

CLaS-Bench: A Cross-Lingual Alignment and Steering Benchmark

Daniil Gurgurov^{1,2*} Yusser Al Ghussin^{1,2*} Tanja Bäumel^{1,2,3} Cheng-Ting Chou⁴
Patrick Schramowski^{2,5,6} Marius Mosbach^{7,8} Josef van Genabith^{1,2} Simon Ostermann^{1,2,3}

¹ Saarland University ² German Research Center for Artificial Intelligence (DFKI)
³ Centre for European Research in Trusted AI (CERTAIN) ⁴ University of Illinois Urbana-Champaign ⁵ TU Darmstadt
⁶ hessian.AI ⁷ Mila - Quebec Artificial Intelligence Institute ⁸ McGill University
daniil.gurgurov@dfki.de

Abstract

Understanding and controlling the behavior of large language models (LLMs) is an increasingly important topic in multilingual NLP. Beyond prompting or fine-tuning, *language steering*, i.e., manipulating internal representations during inference, has emerged as a more efficient and interpretable technique for adapting models to a target language. Yet, no dedicated benchmarks or evaluation protocols exist to quantify the effectiveness of steering techniques. We introduce CLaS-Bench, a lightweight parallel-question benchmark for evaluating language-forcing behavior in LLMs across 32 languages, enabling systematic evaluation of multilingual steering methods. We evaluate a broad array of steering techniques, including residual-stream DiffMean interventions, probe-derived directions, language-specific neurons, PCA/LDA vectors, Sparse Autoencoders, and prompting baselines. Steering performance is measured along two axes: language control and semantic relevance, combined into a single harmonic-mean steering score. We find that across languages, the simple residual-based DiffMean method consistently outperforms all other methods. Moreover, a layer-wise analysis reveals that language-specific structure emerges predominantly in later layers and steering directions cluster based on language family. CLaS-Bench is the first standardized benchmark for multilingual steering, enabling both rigorous scientific analysis of language representations and practical evaluation of steering as a low-cost adaptation alternative.

1 Introduction

As our understanding of the internal mechanisms of large language models (LLMs) advances, increasing attention is given to methods that exploit these internal mechanisms to control model behavior. This research direction, often referred

*: Equal contribution.

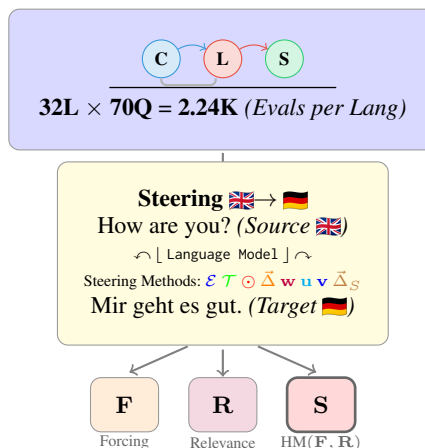


Figure 1: CLaS-Bench pipeline: Multilingual inputs consisting of 70 parallel questions (**Q**) across 32 languages (**L**) are evaluated per target language. Each input is passed to an LLM, which is steered with a selected method. The steered model output is evaluated along two axes: **language forcing F** (whether generation switches to the intended target language) and **output relevance R** (whether response is related to the input). These metrics are combined via a harmonic mean into a single **steering score S**.

to as *actionable interpretability* (Mosbach et al., 2024), has increasingly incorporated techniques collectively called *steering*, i.e., the manipulation of model weights or activations to guide models toward desired outputs (Subramani et al., 2022; Turner et al., 2023; Marks and Tegmark, 2024). Unlike related techniques such as fine-tuning, steering methods are typically applied *at inference time*, positioning them as a more lightweight, direct alternative for adapting models without re-training. Representation-based steering, in particular, intervenes directly on a model’s hidden representations (e.g., residual streams, or latent features) to induce desired behaviors and has been applied successfully to mitigate sycophancy (Rimsky et al., 2024), improve truthfulness (Li et al., 2023), and reduce toxicity (Suau et al., 2024), showing that internal

representations can be used for controllability.

A prominent application of representation steering is control over the language of generation. Recent work probing LLMs has revealed language-specific features and neurons (Tang et al., 2024; Zhao et al., 2024; Kojima et al., 2024), and these insights have been used both to improve downstream performance and control cross-lingual behavior (Gurgurov et al., 2025a; Chou et al., 2025). Steering is particularly promising for multilingual adaptation, enabling targeted language control without requiring costly retraining. However, despite such advances, there is still **no standard evaluation** framework for language steering in language models. Existing benchmarks (Wu et al., 2025; Mueller et al., 2025) focus exclusively on conceptual steering tasks in English, leaving multilingual and cross-lingual settings unexplored.

This gap motivates two central research questions. First, how effective is steering for controlling output language compared to established approaches? Second, does steering perform equally well across languages, given that most LLMs are largely pretrained on English? More broadly, mechanistic interpretability research remains predominantly English-centric: most analyses, circuits, and intervention techniques are developed for and evaluated only in English. Extending these to a broad set of languages is thus essential for scientific understanding and enabling actionable interpretability in truly multilingual settings.

To close this gap, we introduce CLaS-Bench, a lightweight benchmark for evaluating multilingual and cross-lingual language steering. CLaS-Bench covers 32 languages and 70 parallel questions with answers per language (i.e., over **71,680** potential cross-lingual question-answer pairs), enabling controlled, language-by-language evaluation of *steering*, i.e., the informed manipulation of language components, such as neurons or latent directions. Crucially, our evaluation emphasizes *cross-lingual* steering, where the prompt (source) and desired output (target) languages differ—capturing interesting multilingual use cases and being more demanding than a simple case of steering from English.

We use CLaS-Bench to compare a broad set of steering approaches, including neuron-level interventions (Tang et al., 2024), residual-stream difference-in-means vectors (Marks and Tegmark, 2024), probe-derived directions (Li et al., 2023), and vectors from LDA (Balakrishnama and Ganapathiraju, 1998), PCA (Abdi and Williams, 2010),

and Sparse Autoencoders (Bricken et al., 2023), against prompting baselines across several LLMs. We measure steering success along two orthogonal dimensions: (1) **Forcing**, whether the model produces text in the intended language, and (2) **Relevance**, whether the output remains conceptually appropriate; these are combined into an overall steering score via the harmonic mean.

Our experiments reveal three key findings. First, representation-based steering, particularly DiffMean on residual activations, consistently outperforms all other tested methods, including prompting baselines, across most evaluated languages. Second, prompting exhibits failures for specific languages, while DiffMean succeeds across most languages; moreover, we find steering earlier layers is effective with low intervention strengths, whereas later layers require higher strengths. Third, language-specific representations concentrate in later layers (roughly layers 16–32), and typologically related languages cluster geometrically in representation space.

2 Benchmark Design

Languages. CLaS-Bench covers a typologically and geographically diverse subset of 32 languages. We provide details on the languages covered in Appendix C. The selection aims to balance high-resource and low-resource languages while spanning different language families and scripts, which enables evaluation across a broad linguistic spectrum. This diversity ensures that CLaS-Bench evaluates language steering across a wide range of typological phenomena, scripts, and resource levels. The diversity also enables focusing on challenging cross-lingual settings, particularly for low-resource and non-Latin-script languages, where LLMs often struggle (Joshi et al., 2020; Gurgurov et al., 2025b).

Evaluation data. We use a curated subset of 70 diverse open-ended questions from the Vicuna dataset (Chiang et al., 2023), originally introduced by Tang et al. (2024). These questions (see Table 1 for examples) cover a wide range of conversational domains, which we identify and label manually (reasoning, knowledge, personal opinion, creative, and professional writing), and are designed to elicit multi-sentence outputs.

We translate all English questions into the remaining 31 languages using the Google Translate API (Wu et al., 2016), resulting in a parallel dataset of 70 questions per language across 32 languages.

Domain	English Prompt	Translations (Sampling)
Knowledge	“Can you explain the basics of quantum computing?”	Spanish: “¿Puedes explicar los conceptos básicos de la computación cuántica?” Norwegian: “Kan du forklare grunnleggende kvanteberegning?”
Reasoning	“How many words are spoken daily on Earth? Try to explain your answer. Your explanation should take the reader through your reasoning step-by-step.”	German: “Wie viele Wörter werden täglich auf Erden gesprochen? Versuche, deine Antwort zu erklären. Deine Erklärung sollte den Leser Schritt für Schritt durch dein Denken führen.” Japanese: “地球上で毎日いくつの単語が発せられていますか？あなたの答えを説明してみてください。あなたの説明は読み手をあなたの思考回路に一步一步案内していくようなものであるべきです。”
Creative	“How would you introduce yourself as a medieval knight at a royal banquet?”	French: “Comment vous présenteriez-vous en tant que chevalier médiéval dans un banquet royal?” Korean: “중세 기사로서 왕실 연회에서 자신을 어떻게 소개하시겠습니까?”
Opinion	“Do we have a moral obligation to explore space, or should we focus on solving Earth’s problems first?”	Dutch: “Har vi en moralsk forpligtelse til at udforske rummet, eller skal vi først fokusere på at løse Jordens problemer?” Polish: “Czy mamy moralny obowiązek odkrywania przestrzeni kosmicznej, czy powinniśmy najpierw skupić się na rozwiązywaniu problemów Ziemi?”
Writing	“Can you help me write a formal email to a potential business partner proposing a joint venture?”	Chinese: “你能帮我写一封正式邮件，向潜在的商业伙伴提议合资吗？” Russian: “Можете ли вы помочь мне написать формальное письмо потенциальному бизнес-партнёру с предложением совместного предприятия?”

Table 1: Examples of questions included in the CLaS-Bench benchmark spanning all domains.

All translations are proofread and corrected by native speakers to ensure fluency, idiomaticity, and semantic fidelity to the English source prompts (see Appendix A for details on quality assurance and the proofreading protocol). For evaluation, each question in each source language can be paired with an answer in any target language, yielding $70 \times 32 = 2,240$ instances per target language.

Task definition. Let \mathcal{L} be the set of all languages CLaS-Bench covers. Given a question x_s in source language $s \in \mathcal{L}$, the task is to generate an answer y_t in target language $t \in \mathcal{L}$. $M_\theta(x)$ is a language model with fixed parameters θ and $h_\ell(x) \in \mathbb{R}^{d_\ell}$ is the hidden representation at layer ℓ computed from input x , where d_ℓ is the dimensionality. We use $h_\ell[i]$ to index the i -th element (neuron) of h_ℓ .

A steering method $S(\cdot)$ modifies the generation process either indirectly by changing the input or directly by intervening on hidden representations:

$$\hat{y}_t = M_\theta(S(x_s)).$$

Here, S either transforms the input: $x_s \rightarrow x'_s$, or intervenes on the hidden representations at layer ℓ , i.e., replaces the hidden representation by the intervention δ_ℓ . The parameters θ remain fixed. The goal is to ensure \hat{y}_t is in the target language t while preserving the semantic content of x_s . The overall pipeline is illustrated in Figure 1.

Evaluation metrics. We assess steering effectiveness along two complementary dimensions:

- **Language Forcing Success (LFS).** This measure indicates the overall success of a method to force a specific language. We apply the Fast-Text LID classifier (Joulin et al., 2016) to detect the language of generated outputs, which provides good coverage for the languages in our benchmark. We report both overall success rate and per-language breakdown:

$$LFS = \frac{\# \text{ outputs in target language}}{\text{total \# outputs}} \in [0, 1].$$

- **Output Relevance (OR).** This score measures the semantic fidelity of the answer to the question. We compute this using an LLM-as-a-judge evaluation with Qwen-3-8B (Team, 2025), which demonstrates strong multilingual performance. Each output is scored 0 (unrelated or gibberish), 1 (partially relevant or incomplete), or 2 (clearly relevant and coherent), and we report the normalized average relevance:

$$OR = \frac{1}{N} \sum_{i=1}^N \frac{\text{score}_i}{2} \in [0, 1],$$

where N is the number of evaluated outputs. The judging protocol employed is similar to the one from Wu et al. (2025) and is presented in Appendix B.

We combine these in the **Language Steering Score** which computes the harmonic mean of LFS

and OR:

$$\text{LSS} = \frac{2 \cdot \text{LFS} \cdot \text{OR}}{\text{LFS} + \text{OR}},$$

which penalizes cases where one of the two metrics is very low relative to the other.

3 Experimental Setup

Models. We evaluate CLaS-Bench on two LLMs: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), a widely used mid-sized foundation model, and Aya-Expansive-8B (Dang et al., 2024), a prominent multilingual alternative.

Steering methods. We benchmark a multitude of steering methods, spanning both prompting-based and representation-based interventions. The data for designing representation-based interventions is sourced from CulturaX (Nguyen et al., 2024). Below, α denotes the *steering strength* in all methods.

(I) \mathcal{E} Prompting with Language Specification (Baseline-I). Adding explicit instructions to respond in the target language with the instructions in English, e.g., *Question + "Respond in German"* for steering towards German.

(II) \mathcal{T} Prompting with Language Specification (Baseline-II). Adding explicit instructions to respond in the target language with instructions in the target language, e.g., *Question + "Antworte auf Deutsch"* for steering towards German.

(III) \odot Neuron-Based Steering (LAPE). Identifying language-sensitive neurons (Tang et al., 2024) by analyzing activation patterns across 10M tokens per language. We compute activation probabilities $p_{\ell,h}^{\text{lang}}$ for each neuron $h_{\ell}[i]$ in layer ℓ across languages, then apply entropy filtering to select language sensitive-neurons $\mathcal{N}_{\text{selected}}$ with low cross-lingual entropy (high language selectivity). The intervention is defined as

$$\delta_{\ell} = \sum_{i \in \mathcal{N}_{\text{selected}}} \delta_{\ell,i} \cdot \mathbf{e}_i,$$

where \mathbf{e}_i is the standard basis vector. Let $\bar{a}_{\ell,h}^{\text{lang}}$ be the average activation of neuron $h_{\ell}[i]$ in layer ℓ for the target language. Selected neurons are manipulated via two intervention mechanisms:

1. *additive*: $\delta_{\ell,i} = \alpha \cdot \bar{a}_{\ell,i}^{\text{target}} + h_{\ell}[i]$
2. *replacement*: $\delta_{\ell,i} = \alpha \cdot \bar{a}_{\ell,i}^{\text{target}}$

Non-target language neurons are optionally deactivated by zeroing them out.

(IV) $\vec{\Delta}$ DiffMean Steering Vectors on Residual Activations. Computing language-specific average activations across the residual stream (Marks and Tegmark, 2024) for 10M tokens per language. We define the hidden intervention as

$$\delta_{\ell} = h_{\ell} + \alpha \cdot \frac{\vec{\Delta}_{\ell}}{\|\vec{\Delta}_{\ell}\|_2},^1$$

where $\vec{\Delta}_{\ell} = \bar{h}_{\ell}^{\text{target}} - \bar{h}_{\ell}^{\text{source}}$ and $\bar{h}_{\ell}^{\text{lang}}$ is the average activation at layer ℓ for language *lang*.

(V) \mathbf{w} Probe-based Steering Vectors on Residual Streams. Training linear probes (Belinkov, 2022; Li et al., 2023) to classify target language representations against negative languages. For each layer ℓ , we train a binary classifier $\text{Probe}_{\ell} : \mathbb{R}^d \rightarrow [0, 1]$ on balanced datasets of target language activations (positive class) and negative language activations (negative class), each consisting of 100K samples, optimizing binary cross-entropy loss. The probe weight vector $\mathbf{w}_{\ell} \in \mathbb{R}^d$ encodes the direction in the residual stream that discriminates the target language. We then define the intervention as

$$\delta_{\ell} = h_{\ell} + \alpha \cdot \frac{\mathbf{w}_{\ell}}{\|\mathbf{w}_{\ell}\|_2}.$$

(VI) \mathbf{u} PCA-based Steering Vectors on Residual Streams. Computing language-specific subspaces through Principal Component Analysis (PCA) (Abdi and Williams, 2010) on residual stream activations. We collect activations from each target language across 500K tokens and center them by subtracting the mean. For each layer ℓ , we apply PCA to obtain the top $k = 20$ principal components $U_{\ell} \in \mathbb{R}^{k \times d}$, which span the subspace of maximum variance for that language. During inference, given a hidden state $h_{\ell} \in \mathbb{R}^d$, we project onto the language subspace via $\text{proj}_{\ell} = U_{\ell} h_{\ell}^T \in \mathbb{R}^k$, then reconstruct in the original space: $\mathbf{u}_{\ell} = U_{\ell}^T \text{proj}_{\ell} \in \mathbb{R}^d$. This reconstructed component \mathbf{v}_{ℓ} captures the language-relevant information encoded in the principal directions. The intervention is then defined as

$$\delta_{\ell} = h_{\ell} + \alpha \cdot \frac{\mathbf{u}_{\ell}}{\|\mathbf{u}_{\ell}\|_2}.$$

(VII) \mathbf{v} LDA-based Steering Vectors on Residual Streams. Computing language-discriminative steering vectors through Linear Discriminant Analysis (LDA) (Balakrishnama and

¹All steering vectors are normalized to decouple steering magnitude from component strength.

Lang.	\mathcal{E} Base.-I	\mathcal{T} Base.-II	$\vec{\Delta}$ DiffM.	w Probe	u PCA	$\vec{\Delta}$ SAE-DM.	v LDA	\odot LAPE
ar	62.9	54.0	88.8	14.5	16.3	49.5	34.9	77.2
bo	33.5	38.2	8.5	5.6	6.7	6.1	4.8	7.6
cs	69.1	75.2	92.3	28.6	16.6	39.5	27.3	89.1
da	80.6	47.7	90.1	35.8	17.4	32.7	8.8	53.7
de	70.6	56.6	95.4	83.4	14.7	51.2	0.0	96.5
el	82.3	76.4	88.4	9.0	15.5	39.0	6.1	87.1
en	7.1	7.1	95.9	95.7	11.0	51.2	51.3	98.8
es	74.7	65.3	94.7	78.7	20.6	57.7	47.6	94.0
fa	38.2	53.3	79.0	13.2	13.9	47.4	10.3	85.1
fr	66.8	66.6	96.4	89.7	19.3	64.6	23.3	96.5
hi	52.5	68.0	90.5	13.4	11.8	53.8	0.0	93.2
id	65.7	76.8	94.7	59.0	18.3	17.9	0.0	94.7
it	66.2	71.1	94.7	77.5	23.4	53.4	15.8	95.9
ja	48.0	72.6	84.6	35.5	14.3	10.0	16.9	83.5
ka	80.3	67.9	32.9	6.1	11.8	6.2	3.8	27.4
kk	63.7	56.6	42.1	6.3	6.1	8.8	6.2	28.7
ko	42.7	66.4	89.8	39.8	16.9	35.6	20.9	79.8
mt	68.7	64.5	15.6	5.8	7.3	6.2	5.7	20.8
nl	67.7	66.8	95.3	45.9	13.8	43.5	7.7	95.0
no	63.4	55.8	88.8	29.3	18.0	38.7	4.4	69.5
pl	53.9	53.9	92.5	33.9	10.7	41.4	46.5	89.9
pt	61.1	78.9	95.8	84.6	18.9	48.3	55.9	95.6
ro	70.2	83.7	93.0	57.9	13.8	35.0	6.4	77.3
ru	57.4	69.2	97.0	82.5	16.7	67.6	56.8	95.5
sk	83.5	52.9	77.7	31.8	7.8	17.9	10.9	65.9
sv	63.7	63.6	93.7	47.6	17.0	39.3	60.2	89.8
sw	83.8	79.6	49.8	8.4	2.1	6.2	6.2	52.7
th	73.4	86.0	91.2	24.2	12.4	48.0	9.8	84.5
tr	77.5	73.0	84.5	23.3	14.7	19.8	7.2	84.1
uk	67.3	80.7	92.9	13.5	15.4	69.7	16.1	91.6
vi	79.9	88.4	96.9	56.1	14.7	11.3	11.1	95.0
zh	60.4	76.6	97.6	92.3	12.2	85.4	87.6	96.2
Avg.	67.7	67.3	84.5	48.6	15.1	42.3	23.6	80.1

Table 2: Language steering scores across all methods for 32 languages for Llama-3.1-8B-Instruct. Individual scores for language forcing and output relevance metrics are in Appendix K. Gray columns correspond to baselines. Yellow and orange to the best performing methods.

Ganapathiraju, 1998) on residual stream activations. We collect activations from the target language and multiple negative languages across 100K tokens each. For each layer ℓ , we formulate a binary classification problem where the positive class consists of target language activations and the negative class consists of equal balanced samples from all other languages. LDA finds the optimal linear direction that maximizes class separability by computing: $\mathbf{v}_\ell = \Sigma_w^{-1}(\boldsymbol{\mu}_{\text{tgt}} - \boldsymbol{\mu}_{\text{other}})$, where Σ_w is the within-class covariance matrix and $\boldsymbol{\mu}_{\text{tgt}}, \boldsymbol{\mu}_{\text{other}}$ are the class means. The intervention is then defined as

$$\delta_\ell = h_\ell + \alpha \cdot \frac{\mathbf{v}_\ell}{\|\mathbf{v}_\ell\|_2}.$$

(VIII) $\vec{\Delta}_S$ **DiffMean on Sparse Autoencoder Layer.** Computing language-specific average activations in the sparse autoencoder (SAE) latent space for 10M tokens per language. We utilize pre-trained SAEs from Li et al. (2025) for Llama-3.1-8B-Instruct. For each SAE layer $\ell \in \mathcal{L}$ (where

\mathcal{L} indexes the subset of model layers with trained SAEs), we encode residual stream activations h_ℓ into sparse representations via a JumpReLU encoder: $f_\ell = \text{JumpReLU}(W_{\text{enc}}h_\ell + b_{\text{enc}}) \in \mathbb{R}^{d_{\text{SAE}}}$, where $\text{JumpReLU}(z) = z \cdot \mathbf{1}[z > \theta]$ with learned threshold θ . The steering vector is computed as the difference between target and source language means in this sparse space: $\vec{\Delta}_\ell = \bar{f}_\ell^{\text{target}} - \bar{f}_\ell^{\text{source}}$. During inference, we hook into the input to intercept the residual stream from layer ℓ . The combined hidden state is encoded, steered in SAE latent space with ℓ_2 -normalized strength α , then decoded back:

$$\delta_\ell = W_{\text{dec}} \left(f_\ell + \alpha \cdot \frac{\vec{\Delta}_\ell}{\|\vec{\Delta}_\ell\|_2} \right) + \epsilon.$$

Here, ϵ is a reconstruction error correction term that preserves information not captured by the SAE.²

²Bias terms are left out for the sake of readability.

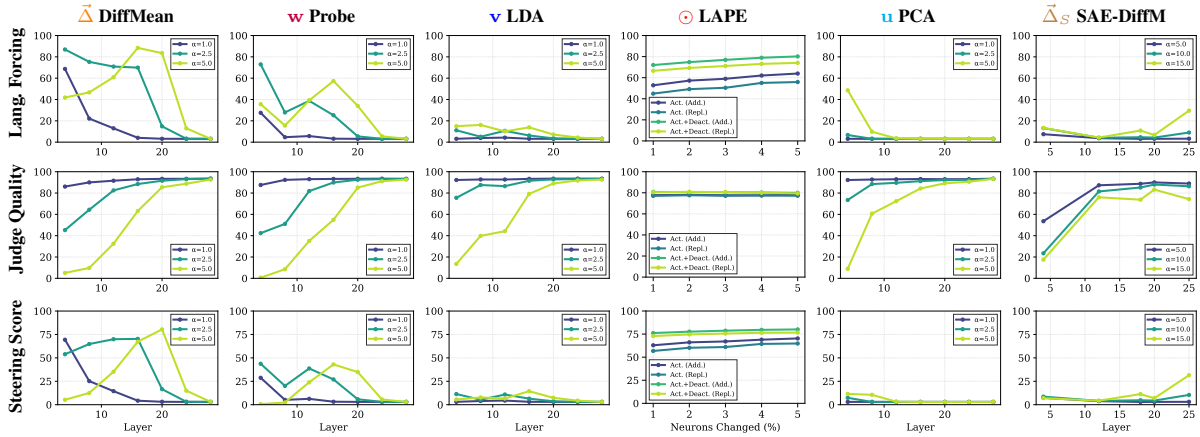


Figure 2: Analysis of steering methods across evaluation metrics for Llama-3.1-8B-Instruct. Columns show different methods. Rows represent: forcing success rate, judge relevance, and overall steering score. Line colors indicate different intervention configurations, including steering strengths and settings.

4 Steering Results

We present the results of our cross-lingual steering evaluation in Table 2, reporting the steering score across 32 typologically diverse languages for Llama-3.1-8B-Instruct (results and discussion for Aya-Expanse-8B are in Appendix K). The results for each method are reported for the best steering configurations, as identified in Section 5 and specified in Appendix J.

Residual-based steering outperforms prompting. DiffMean $\bar{\Delta}$ steering on residual activations achieves the highest average steering score (84.5%), substantially outperforming both prompting baselines \mathcal{E} (67.7%) and \mathcal{T} (67.3%). DiffMean maintains scores exceeding 90% for 19 out of 32 languages, indicating robust cross-lingual generalization. LAPE \odot achieves the second-highest average (80.1%), demonstrating that language-specific neurons can effectively control output language, though with greater variability across languages. In contrast, the prompting baselines reveal interesting inconsistencies: \mathcal{E} , for example, fails for English, Tibetan, and Farsi, while \mathcal{T} mostly struggles with English as a target. These failures highlight a potential limitation—models ignore or misinterpret explicit language directives, particularly for lower-resource languages. Table 3 illustrates this concretely: given a Russian prompt with an explicit “Respond in English” instruction, the prompt-based method produces Russian output, while DiffMean successfully generates English.

SAE-based steering lags behind. DiffMean steering in sparse autoencoder latent space ($\bar{\Delta}$ SAE-

Language Instruction \mathcal{E}	DiffMean Steering $\bar{\Delta}$
Input: Как я могу развить свои навыки критического мышления?	
Развитие навыков критического мышления требует времени, усилий и практики. Вот несколько советов...	Developing your critical thinking skills is a valuable investment in your personal and professional growth. Here are some tips...
Russian	English

Table 3: Prompt-based instruction (left) outputs Russian despite explicit “Respond in English” directive. Activation-based steering (right) produces English without prompt modification. This holds for many source and target languages as indicated in Table 2. More examples are provided in Appendix M.

DM.) achieves moderate performance (42.3% average), outperforming several residual-stream methods but falling short of residual DiffMean steering. The method performs well for high-resource languages like German, Spanish, and Hindi, but struggles with Japanese and Slovak. This gap may stem from reconstruction error inherent in SAE encoding or from SAEs’ training data not covering all languages extensively.

Supervised methods underperform unsupervised approaches. Surprisingly, supervised methods such as probe-based \mathbf{w} steering (48.6%) and LDA \mathbf{v} (23.6%) underperform unsupervised DiffMean and LAPE approaches. Probe-based steering exhibits extreme variance, achieving 95.7% for English but only 9.0% for Greek and 14.5% for Arabic. PCA-based \mathbf{u} methods perform the worst, averaging only 15.1%. These results suggest that supervised objectives may overfit to language-specific characteristics in training

data, while the unsupervised difference-of-means captures more generalizable language directions.

5 Ablation Analysis

We conduct systematic ablations to understand how steering effectiveness varies with intervention layer, steering strength, and method-specific parameters. Figure 2 presents results for Llama-3.1-8B-Instruct across three evaluation dimensions: language forcing success, output relevance, and overall steering score. The results for Aya-Expanse-8B are in Appendix I.

Layer and steering strength interact. A key finding is that optimal steering strength depends on intervention depth. For DiffMean, low strength ($\alpha = 1.0$) suffices at early layers, achieving over 80% steering score, while later layers require progressively higher strengths to be effective: $\alpha = 5.0$ peaks around layer 20. Crucially, output quality remains stable across late-layer interventions even with high steering strengths, whereas early-layer interventions with strong α severely degrade coherence. This suggests that later layers encode language information in a more modular fashion, allowing targeted manipulation without disrupting other generation capabilities.

Probe-based steering shows a similar but more pronounced pattern, with $\alpha = \{1.0, 2.5\}$ effective only at very early layers and higher strengths required beyond layer 10. LDA exhibits weak steering regardless of layer or strength, once exceeding 15% success. For LAPE, the combined activation-plus-deactivation outperforms activation-only intervention strategy, and performance increases slightly when intervening on more neurons (from 1% to 5%). PCA shows modest steering with higher strengths ($\alpha = 2.5-5.0$) in early layers but remains ineffective in other layers, likely due to those layers capturing less variance with the selected principal components. SAE-based steering, operating at higher strengths ($\alpha \in \{5.0, 10.0, 15.0\}$), shows a distinctive pattern: steering performance is best at layer 25 with alpha $\alpha = 15.0$, indicating better language control in the sparse activation space of higher layers.

6 Interpretability Insights

Beyond evaluating steering performance, CLaS-Bench motivates investigation into how multilingual representations are organized within LLMs. We analyze the structural properties of

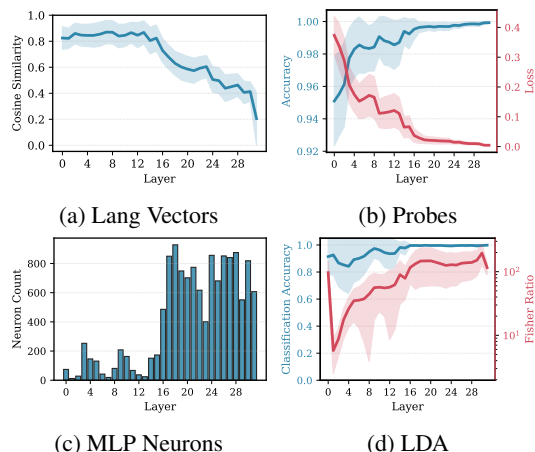


Figure 3: Insights into language-specific components across interpretation tools for Llama-3.1-8B-Instruct. (a) reveals average cosine similarity patterns across all language vectors. (b) demonstrates probe learning dynamics through loss and accuracy trajectories. (c) identifies the distribution of language-specific neurons across layers. (d) provides LDA classification accuracy and Fisher Ratio (the degree of separability between classes).

language-specific components discovered through various methods, revealing consistent patterns.

Language-specific information concentrates in later layers. Converging evidence from multiple analysis methods indicates that language-specific representations emerge predominantly in layers 16–32. Figure 3a shows that cosine similarity between residual-based language vectors across language pairs decreases monotonically through the network, reaching minimum values (maximum separability) in layers 22–32. Linear probes (Figure 3b) achieve >99% language classification accuracy from layer 14 onward, with the probe loss also reaching its minimum in the deeper layers. LAPE-identified language-specific neurons (Figure 3c) cluster predominantly in layers 24–28, with counts increasing sharply from layer 16. Finally, LDA classification accuracy and Fisher ratio (Fisher, 1936) (Figure 3d) both peak in late layers. This convergence suggests a hierarchical processing view: early layers encode language-agnostic features, while later layers encode language-specific generation patterns.

Language families exhibit geometric clustering. Analysis of steering vector similarities (Appendix Figures 8–13) reveals that typologically related languages cluster in representation space. Romance languages (Spanish, French, Portuguese, Italian, Romanian) exhibit high mutual cosine similarity,

as do Germanic (German, Dutch, Swedish, Danish, Norwegian) and Slavic languages (Russian, Polish, Ukrainian, Czech, Slovak). Figure 6 further shows that LAPE-identified neurons are largely language-specific, with substantial overlap only within language families. This geometric structure has practical implications: steering between related languages might require smaller interventions, while cross-family steering (e.g., Japanese to Arabic) could demand larger modifications and be more susceptible to quality degradation.

Implications for multilingual interpretability. Our findings support the view that language control operates through geometrically structured, linearly accessible representations. The success of DiffMean over statistically sophisticated methods (LDA, PCA) suggests that language directions are well-approximated by simple difference vectors. This is encouraging for interpretability: language control does not require modeling complex nonlinear interactions, but rather identifying appropriate linear subspaces in relevant layers. CLaS-Bench thus provides a foundation for systematic investigation of multilingual representations, enabling researchers to probe not only *whether* steering works, but *why* and *where* it succeeds or fails.

7 Related Work

7.1 Representation-based Steering

A broad line of work explores steering language models by directly manipulating their hidden representations, rather than relying solely on prompts or fine-tuning. Typical approaches include adding fixed vectors to activations, selectively activating neurons, or constraining intermediate states.

Several works have explored representation-based paradigm across different tasks. For instance, Rinsky et al. (2024) and Turner et al. (2023) demonstrate that simple vector-based interventions can steer models toward more truthful or less sycophantic behaviors. Marks and Tegmark (2024) formalize geometric methods such as DiffMean, enabling systematic manipulation of residual-stream activations. Sparse latent-space methods (Huben et al., 2023) use interpretable autoencoder directions for controllability. Liu et al. (2023) combine representation interventions with in-context learning to guide semantic properties. Collectively, these studies show that internal representations are a viable interface for direct behavioral control.

7.2 Language-specific Dimensions

Recent work has specifically focused on the identification and steerability of language-specific components in LLMs, with two main lines of research: neuron-based and SAE-based methods.

Neuron-based methods. Neuron-based approaches focus on detecting neurons sensitive to particular languages and testing their causal role through interventions. For example, Zhao et al. (2024) identify language-sensitive neurons in both attention and MLP layers via ablation and hidden state perturbation analysis, showing that setting these neurons to zero suppresses the corresponding language. Kojima et al. (2024) instead train binary classifiers on neuron activations to rank neurons by their discriminative power and propose replacement-based manipulation to steer model outputs. Similarly, Tang et al. (2024) introduce the Language Activation Probability Entropy (LAPE) method, demonstrating language steering by deactivating source neurons (zeroing) and activating target neurons (setting to an average).

SAE-based methods. SAEs have emerged as a powerful tool for uncovering interpretable, language-specific features and manipulating model behavior through interventions in sparse latent spaces. Deng et al. (2025) introduce a metric for monolinguality of SAE features and demonstrate that ablating language-specific features selectively impairs performance in one language and that these features can enhance the construction of steering vectors to control language generation. Chou et al. (2025) propose leveraging SAE features to achieve causal language control in LLMs; by modulating the activation of a single SAE feature in mid-to-late transformer layers, they steer generation toward target languages with up to 90% success, while preserving semantic fidelity.

7.3 Benchmarking Steering

Several benchmarks have been proposed to evaluate steering and interpretable control in LLMs. AxBench (Wu et al., 2025) provides a large-scale benchmark for steering and concept detection in English, comparing prompting, fine-tuning, and representation-based methods (e.g., SAEs, DiffMean, ReFT-r1), and finds that prompting generally outperforms existing approaches. MIB (Mueller et al., 2025) assesses mechanistic interpretability through circuit and causal variable lo-

calization tasks, showing that attribution and supervised distributed alignment search (DAS) (Geiger et al., 2024) methods outperform SAE-based features for recovering causal components. Steer-Bench (Chen et al., 2025) evaluates population-specific steerability with in-context learning and fine-tuning by testing whether LLMs can adapt outputs to align with the norms, perspectives, and communication styles of 30 contrasting subreddit pairs.

While these benchmarks advance evaluation of English-language steering and interpretability, **neither assesses steering in multilingual or cross-lingual settings**. This leaves unanswered how well steering generalizes across languages, how methods perform on low-resource or typologically distant languages, and how multilingual representations can be systematically probed. CLaS-Bench fills this gap in the literature.

8 Conclusion

We introduce CLaS-Bench, the first benchmark for standardized evaluation of cross-lingual language steering in LLMs.³ Covering 32 diverse languages with 70 parallel high-quality open-ended questions each, CLaS-Bench establishes a structured framework for measuring the effectiveness of steering methods in controlling output language. Unlike prior work that primarily focuses on English and conceptual attributes, our benchmark positions multilingualism at the center, highlighting both the strengths and limitations of existing approaches.

Our evaluation setup enables cross-lingual experiments, revealing whether steering works consistently across languages and how it compares to prompting. Designed to be lightweight and easily extendable, the benchmark allows new languages to be incorporated simply by translating the questions and applying the same evaluation protocol. While our current focus is on 32 languages, CLaS-Bench can naturally grow into a broader multilingual resource. By providing a common ground for comparing steering methods, we aim to accelerate research at the intersection of interpretability and multilingual NLP, ultimately advancing our understanding of how LLMs represent language and supporting the development of user-adaptive multilingual systems that operate reliably across diverse linguistic contexts.

³Code and data are available at <https://github.com/d-gurgurov/CLaS-Bench> and <https://huggingface.co/datasets/DGurgurov/CLaS-Bench>, respectively.

Limitations

Our work has several limitations. First, due to computational constraints, we use varying amounts of data across methods: DiffMean and LAPE process 10M tokens per language, while PCA and LDA use 500K and 100K tokens respectively, as these methods require computing covariance matrices that scale quadratically with sample size. We follow established practices for each method (Marks and Tegmark, 2024; Tang et al., 2024), but this variation may affect comparability. Second, SAE-based steering is limited to layers with publicly available pretrained SAEs (layers 4, 12, 18, 20, 25 for Llama-3.1-8B-Instruct), preventing exhaustive layer-wise analysis. Additionally, we were unable to evaluate SAE-based steering for Aya-Expansive-8B due to the absence of publicly available pretrained SAEs for this model. Third, while CLaS-Bench covers 32 typologically diverse languages, many of the world’s languages remain unrepresented, particularly those with limited digital resources. Finally, we evaluate only instruction-tuned models; base models may exhibit different steering dynamics.

Acknowledgments

This research was supported by *lorAI - Low Resource Artificial Intelligence*, a project funded by the European Union under [GA No.101136646](#), and by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005). We also thank Masha Fedzechkina for her valuable feedback on an early draft of the paper.

References

- Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.
- Suresh Balakrishnama and Aravind Ganapathiraju. 1998. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8.
- Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac

- Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Kai Chen, Zihao He, Taiwei Shi, and Kristina Lerman. 2025. Steer-bench: A benchmark for evaluating the steerability of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18338–18366.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality*.
- Cheng-Ting Chou, George Liu, Jessica Sun, Cole Blondin, Kevin Zhu, Vasu Sharma, and Sean O’Brien. 2025. Causal language control in multilingual transformers via sparse feature steering. In *ACL 2025 Student Research Workshop*.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. *Aya expande: Combining research breakthroughs for a new multilingual frontier*. Preprint, arXiv:2412.04261.
- Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.
- Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.
- Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Daniil Gurgurov, Katharina Trnley, Yusser Al Ghussin, Tanja Bäümel, Josef van Genabith, and Simon Ostermann. 2025a. Language arithmetics: Towards systematic language neuron identification and manipulation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2911–2937.
- Daniil Gurgurov, Ivan Vykopal, Josef van Genabith, and Simon Ostermann. 2025b. Small models, big impact: efficient corpus and graph-based adaptation of small multilingual language models for low-resource languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 355–395.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.0. <https://glottolog.org>. Max Planck Institute for Evolutionary Anthropology.
- Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. *The state and fate of linguistic diversity and inclusion in the NLP world*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
- Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. *On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.
- Jiaming Li, Haoran Ye, Yukun Chen, Xinyue Li, Lei Zhang, Hamid Alinejad-Rokny, Jimmy Chih-Hsien Peng, and Min Yang. 2025. *Training superior sparse autoencoders for instruct models*. Preprint, arXiv:2506.07691.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.
- Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2023. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.
- Samuel Marks and Max Tegmark. 2024. *The geometry of truth: Emergent linear structure in large language model representations of true/false datasets*. In *First Conference on Language Modeling*.

- Marius Mosbach, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva. 2024. [From insights to actions: The impact of interpretability and analysis research on NLP](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3105, Miami, Florida, USA. Association for Computational Linguistics.
- Aaron Mueller, Atticus Geiger, Sarah Wiegrefe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, and 4 others. 2025. [MIB: A mechanistic interpretability benchmark](#). In *Forty-second International Conference on Machine Learning*.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. [CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. [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, Bangkok, Thailand. Association for Computational Linguistics.
- Xavier Suau, Pieter Delobelle, Katherine Metcalf, Armand Joulin, Nicholas Apostoloff, Luca Zappella, and Pau Rodríguez. 2024. Whispering experts: Neural interventions for toxicity mitigation in language models. *arXiv preprint arXiv:2407.12824*.
- Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. [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, Bangkok, Thailand. Association for Computational Linguistics.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and 12 others. 2016. [Google’s neural machine translation system: Bridging the gap between human and machine translation](#). *Preprint*, arXiv:1609.08144.
- Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. [Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders](#). In *Forty-second International Conference on Machine Learning*.
- Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319.

Appendix

A Data Curation and Native Speaker Validation

A.1 Translation and Proofreading Protocol

The initial translation of the 70 English prompts into 34 additional languages was performed using the Google Translate API to ensure consistency and comprehensive coverage. To guarantee semantic fidelity, fluency, and idiomaticity, all translations underwent a systematic proofreading process conducted by native speakers of the target languages.

We recruited volunteer native speakers from our institution’s campus community, representing all 34 target languages. Participants were provided with access to a dedicated web interface displaying the English source prompts alongside their machine-translated versions. The interface allowed annotators to review, correct, and refine translations while maintaining semantic equivalence with the original English questions. Proofreaders were instructed to prioritize:

- **Semantic fidelity:** Ensuring the translated prompts retained the intended meaning and conversational intent of the English source
- **Fluency and idiomaticity:** Correcting grammatical errors and replacing awkward phrasing with natural, idiomatic expressions appropriate for native speakers
- **Domain consistency:** Maintaining the conversational tone and style across all linguistic domains (reasoning, knowledge, personal opinions, creative, and professional writing)

Each native speaker volunteer spent less than one hour completing the proofreading task for their respective language(s). No compensation was offered, and participation was entirely voluntary.

A.2 Annotator Background

All proofreaders were native speakers of their respective target languages with fluency in English, enabling them to accurately assess translation quality. The majority of participants had backgrounds in linguistics, computer science, or language technology. Annotator information was collected and stored anonymously to protect participant privacy.

A.3 Ethical Considerations

Prior to participation, all volunteers were informed about the purpose of the data curation task and provided explicit consent for the corrected translations to be used in subsequent research and made available for public release (with appropriate anonymization of annotator identities). No ethics review board approval was sought, as the proofreading task did not fall under institutional requirements for formal ethical review. The study involved minimal risk to participants, consisted of standard proofreading activities, and did not require collection of sensitive personal information beyond basic language background.

B Judge Prompt

Judge Prompt. You are an impartial judge evaluating multilingual responses from a language steering experiment. A question was asked in *{source_language}*, and the model was steered to respond in *{target_language}*.

Important: Do NOT evaluate which language the response is in or whether steering succeeded. Ignore the target language entirely for scoring—evaluate the final response solely on its content, regardless of the language used.

Evaluation Criteria:

- (1) **Relevance:** Is the response a relevant answer to the question?
- (2) **Coherence:** Is the response coherent text, or is it gibberish, repetitive, or useless?

Scoring Rubric:

- **0:** Completely unrelated or gibberish/useless text.
- **1:** Somewhat related but incomplete, partially off-target, or minimally useful.
- **2:** Clearly related to the question, coherent, and useful text.

Output format: Rating: [[score]]

C Selected Languages

Language	ISO	Glottolog Family	Script	Resource Level
Tibetan	bo	Sino-Tibetan (Bodic)	Tibetan	1
Maltese	mt	Afro-Asiatic (Semitic)	Latin	2
Italian	it	Indo-European (Romance)	Latin	4
Spanish	es	Indo-European (Romance)	Latin	5
German	de	Indo-European (Germanic)	Latin	5
Japanese	ja	Japonic	Japanese	5
Arabic	ar	Afro-Asiatic (Semitic)	Arabic	5
Chinese	zh	Sino-Tibetan (Chinese)	Han	5
Dutch	nl	Indo-European (Germanic)	Latin	4
French	fr	Indo-European (Romance)	Latin	5
Portuguese	pt	Indo-European (Romance)	Latin	4
Russian	ru	Indo-European (Slavic)	Cyrillic	4
Korean	ko	Koreanic	Hangul	4
Hindi	hi	Indo-European (Indo-Aryan)	Devanagari	4
Turkish	tr	Turkic	Latin	4
Polish	pl	Indo-European (Slavic)	Latin	4
Swedish	sv	Indo-European (Germanic)	Latin	4
Danish	da	Indo-European (Germanic)	Latin	3
Norwegian	no	Indo-European (Germanic)	Latin	1
English	en	Indo-European (Germanic)	Latin	5
Slovak	sk	Indo-European (Slavic)	Latin	3
Greek	el	Indo-European (Hellenic)	Greek	3
Swahili	sw	Atlantic-Congo (Bantu)	Latin	2
Kazakh	kk	Turkic	Cyrillic	3
Georgian	ka	Kartvelian	Georgian	2
Ukrainian	uk	Indo-European (Slavic)	Cyrillic	3
Persian	fa	Indo-European (Iranian)	Arabic-Persian	4
Thai	th	Kra-Dai	Thai	3
Indonesian	id	Austronesian	Latin	3
Vietnamese	vi	Austroasiatic	Latin	4
Czech	cs	Indo-European (Slavic)	Latin	4
Romanian	ro	Indo-European (Romance)	Latin	3

Table 4: Languages included in CLaS-Bench, with ISO codes, Glottolog family assignments (Hammarström et al., 2024), writing systems, and resource levels (Joshi et al., 2020).

D Language Neurons from MLPs

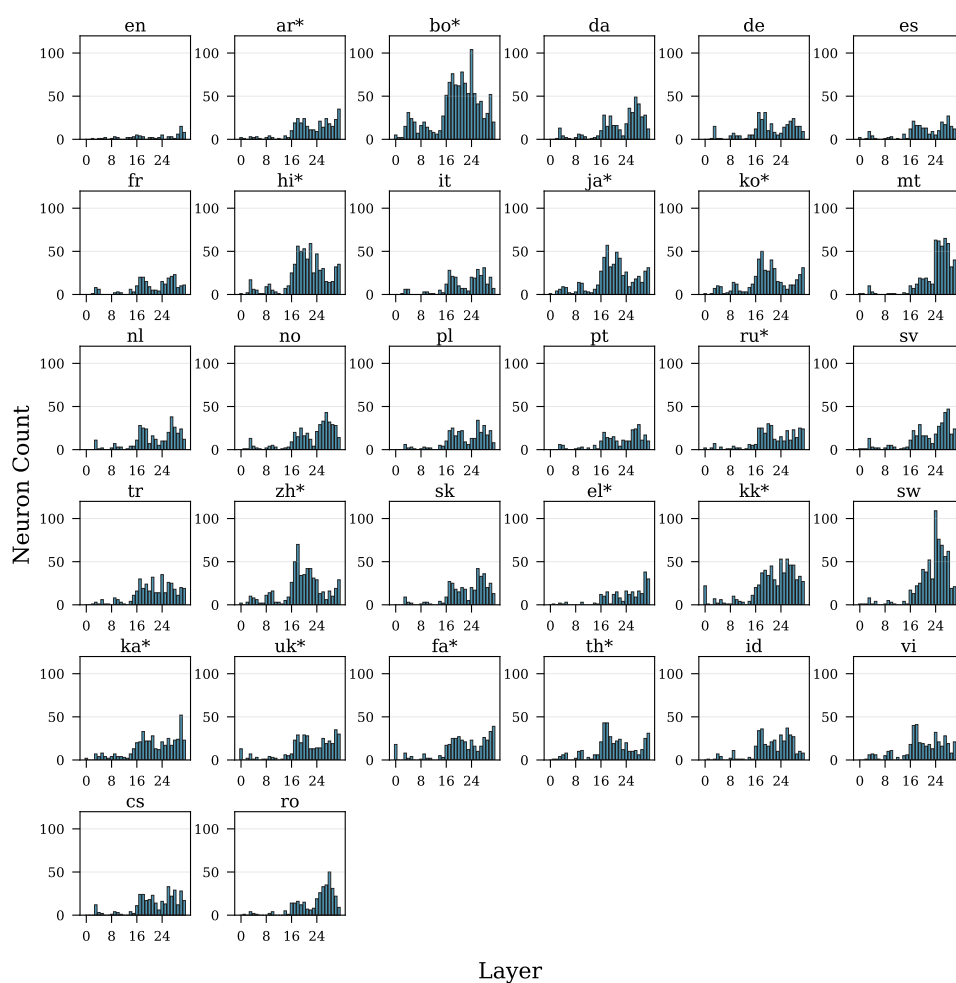


Figure 4: **Distribution** of LAPE identified **language-specific neurons** over layers in Llama-3.1-Instruct for all 32 languages.

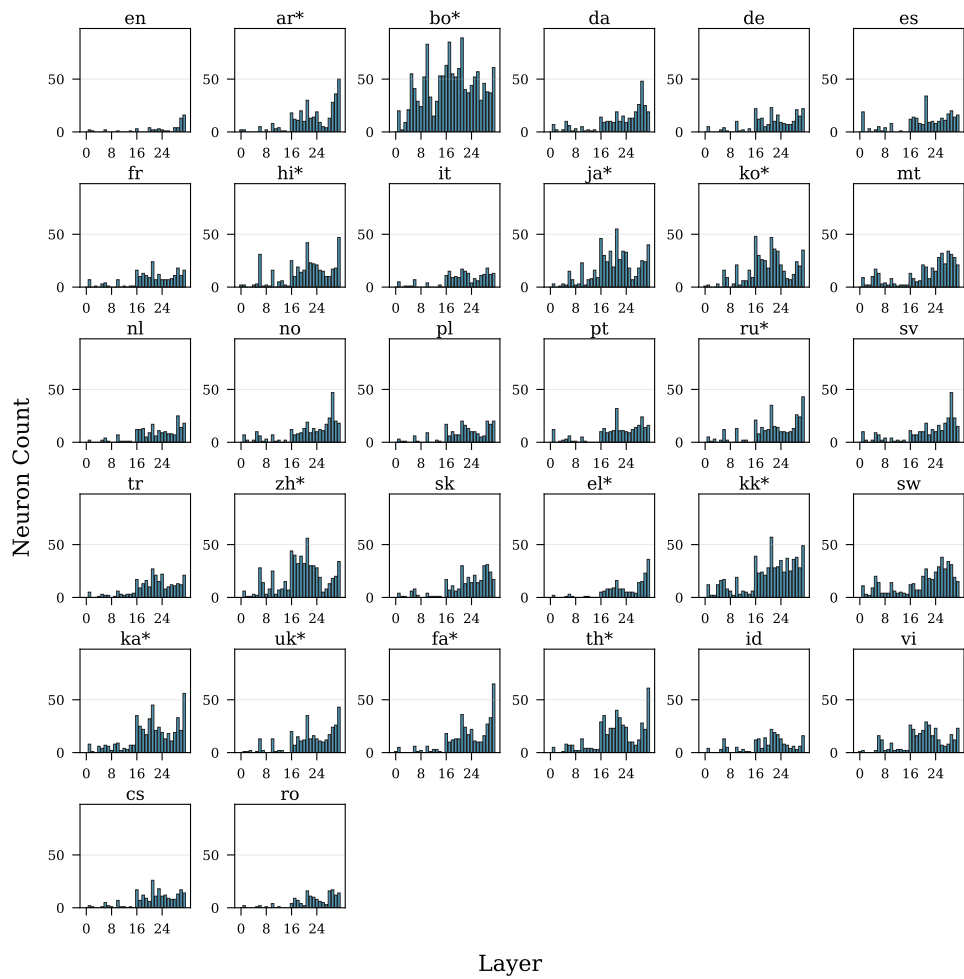


Figure 5: **Distribution** of LAPE identified **language-specific neurons** over layers in Aya-Expansive-8B for all 32 languages.

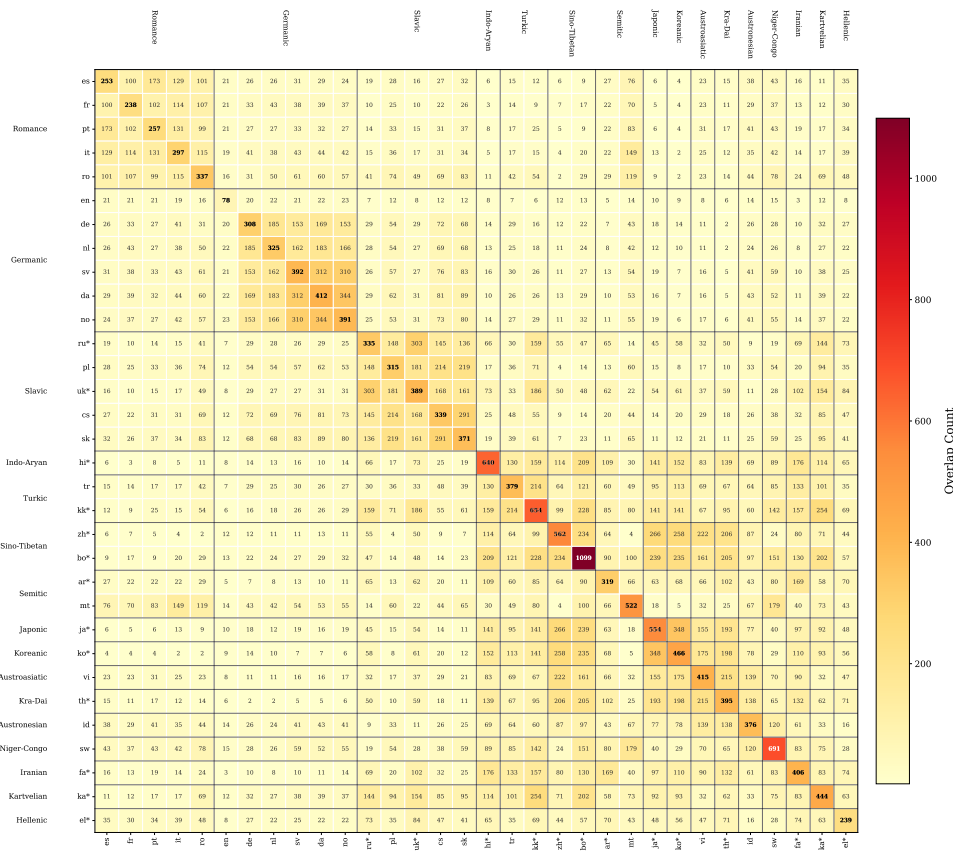


Figure 6: **Overlap** of all LAPE identified **language-specific neurons** in Llama-3.1-Instruct for the selected 32 languages.

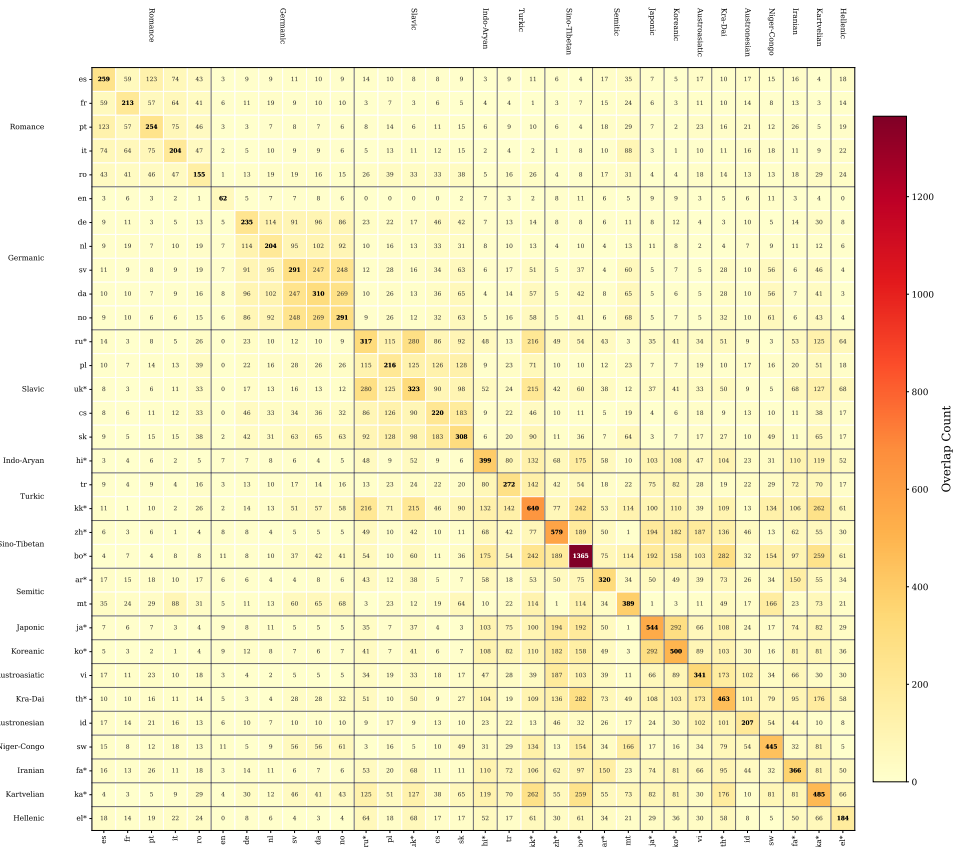


Figure 7: **Overlap** of all LAPE identified **language-specific neurons** in Aya-ExpansE-8B for the selected 32 languages.

E Language Vectors from Residuals

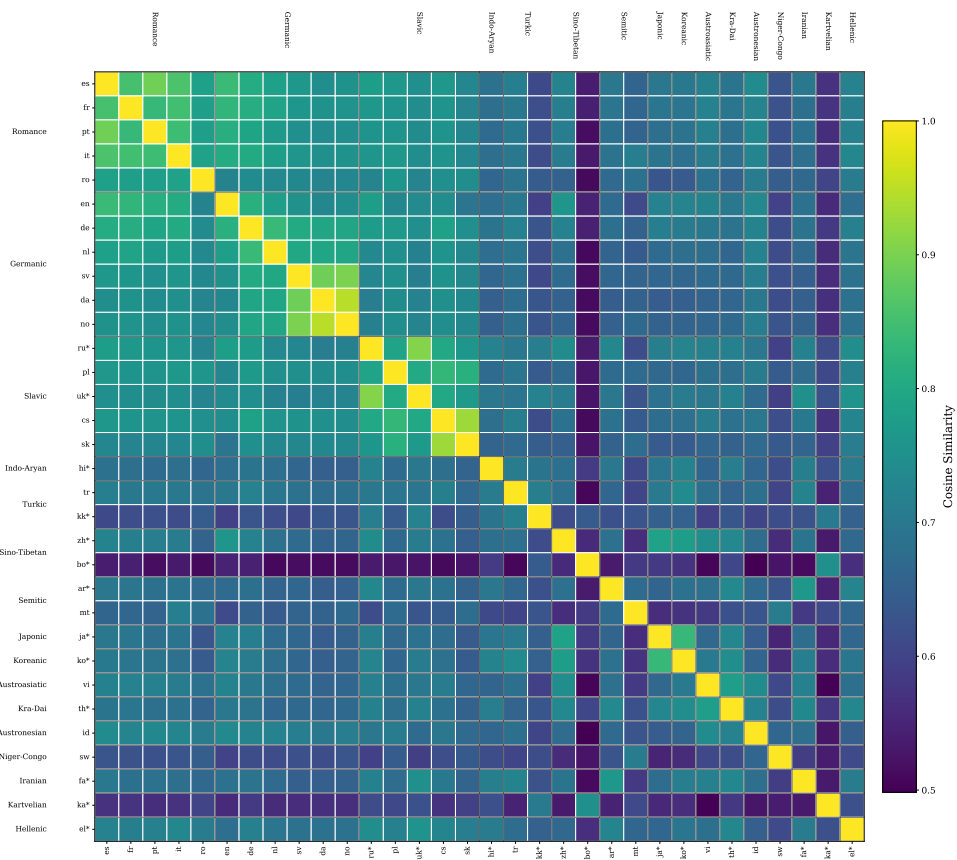


Figure 8: **Cosine similarity** between the **residual-based vectors** for all 32 selected languages in Llama-3.1-Instruct averaged over all layers.

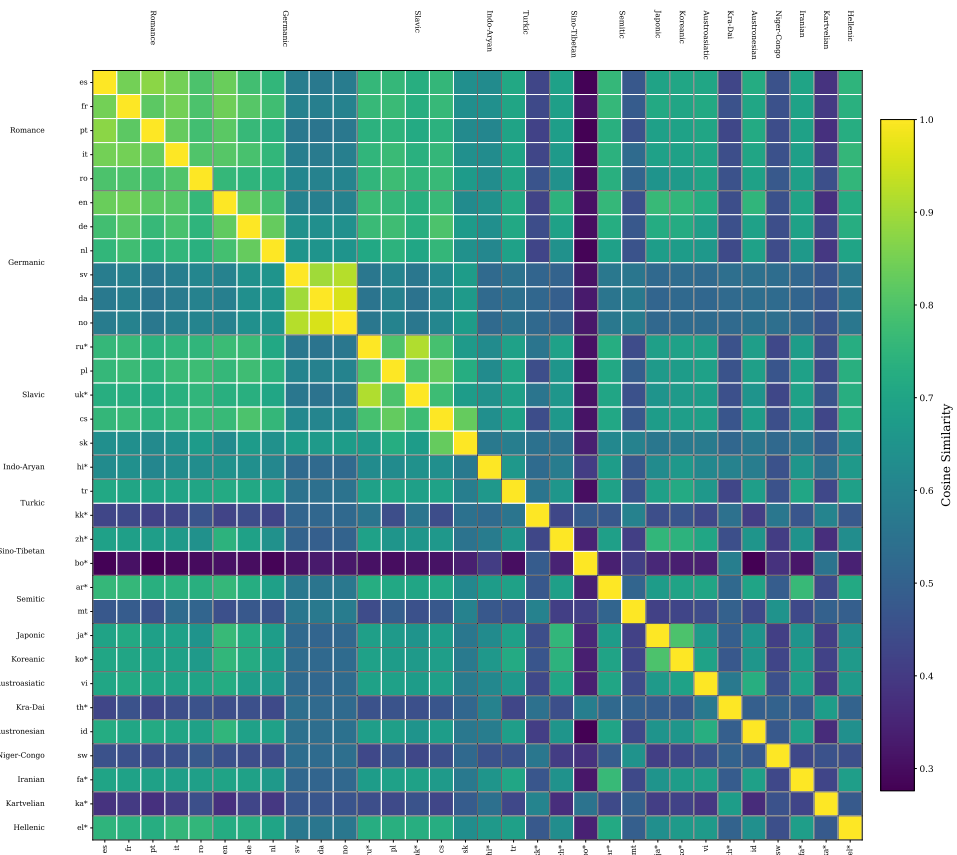


Figure 9: **Cosine similarity** between the **residual-based vectors** for all 32 selected languages in Aya-Expanse-8B averaged over all layers.

F Language Vectors from Probes

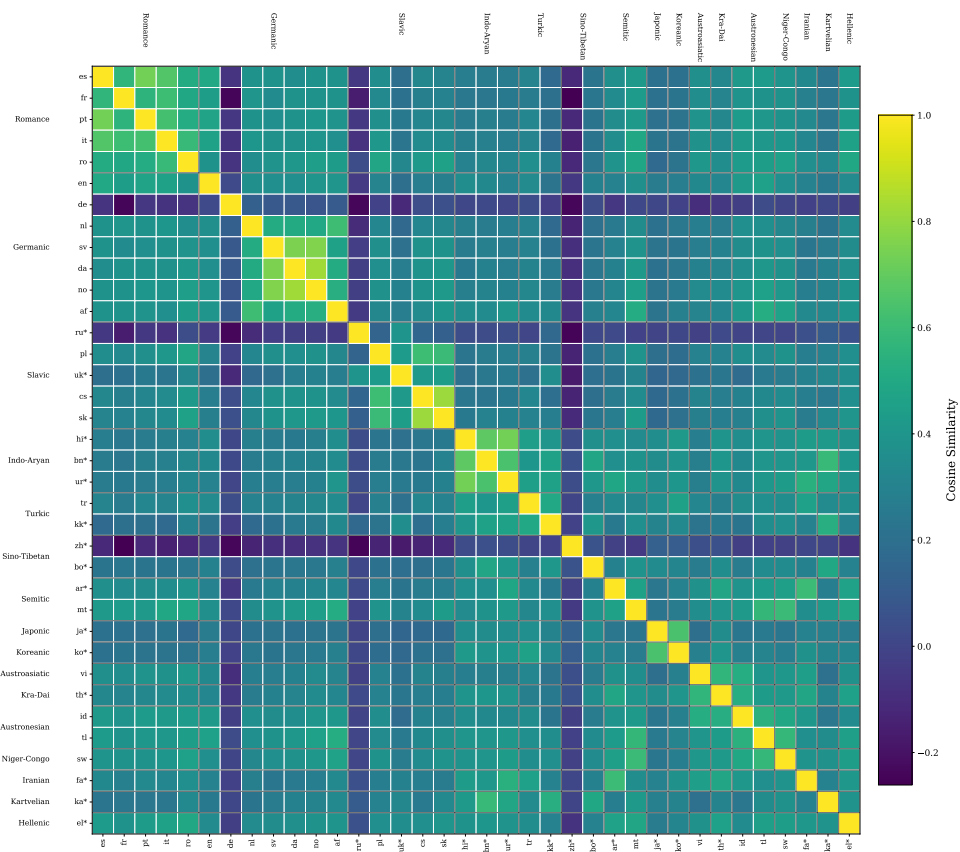


Figure 10: **Cosine similarity** between the **probe-based vectors** for all 32 selected languages in Llama-3.1-Instruct averaged over all layers.

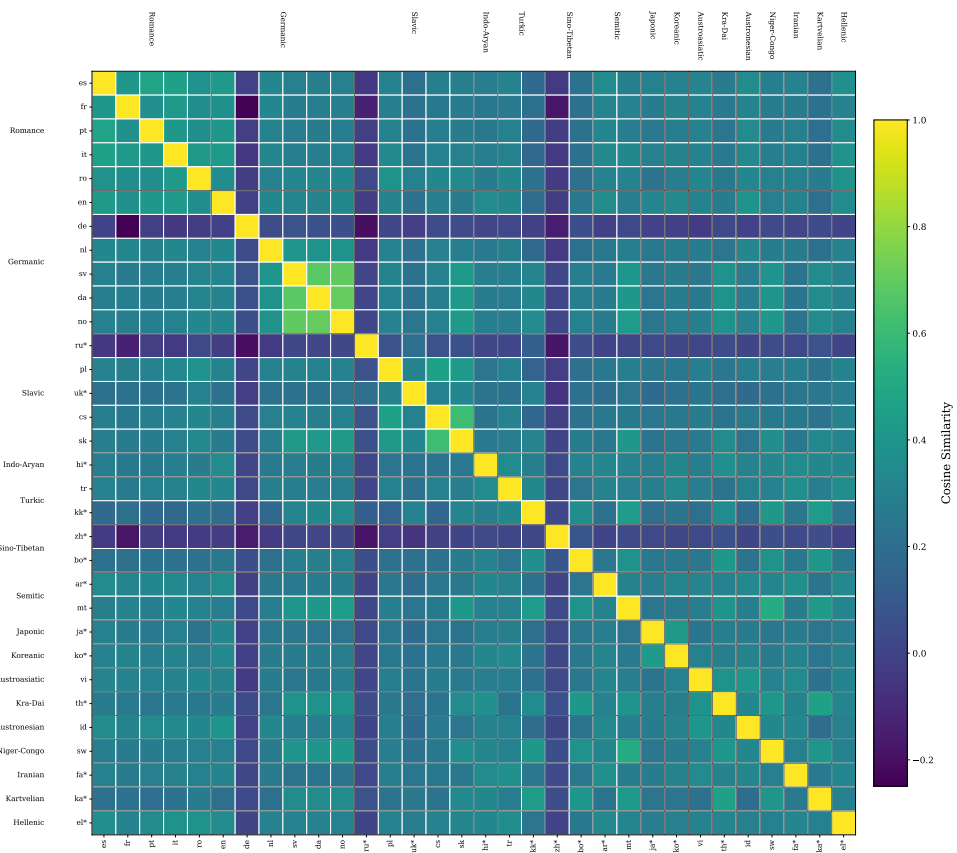


Figure 11: Cosine similarity between the probe-based vectors for all 32 selected languages in Aya-Expanse-8B averaged over all layers.

G Language Vectors from LDA

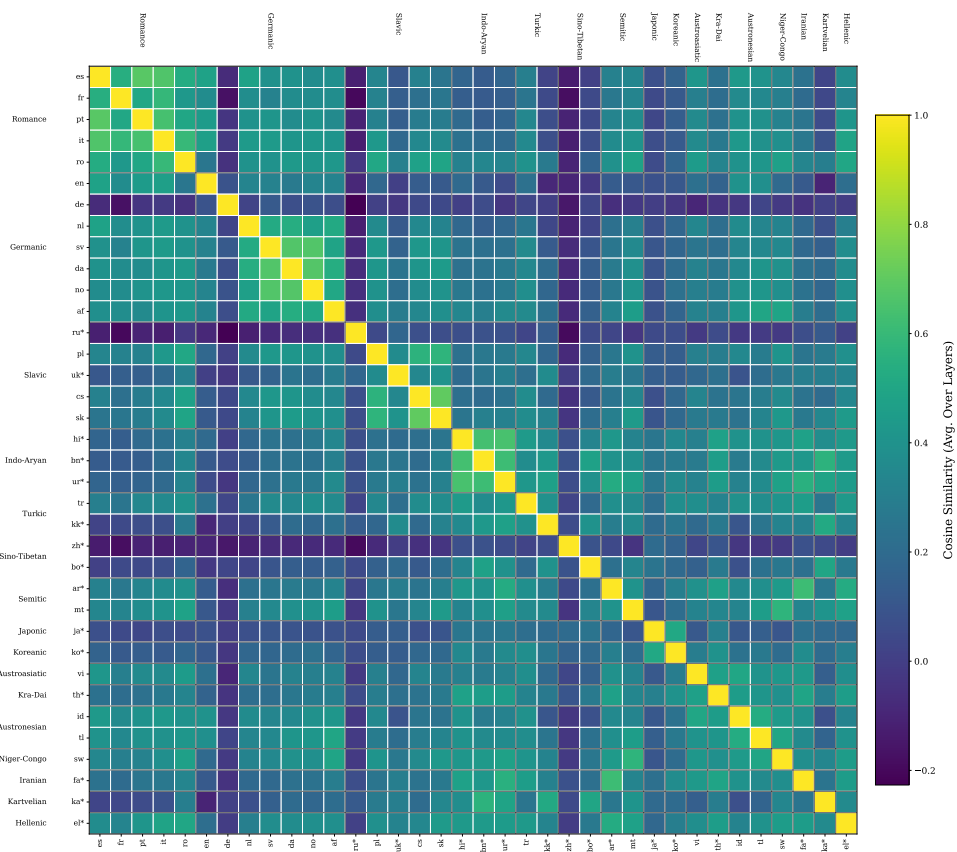


Figure 12: Cosine similarity between the LDA-based vectors for all 32 selected languages in Llama-3.1-Instruct averaged over all layers.

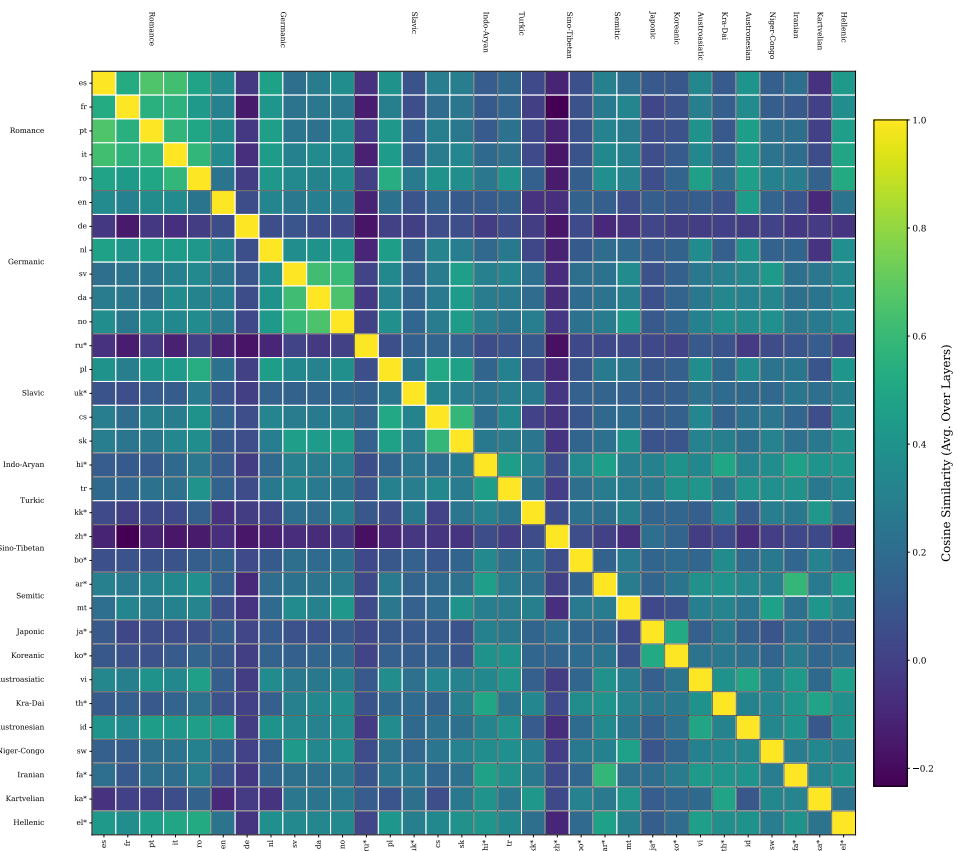


Figure 13: Cosine similarity between the LDA-based vectors for all 32 selected languages in Aya-Expanse-8B averaged over all layers.

H Mechanistic Insights into Language-specific Components

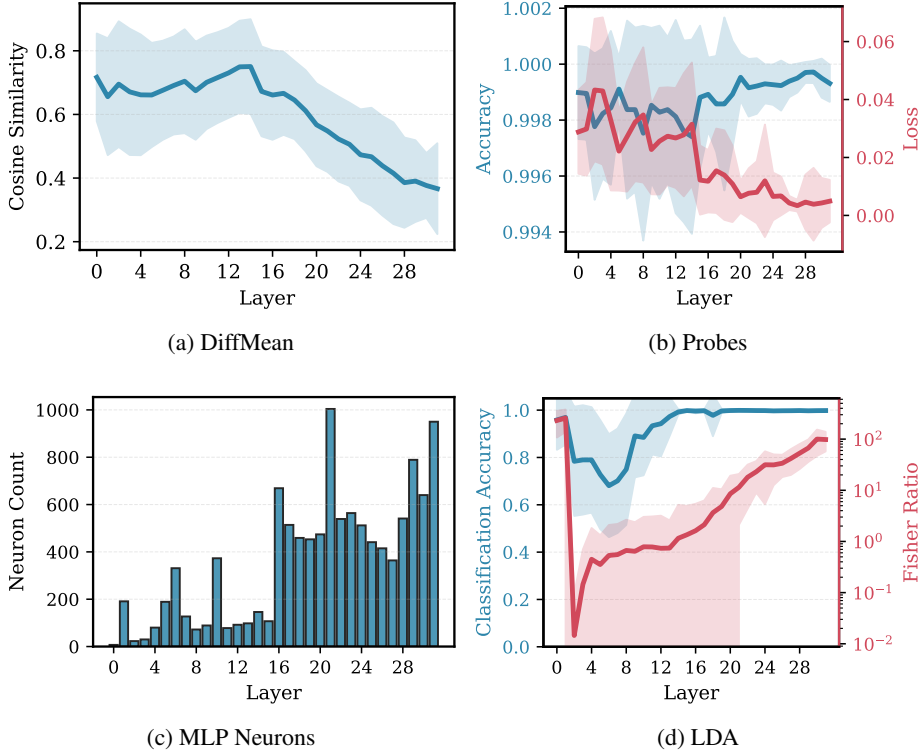


Figure 14: Mechanistic insights into language-specific components across interpretation tools for Aya-Expansive-8B. (a) DiffMean reveals average cosine similarity patterns across all languages. (b) Probes demonstrate learning dynamics through loss and accuracy trajectories. (c) LDA provides classification accuracy and Fisher Ratio (the degree of separability between two classes considered for LDA). (d) LAPE identifies the distribution of language-specific neurons across layers. Across all four methods, a consistent pattern emerges: language specificity concentrates in later layers, suggesting that higher-level representations encode language-dependent information.

I Ablation Results

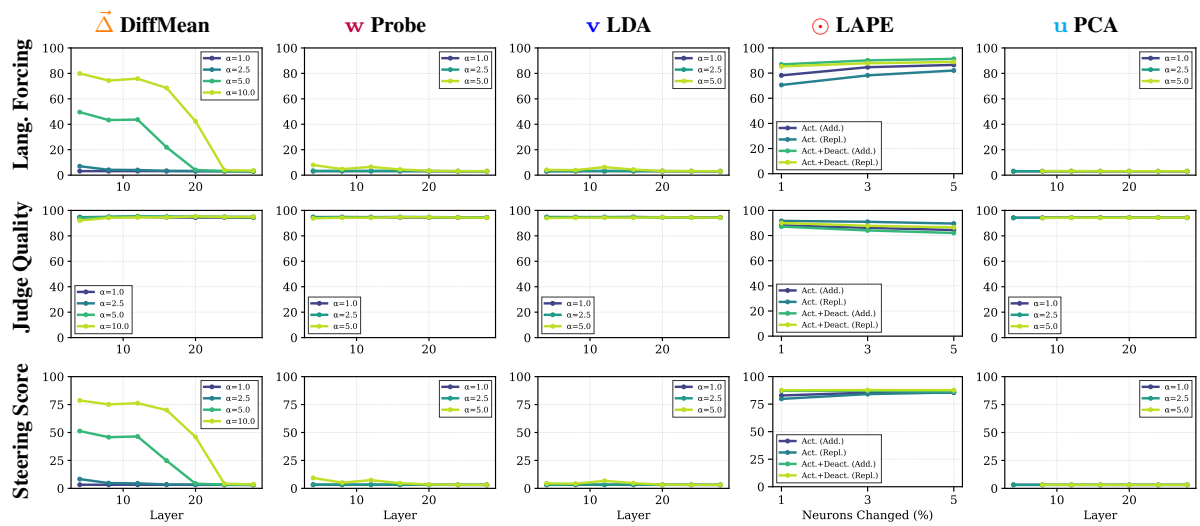


Figure 15: Comparative analysis of steering methods across evaluation metrics for Aya-Expansive-8B. Columns show different methods (DiffMean, Probe, LDA, LAPE, PCA). Rows represent: language forcing success rate, judge relevance quality, and overall steering score.

J Selected Layers and Intervention Strengths

Method	Selected Layer	Alpha Strength
\mathcal{E} Base.-I	-	-
\mathcal{E} Base.-II	-	-
$\vec{\Delta}$ DiffM.	20	5.0
\mathbf{w} Probe	20	5.0
\mathbf{u} PCA	4	5.0
$\vec{\Delta}$ SAE-DM.	25	5.0
\mathbf{v} LDA	16	5.0
\odot LAPE	-	1.0

Table 5: Hyperparameters–selected layer and alpha strength–for each steering method, based on the ablation results in Section 5, for Llama-3.1-8B.

Method	Selected Layer	Alpha Strength
\mathcal{E} Base.-I	-	-
\mathcal{E} Base.-II	-	-
$\vec{\Delta}$ DiffM.	4	10.0
\mathbf{w} Probe	4	5.0
\mathbf{u} PCA	12	5.0
\mathbf{v} LDA	12	5.0
\odot LAPE	-	1.0

Table 6: Hyperparameters–selected layer and alpha strength–for each steering method, based on the ablation results in Appendix I, for Aya-Expanse-8B.

K Per-language Forcing and Judge Scores

K.1 Llama-3.1-8B

Lang.	\mathcal{E} Base.-I	\mathcal{T} Base.-II	$\bar{\Delta}$ DiffM.	w Probe	u PCA	$\bar{\Delta}$ SAE-DM.	v LDA	\odot LAPE
ar	47.4	38.2	97.1	7.9	35.0	51.1	23.1	92.9
bo	64.6	26.7	4.5	2.9	6.5	3.2	2.5	4.2
cs	54.2	62.7	94.7	17.3	76.3	27.2	16.8	89.3
da	69.0	32.1	92.8	22.9	85.5	20.9	4.7	38.2
de	55.7	40.6	97.4	82.0	75.8	38.7	-	96.6
el	73.3	65.3	96.0	4.8	31.5	29.4	3.2	94.7
en	3.7	3.7	96.3	96.0	31.5	36.1	48.3	99.5
es	61.1	49.8	96.5	71.7	64.7	44.7	32.9	91.9
fa	24.0	37.2	72.4	7.2	15.5	39.1	5.5	84.4
fr	51.3	51.8	98.2	89.0	62.9	53.7	13.6	96.6
hi	36.5	53.2	96.4	7.3	35.8	48.0	-	97.3
id	49.8	64.3	97.0	46.0	33.2	10.1	-	95.1
it	50.8	56.8	97.5	73.6	77.7	41.5	8.8	96.3
ja	32.4	59.1	85.9	22.5	32.8	5.3	9.4	92.4
ka	91.9	59.1	21.3	3.2	11.2	3.2	2.0	31.3
kk	75.2	45.6	32.3	3.3	17.2	4.7	3.2	71.4
ko	27.7	51.8	99.5	26.0	50.7	23.1	11.9	94.7
mt	94.8	61.7	8.7	3.0	8.6	3.2	3.0	12.9
nl	52.3	51.8	97.7	31.6	64.9	30.4	4.1	96.0
no	47.6	39.7	90.3	17.8	83.0	27.3	2.3	56.9
pl	37.6	38.0	95.9	21.3	79.2	30.1	33.6	91.2
pt	44.8	67.1	97.7	81.9	48.3	34.5	41.2	95.9
ro	55.5	75.5	96.5	44.0	70.9	23.2	3.3	71.3
ru	41.3	54.8	99.2	78.3	55.2	62.6	43.1	97.6
sk	75.9	37.1	77.9	19.9	82.0	10.4	5.9	57.8
sv	47.7	48.0	95.9	33.4	78.2	26.5	46.1	86.3
sw	95.8	83.3	51.7	4.4	60.3	3.2	3.2	49.3
th	60.5	80.6	99.1	14.1	37.8	36.6	5.2	96.3
tr	66.6	60.8	93.3	13.6	32.4	11.4	3.8	94.7
uk	52.2	70.7	93.2	7.3	63.8	76.4	8.9	98.7
vi	68.8	82.7	99.3	42.8	31.8	6.0	5.9	96.2
zh	44.1	63.5	99.4	92.3	16.1	83.5	94.4	97.4
Avg.	54.8	53.5	83.5	34.0	48.6	29.5	13.9	80.2

Table 7: Language forcing scores across all methods for 32 ablation languages for Llama-3.1-8B-Instruct.

Lang.	\mathcal{E} Base.-I	\mathcal{T} Base.-II	Δ DiffM.	w Probe	u PCA	Δ SAE-DM.	v LDA	\odot LAPE
ar	93.5	91.7	81.8	85.2	10.6	47.9	71.2	66.0
bo	22.6	67.3	79.1	84.2	6.8	85.4	74.0	43.1
cs	95.4	93.8	90.2	82.7	9.3	71.9	73.9	88.8
da	96.8	92.5	87.5	82.6	9.7	75.8	77.3	90.1
de	96.3	93.3	93.5	84.8	8.2	75.8	88.0	96.3
el	93.9	92.0	81.9	87.9	10.3	57.7	78.0	80.7
en	92.6	92.5	95.4	95.5	6.7	87.9	54.5	98.1
es	96.1	94.9	92.9	87.2	12.2	81.3	85.8	96.1
fa	93.9	93.8	86.9	84.3	12.6	60.2	77.0	85.9
fr	95.5	93.5	94.8	90.4	11.4	81.1	82.8	96.4
hi	93.0	94.0	85.2	84.8	7.0	61.1	94.3	89.4
id	96.4	95.3	92.5	82.3	12.7	81.7	95.7	94.2
it	94.9	95.0	92.2	81.8	13.8	74.8	79.1	95.5
ja	92.4	94.0	83.4	84.7	9.1	83.1	82.1	76.1
ka	71.2	79.8	71.9	86.5	12.4	85.7	45.4	24.3
kk	55.3	74.5	60.7	87.3	3.7	69.0	81.4	17.9
ko	93.2	92.6	81.8	85.0	10.2	77.2	82.6	69.0
mt	53.8	67.6	72.9	85.2	6.4	79.4	69.7	53.7
nl	96.0	94.2	93.0	83.7	7.7	76.2	84.1	94.0
no	94.7	93.8	87.5	81.8	10.1	66.1	86.9	89.3
pl	95.2	92.6	89.3	83.1	5.8	66.3	75.2	88.7
pt	95.8	95.7	94.0	87.5	11.7	80.9	86.5	95.3
ro	95.4	93.8	89.8	84.6	7.6	71.1	77.6	84.3
ru	94.3	94.0	94.8	87.1	9.9	73.6	83.3	93.4
sk	92.8	92.2	77.5	79.1	4.1	65.4	70.6	76.5
sv	95.8	94.2	91.6	83.3	9.5	75.9	86.9	93.6
sw	74.5	76.3	48.0	83.6	1.1	84.6	77.8	56.6
th	93.4	92.3	84.5	84.6	7.4	69.9	85.9	75.2
tr	92.5	91.4	77.2	82.7	9.5	73.0	83.6	75.6
uk	94.6	93.9	92.6	87.6	8.8	64.1	82.2	85.4
vi	95.3	94.9	94.6	81.4	9.6	85.5	81.3	93.8
zh	95.9	96.6	95.9	92.4	9.8	87.4	81.8	95.1
Avg.	88.5	90.4	85.5	85.2	8.9	74.3	79.3	80.0

Table 8: **Output relevance scores** across all methods for 32 languages for Llama-3.1-8B-Instruct.

K.2 Aya-ExpansE-8B

Lang.	\mathcal{E} Base.-I	\mathcal{T} Base.-II	$\bar{\Delta}$ DiffM.	w Probe	u PCA	v LDA	\odot LAPE
ar	100.0	100.0	90.4	15.0	3.2	8.4	99.6
bo	5.8	4.6	9.2	4.6	-	-	1.8
cs	99.9	99.5	96.9	43.0	1.7	1.3	100.0
da	99.5	96.4	68.8	3.6	3.5	3.0	99.7
de	100.0	100.0	97.3	7.4	3.2	4.6	100.0
el	100.0	100.0	96.9	8.6	3.2	3.2	100.0
en	89.1	89.1	96.7	6.5	5.0	77.8	100.0
es	99.5	100.0	97.0	8.0	3.2	3.3	100.0
fa	99.4	98.8	98.4	8.7	3.1	3.2	99.3
fr	100.0	100.0	99.2	8.9	3.2	3.2	100.0
hi	99.9	99.4	99.7	6.7	5.8	6.5	99.8
id	99.8	100.0	98.8	4.5	3.4	3.3	100.0
it	99.9	100.0	97.1	5.8	5.1	4.9	100.0
ja	96.0	95.8	96.0	9.9	3.2	10.2	97.8
ka	98.2	80.5	19.1	3.2	3.2	3.2	62.0
kk	69.1	46.1	13.8	13.8	13.8	13.8	1.2
ko	100.0	100.0	100.0	12.9	3.2	4.7	100.0
mt	93.5	11.9	30.6	3.0	2.6	2.6	78.6
nl	100.0	100.0	99.5	13.4	3.2	3.2	100.0
no	99.5	99.2	70.0	3.6	3.7	3.5	85.1
pl	99.8	100.0	96.7	20.6	3.2	3.2	100.0
pt	99.8	100.0	97.0	5.1	3.2	3.2	100.0
ro	100.0	99.7	97.0	10.7	3.2	3.2	99.9
ru	99.7	99.1	96.5	5.8	3.9	4.6	100.0
sk	85.0	95.2	60.5	1.5	2.7	2.6	98.5
sv	99.9	99.9	94.9	3.6	3.4	3.3	99.8
sw	98.8	83.7	34.1	3.2	3.1	3.2	99.5
th	99.9	99.8	45.9	3.4	3.2	3.3	100.0
tr	100.0	99.7	96.3	7.3	3.2	3.3	100.0
uk	100.0	100.0	96.9	6.9	3.3	3.3	100.0
vi	99.9	99.9	95.9	8.3	3.2	3.3	100.0
zh	99.8	99.7	94.2	6.6	3.4	16.0	100.0
Avg.	94.7	90.6	80.7	8.6	3.7	7.0	91.3

Table 9: **Language forcing scores** across all methods for 32 ablation languages for Aya-ExpansE-8B.

Lang.	\mathcal{E} Base.-I	\mathcal{T} Base.-II	$\tilde{\Delta}$ DiffM.	w Probe	u PCA	v LDA	\odot LAPE
ar	96.3	98.2	95.1	94.2	94.6	94.6	94.9
bo	11.6	95.0	89.3	94.0	94.4	94.7	45.2
cs	97.0	98.3	94.8	94.9	94.8	94.2	94.3
da	96.3	96.8	93.5	93.5	94.8	95.2	87.9
de	97.0	98.2	94.8	94.0	94.5	94.8	95.1
el	95.8	98.0	94.4	94.7	94.5	94.4	94.1
en	98.2	98.3	95.8	94.7	94.4	95.0	96.7
es	96.0	97.9	95.5	94.3	94.6	95.0	95.5
fa	96.1	97.6	94.4	94.3	94.3	94.7	93.3
fr	97.1	98.5	95.1	93.1	93.7	95.3	95.2
hi	96.2	97.2	94.7	93.9	93.8	94.3	94.1
id	96.3	98.3	94.9	93.8	93.9	94.9	95.3
it	97.1	97.4	94.8	94.4	94.5	94.5	94.1
ja	95.8	96.2	93.9	93.8	95.0	93.8	94.5
ka	37.2	46.7	81.7	92.9	94.7	93.7	31.7
kk	47.8	55.5	75.8	92.9	94.3	93.6	13.8
ko	96.0	96.2	94.0	93.5	93.5	93.2	93.8
mt	73.9	92.5	81.2	92.8	93.3	93.7	42.0
nl	97.0	98.3	95.8	93.5	94.2	94.3	95.3
no	96.4	95.3	92.3	92.2	93.9	93.8	85.5
pl	96.8	97.5	95.5	94.3	94.4	93.7	93.5
pt	96.6	98.0	95.6	93.7	94.1	94.7	95.5
ro	96.6	98.0	95.6	94.1	93.9	93.0	94.1
ru	96.5	98.1	95.0	93.9	94.0	94.2	94.3
sk	96.0	96.9	94.7	94.6	95.0	93.9	89.7
sv	96.7	97.5	94.6	93.0	93.9	94.0	87.4
sw	48.0	60.9	77.5	92.5	93.8	94.8	14.2
th	73.3	71.9	78.3	92.5	94.7	93.5	48.5
tr	96.2	96.4	93.4	94.3	94.4	94.3	93.8
uk	96.5	97.4	94.7	95.1	95.2	94.3	94.4
vi	96.5	97.5	94.1	94.6	95.0	94.5	94.1
zh	96.4	96.2	94.7	93.6	93.9	94.3	94.7
Avg.	87.5	92.4	92.1	93.8	94.3	94.3	82.1

Table 10: Output relevance scores across all methods for 32 languages for Aya-Expanse-8B.

Lang.	\mathcal{E} Base.-I	\mathcal{T} Base.-II	$\tilde{\Delta}$ DiffM.	w Probe	u PCA	v LDA	\odot LAPE
ar	98.1	99.0	92.7	25.9	6.2	15.4	97.2
bo	7.8	8.8	16.7	8.8	-	-	3.5
cs	98.4	98.9	95.8	59.1	3.4	2.5	97.0
da	97.9	96.6	79.2	6.9	6.8	5.9	93.4
de	98.5	99.1	96.0	13.7	6.2	8.7	97.5
el	97.8	99.0	95.6	15.8	6.2	6.2	97.0
en	93.4	93.5	96.2	12.2	9.5	85.6	98.3
es	97.7	99.0	96.3	14.8	6.2	6.3	97.7
fa	97.7	98.2	96.3	15.9	6.1	6.2	96.2
fr	98.5	99.2	97.1	16.3	6.2	6.2	97.5
hi	98.0	98.3	97.2	12.5	10.8	12.1	96.8
id	98.0	99.2	96.8	8.5	6.5	6.3	97.6
it	98.5	98.7	96.0	11.0	9.7	9.4	97.0
ja	95.9	96.0	95.0	17.9	6.2	18.4	96.1
ka	53.9	59.1	30.9	6.2	6.2	6.2	42.0
kk	56.5	50.4	23.4	24.1	24.1	24.1	2.2
ko	98.0	98.0	96.9	22.6	6.2	8.9	96.8
mt	82.5	21.1	44.4	5.7	5.1	5.1	54.8
nl	98.5	99.2	97.6	23.4	6.2	6.2	97.6
no	97.9	97.2	79.6	6.9	7.2	6.8	85.3
pl	98.3	98.7	96.1	33.8	6.2	6.2	96.6
pt	98.2	99.0	96.3	9.6	6.2	6.2	97.7
ro	98.3	98.8	96.3	19.3	6.2	6.2	96.9
ru	98.1	98.6	95.8	11.0	7.5	8.7	97.0
sk	90.2	96.0	73.8	3.0	5.2	5.0	93.9
sv	98.3	98.7	94.8	6.9	6.6	6.4	93.2
sw	64.6	70.5	47.4	6.2	6.1	6.2	24.8
th	84.6	83.6	57.9	6.6	6.2	6.3	65.3
tr	98.0	98.0	94.8	13.5	6.2	6.3	96.8
uk	98.2	98.7	95.8	12.9	6.3	6.3	97.1
vi	98.2	98.7	95.0	15.2	6.2	6.3	96.9
zh	98.1	97.9	94.5	12.4	6.5	27.3	97.3
Avg.	90.2	88.9	83.1	15.0	7.1	11.1	84.2

Table 11: Language steering scores (i.e. harmonic means of language forcing and output relevance scores) across all methods for 32 ablation languages for Aya-Expanse-8B. Language steering score is a harmonic mean of language forcing success and output relevance.

L Between-language Forcing Results

L.1 Llama-3.1-8B



Figure 16: Between-language forcing scores for **Baseline-I** across 32 languages in Llama-3.1-Instruct. The matrix structure allows for tracing steerability in both directions: which languages are most amenable to being steered away from (rows) and which are most readily steered into (columns).

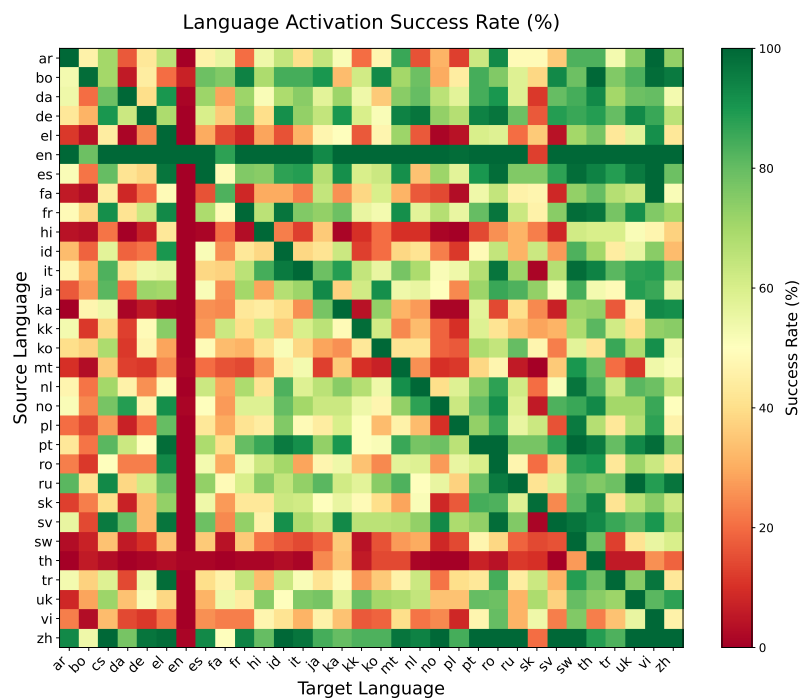


Figure 17: Between-language forcing scores for **Baseline-II** across 32 languages in Llama-3.1-Instruct. The matrix structure allows for tracing steerability in both directions: which languages are most amenable to being steered away from (rows) and which are most readily steered into (columns).

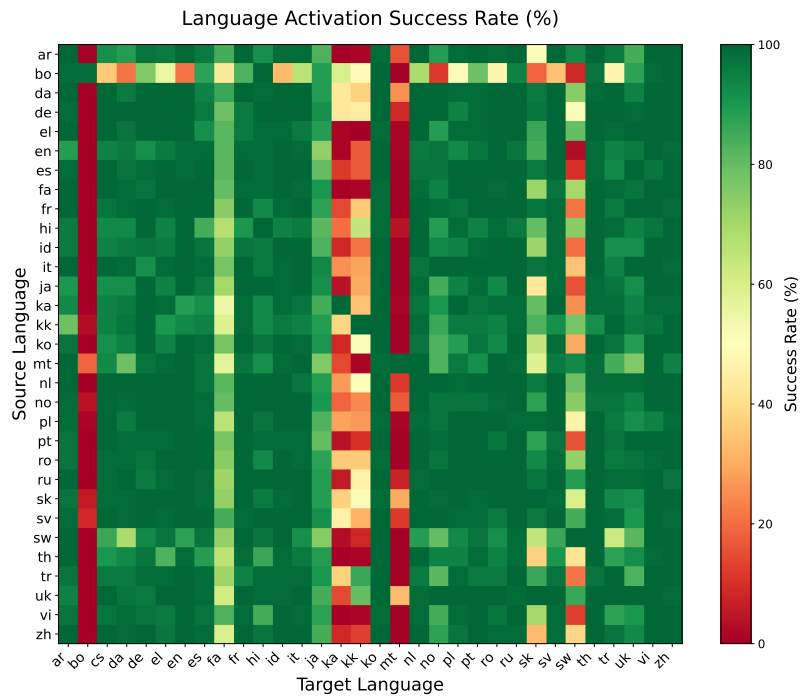


Figure 18: Between-language forcing scores for **DiffMean** across 32 languages in Llama-3.1-Instruct. The matrix structure allows for tracing steerability in both directions: which languages are most amenable to being steered away from (rows) and which are most readily steered into (columns).

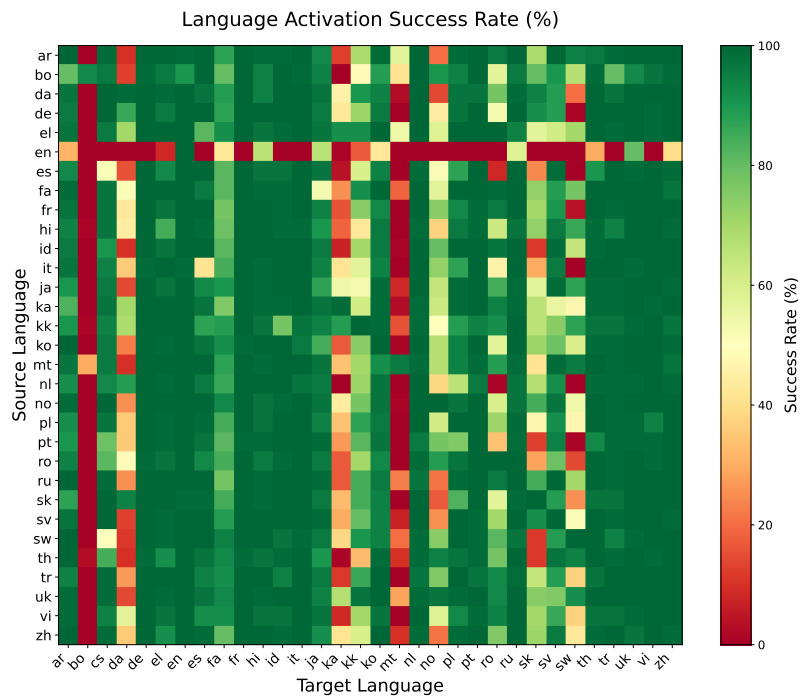


Figure 19: Between-language forcing scores for **LAPE** across 32 languages in Llama-3.1-Instruct. The matrix structure allows for tracing steerability in both directions: which languages are most amenable to being steered away from (rows) and which are most readily steered into (columns).

L.2 Aya-Expense-8B

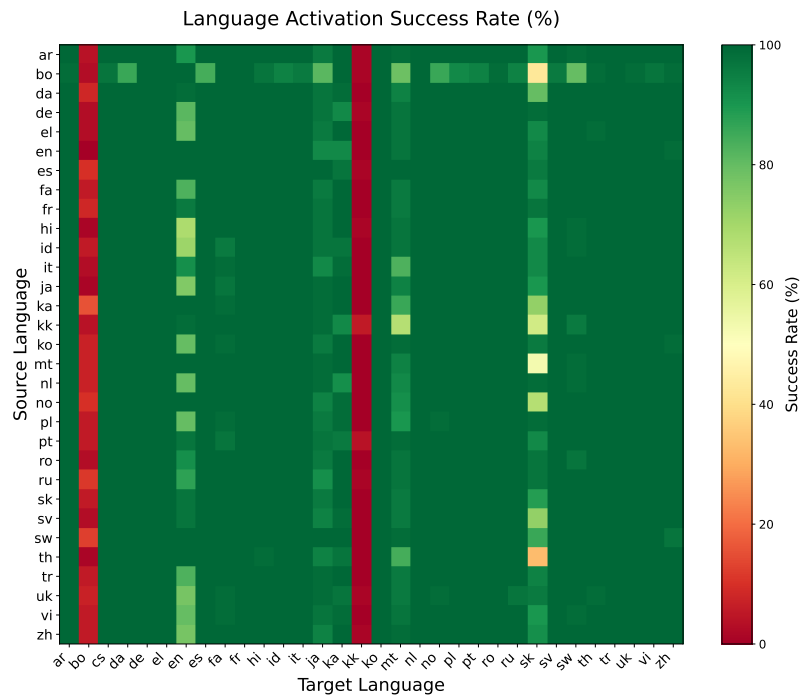


Figure 20: Between-language forcing scores for **Baseline-I** across 32 languages in Aya-Expense-8B. The matrix structure allows for tracing steerability in both directions: which languages are most amenable to being steered away from (rows) and which are most readily steered into (columns).

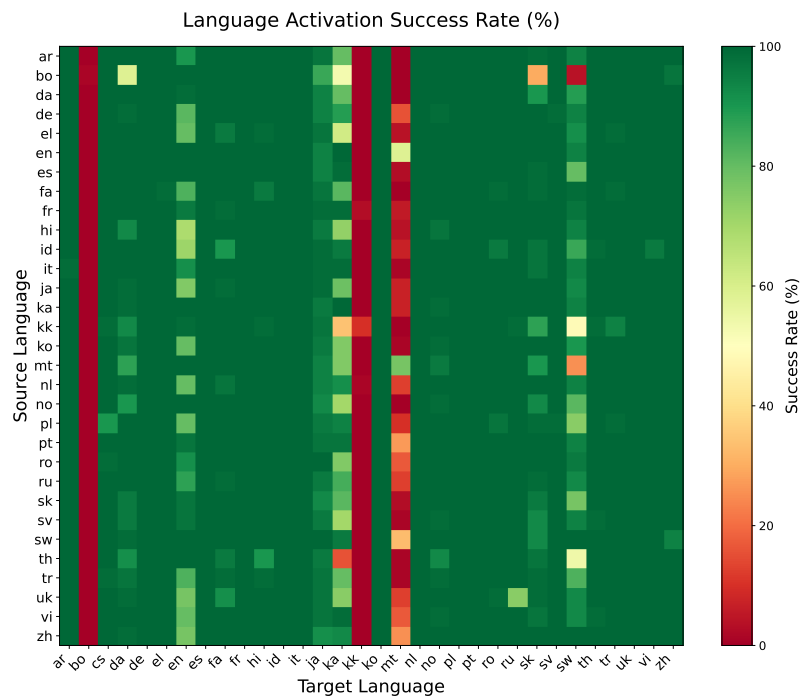


Figure 21: Between-language forcing scores for **Baseline-II** across 32 languages in Aya-Expense-8B. The matrix structure allows for tracing steerability in both directions: which languages are most amenable to being steered away from (rows) and which are most readily steered into (columns).

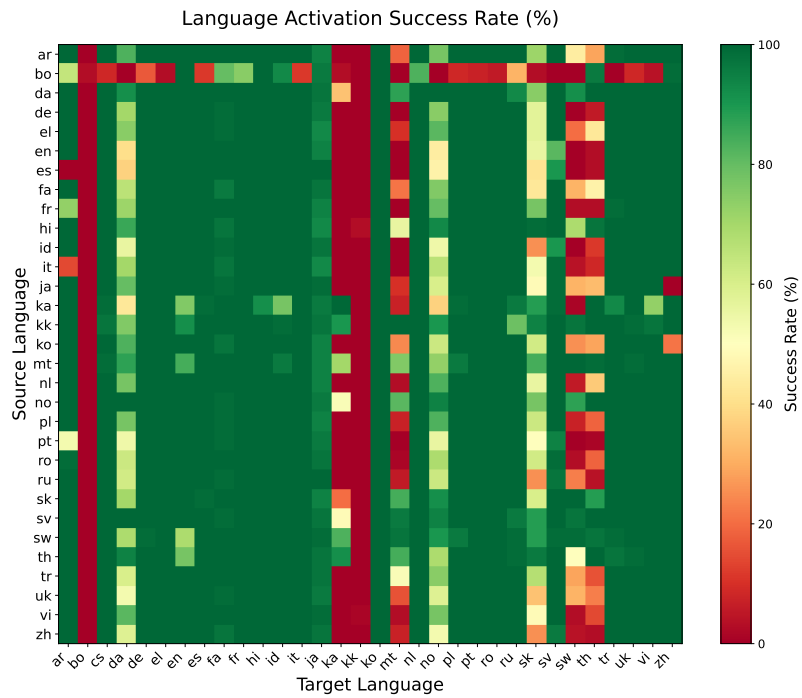


Figure 22: Between-language forcing scores for **DiffMean** across 32 languages in Aya-Expansive-8B. The matrix structure allows for tracing steerability in both directions: which languages are most amenable to being steered away from (rows) and which are most readily steered into (columns).

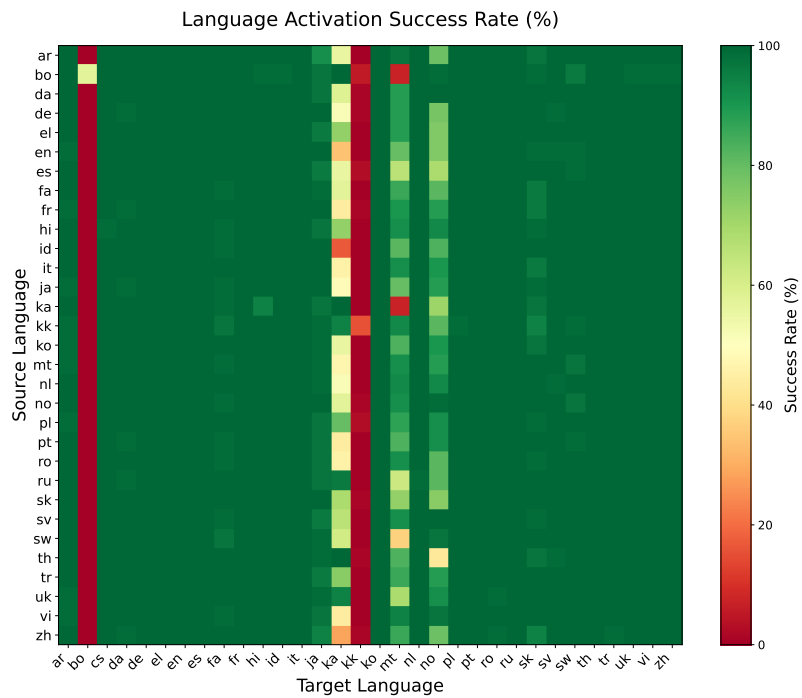


Figure 23: Between-language forcing scores for **LAPE** across 32 languages in Aya-Expansive-8B. The matrix structure allows for tracing steerability in both directions: which languages are most amenable to being steered away from (rows) and which are most readily steered into (columns).

M Sample Generations

M.1 Llama-3.1-8B

Method	Generated Output
Input (Spanish): ¿Cuáles son las formas más efectivas de lidiar con el estrés?	Target: German
\mathcal{E} Baseline-I	Existen varias formas efectivas de lidiar con el estrés. Aquí te presento algunas de ellas: 1. **Meditación y respiración profunda** : La meditación...
\mathcal{E} Baseline-II	Existen varias formas efectivas de lidiar con el estrés. Aquí te presento algunas de ellas: 1. **Meditación y respiración profunda** : La meditación...
$\vec{\Delta}$ DiffMean	Es ist wichtig, dass es verschiedene Strategien gibt, um mit Stress umzugehen. Hier sind einige der effektivsten Methoden: 1. **Atmungsübungen** : Die Atmung ist ein wichtiger Teil des Stress-Management-Systems...
w Probe	Existen varios Methodos efectivos para lidiar con el estrés. Aquí te presento algunos de los más comunes: 1. **Meditación und Yoga** ...
u PCA	Existen varias formas efectivas de lidiar con el estrés, algunas de las cuales incluyen: 1. **Prácticas de mindfulness** :...
$\vec{\Delta}$ SAE-DM	Lidiar con el estrés puede ser un desafío, pero hay varias formas efectivas de manejarlo. Aquí están algunas de las mejores estrategias: 1. **La respiración profunda** :...
v LDA	Es gibt verschiedene Möglichkeiten, mit Stress umzugehen. Hier sind einige der effektivsten Methoden: 1. **Meditation und tiefe Atmung** :...

Table 12: Sample outputs for steering from Spanish to German. Methods highlighted in **green** successfully produce German, and **red** indicates failure to switch from the source language.

Method	Generated Output
Input (Korean): 시간 관리 능력을 향상시키는 방법은 무엇입니까?	Target: English
\mathcal{E} Baseline-I	시간 관리 능력을 향상시키는 방법은 다음과 같습니다. 1. **목표 설정** :...
\mathcal{E} Baseline-II	시간 관리 능력을 향상시키는 방법은 다음과 같습니다. 1. **목표 설정** :...
$\vec{\Delta}$ DiffMean	Time management skills to improve: 1. **Goal setting** : Clear goals help you focus on what's important and allocate your time accordingly. Set specific, measurable, achievable, relevant, and time-bound (SMART) goals...
w Probe	time management skills to improve. Here are some tips to help you improve your time management skills: 1. **Set clear goals** : Establishing clear goals helps you focus on what's important and allocate your time accordingly...
u PCA	시간 관리는 시간을 관리하는 방법을 향상하는 방법을 향상하는 방법을...
$\vec{\Delta}$ SAE-DM	시간 관리 능력을 향상시키는 방법은 다음과 같습니다. 1. **목표 설정** :...
v LDA	시간 관리 ability 향상시키는 방법은 다음과: 1. **Prioritize** : 할 일 목록을 만들고...
\odot LAPE	Improving time management skills can be achieved through several strategies. Here are some effective methods: 1. **Set clear goals** :...

Table 13: Sample outputs for steering from Korean to English. Methods highlighted in **green** successfully produce English, and **red** indicates failure to switch from the source language.

M.2 Aya-Expans-8B

Method	Generated Output
	Input (French): Écrivez une revue de concert symphonique, en discutant de la performance de l'orchestre et de l'expérience globale du public. Target: Slovak
ε Baseline-I	Koncert symfonického orchestru: Hudobná cesta plná emocií. Minulý večer som málokoného zažil tak silný a zapamätateľný koncert ako bol ten, ktorý predvedol symfonický orchestr v našom meste...
ε Baseline-II	**Recenzia symfonického koncertu: Hudobná podívaná plná emocií** Minulý večer som mala príjemnosť navštíviť symfonický koncert, ktorý sa stal nezabudnuteľným zážitkom pre všetkých prítomných...
Δ DiffMean	**Koncert symfonického orchestru: Hudba, ktorá uchvátila srdce** Minulý večer sa v našom meste konal neobyčajný hudobný zážitok, ke ktorému sa sjednotili milovníci klasickej hudby z celého regionu...
w Probe	**Revue du Concert Symphonique : Une Soirée Musicale Inoubliable** Le week-end dernier, j'ai eu le privilège d'assister à un concert symphonique exceptionnel donné par l'Orchestre Philharmonique de la ville, qui a rempli la salle de concert jusqu'à sa capacité maximale...
u PCA	**Une soirée enivrante avec l'Orchestre Symphonique : Une fusion parfaite de la musique et de l'émotion** Le concert symphonique de la semaine dernière a été une expérience musicale inoubliable, offrant au public une performance exceptionnelle de l'Orchestre Symphonique local...
v LDA	**Revue du Concert Symphonique : Une Soirée Musicale Inoubliable** Le week-end dernier, j'ai eu le privilège d'assister à un concert symphonique exceptionnel donné par l'Orchestre Symphonique National dans le cadre de leur tournée annuelle...
⊙ LAPE	**Revú koncert symfonický: Úžasná harmonia a vzrušujúce zážok pre všetkých** Posledný večer bol pre mne nepopírateľným zážitkom, ke jý si pamätám na dlžšie. Mám tu št'astnou príležitosť pozývať na koncert symfonického orchestru v nášho meste, a musím povedať, že to bolo úžasné prežívanie umu a hudby...

Table 14: Sample outputs for steering from French to Slovak. Methods highlighted in green successfully produce Slovak, and red indicates failure to switch from the source language.

Method	Generated Output
	Input (Italian): Struttura un copione per un episodio di podcast che discuta l'influenza delle piattaforme di streaming sull'industria musicale. Target: Romanian
\mathcal{E} Baseline-I	Titlu: "Revoluția Streamingului: Cum modelează platformele digitală viitorul muzicii" Introducere (0:00 - 2:00) **Gazda:** Bună și bine ați venit la un nou episod al podcast-ului nostru, unde explorăm impactul tehnologiei asupra artei...
\mathcal{E} Baseline-II	**Titlu: Revoluția Muzicală: Cum Platformele de Streaming Redescresc Industria ** **Introducere (0:00 - 2:00)** **Prezentator:** Bună și bine ați venit la un nou episod al podcast-ului nostru, unde explorăm impactul tehnologiei asupra artei...
$\vec{\Delta}$ DiffMean	**Titlu: Revoluția Streamingului: Cum Platformele Online Redefinesc Industria Muzicală ** **Introducere (0:00 - 2:00)** - **Prezentator (P):** Bună și bine ați venit la un nou episod al podcast-ului nostru, unde explorăm impactul tehnologiei asupra artei și culturii...
w Probe	**Titolo: La Rivoluzione dello Streaming: Come le Piattaforme Digitali Stanno Rimodellando la Musica ** **Introduzione:** **Narratore:** Benvenuti in un nuovo episodio del nostro podcast, dove esploriamo l'impatto della tecnologia sulla musica...
u PCA	Titolo: "La Rivoluzione dello Streaming: Come le Piattaforme Online stanno Trasformando la Musica" Introduzione (0:00 - 2:00) **Presentatore:** Benvenuti a un nuovo episodio del nostro podcast, dove esploriamo l'impatto della tecnologia sulla musica...
v LDA	Titolo: "La Rivoluzione dello Streaming: Come le Piattaforme Online stanno Trasformando la Musica" Introduzione (0:00 - 2:00) **Presentatore:** Benvenuti a un nuovo episodio del nostro podcast, dove esploriamo l'impatto della tecnologia sulla musica...
⊙ LAPE	Titlu: "Revoluția Streamingului: Cum modelează platformele digitală viitorul muzicii" Introducere (0:00 - 2:00) **Gazda:** Bună și bine ați venit la un nou episod al podcast-ului nostru, unde explorăm impactul tehnologiei asupra artei...

Table 15: Sample outputs for steering from Italian to Romanian. Methods highlighted in **green** successfully produce Romanian, and **red** indicates failure to switch from the source language.