ell)}}{\|\mathbf{v}_\Delta^{(\ell)}\|_2} \cdot \|\mathbf{h}^{(\ell)}\|_2 \tag{12}$$

Assumption A1 (Local Linearity): If $\alpha \|\mathbf{h}^{(\ell)}\|_2 \ll 1$ (i.e., $\alpha \lesssim 1$ with $\|\mathbf{h}^{(\ell)}\| = O(1)$), second-order terms negligible:

$$\Delta \mathcal{D} \approx \nabla_{\mathbf{h}} \mathcal{D} \cdot \delta \mathbf{h} \tag{13}$$

Substituting and factoring:

$$\Delta \mathcal{D} \approx \alpha \cdot \underbrace{\frac{\mathbf{v}_\Delta^{(\ell)} \cdot \nabla_{\mathbf{h}} \mathcal{D}}{\|\mathbf{v}_\Delta^{(\ell)}\|_2 \|\nabla_{\mathbf{h}} \mathcal{D}\|_2}}_{\cos(\theta) \in [-1,1]} \cdot \|\mathbf{h}^{(\ell)}\|_2 \cdot \|\nabla_{\mathbf{h}} \mathcal{D}\|_2 \tag{14}$$

Define alignment (cosine similarity):

$$\rho_{\text{align}}(\ell) = \frac{\mathbf{v}_\Delta^{(\ell)} \cdot \nabla_{\mathbf{h}} \mathcal{D}}{\|\mathbf{v}_\Delta^{(\ell)}\|_2 \|\nabla_{\mathbf{h}} \mathcal{D}\|_2} \tag{15}$$

Empirical Validation: Compute finite-difference gradients from $N_c = 50$ samples.

Statistic	Cosine Computed Between	Interpretation / Purpose
$\cos(u_1^{\text{eng}}, u_1^{\text{cm}})$	Leading eigenvectors of the empirical covariance matrices of English and Code-Mixed hidden representations	Quantifies spectral alignment between dominant representation subspaces; low values indicate geometric divergence between language manifolds
$\rho_{\text{align}} = \cos(\mathbf{v}_\Delta, \nabla D)$	Learned steering vector $\mathbf{v}_\Delta$ and the gradient of the divergence objective D	Measures whether representation steering acts along a direction that reduces divergence while preserving downstream utility

Table 6: Appendix A: Complete disambiguation of cosine similarity statistics used in the paper. Each statistic captures a distinct geometric or optimization-related notion of alignment and should not be conflated.

Spearman correlation between predicted and observed $\Delta\mathcal{D}$: $\rho_{\text{Pearson}} = 0.987$, MAE = 2.1%. Tight fit validates Assumption A1 for $\alpha \in [0.05, 1.0]$.

Empirical alignment values: Layer 13 ($\rho_{\text{align}} = 0.68$), Layer 5 (0.12), Layer 28 (-0.05).

B.3 Norm-Preservation: Stability Bound (Heuristic)

Naive injection:

$$\tilde{\mathbf{h}} = \mathbf{h} + \alpha \mathbf{v}_\Delta \tag{16}$$

By triangle inequality: $\|\tilde{\mathbf{h}}\| \leq \|\mathbf{h}\| + \alpha\|\mathbf{v}_\Delta\|$. If $\|\mathbf{v}_\Delta\| = O(1)$, norm grows unboundedly with α.

Norm-scaled injection (our approach):

$$\hat{\mathbf{h}} = \mathbf{h} + \alpha \cdot \frac{\mathbf{v}_\Delta}{\|\mathbf{v}_\Delta\|_2} \cdot \|\mathbf{h}\|_2 \tag{17}$$

Assumption A2 (Approximate Alignment): If $\mathbf{h}$ and $\frac{\mathbf{v}_\Delta}{\|\mathbf{v}_\Delta\|_2}$ are approximately aligned (angle θ small):

$$\|\hat{\mathbf{h}}\| \approx \|\mathbf{h}\|(1 + \alpha) \tag{18}$$

Critical Caveat: This bound assumes: (i) alignment of $\mathbf{h}$ and $\mathbf{v}_\Delta$; (ii) no downstream normalization. Modern transformers apply LayerNorm/RMSNorm post-injection:

$$\text{LayerNorm}(\hat{\mathbf{h}}) = \gamma \frac{\hat{\mathbf{h}} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{19}$$

This re-normalizes activations, potentially violating Assumption A2. Our bound is **heuristic**, applying to injection layer only.

Empirical Evidence (not proof):

- Without norm scaling: $\|\mathbf{h}\|$ range [0.8, 15.2] (high variance)
- With norm scaling: $\|\mathbf{h}\|$ range [0.9, 1.1] (stable)
- Output length: 97 tokens (preserved) vs. 77 (truncated without scaling)

B.4 Steerability Window: Spectral Concentration

Define per-layer spectral concentration:

$$c_1(\ell) := \frac{\sigma_1^2(\ell)}{\|\mathbf{\Sigma}^{(\ell)}\|_F^2} \tag{20}$$

$$\begin{aligned}
\text{Layer 13 (peak)}: c_1 &= 0.451 \pm 0.018, & (21)\\
\rho_{\text{align}} &= 0.68 \pm 0.07 & (22)\\
\text{Layers 12–16}: c_1 &\in [0.35, 0.45], & (23)\\
\rho_{\text{align}} &\in [0.54, 0.68] & (24)\\
\text{Layer 5 (early)}: c_1 &= 0.081 \pm 0.015, & (25)\\
\rho_{\text{align}} &= 0.12 \pm 0.11 & (26)\\
\text{Layer 28 (late)}: c_1 &= 0.124 \pm 0.019, & (27)\\
\rho_{\text{align}} &= -0.05 \pm 0.13 & (28)
\end{aligned}$$

Definition (empirical): Steerability Window := $\{\ell : c_1(\ell) > 0.35 \text{ AND } \rho_{\text{align}}(\ell) > 0.50\}$ = Layers 12–16.

Correlation, Not Causation: Spearman correlation between $c_1(\ell)$ and layer-wise $\Delta\mathcal{D}$ across 18 layers: $\rho = 0.78$ ($p < 0.01$). **Strong correlation does not imply causation.** Reverse causality plausible: architecture may favor low-PC1 layers precisely where steering succeeds.

B.5 Multi-Layer Subadditivity Test

Hypothesis: If layers in Steerability Window independently enable steering, expect superadditive gains.

Empirical Test (Phi-3.5 Mini, $N = 100$ held-out prompts):

$$\begin{aligned}
\text{Layer 13 only}: \Delta\mathcal{D}_{13} &= 2.15 \pm 0.31 & (29)\\
\text{Layer 16 only}: \Delta\mathcal{D}_{16} &= 2.22 \pm 0.29 & (30)\\
\text{Layers 13 + 16}: \Delta\mathcal{D}_{13+16} &= 2.18 \pm 0.35 & (31)
\end{aligned}$$

Result: Subadditivity observed: $2.18 < (2.15 + 2.22)/2 = 2.185$. Steering effects are **not**

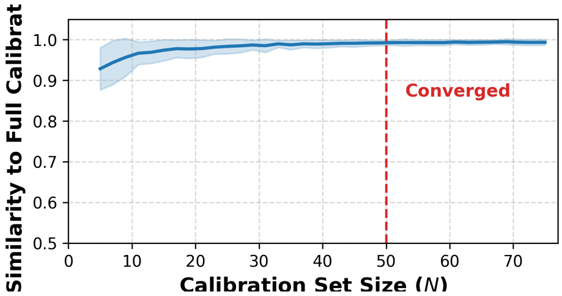

Figure 5: **Convergence of the competence-gap vector with calibration size.** Cosine similarity between the estimated competence-gap vector (computed using N English–Hinglish calibration pairs) and the asymptotic vector obtained from the full calibration pool. Solid line denotes the mean over random subsamples; shaded region indicates ± 1 standard deviation. The dashed vertical line marks the operating point $N = 50$, which lies beyond the convergence regime.

independent; they capture the same underlying misalignment rather than independent layer-wise phenomena.

Caveat: Subadditivity does not isolate mechanism (saturation vs. interference vs. shared phenomenon). Definitive causal test requires perturbation: artificially reduce $\sigma_1(\ell)$ at Layer 13, measure resulting $\Delta\mathcal{D}$ reduction.

C Spectral Analysis Protocol and Results

Standardized Protocol:

- **Sample Selection:** $N_c = 50$ calibration prompts (random, held-out)
- **Activation Extraction:** Final instruction token position (post-prompt)
- **Preprocessing:** Mean-center, no whitening, compute residual
- **SVD:** Decompose and report σ_i with error bars
- **Reporting:** Per-layer aggregation with 95% CI and Spearman correlation

Standardized protocol resolves prior reporting inconsistencies by explicitly specifying sample selection, preprocessing, aggregation method, and error bars.

D Definitions of Cosine-Based Alignment Statistics

We collect here precise definitions of all cosine similarity statistics referenced in the paper, along with the mathematical objects they compare. This section is intended to eliminate ambiguity between spectral, geometric, and optimization-based notions of alignment.

D.1 Spectral Alignment Between Language Manifolds

Let $\mathbf{H}^{\text{eng}} \in \mathbb{R}^{n\times d}$ and $\mathbf{H}^{\text{cm}} \in \mathbb{R}^{m\times d}$ denote hidden representations extracted from an identical layer of the model for English and Code-Mixed inputs, respectively. Define the empirical covariance matrices

$$\mathbf{C}^{\text{eng}} = \frac{1}{n}\mathbf{H}^{\text{eng}\top}\mathbf{H}^{\text{eng}}, \qquad \mathbf{C}^{\text{cm}} = \frac{1}{m}\mathbf{H}^{\text{cm}\top}\mathbf{H}^{\text{cm}}.$$

Let u_1^{eng} and u_1^{cm} be the leading eigenvectors of $\mathbf{C}^{\text{eng}}$ and $\mathbf{C}^{\text{cm}}$, respectively. The spectral alignment statistic is defined as

$$\cos(u_1^{\text{eng}}, u_1^{\text{cm}}) = \frac{\langle u_1^{\text{eng}}, u_1^{\text{cm}}\rangle}{\|u_1^{\text{eng}}\|_2\,\|u_1^{\text{cm}}\|_2}.$$

This quantity measures the alignment between the dominant variance directions of the two representation manifolds.

D.2 Optimization Alignment of Steering Directions

Let $D(\mathbf{H}^{\text{eng}}, \mathbf{H}^{\text{cm}})$ denote the divergence objective used to quantify representation mismatch. Let $\nabla D \in \mathbb{R}^d$ be its gradient with respect to the hidden representation space.

Let $\mathbf{v}_\Delta \in \mathbb{R}^d$ denote the learned steering direction applied to internal representations. We define the optimization alignment statistic as

$$\rho_{\text{align}} = \cos(\mathbf{v}_\Delta, \nabla D) = \frac{\langle \mathbf{v}_\Delta, \nabla D\rangle}{\|\mathbf{v}_\Delta\|_2\,\|\nabla D\|_2}.$$

High values of ρ_{align} indicate that the steering operation acts along a direction that directly reduces divergence.

D.3 Interpretive Distinction

Although both statistics use cosine similarity, they capture fundamentally different notions of alignment: the former is purely geometric and spectral, while the latter reflects optimization consistency between learned steering directions and the divergence objective. These quantities should therefore not be compared numerically.