LangFIR: Discovering Sparse Language-Specific Features from Monolingual Data for Language Steering

Sing Hieng Wong[1], **Hassan Sajjad**[2], **A.B. Siddique**[1]

[1]Department of Computer Science, University of Kentucky, USA
[2]Department of Computer Science, Dalhousie University, Canada
Correspondence: singhieng.wong@uky.edu

Abstract

Large language models (LLMs) show strong multilingual capabilities, yet reliably controlling the language of their outputs remains difficult. Representation-level steering addresses this by adding language-specific vectors to model activations at inference time, but identifying language-specific directions in the residual stream often relies on multilingual or parallel data that can be expensive to obtain. Sparse autoencoders (SAEs) decompose residual activations into interpretable, sparse feature directions and offer a natural basis for this search, yet existing SAE-based approaches face the same data constraint. We introduce **LangFIR** (**Lang**uage **F**eature **I**dentification via **R**andom-token Filtering), a method that discovers language-specific SAE features using only a small amount of monolingual data and random-token sequences. Many SAE features consistently activated by target-language inputs do not encode language identity. Random-token sequences surface these language-agnostic features, allowing LangFIR to filter them out and isolate a sparse set of language-specific features. We show that these features are extremely sparse, highly selective for their target language, and causally important: directional ablation increases cross-entropy loss only for the corresponding language. Using these features to construct steering vectors for the multilingual generation control task, LangFIR achieves the best average accuracy $\times$ BLEU among steering methods across three models (Gemma 3 1B, Gemma 3 4B, and Llama 3.1 8B), three datasets, and twelve target languages, outperforming the strongest monolingual baseline by up to 4.7$\times$ and surpassing methods that use parallel data. Our results suggest that language identity in multilingual LLMs is localized in a sparse set of feature directions discoverable with monolingual data. Code is available at https://github.com/JaMussCraft/LangFIR.

1 Introduction

Large language models (LLMs) exhibit impressive capabilities across text generation, instruction following, and reasoning (Grattafiori et al., 2024; Kamath et al., 2025). Considerable effort has also been devoted to improving their multilingual abilities, motivated by the growing need to deploy them across diverse linguistic settings. While multilingual LLMs can generate text in many languages, reliably controlling which language they output (a task known as multilingual generation control) remains a challenge (Marchisio et al., 2024; Gurgurov et al., 2026). Representation-level steering provides a training-free alternative that enables direct intervention on model activations at inference time, while also offering insights into how models encode language information.

Recent representation-level approaches to multilingual control locate language-specific structure in model activations in different ways. Neuron-based methods (Tang et al., 2024; Kojima et al., 2024) detect language-discriminative neurons, but require logging activations over millions of tokens across multiple languages. Sparse autoencoders (SAEs) decompose activations into sparse, interpretable feature directions, providing a natural basis for identifying language-specific directions. Yet SAE-based approaches (Deng et al., 2025; Chou et al., 2025) similarly rely on multilingual data to identify language-specific features.

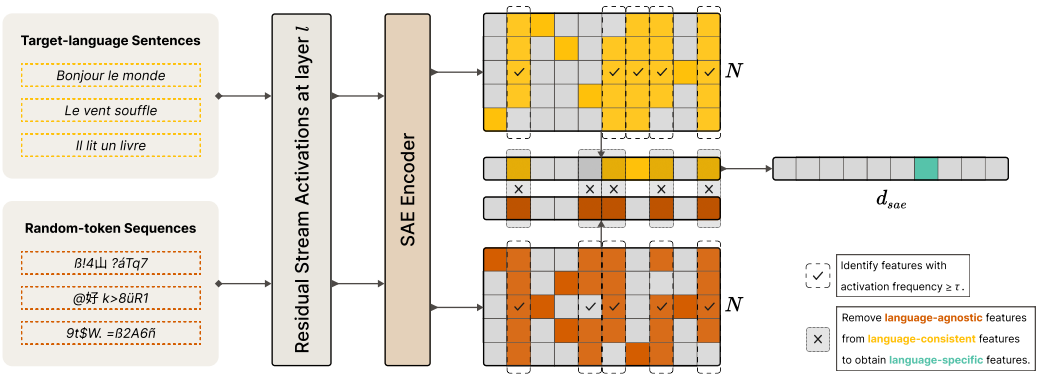


Figure 1: Overview of LangFIR. (1) For each target-language sentence, a random-token sequence is generated by sampling from the tokenizer's vocabulary. (2) Residual-stream activations are extracted at a layer l and encoded through SAE. (3) Sample-wise filtering is performed with activation frequency threshold τ to identify features that activate frequently for target-language sentences (language-consistent) and for random-token sequences (language-agnostic). (4) Removing language-agnostic features from language-consistent features yields a sparse set of language-specific features.

More recently, Zhong et al. (2025) propose a data-efficient method, but its stronger variant still depends on parallel sentence pairs. A key limitation of most existing approaches is their reliance on multilingual data, often parallel corpora, which is restrictive in domains where such resources are unavailable or costly to curate. Identifying language-specific features from small amounts of monolingual data would therefore substantially broaden the applicability of inference-time language control.

We observe that when encoding target-language text through an SAE, many highly activated features are not language-specific: they activate across all languages, responding to non-linguistic tokens such as punctuation patterns and whitespace. We hypothesize that random-token sequences, which lack coherent linguistic content, provide a simple proxy for surfacing such language-agnostic features. Removing them from the features activated consistently by target-language text should leave features that genuinely encode language identity.

Building on this intuition, we introduce **LangFIR** (**Lang**uage **F**eature **I**dentification via **R**andom-token Filtering). Our method collects a set of target-language sentences and generates a set of random-token sequences, encoding their residual-stream activations using an SAE. It then identifies features that activate frequently for the target language while filtering out those that also activate frequently for random tokens. This yields a sparse set of language-specific features.

We show these features activate selectively for their target language, are causally important, as ablating them substantially increases cross-entropy loss only for the target language, and concentrate their effects most strongly in later model layers. We then use language-specific features to construct steering vectors for controlled multilingual generation. Across three models, three datasets, and twelve target languages, steering vectors constructed from these features achieve the best average accuracy $\times$ BLEU. They outperform the strongest monolingual baseline by up to 4.7$\times$ and the strongest overall baseline by up to 2.7 points.

This work makes the following contributions:

- We propose LangFIR, a monolingual-data-only method that identifies sparse language-specific SAE features via random-token filtering.
- We show these features are language-selective and causally important under directional ablation, with effects concentrated in later model layers.
- We demonstrate that steering vectors built from these features outperform strong baselines across models and datasets on the multilingual generation control task.

2 Related Work

2.1 Multilingual Representations in LLMs

Prior work has revealed a layer-wise progression in how LLMs process multilingual input. Wendler et al. (2024) find that in English-centric models such as Llama 2, early layers encode general lexical and semantic features, intermediate layers map representations into an English-aligned concept space, and later layers shift toward target-language token spaces. Dumas et al. (2025) provide further evidence for language-agnostic concept representations in intermediate layers, showing that language and concept information can be separated through activation patching. Similarly, Gurgurov et al. (2026) demonstrate that language-specific structure emerges predominantly in later layers and that steering directions cluster by language family. In a contrastive approach related to LangFIR, AlKhamissi et al. (2025) compare sentences against strings of non-words to localize residual-stream units selective for natural language in general, rather than for a specific target language.

2.2 Language Steering

A broad line of work explores controlling an LLM's behavior by manipulating its internal representations. Turner et al. (2023) and Rimsky et al. (2024) demonstrate that simple vector-based interventions can steer models toward desired behaviors such as improved truthfulness. Language steering applies this paradigm to multilingual generation, using representation interventions to control which language a model outputs. Gurgurov et al. (2026) propose CLaS-Bench, a standardized benchmark for evaluating multilingual language steering, finding that the simple DiffMean method on residual activations (Marks & Tegmark, 2023) consistently outperforms more complex approaches.

Neuron-based methods. Kojima et al. (2024) train binary classifiers on neuron activations to identify language-discriminative neurons and propose replacement-based manipulation. Tang et al. (2024) introduce Language Activation Probability Entropy (LAPE), which identifies language-specific neurons through entropy-based filtering and steers generation by selectively activating and deactivating them. However, neuron-based methods can be unreliable due to superposition (Elhage et al., 2022), where multiple concepts are encoded in a single neuron, and they typically require large-scale multilingual corpora.

Dimension-based methods. Zhong et al. (2025) identify sparse language-specific dimensions in the residual stream and steer multilingual generation by overriding activations in these dimensions. Their approach requires a small amount of monolingual or parallel data. While data-efficient, the method can be sensitive to hyperparameter choices and model architecture, and its stronger variant relies on parallel data.

SAE-based methods. SAEs have been used for uncovering interpretable, language-specific features. Deng et al. (2025) introduce a monolinguality metric for SAE features and show that ablating language-specific features selectively impairs performance in one language while leaving others unaffected. They also demonstrate that these features can enhance steering vectors. Chou et al. (2025) show that modulating a small number of SAE features in mid-to-late layers can steer generation toward target languages with high success rates. However, these methods typically require multilingual parallel corpora for feature identification. In contrast, our method removes this requirement by identifying language-specific SAE features using only monolingual data.

3 Preliminaries

Sparse Autoencoders. Given a residual-stream activation $\mathbf{x} \in \mathbb{R}^{d_{\text{res}}}$ at layer l, an SAE computes a sparse latent activation $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^{d_{\text{sae}}}$ and reconstructs the input as $\hat{\mathbf{x}}$:

$$\mathbf{f}(\mathbf{x}) := \text{ReLU}(W_{\text{enc}}\,\mathbf{x} + \mathbf{b}_{\text{enc}}), \qquad \hat{\mathbf{x}}(\mathbf{f}) := W_{\text{dec}}\,\mathbf{f} + \mathbf{b}_{\text{dec}}. \tag{1}$$

To encourage sparsity, Bricken et al. (2023) and Huben et al. (2023) incorporate an L_1 penalty on $\mathbf{f}(\cdot)$ in the training loss. An alternative is the Top-K SAE (Gao et al., 2024), which enforces

sparsity by retaining only the K largest activations. The JumpReLU variant (Rajamanoharan et al., 2024) replaces the standard ReLU with a learned threshold θ, setting $\text{JumpReLU}(z) = z \cdot 1[z > \theta]$. By projecting activations into a higher-dimensional latent space, SAEs yield features that are significantly more interpretable than individual neurons, capturing specific syntactic patterns (Huben et al., 2023), entity recognition signals (Ferrando et al., 2024), and other semantically coherent concepts (Paulo et al., 2024).

Following Rajamanoharan et al. (2024), we write $W_{\text{dec}} = [\mathbf{d}_1, \ldots, \mathbf{d}_{d_{\text{sae}}}]$, where each column $\mathbf{d}_j \in \mathbb{R}^{d_{\text{res}}}$ is a feature direction. We refer to each such direction as a *feature* throughout this paper. We use residual-stream activations at layer l as the input $\mathbf{x}$, as they are more interpretable than other internal states (Ferrando et al., 2024; Chanin et al., 2024). Most prior steering methods also operate on the residual stream (Gurgurov et al., 2026; Zhong et al., 2025), enabling a fair comparison.

4 LangFIR: Language Feature Identification via Random-Token Filtering

We observed that encoding target-language text through an SAE produces a set of highly activated features, many of which are shared across all languages. Inspection of these features using Neuronpedia (Lin, 2023) revealed that many do not increase the logits of target-language tokens. Instead, they appear to respond to language-universal properties (e.g., punctuation and whitespace) or non-linguistic symbol patterns (see examples in Appendix B). Since these features correlate with generic textual structure rather than language identity, we hypothesize that random-token sequences should activate them consistently. Filtering out such features from the set of features activated consistently by target-language text should leave a set of features that genuinely encodes language identity. The method proceeds in three stages.

Random-Token Sequence Generation. For each target language, we sample N monolingual sentences from a target-language corpus. For each sentence, we generate a random-token sequence with a similar number of tokens by sampling uniformly from the tokenizer's vocabulary (excluding special tokens, control characters, and reserved entries).

Activation Extraction and SAE Encoding. Both target-language sentences and random-token sequences are traced through the LLM's forward pass, and residual-stream activations are collected at layer l. For each sample i, we encode each token's residual activation $\mathbf{x}_{i,t} \in \mathbb{R}^{d_{\text{res}}}$ with the SAE and average the resulting per-token SAE activations across the sequence to obtain a sample-level SAE activation vector $\mathbf{f}_i \in \mathbb{R}^{d_{\text{sae}}}$. This gives per-sample SAE activation vectors for target-language and random-token sequences, $\mathbf{f}_i^{\text{lang}}$ and $\mathbf{f}_i^{\text{rand}}$. Stacking across N samples gives two latent activation matrices, $\mathbf{F}^{\text{lang}}, \mathbf{F}^{\text{rand}} \in \mathbb{R}^{N \times d_{\text{sae}}}$.

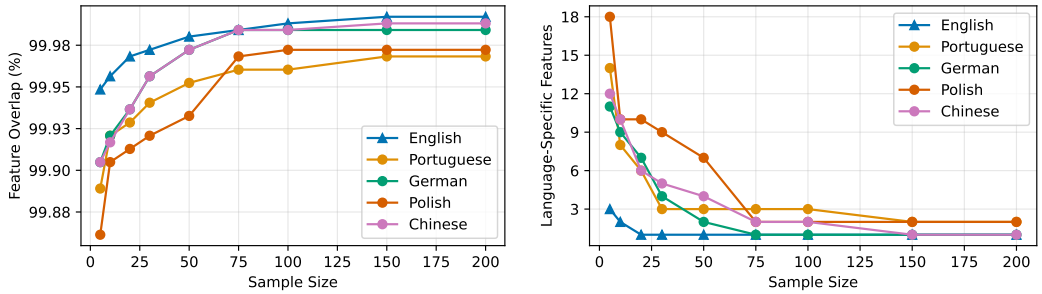
Sample-wise Filtering and Feature Overlap Removal. A feature j is considered high-frequency if it activates (i.e., has nonzero activation) for at least a fraction $\tau \in (0, 1]$ of samples. We define sample-level activation frequencies:

$$\rho_j^{\text{lang}} = \frac{1}{N} \sum_{i=1}^{N} 1\left[\mathbf{F}_{ij}^{\text{lang}} > 0\right], \qquad \rho_j^{\text{rand}} = \frac{1}{N} \sum_{i=1}^{N} 1\left[\mathbf{F}_{ij}^{\text{rand}} > 0\right]. \tag{2}$$

Thresholding these frequencies gives:

- *Language-consistent features*: consistently activated by target-language sentences, $S_{\text{lang}} = \{j : \rho_j^{\text{lang}} \geq \tau\}$.
- *Language-agnostic features*: consistently activated by random-token sequences, $S_{\text{rand}} = \{j : \rho_j^{\text{rand}} \geq \tau\}$.

The *language-specific* feature set, $S_{\text{spec}} = S_{\text{lang}} \setminus S_{\text{rand}}$, is the language-consistent features with language-agnostic features removed.



(a) Feature set overlap vs sample size. (b) Language-specific feature count vs sample size.

Figure 2: Analyses of language-specific feature identification across sample sizes on Llama 3.1 8B. (a) Overlap between language-consistent and language-agnostic feature sets approaches 100%. (b) Fewer than 5 language-specific features remain per language. Both quantities stabilize by $\sim$100 sample size.

5 Analysis of Language-Specific Features

We present analyses on Llama 3.1 8B (Grattafiori et al., 2024) using Llama Scope SAEs (He et al., 2024), with data sampled from FLORES+ (Goyal et al., 2022; Costa-Jussà et al., 2022), a high-quality multilingual parallel translation dataset covering over 200 languages. Unless otherwise noted, we use a default configuration of sample size $N = 100$, layer percentile $= 0.9$, and $\tau = 1.0$. Analyses focus on a set of five languages: English, Portuguese, German, Polish, and Chinese.

5.1 Sparsity and Stability of Feature Identification

Extreme sparsity and stability with minimal monolingual data. Figure 2 shows the two key outcomes of LangFIR's filtering stage across sample sizes. The overlap between the language-consistent and language-agnostic feature sets approaches 100% (Figure 2a), meaning that random-token filtering removes the vast majority of candidate features. What remains is an extremely sparse set of language-specific features, typically fewer than five per language (Figure 2b). Both quantities stabilize by around 100 sentences, providing a reliable identification signal. Convergence curves for the intermediate feature sets are provided in Figure 10.

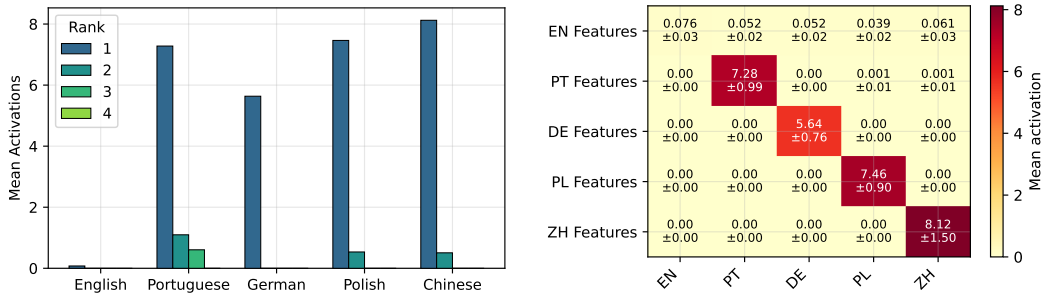
Stability with respect to threshold. The number of language-specific features stabilizes for $\tau \geq 0.8$ (Figure 7). Extremely high thresholds may exclude weaker language-specific features (Appendix F, Figure 13a), but the method is robust across a broad range of high threshold values.

English has few language-specific features. English consistently yields fewer language-specific features than other languages. We can attribute this to the English bias of multilingual LLMs (Wendler et al., 2024): general concepts may be entangled with English representations, making it harder for SAEs to isolate English-specific features.

5.2 Properties of Language-Specific Features

Dominance of the top feature. For all non-English languages, the rank-1 feature dominates in mean activation magnitude, the average activation over target-language samples (Figure 3a). English is the sole exception: its rank-1 feature activates at near-zero magnitude.

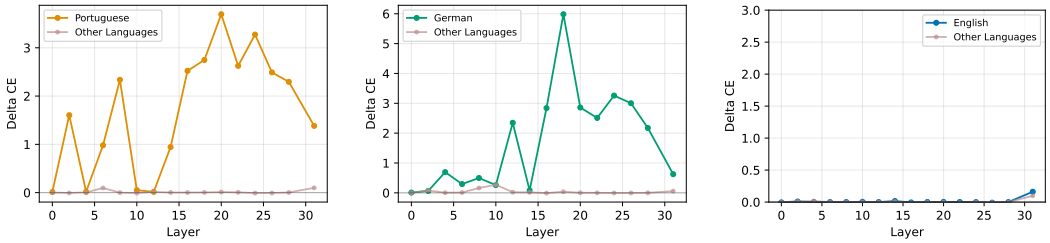
Language-selective activation. Given the dominance of the top-ranked feature, we visualize the mean activation of each language's rank-1 feature across texts in all five languages in Figure 3b. Each feature activates strongly only on its corresponding target language, with near-zero activation for all others. In contrast, features not identified as language-specific for any language activate broadly and uniformly across languages, with English showing a notably higher mean activation (Figure 8).



(a) Activations of top-ranked language-specific features.

(b) Language-specific feature activations across evaluation languages.

Figure 3: Analyses of language-specific feature properties on Llama 3.1 8B. (a) Rank-1 features dominate in mean activation magnitude. (b) Language-specific features activate selectively on their target language and remain near zero on others.



(a) Ablate Portuguese features.

(b) Ablate German features.

(c) Ablate English features.

Figure 4: Change in cross-entropy after ablating the top-2 language-specific features per layer on Llama 3.1 8B. Ablation increases loss selectively for the target language. English is a notable exception.

Late-layer activation dominance. Language-specific feature counts peak at the earliest and latest layers, with moderate counts in between (Figure 9a), consistent with the finding that multilingual LLMs shift from English-aligned intermediate representations to target-language spaces in later layers (Wendler et al., 2024). Moreover, late-layer features activate significantly more strongly than early-layer ones (Figure 9b), suggesting that language-specific information concentrates in later layers (Gurgurov et al., 2026).

5.3 Validating Language-Specific Features

Following Deng et al. (2025), we examine whether the identified features causally influence model behavior by using directional ablation (Arditi et al., 2024; Ferrando et al., 2024) to "zero out" language-specific features. Given a feature direction $\mathbf{d} \in \mathbb{R}^{d_{\text{res}}}$, directional ablation removes its contribution from the residual stream activation $\mathbf{x}$ via $\mathbf{x}' \leftarrow \mathbf{x} - \hat{\mathbf{d}}\,\hat{\mathbf{d}}^\top \mathbf{x}$, where $\hat{\mathbf{d}}$ is the unit vector of $\mathbf{d}$. For each target language and layer in Llama 3.1 8B, we ablate the top-2 language-specific features and measure changes in cross-entropy (CE) loss for texts in the target language versus texts in other languages.

The results reveal three key findings. First, ablating language-specific features substantially increases CE loss only for the target language, with negligible effects on other languages. This confirms that the features are causally important for language-specific processing. Second, changes in CE loss are generally larger in later layers, providing further evidence that language-specific information is concentrated in later model layers. Third, English is a notable exception: ablating its features produces near-zero change in CE across most layers, possibly because English-specific information is entangled with general concepts and thus

diffused across many SAE features, so that the few features our method identifies have minimal impact when ablated. See Appendix D for additional results on other languages.

6 Feature-Guided Language Steering

6.1 LangFIR-guided Steering

Recall that S_{spec} is the set of language-specific SAE feature indices identified in Section 4, and $\mathbf{f}_i^{\text{lang}} \in \mathbb{R}^{d_{\text{sae}}}$ is the SAE encoding of the i-th target-language sample. We first compute the mean SAE activation across all N target-language samples and select the top-k most dominant features:

$$\bar{\mathbf{f}}^{\text{lang}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{f}_i^{\text{lang}}, \qquad S_{\text{spec}}^{(k)} = \text{TopK}\left(\left\{\bar{\mathbf{f}}_j^{\text{lang}} : j \in S_{\text{spec}}\right\}, k\right). \tag{3}$$

We then form a sparse latent vector by masking all non-selected coordinates and decode it back into residual space:

$$\tilde{\mathbf{f}}_j^{\text{lang},(k)} = \begin{cases} \bar{\mathbf{f}}_j^{\text{lang}}, & j \in S_{\text{spec}}^{(k)}, \\ 0, & \text{otherwise}, \end{cases} \qquad \mathbf{v}_{\text{lang}}^{(k)} = W_{\text{dec}}\, \tilde{\mathbf{f}}^{\text{lang},(k)} = \sum_{j \in S_{\text{spec}}^{(k)}} \bar{\mathbf{f}}_j^{\text{lang}}\, \mathbf{d}_j. \tag{4}$$

The resulting steering vector is ℓ_2-normalized and added to the residual stream at a chosen layer with scaling coefficient α.

6.2 Experimental Setup

Task. Following Kojima et al. (2024), we adopt the multilingual generation control task. The model receives a prompt of the form "`Translate an English sentence into a target language. English: {source_text} Target language:`" without specifying the target language, isolating the steering method's ability to control output language.

Models. We evaluate three open-weight LLMs for which pretrained SAEs covering all layers are available: Gemma 3 1B and Gemma 3 4B (Kamath et al., 2025) using Gemma Scope 2 SAEs (Lieberum et al., 2024; McDougall et al., 2025), and Llama 3.1 8B (Grattafiori et al., 2024) using Llama Scope SAEs (He et al., 2024).

Data and languages. We evaluate twelve English-to-X directions across three multilingual datasets: FLORES+ (Goyal et al., 2022; Costa-Jussà et al., 2022), WikiMatrix (Schwenk et al., 2021), and Tatoeba (Tiedemann, 2020). The target languages span twelve languages from multiple language families: Spanish (es), French (fr), German (de), Italian (it), Indonesian (id), Portuguese (pt), Vietnamese (vi), Polish (pl), Czech (cs), Turkish (tr), Chinese (zh), and Dutch (nl). Data is partitioned into three disjoint sets for steering vector construction, hyperparameter tuning, and final evaluation. More details are provided in Appendix A.2.

Evaluation metrics. We use three metrics. **Accuracy (ACC)** is the fraction of outputs in the target language as determined by fastText (Joulin et al., 2017) with confidence threshold 0.5. **BLEU** measures translation quality on successful samples only. **ACC** $\times$ **BLEU** is a composite score reflecting both success rate and quality.

Baselines. We compare against seven baselines. Five use parallel multilingual sentence pairs: LDA (Balakrishnama & Ganapathiraju, 1998), Zhong's Parallel (Zhong et al., 2025), SAE-DiffMean (Gurgurov et al., 2026), Gated-DiffMean (Deng et al., 2025), and DiffMean (Rimsky et al., 2024; Marks & Tegmark, 2023). Two are monolingual: PCA (Abdi & Williams, 2010) and Zhong's Monolingual. All methods produce a steering vector that is normalized and added to the residual stream at a chosen layer with a scaling coefficient α controlling the intervention strength. We also report two non-steering references: No Steering and Prompt Only. Details of all methods and references are provided in Appendix A.4.

Table 1: ACC×BLEU averaged over twelve target languages and three datasets (FLORES+, WikiMatrix, Tatoeba). The best result among steering methods within each model is **bolded**.

	Non-steering		Steering							
Model	No Steering	Prompt Only	PCA	LDA	Zhong (M)	Zhong (P)	SAE-DM	Gated-DM	DiffMean	LangFIR
Gemma 3 1B	0.8	22.2	1.6	2.0	3.3	9.5	14.7	9.0	15.3	**15.5**
Gemma 3 4B	2.6	35.5	3.3	15.2	10.5	13.9	23.2	10.8	23.9	**26.6**
Llama 3.1 8B	1.5	34.7	2.9	22.7	6.6	7.6	4.7	22.6	22.2	**23.4**
Avg	1.6	30.8	2.6	13.3	6.8	10.3	14.2	14.1	20.5	**21.8**

6.3 Results

We focus on ACC×BLEU because a method is only useful if it succeeds on both axes at once: producing the target language (ACC) while preserving translation quality (BLEU). Table 1 reports average ACC×BLEU over twelve languages and three datasets. Per-language scores and the disaggregated ACC and BLEU tables are in Appendix E. Among steering methods, LangFIR achieves the highest average ACC×BLEU on all three models (15.5, 26.6, and 23.4 on Gemma 3 1B, Gemma 3 4B, and Llama 3.1 8B, respectively), outperforming the next-best method, DiffMean (15.3, 23.9, and 22.2). The gain is largest on Gemma 3 4B, where LangFIR leads DiffMean by 2.7 points.

Monolingual advantage. Despite requiring only monolingual target-language data, LangFIR slightly outperforms the strongest multilingual baseline (DiffMean) on average across all three models. Among the other monolingual methods (PCA, Zhong's Monolingual), none exceeds 10.5 average ACC×BLEU on any model, while LangFIR achieves 15.5–26.6. On Gemma 3 1B, LangFIR surpasses the strongest monolingual baseline (Zhong's Monolingual) by 4.7×. LangFIR's effectiveness likely stems from SAE features being more disentangled by construction, which gives it a cleaner representational basis for isolating language-specific directions.

Robustness across datasets. Per-dataset results show that LangFIR maintains its advantage across FLORES+, WikiMatrix, and Tatoeba (Tables 9–11). The occasional weaker performance on certain languages can be attributed to the quality of the pretrained SAE and to hyperparameter tuning on a small subset of four languages, which may not optimally transfer to all twelve target languages.

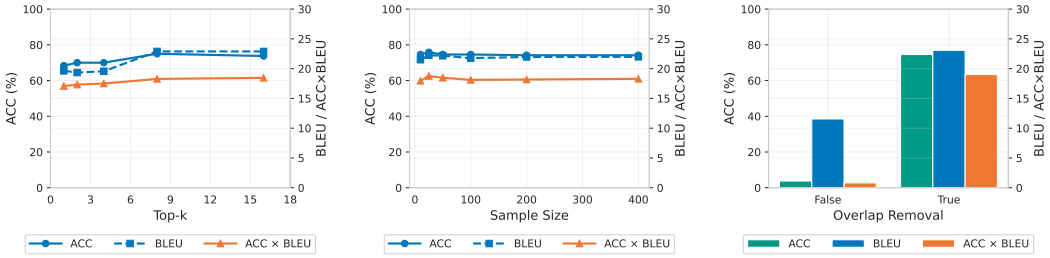
DiffMean-family baselines. Consistent with Gurgurov et al. (2026), DiffMean on residual activations remains a strong baseline, outperforming both LDA and SAE-DiffMean on Gemma models. Gated-DiffMean (Deng et al., 2025) is competitive on Llama 3.1 8B but significantly weaker on the Gemma models.

Prompt Only. Prompt Only attains the highest ACC×BLEU on this task (22.2, 35.5, and 34.7). However, this advantage does not generalize: on the cross-lingual answer generation task (Section 6.5), Prompt Only underperforms the steering methods. Moreover, Prompt Only and steering operate at fundamentally different levels: the former relies on the model interpreting a language directive from the prompt, while the latter acts directly on residual-stream activations regardless of instruction-following behavior. We therefore treat Prompt Only as an informative reference rather than a competing baseline.

6.4 Ablation Study

We ablate the following aspects of LangFIR on Llama 3.1 8B: number of top-k features, sample size, feature-overlap removal, intervention layer, scaling coefficient, frequency threshold, random-token sequence length, random seed, and filter negative set. Each factor is varied independently while holding all others at their validation-selected values. Results on Gemma models are generally consistent. We highlight the most important findings below and report all setup details, additional ablations, and Gemma results in Appendix F.

Number of features. Figure 5a shows that even at $k = 1$, all three metrics are already high, consistent with the top-feature dominance observed in Section 5. Including additional language-specific features yields modest further improvement.



(a) Top-k features. (b) Sample size N. (c) Feature overlap removal.

Figure 5: Ablation of number of top-k language-specific features, sample size, and feature-overlap removal on Llama 3.1 8B. (a) Even k=1 achieves strong performance. (b) Steering metrics are stable with as few as 10 sentences. (c) Removing feature-overlap removal degrades all steering metrics dramatically.

Table 2: Cross-lingual answer generation results: F1 averaged over five target languages (es, de, vi, tr, zh). The best result among steering methods within each model is **bolded**.

	Non-steering		Steering							
Model	No Steering	Prompt Only	PCA	LDA	Zhong (M)	Zhong (P)	SAE-DM	Gated-DM	DiffMean	LangFIR
Gemma 3 1B	12.7	12.3	10.6	12.7	11.1	15.1	13.3	13.6	14.4	**16.7**
Gemma 3 4B	14.6	7.9	15.6	14.7	11.3	13.3	**22.5**	19.3	21.7	20.0
Llama 3.1 8B	12.3	12.2	11.6	11.8	10.6	11.2	11.1	20.8	20.7	**21.6**
Avg	13.2	10.8	12.6	13.0	11.0	13.2	15.6	17.9	18.9	**19.4**

Sample size. Figure 5b shows that all three metrics remain stable across sample sizes. LangFIR achieves comparable performance with as few as 10 monolingual sentences, demonstrating that the method is extremely data-efficient.

Feature overlap removal. The feature-overlap-removal step is essential to LangFIR. Omitting it (using S_{lang} directly instead of $S_{\text{spec}} = S_{\text{lang}} \setminus S_{\text{rand}}$) degrades ACC by approximately 18×, BLEU by 2×, and ACC×BLEU by 19×, even after retuning hyperparameters (Figure 5c). Without this step, the steering vector is dominated by features unrelated to language identity.

6.5 Cross-Lingual Answer Generation

To test whether LangFIR generalizes beyond sentence translation, we evaluate it on an additional cross-lingual answer generation task using XQuAD (Artetxe et al., 2020), covering the five languages that are both in the multilingual generation control experiment and available in XQuAD (es, de, vi, tr, zh). Given an English context and question, the model must generate the answer in the target language. The setup follows our multilingual generation control task, with steering vectors constructed from FLORES+. The prompt is "`Read the following passage and answer the question [in {target language}]. Context: ... Question: ... Answer:`", where the bracketed clause is used only for Prompt Only. We focus on F1, since exact match (EM) may penalize valid paraphrases. The full setup, EM, and per-language results are in Appendix G.

On cross-lingual answer generation, Table 2 shows that the picture inverts relative to sentence translation: all steering methods beat Prompt Only on average (11.0–19.4 vs. 10.8). Among steering methods, LangFIR is the strongest on average despite being monolingual (19.4), slightly ahead of DiffMean (18.9). We speculate that Prompt Only performs poorly here because the large English context creates stronger English-generation momentum than the sentence-translation task does, a momentum that Prompt Only does not reliably overcome. Inspecting Prompt Only's generated outputs confirms this: it almost always generates in English rather than the target language. Activation-level steering methods, by

operating directly on the residual stream rather than relying on the model to interpret a language directive mid-context, appear better suited to overcoming this momentum.

7 Conclusion

We introduced LangFIR, a method for identifying sparse language-specific SAE features using minimal monolingual data and random-token sequences. The key insight is that random-token sequences serve as a simple proxy for surfacing language-agnostic features, and removing them from the set of language-consistent features isolates a clean, sparse set of features that encode language identity. Our analyses confirm that the identified features are language-selective, causally meaningful under directional ablation, and concentrated in later model layers. Steering vectors built from LangFIR features consistently outperform strong residual-stream baselines across three models, three datasets, and twelve target languages on the multilingual generation control task, and the advantage carries over to cross-lingual answer generation on XQuAD.

These results show that language steering can be achieved without parallel or multilingual data, enabling steering in settings where such data is unavailable, and that language identity in multilingual LLMs is concentrated in a small number of sparse, interpretable, and manipulable feature directions. Future work could extend the underlying framework, contrasting a target-concept corpus against a concept-agnostic proxy, to non-language concepts. A deeper analysis of how language-specific features organize across language families and model layers could further illuminate how language information is structured in multilingual models.

Limitations

Our method depends on the availability and quality of pretrained SAEs, which restricts our evaluation to three models, all under 10B parameters and from two English-centric model families (Gemma and Llama). The languages we evaluate are also primarily mid- to high-resource. Whether our findings hold for larger models, for models pretrained with a more balanced language mix, or for low-resource languages remains open.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under Grant IIS-2401685 and by the Humanities and AI Virtual Institute (HAVI) program of Schmidt Sciences. This work used the Delta system at the National Center for Supercomputing Applications [award OAC 2005572] through allocation CIS250532 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program (Boerner et al., 2023), which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. We also thank the National Deep Inference Fabric (NDIF) (Fiotto-Kaufman et al., 2025) for providing remote inference access through its pilot program.

LLM Usage Disclosure

We disclose that LLMs were used for editing and polishing the language of this manuscript. All scientific content, experimental design, and analysis were conducted by the authors.

References

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The llm language network: A neuroscientific approach for identifying causally task-relevant units. In

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 10887–10911, 2025.

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

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4623–4637, 2020.

Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8, 1998.

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. Saelens. https://github.com/decoderesearch/SAELens, 2024.

Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, PEARC '23, pp. 173–176, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399852. doi: 10.1145/3569951.3597559. URL https://doi.org/10.1145/3569951.3597559.

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, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Cheng-Ting Chou, George Liu, Jessica Sun, Cole Blondin, Kevin Zhu, Vasu Sharma, and Sean O'Brien. Causal language control in multilingual transformers via sparse feature steering. *arXiv preprint arXiv:2507.13410*, 2025.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. Unveiling language-specific features in large language models via sparse autoencoders. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4563–4608, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.229. URL https://aclanthology.org/2025.acl-long.229/.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31822–31841, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1536. URL https://aclanthology.org/2025.acl-long.1536/.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. *arXiv preprint arXiv:2411.14257*, 2024.

Jaden Fried Fiotto-Kaufman, Alexander Russell Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla E. Brodley, Arjun Guha, Jonathan Bell, Byron C Wallace, and David Bau. NNsight and NDIF: Democratizing access to foundation model internals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MxbEiFRf39.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daniil Gurgurov, Yusser Al Ghussin, Tanja Baeumel, Cheng-Ting Chou, Patrick Schramowski, Marius Mosbach, Josef van Genabith, and Simon Ostermann. Clas-bench: A cross-lingual alignment and steering benchmark. *arXiv preprint arXiv:2601.08331*, 2026.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers*, pp. 427–431, 2017.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Istvan Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, Andr'as Gyorgy, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Boxi Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, Cj Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga,

Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, J. Michael Wieting, Jonathan Lai, Jordi Orbay, Joe Fernandez, Joshua Newlan, Junsong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Ping mei Xu, Piotr Stańczyk, Pouya Dehghani Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Ardeshir Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vladimir Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab S. Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and L'eonard Hussenot. Gemma 3 technical report. *ArXiv*, abs/2503.19786, 2025. URL https://api.semanticscholar.org/CorpusID:277313563.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 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)*, pp. 6919–6971, 2024.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Johnny Lin. Neuronpedia: Interactive reference and tooling for analyzing neural networks, 2023. URL https://www.neuronpedia.org. Software available from neuronpedia.org.

Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. Understanding and mitigating language confusion in llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6653–6677, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Callum Stuart McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, Neel Nanda, and Google. Gemma scope 2 technical paper. Technical report, Google, September 2025. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/gemma-scope-2-helping-the-ai-safety-community-deepen-understanding-of-complex-language-model-behavior/Gemma_Scope_2_Technical_Paper.pdf.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*, 2024.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

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

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1351–1361, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.115. URL https://aclanthology.org/2021.eacl-main.115/.

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

Jörg Tiedemann. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri (eds.), *Proceedings of the Fifth Conference on Machine Translation*, pp. 1174–1182, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.wmt-1.139. URL https://aclanthology.org/2020.wmt-1.139/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Yugo Murawaki, Chenhui Chu, and Sadao Kurohashi. Language lives in sparse dimensions: Toward interpretable and efficient multilingual control for large language models. *arXiv preprint arXiv:2510.07213*, 2025.

A Experimental Details

A.1 Models and SAEs

We experiment with three pretrained language models paired with publicly available SAEs that cover all residual-stream layers. All SAEs are loaded from the SAELens library (Bloom et al., 2024). All model interventions are implemented using NNsight (Fiotto-Kaufman et al., 2025). Table 3 summarizes the configurations.

Table 3: Models and SAEs used in our experiments.

Model ID	SAE Release	SAE ID
google/gemma-3-1b-pt	gemma-scope-2-1b-pt-res-all	layer_{layer}_width_16k_l0_small
google/gemma-3-4b-pt	gemma-scope-2-4b-pt-res-all	layer_{layer}_width_16k_l0_small
meta-llama/Llama-3.1-8B	llama_scope_lxr_8x	l{layer}r_8x

A.2 Datasets and Preprocessing

We evaluate on three multilingual datasets. **FLORES+** (Goyal et al., 2022; Costa-Jussà et al., 2022) is a high-quality parallel translation dataset covering 200+ languages. **WikiMatrix** (Schwenk et al., 2021) is a Wikipedia-mined corpus providing aligned sentence pairs across 1,600+ language pairs. **Tatoeba** (Tiedemann, 2020) is a community-contributed multi-parallel corpus of short translated sentences spanning 400+ languages.

For Tatoeba, the validation split is used for vector construction and the test split for evaluation. For WikiMatrix, the train split is used for vector construction and the dev split for evaluation. For FLORES+, the dev split is used for construction and disjoint indices from the devtest split are used for evaluation.

Sentences are filtered by length (30 to 300 characters). For each language, 100 parallel sentence pairs are used to construct steering vectors, 20 samples form a validation set for hyperparameter tuning, and 100 test samples are used for final evaluation. All methods use the same splits.

A.3 Steering Setup

All steering interventions are applied at the last token position of each generation step. We use greedy decoding (no sampling) for all experiments.

A.4 Baselines and Non-Steering References

All addition-based methods normalize their steering vectors and add them to the residual stream with a scaling factor α. Hyperparameter tuning details are provided in Appendix A.5.

PCA. Applies Principal Component Analysis (Abdi & Williams, 2010) to residual activations from the target language, projecting hidden states onto a low-dimensional principal subspace.

LDA. Uses Linear Discriminant Analysis (Balakrishnama & Ganapathiraju, 1998) to find a direction maximizing separation between target and non-target language representations.

Zhong's Monolingual. Identifies language-specific dimensions by comparing target-language representations between an intermediate and the final layer, then overwrites these dimensions during inference (Zhong et al., 2025).

Zhong's Parallel. Identifies language-specific dimensions by comparing parallel English/target sentence pairs at the final layer, then overwrites these dimensions during inference (Zhong et al., 2025).

SAE-DiffMean. Extends DiffMean to the SAE latent space, computing the steering direction as the difference between mean latent activations of the target and source languages, then decoding back to model space.

Gated-DiffMean. Deng et al. (2025) extend DiffMean with conditional application: the steering vector is added only at generation steps where the source language's top-2 language-specific features are active. The original method intervenes on three layers. For fair comparison, we apply it at a single layer.

DiffMean. Computes a steering direction as the difference between average residual-stream activations of the target and source languages (Rimsky et al., 2024; Marks & Tegmark, 2023). This is the strongest baseline on CLaS-Bench.

We also report two non-steering references that apply no activation-level intervention. **No Steering** runs the base model on the task prompt as is. **Prompt Only** augments the task prompt with an explicit instruction to generate in the target language.

A.5 Hyperparameter Search Space

Hyperparameters are tuned as fairly as possible across methods using a subset of four languages (Chinese, Spanish, French, and German) from the twelve evaluated. Per dataset and model combination, a validation run is conducted. For all addition-based steering methods (PCA, LDA, SAE-DiffMean, Gated-DiffMean, DiffMean, LangFIR), we sweep over the same set of late layers, motivated by the finding that language-specific structure emerges predominantly in later layers (Gurgurov et al., 2026), and a model-specific set of intervention scaling coefficients α. For Zhong's replacement-based methods, we use the hyperparameter search space recommended by the original paper (Zhong et al., 2025). We select hyperparameters based on ACC$\times$BLEU averaged across the four languages on a 20-sample validation set. Table 4 lists the full search space for each method.

Table 4: Hyperparameter search space. Tuning is conducted on Chinese, Spanish, French, and German for every dataset and model combination.

Hyperparameter	Values
Intervention layer percentile	0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
Steering scalar α (Zhong Mono/Para)	0.2, 0.6, 1.0, 1.4, 1.8, 2.0
Steering scalar α (Other methods, Llama)	0.1, 0.5, 1, 2.5, 5, 10
Steering scalar α (Other methods, Gemma)	0.1, 1, 10, 100, 1000, 10000, 100000
PCA top-k components	1, 4, 8, 16, 32
Zhong top-k dimensions	200, 400, 600
Zhong anchor layer percentile	0.375, 0.5, 0.625
LangFIR activation threshold τ	0.8, 0.9, 1.0
LangFIR top-k SAE features	1, 2, 100 (effectively selects all features)

A.6 Validation-Tuned Hyperparameters

Tables 5–7 report the hyperparameters selected by validation for each method, dataset, and model.

Table 5: Selected hyperparameters for Gemma-3-1B.

Dataset	Method	Selected hyperparameters
FLORES+	PCA	layer 25, α=1000, top-k=1
	LDA	layer 22, α=1000
	Zhong (Mono)	layer 22, α=0.6, top-k=200, anchor=15
	Zhong (Para)	layer 20, α=0.6, top-k=200
	SAE-DM	layer 17, α=1000
	Gated-DM	layer 10, α=1000
	DiffMean	layer 17, α=1000
	LangFIR	layer 17, α=1000, τ=0.8, top-k=100
WikiMatrix	PCA	layer 20, α=10000, top-k=16
	LDA	layer 17, α=1000
	Zhong (Mono)	layer 20, α=0.6, top-k=200, anchor=9
	Zhong (Para)	layer 20, α=0.6, top-k=200
	SAE-DM	layer 17, α=1000
	Gated-DM	layer 10, α=1000
	DiffMean	layer 17, α=1000
	LangFIR	layer 17, α=1000, τ=0.9, top-k=100
Tatoeba	PCA	layer 22, α=10000, top-k=8
	LDA	layer 12, α=100
	Zhong (Mono)	layer 22, α=0.6, top-k=200, anchor=9
	Zhong (Para)	layer 20, α=0.6, top-k=200
	SAE-DM	layer 17, α=1000
	Gated-DM	layer 10, α=1000
	DiffMean	layer 17, α=1000
	LangFIR	layer 17, α=1000, τ=0.8, top-k=1

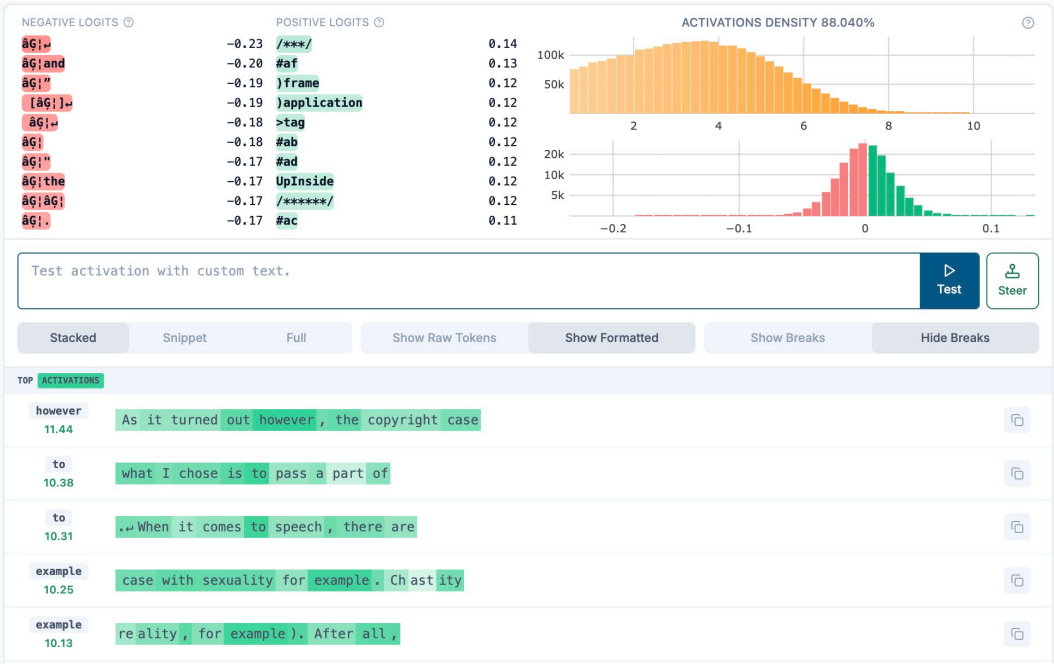
Table 6: Selected hyperparameters for Gemma-3-4B.

Dataset	Method	Selected hyperparameters
FLORES+	PCA	layer 29, α=100000, top-k=16
	LDA	layer 19, α=1000
	Zhong (Mono)	layer 29, α=0.6, top-k=200, anchor=12
	Zhong (Para)	layer 29, α=0.6, top-k=200
	SAE-DM	layer 29, α=10000
	Gated-DM	layer 13, α=10000
	DiffMean	layer 23, α=10000
	LangFIR	layer 19, α=1000, τ=0.8, top-k=100
WikiMatrix	PCA	layer 29, α=100000, top-k=8
	LDA	layer 19, α=1000
	Zhong (Mono)	layer 29, α=1.0, top-k=200, anchor=12
	Zhong (Para)	layer 26, α=0.6, top-k=200
	SAE-DM	layer 23, α=10000
	Gated-DM	layer 16, α=10000
	DiffMean	layer 23, α=10000
	LangFIR	layer 19, α=1000, τ=0.8, top-k=2
Tatoeba	PCA	layer 26, α=10000, top-k=16
	LDA	layer 13, α=100
	Zhong (Mono)	layer 29, α=1.0, top-k=200, anchor=16
	Zhong (Para)	layer 29, α=1.0, top-k=200
	SAE-DM	layer 13, α=10000
	Gated-DM	layer 16, α=10000
	DiffMean	layer 23, α=10000
	LangFIR	layer 19, α=1000, τ=0.9, top-k=2

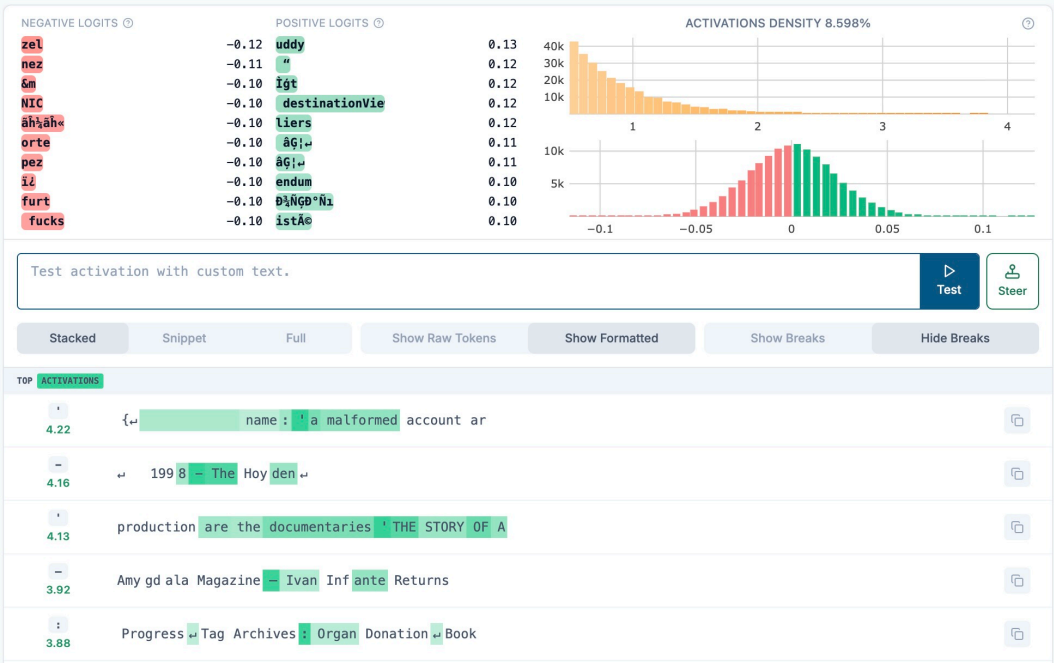
Table 7: Selected hyperparameters for Llama-3.1-8B.

Dataset	Method	Selected hyperparameters
FLORES+	PCA	layer 31, α=10, top-k=1
	LDA	layer 15, α=2.5
	Zhong (Mono)	layer 31, α=1.0, top-k=200, anchor=15
	Zhong (Para)	layer 31, α=1.0, top-k=200
	SAE-DM	layer 24, α=5
	Gated-DM	layer 21, α=10
	DiffMean	layer 18, α=10
	LangFIR	layer 18, α=5, τ=0.8, top-k=100
WikiMatrix	PCA	layer 31, α=10, top-k=16
	LDA	layer 12, α=2.5
	Zhong (Mono)	layer 27, α=0.2, top-k=400, anchor=19
	Zhong (Para)	layer 21, α=0.2, top-k=400
	SAE-DM	layer 31, α=10
	Gated-DM	layer 21, α=10
	DiffMean	layer 18, α=10
	LangFIR	layer 18, α=5, τ=0.8, top-k=1
Tatoeba	PCA	layer 31, α=10, top-k=16
	LDA	layer 15, α=2.5
	Zhong (Mono)	layer 21, α=0.2, top-k=200, anchor=19
	Zhong (Para)	layer 21, α=0.2, top-k=200
	SAE-DM	layer 31, α=10
	Gated-DM	layer 24, α=10
	DiffMean	layer 21, α=10
	LangFIR	layer 12, α=2.5, τ=1.0, top-k=1

B Examples of Language-Agnostic Features



(a) Feature 2618: activates near-indiscriminately at very high density (88.0%).



(b) Feature 3157: activates on surface patterns including formatted headings and punctuations.

Figure 6: Two language-agnostic features from Llama 3.1 8B at layer 18 using the `llamascope-res-32k` SAE. Panels show logit effects, activation density, and top activating examples from Neuronpedia (Lin, 2023). Neither increases target-language token logits, and both track generic textual structure rather than language identity. Feature pages: 2618, 3157.

C Analysis Details

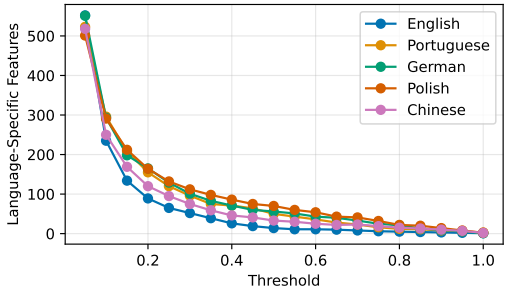


Figure 7: Language-specific feature count versus frequency threshold τ. Feature counts stabilize for $\tau \geq 0.8$.

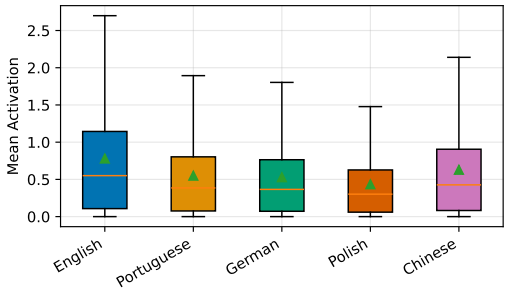
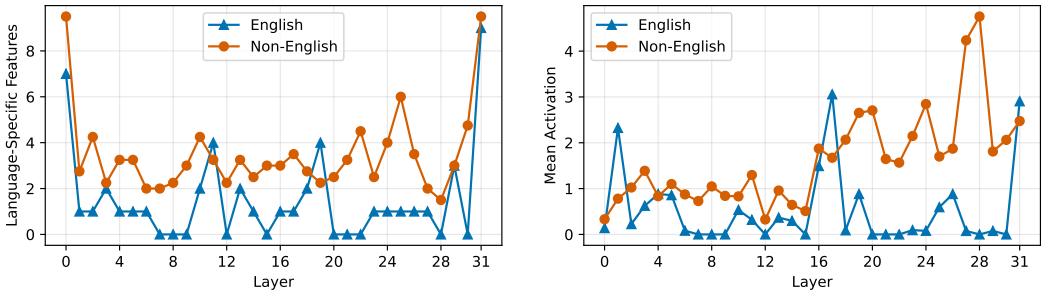


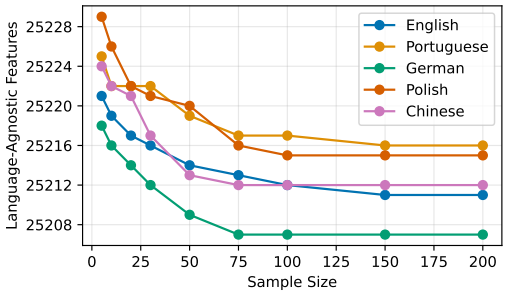
Figure 8: Non-language-specific feature activations across evaluation languages. These features activate broadly and uniformly across languages, with English slightly elevated.



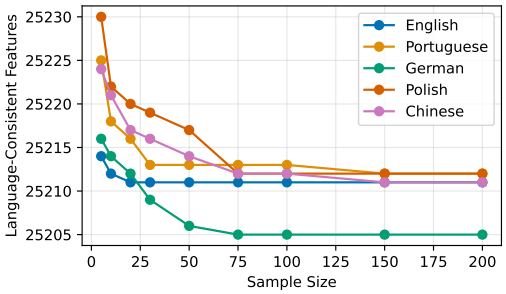
(a) Feature count across layers.

(b) Mean feature activation across layers.

Figure 9: Layer-wise analyses on Llama 3.1 8B. (a) Feature density dips in middle layers and recovers in later layers. (b) Activation magnitudes peak in later layers, indicating stronger language specialization near the output.



(a) Language-agnostic feature count vs sample size.



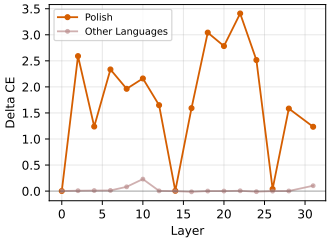
(b) Language-consistent feature count vs sample size.

Figure 10: Convergence of intermediate feature sets (before feature overlap removal) with increasing sample size on Llama 3.1 8B. (a) Language-agnostic features (identified from random-token sequences) and (b) language-consistent features (identified from target-language sentences) both stabilize by ∼100 sentences, confirming that LangFIR's inputs to the filtering step are themselves stable with limited data.

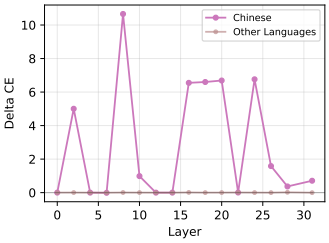
D Directional Ablation Details

Additional results. Figure 11 shows additional directional ablation results for Polish and Chinese on Llama 3.1 8B. Both languages exhibit similar patterns to Portuguese and German in Figure 4, though the late-layer effect is less pronounced for Chinese.

Setup. We use 100 evaluation samples per language from the FLORES+ validation split, disjoint from the 100 samples used for feature identification. When multiple features are ablated, their projections are subtracted from the residual stream sequentially.



(a) Ablate Polish features.



(b) Ablate Chinese features.

Figure 11: Change in cross-entropy after ablating the top-2 language-specific features per layer on Llama 3.1 8B. Consistent with Figure 4, ablation selectively increases loss for the target language.

E Results

This appendix provides the full per-language steering results of the multilingual generation control experiment, organized by metric. For each metric we report an average over the three datasets, followed by per-dataset breakdowns.

E.1 ACC×BLEU

Table 8 gives the per-language ACC×BLEU averaged over the three datasets, and Tables 9–11 give the per-dataset breakdowns. Per language, LangFIR achieves the best ACC×BLEU among steering methods on 7, 10, and 8 of 12 languages on Gemma 3 1B, Gemma 3 4B, and Llama 3.1 8B, respectively.

Table 8: ACC×BLEU averaged over FLORES+, WikiMatrix, and Tatoeba for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	0.9	1.1	4.9	0.8	0.1	0.3	0.4	0.2	0.4	0.5	0.2	0.3	0.8
	Prompt Only	20.3	23.7	23.8	27.0	29.1	36.8	25.6	14.7	15.9	7.8	15.5	26.4	22.2
	PCA	1.0	1.5	8.3	1.3	0.4	0.2	0.6	1.7	2.1	0.1	0.6	0.9	1.6
	LDA	3.2	5.0	8.3	1.8	1.1	1.7	0.3	0.2	0.9	0.7	0.1	1.1	2.0
	Zhong (Mono)	2.3	4.2	6.5	3.4	1.8	6.5	4.5	0.7	2.1	0.9	2.4	4.2	3.3
	Zhong (Para)	14.2	14.9	12.7	9.1	10.2	21.1	12.0	0.7	2.9	1.2	4.1	10.6	9.5
	SAE-DM	27.0	26.9	21.9	12.4	23.3	25.2	**17.8**	1.5	3.0	2.2	11.4	3.6	14.7
	Gated-DM	18.0	10.5	16.7	8.6	14.5	13.8	11.6	1.3	1.6	2.9	3.6	5.2	9.0
	DiffMean	**28.2**	28.3	22.0	13.9	**24.6**	**26.1**	17.4	1.5	2.0	4.4	**11.9**	3.8	15.3
	LangFIR	25.3	**30.4**	**22.8**	**16.8**	19.1	15.8	16.8	**3.6**	**6.0**	**4.8**	6.4	**17.8**	**15.5**
Gemma 3 4B	No Steering	14.2	6.6	2.0	1.1	0.9	0.8	0.5	0.5	1.2	0.3	2.2	0.5	2.6
	Prompt Only	41.1	46.9	35.3	42.2	34.4	50.6	36.1	25.5	28.6	19.4	27.7	37.9	35.5
	PCA	12.5	9.8	2.8	1.1	1.5	2.4	0.8	2.3	2.8	0.1	2.3	0.6	3.3
	LDA	31.5	27.7	21.0	18.5	17.3	25.3	17.4	2.3	4.7	1.8	5.2	9.8	15.2
	Zhong (Mono)	14.0	15.9	12.0	11.8	11.6	21.7	7.4	4.3	4.4	2.8	9.8	10.7	10.5
	Zhong (Para)	12.6	22.5	21.6	7.7	20.5	22.5	8.5	7.2	8.9	6.5	11.9	16.8	13.9
	SAE-DM	32.2	36.0	27.4	28.6	22.6	35.7	**29.0**	12.5	12.0	10.6	9.7	22.4	23.2
	Gated-DM	22.0	16.3	13.5	10.8	12.6	14.0	11.0	4.6	4.8	5.2	10.4	4.8	10.8
	DiffMean	32.5	38.9	27.7	32.8	23.3	38.9	26.9	14.0	**15.2**	8.8	6.3	21.5	23.9
	LangFIR	**38.3**	**40.9**	**30.7**	**33.1**	**26.0**	**40.3**	27.0	**14.9**	9.7	**11.9**	**16.4**	**30.5**	**26.6**
Llama 3.1 8B	No Steering	3.3	4.3	1.9	0.8	0.0	1.9	0.5	1.3	2.1	0.6	0.6	0.2	1.5
	Prompt Only	42.1	46.0	35.5	42.1	33.9	48.8	33.2	24.4	29.2	17.5	27.2	36.3	34.7
	PCA	5.1	7.1	3.7	1.9	0.2	5.0	0.7	2.3	3.8	2.6	1.1	1.2	2.9
	LDA	37.5	39.6	**30.9**	29.8	19.5	29.0	15.8	12.5	13.4	7.4	12.0	**25.0**	22.7
	Zhong (Mono)	9.4	14.0	3.7	5.2	1.3	12.5	4.5	4.6	4.0	1.4	8.2	10.5	6.6
	Zhong (Para)	10.8	14.4	6.6	5.7	1.1	13.0	3.4	4.6	3.6	1.5	7.1	19.2	7.6
	SAE-DM	9.2	12.4	7.7	2.3	1.2	9.2	0.7	3.0	4.5	3.0	2.1	1.2	4.7
	Gated-DM	33.2	36.0	28.1	26.9	16.7	**34.2**	**21.2**	6.8	12.8	5.6	25.6	24.5	22.6
	DiffMean	36.0	37.3	23.6	28.5	15.6	33.7	19.7	6.9	12.0	5.2	24.4	23.7	22.2
	LangFIR	**38.4**	**42.6**	30.8	**32.3**	**24.1**	30.6	4.5	**16.1**	**19.9**	**9.3**	**29.2**	2.7	**23.4**

Table 9: ACC×BLEU scores on FLORES+ for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Prompt Only	17.9	25.3	16.8	14.1	29.1	26.8	26.3	7.5	10.4	7.3	18.0	15.1	17.9
	PCA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LDA	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Zhong (Mono)	2.4	3.6	1.8	1.7	0.9	5.5	5.1	0.7	0.7	0.3	3.0	2.3	2.3
	Zhong (Para)	7.4	10.2	9.1	6.4	7.3	15.9	9.4	0.3	1.4	0.6	4.2	4.7	6.4
	SAE-DM	14.4	23.6	13.8	5.9	18.6	17.9	18.4	0.0	0.4	1.0	**14.3**	1.4	10.8
	Gated-DM	10.4	10.1	10.2	5.5	12.1	11.8	16.5	0.5	0.3	**2.2**	9.9	5.6	7.9
	DiffMean	15.8	23.9	13.8	7.0	**20.4**	18.5	**18.4**	0.0	0.4	1.2	14.2	1.5	11.3
	LangFIR	**15.8**	**25.2**	**16.2**	**12.2**	14.3	**19.2**	12.4	**3.3**	**3.1**	1.9	9.2	**10.1**	**11.9**
Gemma 3 4B	No Steering	11.0	5.3	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	1.1	0.0	1.5
	Prompt Only	25.4	41.5	31.4	27.8	37.2	45.0	38.2	17.5	25.6	17.6	34.3	22.4	30.3
	PCA	10.2	9.4	0.7	1.3	0.6	0.0	0.0	0.0	0.0	0.0	1.5	0.0	2.0
	LDA	18.2	25.8	21.9	15.1	21.1	27.7	25.6	0.7	4.0	0.6	8.5	6.2	14.6
	Zhong (Mono)	12.9	19.2	10.0	11.6	8.2	21.2	8.4	1.4	2.1	2.1	8.3	3.8	9.1
	Zhong (Para)	14.1	14.7	19.1	9.1	15.1	17.4	4.9	2.8	7.0	3.9	10.1	10.4	10.7
	SAE-DM	21.4	30.2	20.6	18.6	25.3	31.2	26.1	8.3	6.4	7.2	19.8	**14.9**	19.2
	Gated-DM	20.0	22.5	19.6	13.6	22.2	25.8	25.5	8.6	8.0	9.8	9.1	8.2	16.1
	DiffMean	22.6	35.2	20.3	**20.7**	26.5	**31.4**	24.0	**9.6**	**12.1**	8.2	4.9	10.7	18.8
	LangFIR	**23.9**	**36.7**	**24.7**	20.3	**35.9**	24.3	**36.3**	1.3	0.3	**10.2**	**27.2**	11.3	**21.0**
Llama 3.1 8B	No Steering	2.5	1.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.4
	Prompt Only	25.3	42.2	30.3	27.6	36.3	41.9	35.7	17.6	24.4	16.9	32.6	22.7	29.5
	PCA	3.2	2.7	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.6
	LDA	**24.0**	37.3	26.3	20.7	12.4	19.4	17.2	6.3	4.4	4.6	10.9	12.9	16.4
	Zhong (Mono)	3.9	4.0	1.0	1.2	1.6	5.5	12.3	1.2	2.2	0.2	23.6	1.6	4.9
	Zhong (Para)	4.6	4.8	1.1	0.3	2.0	3.4	8.8	1.9	2.2	0.5	20.5	1.4	4.3
	SAE-DM	4.8	5.6	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.9
	Gated-DM	20.7	34.3	21.7	18.1	24.4	25.8	**29.4**	3.5	11.2	5.0	**30.6**	14.0	19.9
	DiffMean	23.6	30.0	20.2	22.2	21.7	33.3	27.5	5.5	12.0	3.4	26.1	**14.7**	20.0
	LangFIR	23.7	**40.0**	**28.8**	**25.5**	**27.8**	**37.9**	11.8	**10.7**	**15.9**	**8.0**	30.2	5.2	**22.1**

Table 10: ACC×BLEU scores on WikiMatrix for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	2.6	2.4	9.4	2.3	0.4	0.8	0.6	0.6	1.2	1.5	0.6	0.9	1.9
	Prompt Only	34.4	34.1	29.6	28.8	32.3	39.0	25.4	17.4	22.2	8.2	15.3	33.1	26.6
	PCA	1.8	2.3	9.8	1.3	0.4	0.2	0.0	1.7	2.1	0.1	0.6	0.9	1.8
	LDA	1.3	2.1	3.7	2.1	1.4	0.1	0.0	0.2	0.9	0.7	0.1	1.7	1.2
	Zhong (Mono)	2.8	5.7	9.2	3.4	3.5	8.6	3.0	0.8	5.5	2.1	3.0	5.1	4.4
	Zhong (Para)	16.0	13.6	13.6	9.4	9.1	22.6	14.0	0.8	5.8	1.8	5.8	9.5	10.2
	SAE-DM	34.8	25.7	27.7	18.6	29.0	21.9	16.2	**2.4**	5.6	3.6	11.8	4.5	16.8
	Gated-DM	15.1	11.5	21.0	11.0	18.4	13.7	2.4	1.0	3.9	2.8	0.6	3.4	8.7
	DiffMean	**34.9**	26.9	**28.1**	18.7	**29.9**	22.3	15.7	**2.4**	5.6	4.5	**12.5**	4.8	17.2
	LangFIR	32.5	**30.5**	27.4	**23.4**	21.5	**28.0**	**17.8**	1.3	**11.4**	**4.8**	10.0	**20.4**	**19.1**
Gemma 3 4B	No Steering	11.7	8.2	5.0	2.5	2.7	1.8	0.6	1.5	3.7	1.0	2.9	1.4	3.6
	Prompt Only	45.0	48.4	37.5	45.3	37.0	53.4	36.6	30.7	34.6	19.8	28.6	48.2	38.8
	PCA	10.1	13.5	6.6	1.1	2.5	4.0	0.6	2.3	2.8	0.1	3.0	0.6	3.9
	LDA	36.0	31.2	18.9	26.0	21.7	38.9	17.6	5.9	9.4	4.3	3.3	22.2	19.6
	Zhong (Mono)	18.7	21.7	14.0	15.4	17.2	28.0	7.9	9.1	6.4	4.7	13.9	18.9	14.7
	Zhong (Para)	12.1	26.5	28.5	11.0	27.2	20.0	12.1	11.8	14.3	9.5	**19.7**	22.7	18.0
	SAE-DM	40.2	39.0	30.5	26.2	29.6	44.1	29.6	15.7	19.7	12.3	9.1	20.9	26.4
	Gated-DM	18.6	12.4	12.4	9.1	10.7	7.0	3.0	4.5	3.9	2.5	16.8	4.2	8.8
	DiffMean	38.2	37.8	30.7	**36.1**	29.1	47.1	**30.5**	17.0	**23.3**	10.6	10.8	24.1	27.9
	LangFIR	**40.6**	**40.1**	**31.9**	33.2	**31.8**	**50.7**	24.0	**22.7**	18.3	**15.4**	19.3	**42.2**	**30.9**
Llama 3.1 8B	No Steering	4.5	6.9	5.5	2.2	0.1	3.7	0.6	3.7	6.3	1.9	0.5	0.4	3.0
	Prompt Only	47.1	48.9	37.9	49.0	34.7	54.9	32.6	29.8	37.0	15.3	26.8	46.7	38.4
	PCA	5.7	10.4	6.5	3.9	0.2	4.6	0.6	3.6	7.4	2.6	1.1	1.2	4.0
	LDA	39.5	42.6	31.7	26.7	24.4	30.9	7.1	16.5	**25.5**	**13.7**	12.1	**36.0**	25.6
	Zhong (Mono)	10.6	13.8	9.3	3.1	2.1	10.9	0.6	4.7	9.3	2.6	0.3	1.9	5.8
	Zhong (Para)	11.0	19.6	11.0	8.0	1.0	17.4	0.6	6.1	8.2	2.6	0.6	28.7	9.6
	SAE-DM	9.7	15.6	11.4	3.0	1.7	9.0	0.6	4.8	8.4	3.0	2.0	1.5	5.9
	Gated-DM	37.7	39.7	**32.8**	28.0	12.8	38.7	**15.8**	10.2	20.3	10.4	**28.5**	34.8	25.8
	DiffMean	**44.2**	43.1	25.3	31.5	13.1	32.6	12.3	6.9	17.6	9.3	26.6	28.4	24.2
	LangFIR	42.5	**48.7**	32.2	**41.2**	**26.9**	**52.9**	1.0	**18.9**	24.0	10.6	28.2	0.2	**27.3**

Table 11: ACC×BLEU scores on Tatoeba for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	0.0	1.0	5.4	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.6
	Prompt Only	8.5	11.7	24.9	38.0	25.7	44.6	25.1	19.2	15.2	7.9	13.3	30.9	22.1
	PCA	0.1	0.8	6.8	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.7
	LDA	5.0	12.6	13.0	1.4	0.7	3.4	0.6	0.0	0.0	0.0	0.0	0.6	3.1
	Zhong (Mono)	1.7	3.4	8.6	5.1	1.1	5.3	5.4	0.5	0.1	0.2	1.1	5.1	3.1
	Zhong (Para)	19.2	20.9	15.4	11.4	14.1	24.7	12.5	1.1	1.6	1.1	2.3	17.7	11.8
	SAE-DM	32.0	31.3	24.2	12.8	22.2	35.8	18.9	0.7	0.0	2.0	8.1	4.8	16.1
	Gated-DM	28.6	9.8	19.1	9.2	13.0	15.9	16.0	2.3	0.6	3.8	0.3	6.5	10.4
	DiffMean	**33.9**	34.0	24.0	**16.1**	**23.5**	**37.4**	17.9	0.7	0.1	7.5	**9.1**	5.2	**17.4**
	LangFIR	27.7	**35.5**	**24.7**	14.9	21.4	0.3	**20.2**	**6.2**	**3.3**	**7.8**	0.1	**23.0**	15.4
Gemma 3 4B	No Steering	19.9	6.4	0.4	0.6	0.0	0.7	1.0	0.0	0.0	0.0	2.7	0.0	2.6
	Prompt Only	52.9	50.9	37.0	53.3	29.1	53.5	33.6	28.4	25.5	20.9	20.3	43.0	37.4
	PCA	17.4	6.5	1.0	0.8	0.0	0.7	1.0	0.0	0.0	0.0	2.4	0.0	2.5
	LDA	40.3	26.2	22.2	14.4	9.0	9.3	9.0	0.2	0.6	0.6	3.8	0.9	11.4
	Zhong (Mono)	10.3	6.8	11.9	8.4	9.6	16.0	5.8	2.5	4.7	1.5	**7.1**	9.3	7.8
	Zhong (Para)	11.6	26.4	17.2	3.2	**19.2**	30.1	8.6	7.0	5.4	6.1	5.8	17.3	13.2
	SAE-DM	35.0	38.7	31.0	40.8	12.8	31.7	**31.2**	13.4	9.9	**12.3**	0.0	31.5	24.0
	Gated-DM	27.3	13.9	8.4	9.7	4.8	9.3	4.3	0.5	2.6	3.4	5.2	2.1	7.6
	DiffMean	36.6	43.7	32.2	41.5	14.3	38.2	26.4	15.4	10.3	7.5	3.1	29.8	24.9
	LangFIR	**50.3**	**46.0**	**35.5**	**45.9**	10.3	**46.1**	20.7	**20.7**	**10.5**	10.1	2.6	**38.0**	**28.0**
Llama 3.1 8B	No Steering	2.9	4.7	0.1	0.2	0.0	2.0	0.8	0.2	0.1	0.0	1.2	0.0	1.0
	Prompt Only	53.9	46.7	38.2	49.7	30.8	49.7	31.3	25.7	26.2	20.3	22.2	39.6	36.2
	PCA	6.4	8.2	0.8	1.5	0.0	5.4	0.8	0.9	0.1	0.0	1.4	0.0	2.1
	LDA	49.0	38.9	**34.7**	**42.1**	**21.6**	36.7	**22.9**	14.7	**10.3**	**3.9**	12.9	26.3	**26.2**
	Zhong (Mono)	13.7	24.1	0.8	11.3	0.3	21.0	0.8	7.9	0.6	0.0	0.6	28.0	9.1
	Zhong (Para)	16.9	18.9	7.6	8.9	0.3	18.2	0.9	5.8	0.4	0.0	0.2	27.4	8.8
	SAE-DM	13.0	15.9	4.0	3.1	0.7	9.3	0.9	1.2	0.6	0.0	4.2	1.7	4.6
	Gated-DM	41.3	33.9	29.8	34.6	12.8	**38.1**	18.5	6.6	6.7	1.4	17.8	24.6	22.2
	DiffMean	40.1	38.9	25.2	31.8	12.0	35.3	19.4	8.4	6.3	2.8	**20.5**	**28.2**	22.4
	LangFIR	**49.1**	**39.0**	31.5	30.3	17.6	1.0	0.8	**18.8**	0.0	0.0	0.0	0.0	15.7

E.2 ACC

We analyze the dataset-averaged table (Table 12). The per-dataset tables (Tables 13–15) are mostly consistent with it. Prompt Only achieves substantially higher ACC than all steering methods across all models (82.4%, 94.0%, and 95.2% on Gemma 3 1B, Gemma 3 4B, and Llama 3.1 8B, respectively). Among steering methods, LangFIR achieves the best accuracy on Gemma 3 1B (62.2%), but trails DiffMean on Gemma 3 4B (72.2% vs. 85.3%) and Llama 3.1 8B (67.7% vs. 74.2%). PCA has the lowest ACC across all models (4.2%, 9.7%, and 5.8%).

Table 12: ACC scores (%) averaged over FLORES+, WikiMatrix, and Tatoeba for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	1.7	2.7	13.7	1.3	0.7	0.3	1.0	0.7	1.0	1.0	0.3	0.3	2.1
	Prompt Only	69.3	71.0	92.3	81.7	84.3	88.0	94.3	87.3	77.3	79.3	74.7	89.7	82.4
	PCA	3.0	3.0	24.0	3.0	2.0	1.0	2.0	4.0	4.0	2.0	1.0	1.0	4.2
	LDA	14.0	12.7	43.5	4.5	3.5	5.5	3.0	2.0	3.0	3.0	1.0	6.0	8.5
	Zhong (Mono)	35.0	33.7	51.7	19.3	14.0	37.7	35.7	23.3	12.7	14.3	30.3	39.0	28.9
	Zhong (Para)	67.3	72.7	73.0	54.0	51.3	**69.7**	69.0	19.3	**39.3**	24.3	37.0	61.0	53.2
	SAE-DM	84.7	77.0	83.7	45.0	75.7	68.3	75.7	5.5	5.5	22.7	59.0	9.3	51.0
	Gated-DM	64.7	34.7	80.3	34.0	56.3	49.0	47.3	11.7	4.7	26.3	27.3	25.7	38.5
	DiffMean	85.3	80.0	84.7	53.0	**79.0**	**69.7**	**77.3**	5.5	4.0	31.3	**60.7**	12.0	53.5
	LangFIR	**87.7**	**90.0**	**90.0**	**70.7**	59.3	51.3	70.0	**40.0**	29.0	**57.3**	31.0	**70.3**	**62.2**
Gemma 3 4B	No Steering	34.0	13.7	5.7	3.3	2.3	1.3	1.0	0.7	1.7	1.7	7.0	0.7	6.1
	Prompt Only	94.3	97.7	97.0	96.0	85.3	95.7	95.7	93.3	89.3	92.0	94.7	97.0	94.0
	PCA	40.0	28.0	10.7	5.0	3.0	4.0	1.5	5.0	5.0	3.0	8.7	2.0	9.7
	LDA	77.3	65.3	69.7	50.7	49.0	56.3	50.3	12.7	16.0	14.3	19.0	35.0	43.0
	Zhong (Mono)	45.3	47.7	48.7	43.7	36.0	58.0	38.7	22.0	17.3	34.0	53.3	38.3	40.2
	Zhong (Para)	54.3	62.3	87.0	53.7	68.3	57.0	55.0	40.3	40.0	55.7	60.7	72.7	58.9
	SAE-DM	94.7	91.0	**96.0**	87.0	78.3	87.0	92.7	68.7	49.3	80.7	**67.0**	82.0	81.2
	Gated-DM	61.7	44.0	45.0	34.0	40.3	35.7	37.7	25.3	16.7	33.7	57.7	20.3	37.7
	DiffMean	94.7	**95.3**	95.7	**93.0**	**88.0**	**91.7**	**96.0**	**81.0**	**66.3**	**84.0**	54.7	**83.0**	**85.3**
	LangFIR	**95.3**	92.3	92.0	85.0	62.0	85.3	72.3	55.3	30.0	67.3	50.3	79.3	72.2
Llama 3.1 8B	No Steering	8.3	8.0	4.3	1.3	0.3	2.7	1.0	3.0	3.0	1.0	2.0	0.7	3.0
	Prompt Only	96.3	98.7	97.0	98.0	86.0	96.7	96.3	98.0	96.0	96.3	89.0	94.3	95.2
	PCA	12.3	13.7	9.5	3.7	1.0	8.5	1.5	5.0	5.0	4.0	2.7	3.0	5.8
	LDA	87.3	90.7	88.3	71.3	53.0	64.0	48.0	54.3	43.0	49.7	34.0	64.0	62.3
	Zhong (Mono)	32.0	35.7	16.7	17.7	5.3	33.7	30.0	29.3	14.0	12.0	30.7	35.7	24.4
	Zhong (Para)	35.3	39.3	20.7	17.7	5.7	32.3	28.0	33.3	13.7	12.0	28.0	52.3	26.5
	SAE-DM	26.3	26.0	21.5	6.7	4.0	18.5	2.0	7.5	7.0	5.0	6.7	2.7	11.2
	Gated-DM	87.7	88.7	83.0	71.7	47.3	79.7	69.0	32.7	50.3	41.3	84.0	75.7	67.6
	DiffMean	90.3	91.0	86.3	81.7	50.7	**81.0**	**73.7**	47.7	57.3	**61.3**	**88.3**	**80.7**	**74.2**
	LangFIR	**95.3**	**94.3**	**94.0**	**86.0**	**65.0**	64.3	12.7	**70.3**	**75.0**	61.0	81.5	12.5	67.7

Table 13: ACC scores (%) on FLORES+ for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Prompt Only	92.0	93.0	87.0	76.0	90.0	86.0	95.0	83.0	80.0	76.0	78.0	90.0	85.5
	PCA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LDA	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	Zhong (Mono)	59.0	34.0	20.0	30.0	15.0	47.0	36.0	42.0	24.0	24.0	46.0	34.0	34.2
	Zhong (Para)	77.0	62.0	66.0	57.0	44.0	72.0	53.0	23.0	**41.0**	32.0	39.0	54.0	51.7
	SAE-DM	80.0	72.0	68.0	33.0	69.0	69.0	78.0	0.0	1.0	12.0	68.0	6.0	46.3
	Gated-DM	59.0	46.0	68.0	38.0	62.0	53.0	72.0	10.0	4.0	23.0	**79.0**	34.0	45.7
	DiffMean	82.0	74.0	69.0	39.0	**75.0**	71.0	**80.0**	0.0	1.0	21.0	68.0	11.0	49.2
	LangFIR	**90.0**	**89.0**	**84.0**	**66.0**	49.0	**77.0**	52.0	**43.0**	34.0	**40.0**	49.0	**69.0**	**61.8**
Gemma 3 4B	No Steering	44.0	13.0	2.0	3.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	5.4
	Prompt Only	99.0	99.0	100.0	100.0	92.0	100.0	98.0	96.0	98.0	91.0	99.0	100.0	97.7
	PCA	54.0	39.0	4.0	6.0	1.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	9.2
	LDA	86.0	77.0	83.0	63.0	56.0	78.0	80.0	13.0	22.0	13.0	28.0	44.0	53.6
	Zhong (Mono)	59.0	63.0	45.0	60.0	30.0	57.0	47.0	16.0	14.0	36.0	52.0	25.0	42.0
	Zhong (Para)	82.0	49.0	85.0	57.0	60.0	54.0	61.0	21.0	46.0	58.0	48.0	75.0	58.0
	SAE-DM	**99.0**	**97.0**	**99.0**	90.0	85.0	**94.0**	92.0	82.0	49.0	80.0	**88.0**	**93.0**	**87.3**
	Gated-DM	84.0	63.0	66.0	61.0	64.0	64.0	76.0	64.0	36.0	65.0	62.0	46.0	62.6
	DiffMean	**99.0**	**97.0**	95.0	**95.0**	89.0	89.0	96.0	**83.0**	**69.0**	**83.0**	42.0	82.0	84.9
	LangFIR	98.0	94.0	87.0	83.0	**90.0**	69.0	**97.0**	5.0	2.0	62.0	81.0	58.0	68.8
Llama 3.1 8B	No Steering	9.0	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.3
	Prompt Only	99.0	100.0	99.0	100.0	95.0	100.0	100.0	100.0	100.0	96.0	97.0	98.0	98.7
	PCA	12.0	5.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	1.7
	LDA	90.0	91.0	85.0	77.0	38.0	57.0	45.0	41.0	21.0	33.0	30.0	57.0	55.4
	Zhong (Mono)	47.0	32.0	22.0	19.0	10.0	32.0	87.0	44.0	28.0	19.0	89.0	31.0	38.3
	Zhong (Para)	54.0	33.0	11.0	17.0	11.0	28.0	81.0	57.0	29.0	19.0	82.0	21.0	36.9
	SAE-DM	20.0	14.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	3.3
	Gated-DM	98.0	98.0	84.0	81.0	74.0	91.0	95.0	40.0	67.0	63.0	**99.0**	91.0	81.7
	DiffMean	99.0	98.0	98.0	**100.0**	**81.0**	95.0	**98.0**	**71.0**	**83.0**	**86.0**	94.0	**95.0**	**91.5**
	LangFIR	**100.0**	**100.0**	**100.0**	96.0	80.0	**97.0**	34.0	67.0	73.0	57.0	90.0	24.0	76.5

Table 14: ACC scores (%) on WikiMatrix for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	5.0	5.0	21.0	4.0	2.0	1.0	1.0	2.0	3.0	3.0	1.0	1.0	4.1
	Prompt Only	85.0	86.0	91.0	77.0	84.0	84.0	90.0	81.0	71.0	72.0	65.0	85.0	80.9
	PCA	5.0	4.0	23.0	3.0	2.0	1.0	0.0	4.0	4.0	2.0	1.0	1.0	4.2
	LDA	15.0	5.0	32.0	6.0	4.0	4.0	1.0	2.0	3.0	3.0	1.0	9.0	7.1
	Zhong (Mono)	17.0	28.0	61.0	12.0	13.0	26.0	14.0	13.0	12.0	15.0	20.0	24.0	21.3
	Zhong (Para)	58.0	68.0	67.0	47.0	49.0	62.0	**66.0**	**17.0**	**38.0**	19.0	33.0	55.0	48.2
	SAE-DM	88.0	68.0	88.0	46.0	74.0	51.0	57.0	7.0	10.0	17.0	53.0	9.0	47.3
	Gated-DM	61.0	33.0	80.0	32.0	53.0	36.0	6.0	9.0	7.0	7.0	1.0	16.0	28.4
	DiffMean	88.0	73.0	**90.0**	52.0	**77.0**	52.0	58.0	7.0	10.0	22.0	**55.0**	10.0	49.5
	LangFIR	**90.0**	**84.0**	**90.0**	**74.0**	56.0	**75.0**	**66.0**	3.0	31.0	**51.0**	42.0	**65.0**	**60.6**
Gemma 3 4B	No Steering	22.0	13.0	11.0	6.0	6.0	3.0	1.0	2.0	5.0	5.0	7.0	2.0	6.9
	Prompt Only	88.0	95.0	92.0	92.0	82.0	91.0	92.0	93.0	86.0	90.0	89.0	95.0	90.4
	PCA	30.0	29.0	22.0	7.0	5.0	7.0	1.0	5.0	5.0	3.0	9.0	2.0	10.4
	LDA	71.0	66.0	56.0	60.0	65.0	74.0	47.0	24.0	23.0	24.0	12.0	56.0	48.2
	Zhong (Mono)	46.0	60.0	56.0	44.0	45.0	66.0	39.0	42.0	22.0	43.0	52.0	45.0	46.7
	Zhong (Para)	50.0	66.0	86.0	53.0	84.0	50.0	57.0	63.0	54.0	69.0	**73.0**	69.0	64.5
	SAE-DM	**92.0**	85.0	89.0	74.0	81.0	80.0	87.0	64.0	49.0	69.0	61.0	71.0	75.2
	Gated-DM	45.0	33.0	38.0	20.0	31.0	17.0	14.0	10.0	8.0	13.0	66.0	8.0	25.2
	DiffMean	**92.0**	**91.0**	**93.0**	**86.0**	**87.0**	90.0	**92.0**	72.0	**63.0**	**76.0**	70.0	81.0	**82.8**
	LangFIR	89.0	85.0	89.0	79.0	70.0	**94.0**	62.0	**80.0**	43.0	70.0	58.0	**88.0**	75.6
Llama 3.1 8B	No Steering	8.0	11.0	11.0	3.0	1.0	6.0	1.0	8.0	8.0	3.0	1.0	1.0	5.2
	Prompt Only	90.0	97.0	93.0	94.0	79.0	94.0	91.0	94.0	91.0	93.0	75.0	91.0	90.2
	PCA	11.0	19.0	14.0	5.0	1.0	8.0	1.0	8.0	9.0	4.0	3.0	3.0	7.2
	LDA	79.0	87.0	82.0	55.0	53.0	56.0	21.0	60.0	63.0	**65.0**	31.0	72.0	60.3
	Zhong (Mono)	23.0	26.0	23.0	6.0	4.0	19.0	1.0	9.0	12.0	5.0	1.0	5.0	11.2
	Zhong (Para)	21.0	43.0	30.0	17.0	4.0	32.0	1.0	20.0	10.0	5.0	1.0	64.0	20.7
	SAE-DM	27.0	31.0	28.0	6.0	4.0	16.0	1.0	11.0	11.0	5.0	7.0	4.0	12.6
	Gated-DM	79.0	86.0	82.0	61.0	36.0	67.0	**47.0**	33.0	50.0	42.0	**80.0**	**76.0**	61.6
	DiffMean	86.0	89.0	79.0	77.0	37.0	59.0	43.0	40.0	44.0	47.0	78.0	74.0	62.8
	LangFIR	**91.0**	**97.0**	**92.0**	**95.0**	**71.0**	**95.0**	2.0	**67.0**	**77.0**	**65.0**	73.0	1.0	**68.8**

Table 15: ACC scores (%) on Tatoeba for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	0.0	3.0	20.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	2.1
	Prompt Only	31.0	34.0	99.0	92.0	79.0	94.0	98.0	98.0	81.0	90.0	81.0	94.0	80.9
	PCA	1.0	2.0	25.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	2.5
	LDA	13.0	32.0	55.0	3.0	3.0	7.0	5.0	0.0	0.0	0.0	0.0	3.0	10.1
	Zhong (Mono)	29.0	39.0	74.0	16.0	14.0	40.0	57.0	15.0	2.0	4.0	25.0	59.0	31.2
	Zhong (Para)	67.0	88.0	86.0	58.0	61.0	75.0	88.0	18.0	**39.0**	22.0	39.0	74.0	59.6
	SAE-DM	**86.0**	91.0	95.0	56.0	84.0	85.0	92.0	4.0	0.0	39.0	56.0	13.0	58.4
	Gated-DM	74.0	25.0	93.0	32.0	54.0	58.0	64.0	16.0	3.0	49.0	2.0	27.0	41.4
	DiffMean	**86.0**	93.0	95.0	68.0	**85.0**	**86.0**	**94.0**	4.0	1.0	51.0	**59.0**	15.0	61.4
	LangFIR	83.0	**97.0**	**96.0**	**72.0**	73.0	2.0	92.0	**74.0**	22.0	**81.0**	2.0	**77.0**	**64.2**
Gemma 3 4B	No Steering	36.0	15.0	4.0	1.0	0.0	1.0	2.0	0.0	0.0	0.0	12.0	0.0	5.9
	Prompt Only	96.0	99.0	99.0	96.0	82.0	96.0	97.0	91.0	84.0	95.0	96.0	96.0	93.9
	PCA	36.0	16.0	6.0	2.0	0.0	1.0	2.0	0.0	0.0	0.0	11.0	0.0	6.2
	LDA	75.0	53.0	70.0	29.0	26.0	17.0	24.0	1.0	3.0	6.0	17.0	5.0	27.2
	Zhong (Mono)	31.0	20.0	45.0	27.0	33.0	51.0	30.0	8.0	16.0	23.0	56.0	45.0	32.1
	Zhong (Para)	31.0	72.0	90.0	51.0	61.0	67.0	47.0	37.0	20.0	40.0	**61.0**	74.0	54.3
	SAE-DM	93.0	91.0	**100.0**	97.0	69.0	87.0	99.0	60.0	50.0	**93.0**	52.0	82.0	81.1
	Gated-DM	56.0	36.0	31.0	21.0	26.0	26.0	23.0	2.0	6.0	23.0	45.0	7.0	25.2
	DiffMean	93.0	**98.0**	99.0	**98.0**	**88.0**	**96.0**	**100.0**	**88.0**	**67.0**	**93.0**	52.0	86.0	**88.2**
	LangFIR	**99.0**	**98.0**	**100.0**	93.0	26.0	93.0	58.0	81.0	45.0	70.0	12.0	**92.0**	72.3
Llama 3.1 8B	No Steering	8.0	10.0	1.0	1.0	0.0	2.0	2.0	1.0	1.0	0.0	4.0	0.0	2.5
	Prompt Only	100.0	99.0	99.0	100.0	84.0	96.0	98.0	100.0	97.0	100.0	95.0	94.0	96.8
	PCA	14.0	17.0	5.0	5.0	0.0	9.0	2.0	2.0	1.0	0.0	3.0	0.0	4.8
	LDA	93.0	**94.0**	**98.0**	**82.0**	**68.0**	79.0	78.0	62.0	**45.0**	**51.0**	41.0	63.0	**71.2**
	Zhong (Mono)	26.0	49.0	5.0	28.0	2.0	50.0	2.0	35.0	2.0	0.0	2.0	71.0	22.7
	Zhong (Para)	31.0	42.0	21.0	19.0	2.0	37.0	2.0	23.0	2.0	0.0	1.0	72.0	21.0
	SAE-DM	32.0	33.0	15.0	10.0	4.0	21.0	3.0	4.0	3.0	0.0	12.0	3.0	11.7
	Gated-DM	86.0	82.0	83.0	73.0	32.0	81.0	65.0	25.0	34.0	19.0	73.0	60.0	59.4
	DiffMean	86.0	86.0	82.0	68.0	34.0	**89.0**	**80.0**	32.0	**45.0**	**51.0**	**93.0**	**73.0**	68.2
	LangFIR	**95.0**	86.0	90.0	67.0	44.0	1.0	2.0	**77.0**	0.0	0.0	0.0	0.0	38.5

E.3 BLEU

We analyze the dataset-averaged table (Table 16). The per-dataset tables (Tables 17–19) are mostly consistent with it. PCA anomalously leads BLEU across all models (38.9, 37.7, and 44.4). This is a small-sample artifact: BLEU is computed only on successfully steered samples, and PCA's extremely low ACC produces a small, unrepresentative sample. Comparing the methods that steer reliably, LangFIR's BLEU exceeds DiffMean's on Gemma 3 4B (34.9 vs. 27.1) and Llama 3.1 8B (35.3 vs. 29.8), but is lower on Gemma 3 1B (23.5 vs. 28.1).

Table 16: BLEU scores averaged over FLORES+, WikiMatrix, and Tatoeba for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	17.5	27.1	23.8	19.1	7.0	27.0	31.2	9.3	13.8	16.2	19.5	29.6	20.1
	Prompt Only	29.1	33.7	25.7	32.4	34.5	41.7	27.2	16.7	21.0	9.9	21.0	29.5	26.9
	PCA	21.7	**48.3**	**35.0**	**43.1**	21.1	21.4	28.7	**41.7**	**51.7**	6.6	**58.6**	**88.7**	**38.9**
	LDA	23.9	39.4	17.6	40.9	29.5	25.0	7.1	11.2	30.7	**23.4**	5.3	18.6	22.7
	Zhong (Mono)	8.9	13.2	11.9	21.9	13.5	19.4	15.0	3.8	18.5	6.4	8.6	12.3	12.8
	Zhong (Para)	22.0	20.1	17.3	17.0	19.4	30.5	17.7	4.1	7.6	5.4	11.4	16.6	15.8
	SAE-DM	31.5	35.0	25.8	27.0	30.9	37.0	24.2	25.5	48.8	11.4	19.2	36.6	29.4
	Gated-DM	27.0	32.0	20.6	25.8	26.1	29.3	**29.3**	10.2	27.7	19.2	28.6	20.6	24.7
	DiffMean	**32.8**	35.2	25.5	25.9	31.2	**37.5**	23.1	25.5	34.7	13.7	19.7	32.2	28.1
	LangFIR	29.0	33.7	25.2	23.6	**32.3**	25.2	24.3	19.9	20.4	7.9	15.5	25.3	23.5
Gemma 3 4B	No Steering	44.4	48.7	26.6	37.0	18.3	42.1	36.2	24.8	24.5	6.5	39.3	22.8	30.9
	Prompt Only	43.9	48.1	36.5	44.2	40.3	53.1	37.8	27.5	32.2	21.1	29.3	39.3	37.8
	PCA	33.6	37.3	21.8	25.0	**55.3**	**62.5**	**54.4**	**46.8**	**56.2**	3.6	26.5	29.2	**37.7**
	LDA	**41.9**	43.4	30.6	**38.9**	35.2	47.6	35.7	16.8	26.4	10.7	26.7	24.1	31.5
	Zhong (Mono)	31.9	33.5	24.5	28.5	31.5	37.0	19.2	20.6	24.3	7.8	18.5	25.9	25.3
	Zhong (Para)	26.3	35.6	24.9	14.3	29.7	39.1	15.8	17.1	22.8	11.9	19.2	23.4	23.3
	SAE-DM	34.3	39.9	28.7	32.7	28.3	41.6	31.3	19.0	24.4	13.3	12.5	27.9	27.8
	Gated-DM	37.9	37.3	29.8	38.0	29.3	39.0	24.6	28.6	38.1	16.3	17.2	33.3	30.8
	DiffMean	34.6	40.8	29.0	35.4	26.5	42.5	28.2	17.5	23.3	10.6	11.0	25.8	27.1
	LangFIR	40.3	**44.4**	**33.2**	38.6	41.7	46.2	37.3	26.7	26.5	**17.6**	**29.5**	**36.3**	34.9
Llama 3.1 8B	No Steering	40.3	51.5	26.4	32.5	4.4	53.8	32.6	21.0	28.6	21.2	33.4	18.9	30.4
	Prompt Only	43.9	46.6	36.6	43.2	39.6	50.7	34.5	25.0	30.7	18.1	30.9	38.9	36.6
	PCA	41.3	**52.3**	31.4	42.2	23.4	58.4	**48.9**	**46.4**	44.8	**66.2**	**38.8**	38.4	**44.4**
	LDA	**43.1**	43.8	**35.0**	**42.3**	36.8	45.2	33.8	22.2	28.1	14.3	35.6	38.1	34.9
	Zhong (Mono)	35.6	38.2	20.0	32.6	27.4	38.8	36.5	25.9	38.9	26.9	28.1	27.6	31.4
	Zhong (Para)	38.5	35.0	27.6	31.9	19.4	38.6	37.6	19.7	35.6	26.7	34.6	29.9	31.3
	SAE-DM	33.6	46.1	33.6	32.7	30.4	50.3	43.9	36.3	**48.4**	59.4	22.2	**41.9**	39.9
	Gated-DM	38.9	40.8	33.9	38.6	36.3	44.4	31.0	22.0	25.7	13.3	30.3	34.0	32.4
	DiffMean	40.6	41.4	27.8	36.6	32.6	43.3	27.0	17.0	22.8	9.8	28.0	30.8	29.8
	LangFIR	40.7	45.2	32.9	38.4	**37.5**	**64.9**	41.2	22.8	26.5	15.2	36.1	22.4	35.3

Table 17: BLEU scores on FLORES+ for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Prompt Only	19.5	27.2	19.3	18.6	32.4	31.2	27.7	9.0	13.0	9.6	23.1	16.8	20.6
	PCA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LDA	0.0	**38.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2
	Zhong (Mono)	4.1	10.7	9.1	5.5	6.0	11.8	14.2	1.6	2.8	1.3	6.4	6.8	6.7
	Zhong (Para)	9.6	16.5	13.8	11.3	16.7	22.1	17.7	1.3	3.4	1.9	10.8	8.7	11.1
	SAE-DM	18.0	32.7	**20.4**	17.8	27.0	25.9	23.6	0.0	**42.0**	8.1	**21.0**	**22.8**	**21.6**
	Gated-DM	17.6	22.0	15.0	14.4	19.5	22.3	22.9	5.2	6.6	**9.4**	12.6	16.5	15.3
	DiffMean	**19.3**	32.3	20.0	17.9	27.2	**26.0**	23.1	0.0	**42.0**	5.8	20.9	13.6	20.7
	LangFIR	17.6	28.4	19.3	**18.4**	**29.1**	24.9	**23.9**	**7.7**	9.2	4.8	18.8	14.7	18.1
Gemma 3 4B	No Steering	25.0	40.4	25.1	10.8	10.0	0.0	0.0	0.0	0.0	0.0	54.5	0.0	13.8
	Prompt Only	25.6	41.9	31.4	27.8	40.4	45.0	39.0	18.3	26.1	19.3	34.6	22.4	31.0
	PCA	18.8	24.2	18.3	21.6	**61.1**	0.0	0.0	0.0	0.0	0.0	24.3	0.0	14.0
	LDA	21.2	33.5	26.4	23.9	37.7	35.5	32.0	5.1	18.4	4.3	30.5	14.1	23.5
	Zhong (Mono)	21.9	30.5	22.2	19.3	27.2	37.1	17.8	8.8	14.7	5.7	16.0	15.1	19.7
	Zhong (Para)	17.2	30.1	22.5	15.9	25.2	32.2	8.1	13.5	15.1	6.7	21.0	13.9	18.4
	SAE-DM	21.6	31.1	20.8	20.7	29.7	33.2	28.3	10.2	13.0	9.0	22.5	16.0	21.3
	Gated-DM	23.8	35.7	**29.7**	22.3	34.7	**40.4**	33.6	13.5	**22.3**	15.1	14.7	17.7	25.3
	DiffMean	22.8	36.2	21.4	21.8	29.7	35.3	25.0	11.6	17.5	9.8	11.7	13.0	21.3
	LangFIR	**24.4**	**39.1**	28.4	**24.5**	39.9	35.2	**37.4**	**26.3**	13.6	**16.4**	**33.5**	**19.5**	**28.2**
Llama 3.1 8B	No Steering	27.9	44.3	18.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.5	12.6	10.7
	Prompt Only	25.5	42.2	30.6	27.6	38.2	41.9	35.7	17.6	24.4	17.6	33.6	23.2	29.9
	PCA	26.4	**53.9**	0.0	17.7	0.0	0.0	0.0	0.0	0.0	0.0	32.4	0.0	10.9
	LDA	**26.6**	41.0	**31.0**	**26.9**	32.6	34.0	**38.3**	15.5	20.8	14.0	**36.4**	22.5	**28.3**
	Zhong (Mono)	8.4	12.4	4.4	6.2	16.2	17.1	14.1	2.7	7.9	1.0	26.5	5.2	10.2
	Zhong (Para)	8.5	14.7	10.2	1.6	17.8	12.0	10.9	3.3	7.5	2.5	25.0	6.9	10.1
	SAE-DM	24.0	39.7	0.0	17.2	0.0	0.0	0.0	0.0	0.0	0.0	3.5	**29.2**	9.5
	Gated-DM	21.1	35.0	25.8	22.3	33.0	28.4	30.9	8.7	16.8	8.0	30.9	15.4	23.0
	DiffMean	23.8	30.6	20.6	22.2	26.8	35.0	28.0	7.7	14.5	4.0	27.8	15.4	21.4
	LangFIR	23.7	40.0	28.8	26.6	**34.7**	**39.1**	34.6	**15.9**	**21.7**	**14.1**	33.6	21.5	27.9

Table 18: BLEU scores on WikiMatrix for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	52.6	47.6	44.7	57.3	21.1	80.9	57.6	27.9	41.3	48.6	58.6	88.7	52.2
	Prompt Only	40.5	39.6	32.6	37.4	38.5	46.4	28.2	21.5	31.3	11.3	23.5	38.9	32.5
	PCA	36.7	**56.9**	**42.7**	**43.1**	21.1	21.4	0.0	41.7	51.7	6.6	**58.6**	**88.7**	**39.1**
	LDA	9.0	41.0	11.5	35.3	34.7	2.1	1.5	11.2	30.7	23.4	5.3	18.4	18.7
	Zhong (Mono)	16.5	20.3	15.2	28.3	27.0	33.3	21.3	6.2	45.8	14.0	14.8	21.3	22.0
	Zhong (Para)	27.7	20.1	20.3	20.1	18.5	36.5	21.2	4.9	15.2	9.4	17.5	17.2	19.0
	SAE-DM	39.5	37.8	31.5	40.4	**39.2**	42.9	28.4	34.2	**55.6**	20.9	22.3	50.3	36.9
	Gated-DM	24.8	35.0	26.2	34.4	34.6	38.2	**39.8**	10.8	55.0	**40.3**	**58.6**	21.0	34.9
	DiffMean	**39.7**	36.9	31.2	36.0	38.9	**42.9**	27.1	34.2	**55.6**	20.6	22.8	47.9	36.1
	LangFIR	36.1	36.3	30.5	31.7	38.4	37.3	27.0	**43.7**	36.8	9.4	23.7	31.4	31.9
Gemma 3 4B	No Steering	53.1	63.0	45.5	41.5	45.0	58.9	57.6	74.4	73.6	19.5	41.4	68.5	53.5
	Prompt Only	51.2	50.9	40.8	49.3	45.1	58.7	39.8	33.0	40.3	22.0	32.1	50.7	42.8
	PCA	33.7	46.7	30.2	15.8	**49.5**	**57.8**	**57.6**	**46.8**	**56.2**	3.6	**33.5**	29.2	38.4
	LDA	**50.7**	**47.3**	33.7	43.3	33.3	52.6	37.4	24.8	40.9	18.1	27.1	39.6	37.4
	Zhong (Mono)	40.7	36.2	25.0	35.0	38.2	42.4	20.2	21.7	29.1	11.0	26.8	42.0	30.7
	Zhong (Para)	24.2	40.2	33.1	20.8	32.4	40.0	21.1	18.8	26.4	13.8	27.0	33.0	27.6
	SAE-DM	43.7	45.9	34.3	35.4	36.6	55.1	34.1	24.5	40.3	17.8	14.9	29.4	34.3
	Gated-DM	41.3	37.6	32.7	**45.5**	34.6	41.1	21.3	45.2	49.2	19.0	25.4	**52.8**	37.1
	DiffMean	41.6	41.5	33.0	42.0	33.4	52.3	33.1	23.6	37.0	14.0	15.4	29.8	33.1
	LangFIR	45.6	47.2	**35.8**	42.0	45.5	53.9	38.7	28.4	42.6	**22.0**	33.4	48.0	**40.3**
Llama 3.1 8B	No Steering	56.7	63.1	50.1	74.0	13.1	61.5	57.6	45.8	78.9	63.7	46.2	44.1	54.6
	Prompt Only	52.3	50.4	40.7	52.1	44.0	58.4	35.8	31.7	40.7	16.5	35.7	51.3	42.5
	PCA	51.8	**54.6**	**46.7**	**78.5**	23.4	57.0	**57.6**	45.3	**82.8**	**66.2**	37.6	38.4	**53.3**
	LDA	50.0	49.0	38.7	48.6	46.0	55.3	33.8	27.5	40.5	21.0	38.9	**49.9**	41.6
	Zhong (Mono)	45.9	53.0	40.4	50.9	**52.9**	57.4	**57.6**	**52.5**	77.2	52.7	29.2	38.2	50.7
	Zhong (Para)	**52.6**	45.5	36.6	47.2	26.0	54.4	**57.6**	30.6	81.5	51.0	**61.8**	44.8	49.2
	SAE-DM	35.9	50.3	40.7	50.0	42.3	56.3	**57.6**	43.6	76.1	59.4	28.2	38.4	48.3
	Gated-DM	47.8	46.1	40.0	45.9	35.6	**57.7**	33.6	30.8	40.6	24.7	35.7	45.8	40.4
	DiffMean	51.3	48.4	32.1	40.9	35.4	55.2	28.6	17.3	40.0	19.8	34.1	38.3	36.8
	LangFIR	46.7	50.2	35.0	43.4	37.9	55.7	48.8	28.1	31.2	16.4	38.6	23.2	37.9

Table 19: BLEU scores on Tatoeba for Gemma 3 1B, Gemma 3 4B, Llama 3.1 8B. The best result among steering methods per language within each model is **bolded**.

Model	Method	es	fr	de	it	id	pt	vi	pl	cs	tr	zh	nl	Avg
Gemma 3 1B	No Steering	0.0	33.7	26.9	0.0	0.0	0.0	35.9	0.0	0.0	0.0	0.0	0.0	8.0
	Prompt Only	27.4	34.4	25.1	41.3	32.6	47.4	25.6	19.5	18.8	8.8	16.4	32.9	27.5
	PCA	6.6	**39.7**	**27.2**	0.0	0.0	0.0	**28.7**	0.0	0.0	0.0	0.0	0.0	8.5
	LDA	38.8	39.2	23.6	**46.5**	24.3	**47.9**	12.8	0.0	0.0	0.0	0.0	18.7	21.0
	Zhong (Mono)	6.0	8.7	11.6	31.9	7.5	13.2	9.5	3.4	6.8	4.0	4.6	8.7	9.7
	Zhong (Para)	28.7	23.7	17.9	19.7	23.0	32.9	14.2	6.1	4.2	4.8	5.9	23.9	17.1
	SAE-DM	37.2	34.4	25.5	22.8	26.5	42.2	20.6	**16.8**	0.0	5.1	14.4	**36.8**	23.5
	Gated-DM	38.6	39.1	20.5	28.6	24.1	27.4	25.1	14.6	**21.5**	7.7	14.5	24.2	23.8
	DiffMean	**39.4**	36.6	25.3	23.6	27.6	43.5	19.1	**16.8**	6.4	**14.7**	**15.4**	35.0	**25.3**
	LangFIR	33.3	36.6	25.7	20.6	**29.4**	13.4	22.0	8.4	15.1	9.6	3.8	29.9	20.7
Gemma 3 4B	No Steering	55.2	42.6	9.4	58.7	0.0	67.3	51.1	0.0	0.0	0.0	22.1	0.0	25.5
	Prompt Only	55.1	51.4	37.4	55.6	35.5	55.7	34.6	31.2	30.3	22.0	21.2	44.8	39.6
	PCA	48.2	40.9	16.9	37.7	0.0	**67.3**	**51.1**	0.0	0.0	0.0	21.6	0.0	23.6
	LDA	**53.8**	**49.4**	31.8	**49.6**	34.7	54.6	37.6	20.6	19.8	9.7	**22.5**	18.6	33.5
	Zhong (Mono)	33.1	34.0	26.4	31.2	29.1	31.3	19.5	**31.4**	29.2	6.5	12.6	20.7	25.4
	Zhong (Para)	37.5	36.6	19.1	6.2	31.5	45.0	18.2	18.9	26.9	**15.3**	9.5	23.4	24.0
	SAE-DM	37.7	42.6	31.0	42.1	18.5	36.5	31.5	22.3	19.8	13.2	0.1	38.4	27.8
	Gated-DM	48.7	38.5	27.1	46.3	18.5	35.6	18.9	27.1	**42.9**	14.7	11.5	29.5	29.9
	DiffMean	39.3	44.6	32.6	42.3	16.2	39.8	26.4	17.5	15.3	8.0	5.9	34.7	26.9
	LangFIR	50.8	46.9	**35.5**	49.3	**39.7**	49.5	35.7	25.5	23.3	14.4	21.7	**41.3**	**36.1**
Llama 3.1 8B	No Steering	36.3	47.0	10.4	23.5	0.0	100.0	40.1	17.3	6.9	0.0	29.5	0.0	25.9
	Prompt Only	53.9	47.2	38.6	49.7	36.6	51.8	32.0	25.7	27.1	20.3	23.4	42.2	37.4
	PCA	45.7	48.3	16.0	30.5	0.0	59.8	40.1	**47.4**	6.9	0.0	**46.5**	0.0	28.4
	LDA	52.7	41.4	35.4	**51.3**	31.8	46.5	29.4	23.7	23.0	**7.7**	31.5	41.7	**34.7**
	Zhong (Mono)	52.5	**49.3**	15.1	40.5	13.0	42.0	37.7	22.5	**31.5**	0.0	28.7	39.4	31.0
	Zhong (Para)	**54.6**	44.9	**36.0**	47.0	14.4	49.3	**44.3**	25.1	17.7	0.0	17.1	38.0	32.4
	SAE-DM	40.8	48.2	26.4	30.8	18.5	44.4	30.2	29.0	20.7	0.0	35.0	**58.1**	31.8
	Gated-DM	48.0	41.3	35.9	47.4	**40.1**	47.0	28.4	26.5	19.8	7.2	24.4	40.9	33.9
	DiffMean	46.7	45.2	30.7	46.7	35.4	39.6	24.3	26.1	13.9	5.5	22.0	38.6	31.2
	LangFIR	51.7	45.3	35.0	45.3	40.0	**100.0**	40.1	24.4	0.0	0.0	0.0	0.0	31.8

F Ablation Details

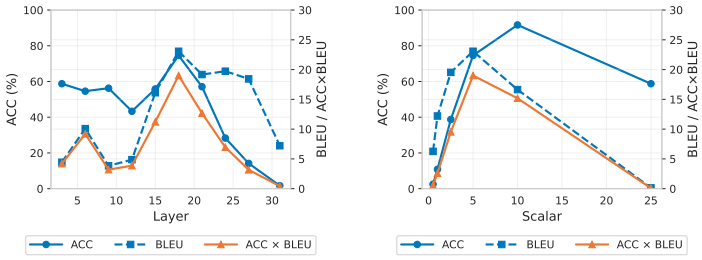
F.1 Ablation Setup

All ablations are conducted using the same twelve languages from the FLORES+ validation split used in the steering experiment. For each hyperparameter under study, all other hyperparameters are held fixed at the validation-selected configuration from the steering experiment, so that the effect of each factor can be measured independently.

F.2 Ablation on Llama 3.1 8B

Layer. Figure 12a shows steering metrics across intervention layers. ACC is highest in the early-to-middle layers, while BLEU increases in the middle-to-late layers, possibly because richer semantic representations in these layers (Wendler et al., 2024) allow steered outputs to better preserve translation quality. The composite ACC×BLEU peaks in the middle layers, where steering successfully redirects language without disrupting content fidelity. In the very late layers, both metrics decline, suggesting that the model has largely committed to its output distribution.

Scaling coefficient. Figure 12b shows a clear peak in all three metrics at an intermediate scalar value. When the scalar is too small, the steering vector has minimal effect on the model's output. When the scalar is too large, BLEU drops sharply while ACC remains relatively high. This indicates that an overly strong steering signal causes the model to produce tokens in the target language that are incoherent, so language classification succeeds but translation quality deteriorates.

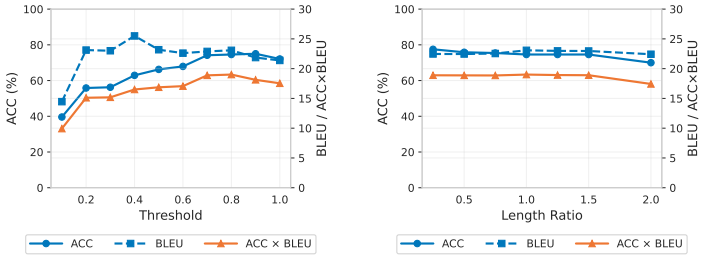


(a) Intervention layer. (b) Scaling coefficient α.

Figure 12: Ablation of intervention layer and scaling coefficient α on Llama 3.1 8B. (a) ACC×BLEU peaks in middle layers, where steering redirects language without disrupting content. (b) Intermediate α balances steering strength and translation quality.

Frequency threshold. Figure 13a shows that performance generally trends upward with increasing τ and stabilizes around $\tau = 0.8$, consistent with the threshold stability reported in Section 5. Performance drops drastically when $\tau < 0.2$, demonstrating that a sufficiently high threshold is needed to isolate a reliable set of language-consistent and language-agnostic features. A slight decline at very high thresholds ($\tau > 0.9$) suggests some weaker language-specific features are excluded.

Random-token sequence length. Figure 13b shows steering metrics as a function of the length ratios between random-token sequences and target-language sentences. Metrics remain stable across ratios from 0.25 to 1.5, with only a slight drop at ratio 2.0. This shows that random-token sequences need only be of similar length to the target-language sentences, and exact length matching is not required, though it is a convenient default.



(a) Frequency threshold τ. (b) Sequence length ratio.

Figure 13: Ablation of frequency threshold τ and length ratio between random-token sequences and target-language sentences on Llama 3.1 8B. (a) Performance trends upward with increasing τ and stabilizes around $\tau = 0.8$. (b) Steering metrics remain stable across length ratios, showing that exact length matching is not required.

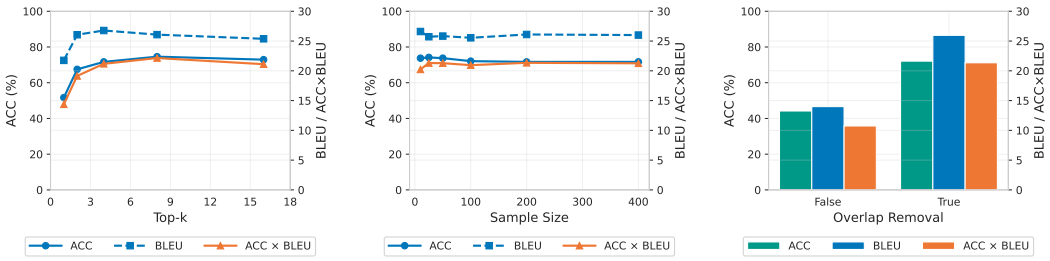
Random seed. Table 20 reports mean and standard deviation of steering metrics across three independent runs with different random seeds used for random-token sequence generation. The standard deviation is near zero for all metrics, demonstrating that LangFIR's performance is robust to random seeds.

Table 20: Steering metrics stability across 3 random seeds on Llama 3.1 8B. Near-zero standard deviations indicate robustness to random seeds.

Metric	Mean	Std
ACC (%)	74.58	0.000
BLEU	23.06	0.036
ACC × BLEU	18.99	0.025

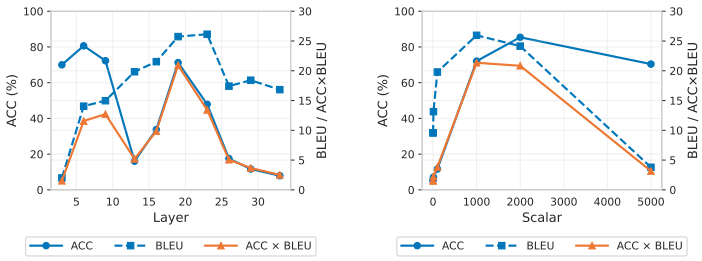
F.3 Ablation on Gemma 3 4B

Figures 14–16 and Table 21 show the ablation results for Gemma 3 4B. Results are generally consistent with Llama 3.1 8B.



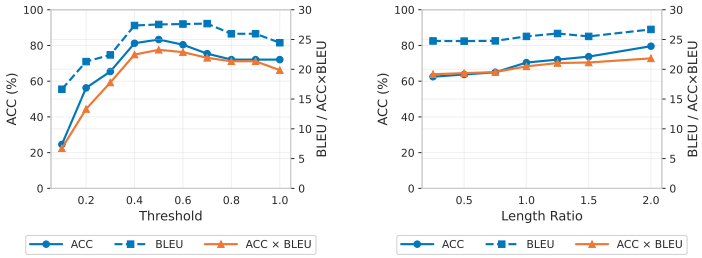
(a) Top-k features. (b) Sample size N. (c) Feature overlap removal.

Figure 14: Ablation of number of top-k language-specific features, sample size, and feature-overlap removal on Gemma 3 4B. (a) Even k=2 achieves strong performance. (b) Steering metrics are stable with as few as 10 sentences. (c) Removing feature-overlap removal degrades all steering metrics significantly.



(a) Intervention layer. (b) Scaling coefficient α.

Figure 15: Ablation of intervention layer and scaling coefficient α on Gemma 3 4B. (a) ACC$\times$BLEU peaks in middle layers, where steering redirects language without disrupting content. (b) Intermediate α balances steering strength and translation quality.



(a) Frequency threshold τ. (b) Sequence length ratio.

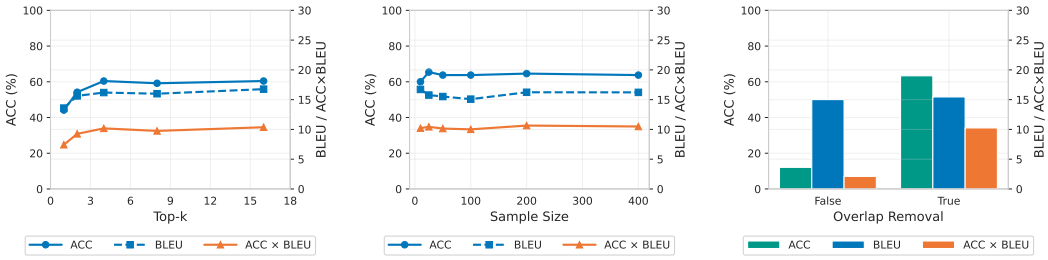
Figure 16: Ablation of frequency threshold τ and length ratio between random-token sequences and target-language sentences on Gemma 3 4B. (a) Performance generally trends upward with increasing τ and stabilizes around $\tau = 0.8$. (b) Steering metrics remain generally stable across length ratios, showing that exact length matching is not required.

Table 21: Steering metrics stability across 3 random seeds on Gemma 3 4B. Small standard deviations indicate robustness to random seeds.

Metric	Mean	Std
ACC (%)	69.86	1.339
BLEU	25.67	0.317
ACC $\times$ BLEU	20.47	0.040

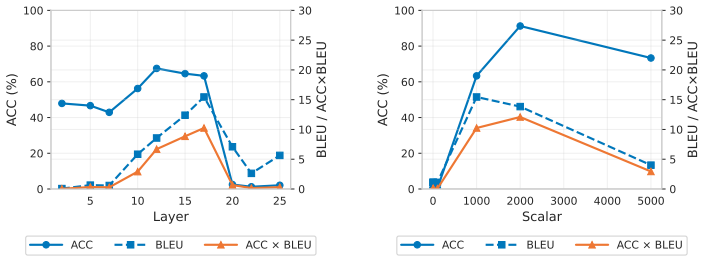
F.4 Ablation on Gemma 3 1B

Figures 17–19 and Table 22 show the ablation results for Gemma 3 1B. Results are generally consistent with Llama 3.1 8B.



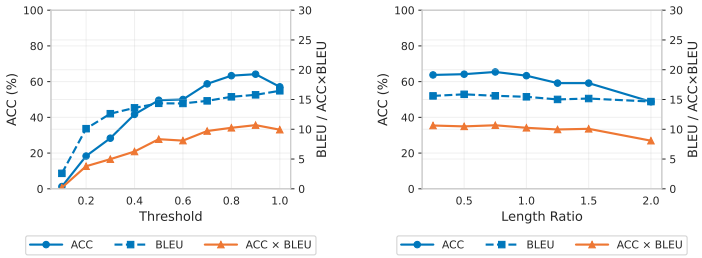
(a) Top-k features. (b) Sample size N. (c) Feature overlap removal.

Figure 17: Ablation of number of top-k language-specific features, sample size, and feature-overlap removal on Gemma 3 1B. (a) Even k=2 achieves strong performance. (b) Steering metrics are stable with as few as 10 sentences. (c) Removing feature-overlap removal degrades all steering metrics dramatically.



(a) Intervention layer. (b) Scaling coefficient α.

Figure 18: Ablation of intervention layer and scaling coefficient α on Gemma 3 1B. (a) ACC$\times$BLEU peaks in middle layers, where steering redirects language without disrupting content. (b) Intermediate α balances steering strength and translation quality.



(a) Frequency threshold τ. (b) Sequence length ratio.

Figure 19: Ablation of frequency threshold τ and length ratio between random-token sequences and target-language sentences on Gemma 3 1B. (a) Performance trends upward with increasing τ and stabilizes around $\tau = 0.8$. (b) Steering metrics remain stable across length ratios, showing that exact length matching is not required.

Table 22: Steering metrics stability across 3 random seeds on Gemma 3 1B. Small standard deviations indicate robustness to random seeds.

Metric	Mean	Std
ACC (%)	63.06	1.273
BLEU	15.77	0.292
ACC × BLEU	10.30	0.066

F.5 Ablation on the Filter Negative Set

LangFIR uses random-token sequences as the filter negative set. We ablate this choice against two alternatives that require additional data: non-parallel text from the other eleven languages (Other Languages) and non-parallel English sentences (Non-parallel English). Table 23 reports FLORES+ results averaged over the twelve languages.

Text from other languages is slightly better than random tokens in ACC × BLEU (+0.5 to +0.8 across the three models), likely because a mixture of languages also surfaces language-agnostic features. Non-parallel English shows model-specific variation: relative to random tokens, it gains 0.7 points on Llama 3.1 8B and 1.0 on Gemma 3 1B, but loses 3.3 on Gemma 3 4B. We attribute this to two competing effects. Random tokens cast a wider net over language-agnostic features, which helps when the noise is diverse but may occasionally over-filter. Non-parallel English is a more conservative filter, but may remove English-biased features that overlap with closely related languages. Consistent with this, we verify on Llama 3.1 8B that random tokens surface a slightly larger set of language-agnostic features than Non-parallel English. English likely works as a negative set because general concepts are entangled with English representations (Section 5), so English sentences can also surface many language-agnostic features. Overall, random tokens remain competitive with both alternatives while requiring no data beyond the target-language corpus.

Table 23: Ablation of filter negative set on Llama 3.1 8B, Gemma 3 4B, and Gemma 3 1B. The best result per metric within each model is **bolded**. Random tokens achieve competitive performance without requiring any additional language data.

Model	Filter Negative Set	ACC (%)	BLEU	ACC × BLEU
Llama 3.1 8B	Random Tokens	78.8	23.2	18.9
	Other Languages	80.0	**25.6**	**19.6**
	Non-parallel English	**81.2**	23.5	**19.6**
Gemma 3 4B	Random Tokens	68.8	**26.1**	19.7
	Other Languages	**72.1**	25.3	**20.5**
	Non-parallel English	54.6	24.3	16.4
Gemma 3 1B	Random Tokens	60.4	**16.2**	9.8
	Other Languages	63.3	15.8	10.3
	Non-parallel English	**63.7**	15.8	**10.8**

G Cross-Lingual Answer Generation

This appendix details the cross-lingual answer generation experiment, covering both the experimental setup and the full steering results.

G.1 Experimental Setup

The setup mirrors the multilingual generation control task (Section 6.2, Appendix A). We describe only the task-specific differences below.

Dataset and Preprocessing. We use XQuAD (Artetxe et al., 2020), an extractive question answering dataset of 1,190 question–context–answer triples that is entirely parallel across 11 languages. For each language, 20 examples form a validation set for hyperparameter tuning, and 100 disjoint examples are used for final evaluation.

Evaluation metrics. We compute F1 and exact match (EM) against the gold target-language answer following the standard SQuAD v1.1 evaluation (Rajpurkar et al., 2016), which normalizes answers by lowercasing, removing punctuation and articles, and collapsing whitespace. We use character-level tokenization for Chinese.

Hyperparameter tuning. Hyperparameters are re-tuned for this task, selecting the configuration with the highest F1 averaged over the five languages. Table 24 reports the selected hyperparameters for each method and model on the cross-lingual answer generation task.

Table 24: Selected hyperparameters for the cross-lingual answer generation task across all three models.

Model	Method	Selected hyperparameters
Gemma-3-1B	PCA	layer 12, α=10000, top-k=1
	LDA	layer 25, α=0.1
	Zhong (Mono)	layer 20, α=0.6, top-k=200, anchor=9
	Zhong (Para)	layer 20, α=0.6, top-k=200
	SAE-DM	layer 10, α=1000
	Gated-DM	layer 10, α=1000
	DiffMean	layer 10, α=1000
	LangFIR	layer 15, α=1000, τ=0.8, top-k=2
Gemma-3-4B	PCA	layer 16, α=10000, top-k=1
	LDA	layer 29, α=1000
	Zhong (Mono)	layer 26, α=0.2, top-k=200, anchor=12
	Zhong (Para)	layer 23, α=0.2, top-k=200
	SAE-DM	layer 26, α=10000
	Gated-DM	layer 13, α=10000
	DiffMean	layer 26, α=10000
	LangFIR	layer 16, α=1000, τ=0.8, top-k=1
Llama-3.1-8B	PCA	layer 12, α=2.5, top-k=32
	LDA	layer 21, α=0.5
	Zhong (Mono)	layer 12, α=0.2, top-k=200, anchor=11
	Zhong (Para)	layer 12, α=0.2, top-k=400
	SAE-DM	layer 21, α=2.5
	Gated-DM	layer 21, α=10
	DiffMean	layer 21, α=10
	LangFIR	layer 18, α=10, τ=1.0, top-k=100

G.2 Results

Table 25 reports average EM for the cross-lingual answer generation experiment, and Tables 26 and 27 report the full per-language F1 and EM results for all three models. The same pattern as F1 holds on EM (Table 25): every steering method beats Prompt Only on average. Among steering methods, Gated-DiffMean leads on average (8.9) while LangFIR remains competitive (8.3) and is best on Gemma 3 4B.

Table 25: Cross-lingual answer generation results: EM averaged over five target languages (es, de, vi, tr, zh). The best result among steering methods within each model is **bolded**.

	Non-steering		Steering							
Model	No Steering	Prompt Only	PCA	LDA	Zhong (M)	Zhong (P)	SAE-DM	Gated-DM	DiffMean	LangFIR
Gemma 3 1B	7.8	7.2	5.8	7.8	6.8	**8.2**	7.0	7.8	7.0	6.8
Gemma 3 4B	8.8	3.8	9.8	8.8	6.2	6.0	9.8	9.0	9.2	**10.0**
Llama 3.1 8B	6.8	6.2	5.8	6.6	5.0	5.2	5.8	**9.8**	**9.8**	8.2
Avg	7.8	5.7	7.1	7.7	6.0	6.5	7.5	**8.9**	8.7	8.3

Table 26: Per-language F1 scores for Gemma 3 1B, Gemma 3 4B, and Llama 3.1 8B. Best result among steering methods within each model is **bolded**.

Model	**Method**	es	de	vi	tr	zh	**Avg**
Gemma 3 1B	No Steering	10.1	14.9	15.5	11.9	10.9	12.7
	Prompt Only	9.9	14.5	15.0	11.1	10.8	12.3
	PCA	8.4	12.8	12.5	10.0	9.1	10.6
	LDA	10.1	14.9	15.5	11.9	10.9	12.7
	Zhong (Mono)	10.5	13.4	9.6	11.0	11.3	11.1
	Zhong (Para)	14.1	**19.4**	12.3	13.4	16.3	15.1
	SAE-DM	14.5	16.1	16.6	12.2	7.2	13.3
	Gated-DM	11.0	15.2	15.9	12.0	14.0	13.6
	DiffMean	13.7	16.0	**20.2**	**13.5**	8.6	14.4
	LangFIR	**16.4**	15.6	15.4	13.3	**22.8**	**16.7**
Gemma 3 4B	No Steering	13.6	17.2	17.5	13.3	11.5	14.6
	Prompt Only	5.5	6.7	8.7	10.2	8.3	7.9
	PCA	14.3	18.4	18.2	14.1	12.7	15.6
	LDA	14.3	16.5	17.6	13.4	11.5	14.7
	Zhong (Mono)	9.9	15.2	11.7	11.3	8.1	11.3
	Zhong (Para)	13.2	13.8	11.6	11.2	16.7	13.3
	SAE-DM	**17.9**	**24.6**	**26.4**	15.9	27.6	**22.5**
	Gated-DM	15.2	18.2	23.6	17.3	22.3	19.3
	DiffMean	16.1	22.3	24.9	**18.0**	27.4	21.7
	LangFIR	13.0	17.0	18.5	15.0	**36.4**	20.0
Llama 3.1 8B	No Steering	11.2	14.7	15.3	11.3	9.0	12.3
	Prompt Only	9.6	14.3	15.3	11.3	9.9	12.1
	PCA	10.5	13.5	15.0	9.4	9.5	11.6
	LDA	10.9	14.6	14.6	10.9	7.9	11.8
	Zhong (Mono)	9.3	10.6	11.4	12.8	8.8	10.6
	Zhong (Para)	9.6	11.1	11.0	13.1	10.9	11.2
	SAE-DM	10.2	13.8	13.8	9.9	7.6	11.1
	Gated-DM	18.6	**23.6**	23.5	14.7	23.4	20.8
	DiffMean	18.9	23.4	**23.6**	14.8	22.8	20.7
	LangFIR	**22.0**	20.4	19.5	**21.1**	**25.3**	**21.6**

Table 27: Per-language EM scores for Gemma 3 1B, Gemma 3 4B, and Llama 3.1 8B. Best result among steering methods within each model is **bolded**.

Model	Method	es	de	vi	tr	zh	Avg
Gemma 3 1B	No Steering	7.0	12.0	9.0	8.0	3.0	7.8
	Prompt Only	6.0	11.0	8.0	8.0	3.0	7.2
	PCA	5.0	9.0	6.0	7.0	2.0	5.8
	LDA	**7.0**	12.0	**9.0**	8.0	**3.0**	7.8
	Zhong (Mono)	**7.0**	11.0	6.0	8.0	2.0	6.8
	Zhong (Para)	**7.0**	**15.0**	7.0	**9.0**	**3.0**	**8.2**
	SAE-DM	**7.0**	11.0	8.0	**9.0**	0.0	7.0
	Gated-DM	**7.0**	12.0	**9.0**	8.0	**3.0**	7.8
	DiffMean	6.0	11.0	8.0	**9.0**	1.0	7.0
	LangFIR	5.0	11.0	7.0	**9.0**	2.0	6.8
Gemma 3 4B	No Steering	9.0	13.0	10.0	9.0	3.0	8.8
	Prompt Only	1.0	1.0	6.0	7.0	4.0	3.8
	PCA	**10.0**	**14.0**	11.0	10.0	4.0	9.8
	LDA	**10.0**	12.0	10.0	9.0	3.0	8.8
	Zhong (Mono)	6.0	9.0	5.0	8.0	3.0	6.2
	Zhong (Para)	7.0	9.0	4.0	7.0	3.0	6.0
	SAE-DM	9.0	**14.0**	**12.0**	10.0	4.0	9.8
	Gated-DM	8.0	12.0	11.0	**11.0**	3.0	9.0
	DiffMean	7.0	13.0	**12.0**	10.0	4.0	9.2
	LangFIR	9.0	13.0	11.0	10.0	**7.0**	**10.0**
Llama 3.1 8B	No Steering	6.0	9.0	8.0	7.0	4.0	6.8
	Prompt Only	4.0	8.0	8.0	7.0	4.0	6.2
	PCA	6.0	8.0	7.0	5.0	3.0	5.8
	LDA	6.0	9.0	7.0	7.0	4.0	6.6
	Zhong (Mono)	5.0	6.0	5.0	8.0	1.0	5.0
	Zhong (Para)	5.0	6.0	5.0	8.0	2.0	5.2
	SAE-DM	5.0	8.0	7.0	6.0	3.0	5.8
	Gated-DM	**10.0**	**15.0**	**10.0**	9.0	**5.0**	**9.8**
	DiffMean	**10.0**	**15.0**	**10.0**	9.0	**5.0**	**9.8**
	LangFIR	8.0	10.0	8.0	**11.0**	4.0	8.2