Steering with Multilingual Lexicons for Sentiment Control in Low-Resource Languages

Lucinda Laurent[1] **Muhammad Dehan Al Kautsar**[2]
Alfan Farizki Wicaksono[1] **Fajri Koto**[2]
[1]Faculty of Computer Science, Universitas Indonesia
[2]Mohamed bin Zayed University of Artificial Intelligence
lucinda.laurent@ui.ac.id, alfan@cs.ui.ac.id, {muhammad.dehan,fajri.koto}@mbzuai.ac.ae

Abstract

Activation steering offers a lightweight mechanism for controlling model behavior, but existing methods depend on sentence-level contrastive data that is unavailable for most of the world's languages. We show that multilingual sentiment lexicons are a viable substitute across three model families: aggregated multilingual lexicons outperform English-derived steering vectors for mid- and low-resource languages, while monolingual lexicons remain stronger for languages a model has specialized in during pre-training. The steering efficacy is mechanistically dictated by both representational density and pre-training priors. Large-magnitude English vectors overwhelm the compact representations of low-resource languages, causing "linguistic collapse." This fragility is mitigated by the norm-dampening of multilingual aggregation. Architectural biases, including bilingual specialization penalty and latent numerical rigidity, also heavily constrain cross-lingual steering responsiveness. Ultimately, our work establishes lexicons as a practical steering signal and offers a comprehensive mechanistic account of cross-lingual steering dynamics.[1]

1 Introduction

Activation engineering has emerged as a lightweight method for controlling model outputs by manipulating internal activation vectors during inference, without modifying model weights or requiring curated training data (Turner et al., 2024; Rimsky et al., 2024; Arditi et al., 2024; Zou et al., 2025). By identifying linear directions associated with specific concepts (e.g., "sentiment"), one can steer the model's behavior directly. In the context of sentiment-controlled generation, several studies have adopted activation engineering as an alternative to fine-tuning-based approaches (Turner et al., 2024; Liu et al., 2024a; Konen et al., 2024). Despite its promise, current sentiment steering methodologies are constrained by a reliance on sentence-level supervision and evaluation benchmarks optimized for English-centric corpora. This dependence creates a significant barrier for low-resource languages, where high-quality sentiment-labeled sentence corpora are scarce or non-existent, leaving most of the world's languages outside the reach of existing steering methods.

The effectiveness of these interventions across diverse languages also remains largely unexplored. This gap is critical because LLMs exhibit varying degrees of cross-lingual isomorphism: the extent to which semantic concepts share a similar geometric representation across languages (Lample et al., 2018). In models with strong "Anglocentric" bias, a sentiment vector derived from English may be orthogonal to, or incongruent with, the sentiment subspace of low-resource languages like Javanese or Hausa, mirroring findings on the lack of isomorphism in low-resource cross-lingual embeddings (Sannigrahi and Read, 2022). Relying solely on monolingual sentence-derived vectors thus constrains the scalability of activation steering in multilingual contexts.

In this work, we propose multilingual sentiment lexicons as a scalable, resource-efficient signal for cross-lingual activation steering. Our contributions are as follows: **(1) A lexicon-based steering framework.** We formalize activation steering via multilingual sentiment lexicons (illustrated in Figure 1), eliminating the need for sentence-level annotations or paired contrastive prompts. **(2) Systematic cross-lingual evaluation.** We benchmark monolingual versus multilingual lexicon steering across 11 typologically diverse languages spanning high-, mid-, and low-resource tiers and three model families. We show that aggregated multilingual lexicons outperform English-derived steering vectors for most mid- and low-resource languages, while monolingual lexicons remain superior in nar-

[1]Code and data are available at https://github.com/lucindalaurent/SteeringWithMultilingualLexicon.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 174–203
November 6-10, 2026 ©2026 Association for Computational Linguistics

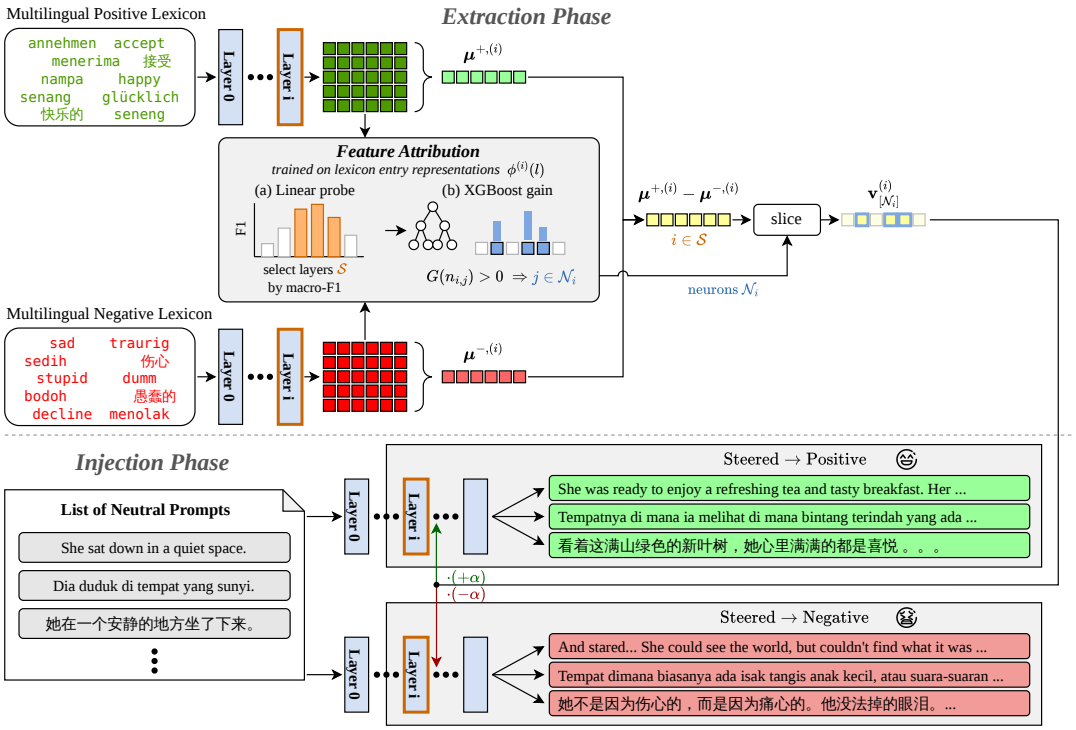


Figure 1: **Workflow of the Cross-Lingual Sentiment Control using Multilingual Lexicon.** In the Extraction Phase, feature attribution selects the steering layers $i \in \mathcal{S}$ (linear-probe macro-F1) and, within each, the neurons $\mathcal{N}_i$ (**XGBoost** gain). At each selected layer, the steering vector is the difference $\mu^{+,(i)} - \mu^{-,(i)}$ between multilingual positive and negative lexicon centroids, sliced to $\mathcal{N}_i$. In the Injection Phase, this vector is applied to neutral prompts with coefficient $+\alpha$ or $-\alpha$, shifting the generated text toward positive or negative sentiment across multiple languages.

rower cases where a model is heavily specialized in a particular language. **(3) Mechanistic analysis of cross-lingual steering dynamics.** We demonstrate that steering success is dictated by a model's pre-training priors and latent scale disparities. We show that large-norm English vectors cause *linguistic collapse* by overwhelming the compact latent subspaces of low-resource languages, whereas multilingual aggregation produces more balanced vectors that eliminate this mismatch and prevent representation collapse. Furthermore, we reveal how architectural biases determine a model's responsiveness to cross-lingual interventions.

2 Background

CTG via Activation Steering. Controllable Text Generation (**CTG**) is the task of generating natural sentences that adhere to certain attributes (Prabhumoye et al., 2020). According to Liang et al. (2024), **CTG** methods can be classified based on the stage at which the intervention occurs: the training stage (e.g. retraining, fine-tuning, reinforcement learning) and the inference stage (e.g. prompt engineering, latent space manipulation, decoding-time intervention). Although training phase methods provide high-fidelity control, they are often constrained by high computational costs and substantial data requirements, making them impractical for low-resource language settings, where the extensive corpora required for effective training are unavailable. In this paper, we utilize the latent space manipulation (also known as activation engineering) method to exploit language model's underlying semantic geometry, potentially offering a more universal and robust framework for cross-lingual control.

We specifically use the contrastive latent vector acquisition method to derive the steering direction (or steering vector). Several existing works have employed this approach for sentiment control, such as Activation Addition (Turner et al., 2024), In-Context Vectors (Liu et al., 2024a), and Style Vectors (Konen et al., 2024). Activation Addition technique derived its steering vector by computing the difference of intermediate activations on a contrast

pair of prompts, such as $h(\text{“Love”}) - h(\text{“Hate”})$, where $h(w)$ represents the activation elicited by the word w. The $h(\text{“Love”}) - h(\text{“Hate”})$ steering direction is then added during forward pass at inference time to shift the activations to generate positive sentiment completions. On the other hand, In-Context Vector (**ICV**) is derived by applying Principal Component Analysis (**PCA**) on the differences between hidden states from multiple example pairs $\{(x_i, y_i)\}_{i=1}^{n}$:

$$\textbf{ICV} = \textbf{PCA}\left(\{h(y_i) - h(x_i)\}_{i=1}^{n}\right) .$$

In the demonstration for sentiment, the example pairs used are pairs of negative and positive text like (“Zero stars, I hate it.”, “Five stars, I love it.”).

Our method for deriving the sentiment steering direction is most similar to the Activation-based Style Vectors technique, in contrast to the other two approaches discussed. This technique extends the concept of Activation Addition by calculating a general style vector for a specific category $s \in S$. This is achieved by computing the difference between the mean activation of sentences in style category s and the mean activation of all sentences in the remaining style categories, $S \setminus s$. In this study, instead of relying on labeled sentiment datasets such as **Yelp** Review (Zhang et al., 2015), we utilize sentiment lexicons consisting of positive and negative words to address the inherent data scarcity in low-resource language settings.

Mechanistic Interpretability and Linear Representations. Rather than modifying model weights through fine-tuning, activation engineering exploits the pre-existing representations within the language model. It relies on the Linear Representation Hypothesis, assuming that semantic attributes are encoded as linear directions, thus allowing for direct steering within the representation space (Park et al., 2024). Previous research by Tigges et al. (2024) has shown that language models linearly represent sentiment, where methods like k-means, linear probing, and Distributed Alignment Search (**DAS**), originally introduced by Geiger et al. (2024), converge on a similar sentiment direction. Di Palma et al. (2025) further corroborates this by unveiling linearly probe-able sentiment and emotion representations specifically within the **Llama** model family. We extend these findings by investigating the cross-lingual universality of sentiment directions; specifically, whether lexicon-derived sentiment directions align in a shared representation space across multiple languages. We primarily utilize the difference-in-means method to derive sentiment directions, complemented by linear probing to evaluate the consistency of sentiment representation across model layers.

The findings by Liu et al. (2024b) suggest that in generative language models, lexical semantics are highly concentrated in the lower layers. By evaluating hidden states on the **Word in Context (WiC)** benchmark, they demonstrate that lower layers manage contextual disambiguation (determining whether identical words convey the same meaning in distinct contexts), while higher layers shift computational focus toward next-token prediction. In contrast, our approach evaluates semantic polarity without surrounding context. We isolate the sentiment axis by extracting global representations directly from uncontextualized lexicon entries, which can range from single words to multi-word phrases. Because semantic encoding varies significantly across network depth (Liu et al., 2024b), mapping these isolated representations layer-by-layer allows us to empirically identify the optimal intervention point for cross-lingual sentiment steering using a multilingual lexicon.

Cross-Lingual Alignment and Universality. While the Linear Representation Hypothesis suggests stable sentiment directions within a single language, recent evaluations highlight that the generalization and reliability of steering vectors can degrade significantly under out-of-distribution shifts (Tan et al., 2025). Because cross-lingual transfer inherently represents a massive distributional shift, the capacity of these vectors to generalize zero-shot across languages remains a critical open question. Ideally, multilingual LLMs would exhibit cross-lingual isomorphism, where semantic spaces across languages share a similar geometric structure (Lample et al., 2018). This phenomenon was first empirically observed in massively multilingual encoders like **mBERT** (Devlin et al., 2019) and **XLM-R** (Conneau et al., 2020), which demonstrated that pre-training on a shared vocabulary across 100+ languages naturally induces a “universal” representation space where semantically similar tokens align across languages without explicit translation supervision (Libovický et al., 2020; Pires et al., 2019). Under this assumption, a sentiment direction derived in English should theoretically transfer to other languages.

Despite the assumption of a universal space, pre-

training on unbalanced corpora frequently produces an "Anglocentric" internal structure (Ahuja et al., 2023; Lauscher et al., 2020). Within this structure, high-resource languages align closely with English subspaces, whereas mid-to-low resource languages suffer from weaker alignment or severe representation degeneration. This representational gap directly challenges cross-lingual steering; if the "sentiment axis" in a low-resource language is disjoint from the English axis, monolingual steering will fail. This geometric misalignment remains intrinsically tied to the model's underlying linguistic proficiency. Multilingual architectures rely on language-specific neurons to map inputs into the shared semantic space and decode them back into target vocabularies (Tang et al., 2024). Without robust neurons for low-resource languages, a severe projection bottleneck emerges, suggesting that the model's limited linguistic proficiency acts as a critical constraint on cross-lingual control, independent of the latent steering mechanism's geometric accuracy.

Operating under these constraints, we propose that aggregation of multilingual lexicons can serve as a geometric bridge. We hypothesize that if the underlying alignment is sufficiently preserved, a multilingual direction can reinforce the shared sentiment direction via coherent reinforcement. Conversely, in cases of extreme representation degeneration, such aggregation may fail to overcome the lack of semantic grounding.

3 Methodology

An overview of our cross-lingual sentiment steering framework is illustrated in Figure 1.

3.1 Sentiment Direction Extraction

Among the various methods for obtaining steering direction, difference-in-means is one of the most simple, yet effective approach (Marks and Tegmark, 2024; Belrose, 2023). To extract the steering direction corresponding to sentiment attribute, we employ the difference-in-means method using two opposing sets of lexicon entries (phrases), denoted as $\mathcal{L}^+$ and $\mathcal{L}^-$.

Definitions. Let $\mathcal{V}$ denote the model's vocabulary consisting of individual tokens. Recognizing that modern language models utilize sub-word tokenization (e.g., Byte-Pair Encoding), a single word or phrase in a lexicon may be decomposed into multiple tokens. We therefore define a *lexicon entry* l as a sequence of tokens $l = (t_1, t_2, \ldots, t_{|l|})$, where each $t_j \in \mathcal{V}$ and $|l| \geq 1$ represents the length of the entry. For example, the word "abhorrence" is tokenized into ("ab", "hor", "rence") using **`Qwen-3`**'s tokenizer. Accordingly, we define two contrastive sets of lexicon entries:

- $\mathcal{L}^+ = \{l_1^+, l_2^+, \ldots, l_N^+\}$ representing the positive lexicon set; and
- $\mathcal{L}^- = \{l_1^-, l_2^-, \ldots, l_M^-\}$ representing the negative lexicon set.

Representation of Lexicon Entries. Let $h^{(i)}(t) \in \mathbb{R}^d$ denote the hidden state vector of a token t at layer i. The vector representation of a lexicon entry l consisting of tokens $(t_1, \ldots, t_{|l|})$ at layer i, denoted as $\phi^{(i)}(l)$, is defined as the average of the hidden states of its constituent tokens at that specific layer:

$$\phi^{(i)}(l) = \frac{1}{|l|} \sum_{j=1}^{|l|} h^{(i)}(t_j)\,.$$

By using these averaged hidden states as input, we ensure that the vector representation directly reflects the model's internal encoding of the sentiment-bearing terms in our lexicon. Furthermore, averaging across constituent tokens avoids over-reliance on the final token, which has been shown to be inconsistently informative for sentiment representation in decoder-only architectures, such as the **`Llama`** family (Di Palma et al., 2025).

Layer-wise Centroids. We compute the centroid for the positive and negative groups by averaging the lexicon entry representations $\phi^{(i)}(l)$ strictly within the context of each layer i:

$$\boldsymbol{\mu}^{+,(i)} = \frac{1}{|\mathcal{L}^+|} \sum_{l \in \mathcal{L}^+} \phi^{(i)}(l)\,,$$

$$\boldsymbol{\mu}^{-,(i)} = \frac{1}{|\mathcal{L}^-|} \sum_{l \in \mathcal{L}^-} \phi^{(i)}(l)\,.$$

Here, $|\mathcal{L}^+|$ and $|\mathcal{L}^-|$ represents the number of lexicon entries in each set.

Layer-specific Steering Direction. Finally, the steering direction is computed independently for each layer i over the full hidden dimension d. The steering direction is determined by the desired target attribute. To steer the model towards the positive attribute (enhancing $\mathcal{L}^+$ traits), the steering

direction is defined as:

$$\mathbf{v}_{\textbf{pos}}^{(i)} = \boldsymbol{\mu}^{+,(i)} - \boldsymbol{\mu}^{-,(i)}\,.$$

Conversely, to steer the model towards the negative attribute (or to suppress $\mathcal{L}^+$ traits), we reverse the direction: $\mathbf{v}_{\textbf{neg}}^{(i)} = -\mathbf{v}_{\textbf{pos}}^{(i)}$.

3.2 Feature Attribution

In the context of activation steering, probing is frequently employed to identify the model components that contribute most significantly to the target attribute, thereby enabling more targeted interventions. For instance, Li et al. (2023) and Wang et al. (2025) utilize probes to select specific attention heads for steering, while Konen et al. (2024) employ probing to determine the optimal layers for intervention. Building upon these approaches, we propose a two-stage feature attribution mechanism: we first identify the layers with the highest sentiment density, and then isolate a sparse subset of neurons within those layers that contribute most significantly to the sentiment representation.

Both stages are trained on the lexicon entry representations defined in Section 3.1. For each layer, we construct a labeled dataset $\mathcal{D}^{(i)} = \{(\phi^{(i)}(l), y_l) : l \in \mathcal{L}^+ \cup \mathcal{L}^-\}$, where $y_l = 1$ if $l \in \mathcal{L}^+$ and 0 otherwise. $\mathcal{D}^{(i)}$ contains the same entries and labels for every layer i and differs only in the layer at which the representations are taken. Since probing in isolation can capture non-causal correlations rather than true task relevance (Ravichander et al., 2021; Saphra and Wiegreffe, 2024), we conduct rigorous intervention experiments against random baselines, fully detailed in the Appendix I.

Layer-wise Probing. We hypothesize that sentiment information is non-uniformly distributed across the model's depth. To identify the most informative layers, we train a linear probe on $\mathcal{D}^{(i)}$ for each layer i to predict the sentiment polarity $y \in \{0, 1\}$ using a logistic regression objective:

$$P(y \mid i, l) = \sigma\left[\mathbf{w}_i^T \phi^{(i)}(l) + b_i\right]\,,$$

where $\mathbf{w}_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$ are the learnable weights and bias of the probe, respectively. The layers are then ranked according to the probes' **`macro-F1`** scores on the held-out portion of $\mathcal{D}^{(i)}$. Since the optimal subset of layers for intervention may vary across architectures and linguistic contexts, we consider three candidate layer configurations: all layers, the top 50% of layers, and the top 5 layers. For each combination of model, target language, and direction type (monolingual vs. multilingual), we use the configuration that yields the highest **Success Rate (SR)** after injection (Tables 1 and 2). We denote the resulting set of selected layers by $\mathcal{S}$.

Sparse Neuron Selection via Non-linear Attribution. Steering all neurons within a selected layer may introduce unnecessary noise. Our preliminary experiments with **`Qwen`** and **`Llama`** indicate that this selection is an empirical necessity; unconstrained steering frequently resulting in incoherent, 'gibberish' outputs. To mitigate this, we perform neuron-level attribution using Gradient Boosted Decision Trees (**`GBDT`**) via **`XGBoost`**. While sparsity-inducing linear probes (e.g., L1-regularized Logistic Regression) can reduce feature counts, they typically retain a broader set of neurons based on linear separability. In contrast, tree-based models perform a more aggressive, non-linear feature selection, isolating a highly compact subset of neurons highly discriminative for sentiment representation.

For each layer i, we train an **`XGBoost`** classifier on $\mathcal{D}^{(i)}$, using the individual neurons of $\phi^{(i)}(l)$, denoted by $\{n_{i,1}, n_{i,2}, \ldots, n_{i,d}\}$, as input features. We then calculate the *gain* of each neuron $n_{i,j}$, denoted by $G(n_{i,j})$, which measures the average improvement in the training objective over all split points in the ensemble that use that neuron:

$$G(n_{i,j}) = \frac{1}{|\mathcal{P}_{i,j}|} \sum_{p \in \mathcal{P}_{i,j}} \Delta\ell_p\,,$$

where $\mathcal{P}_{i,j}$ is the set of split points, across all trees, that split on neuron $n_{i,j}$; and $\Delta\ell_p$ is the improvement in the regularized log-loss objective achieved by split p, as computed by **`XGBoost`**'s split criterion. We set $G(n_{i,j}) = 0$ for neurons that are never used in a split ($\mathcal{P}_{i,j} = \emptyset$). For each selected layer $i \in \mathcal{S}$, we define the active steering set $\mathcal{N}_i$ as the indices of the neurons with a non-zero gain ($G(n_{i,j}) > 0$).

3.3 Sentiment Direction Injection

We apply the derived steering direction during the inference phase by intervening on the model's internal activations. The intervention is performed at every forward pass of the model, specifically targeting the hidden states corresponding to the final token of the current sequence.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_t)$ be the input sequence at the current generation step t, where x_t represents

the last token (the most recently generated or processed token), and let $h^{(i)}(x_t)$ denote the original hidden state at the final token position for layer i. For each selected layer $i \in \mathcal{S}$, we slice the full-dimensional steering direction $\mathbf{v}^{(i)}$, which can be either $\mathbf{v}^{(i)}_{\textbf{pos}}$ or $\mathbf{v}^{(i)}_{\textbf{neg}}$, to the active steering set $\mathcal{N}_i$, and add the scaled result to the corresponding neurons of the hidden state before it is passed to the subsequent layer $i+1$:

$$\tilde{h}^{(i)}_{[\mathcal{N}_i]}(x_t) \leftarrow h^{(i)}_{[\mathcal{N}_i]}(x_t) + \alpha \cdot \mathbf{v}^{(i)}_{[\mathcal{N}_i]},$$

where the subscript $[\mathcal{N}_i]$ denotes the sub-vector indexed by $\mathcal{N}_i$; $\tilde{h}^{(i)}(x_t)$ is the steered hidden state passed to the next layer; and α is the steering coefficient (hyperparameter) controlling the intervention strength. All neurons outside $\mathcal{N}_i$, as well as all layers $i \notin \mathcal{S}$, are left unchanged. By only intervening on the neurons in $\mathcal{N}_i$, we prevent the 'gibberish' output phenomenon observed in unconstrained steering, thereby maintaining the structural integrity of the latent space while maximizing steering efficacy.

4 Experiments

4.1 Experimental Setup

Models and Lexicon. We evaluate three base LLMs: **`Llama-3.2-3B`** (Grattafiori et al., 2024), **`Gemma-3-4B-PT`** (Gemma Team et al., 2025), and **`Qwen3-4B-Base`** (Yang et al., 2025). To derive steering directions, we utilize the **`NRC-VAD`** lexicon (Mohammad, 2018), which provides human-rated valence scores $[0, 1]$. We apply strict thresholds (> 0.8 for $\mathcal{L}^+$ and < 0.2 for $\mathcal{L}^-$) to maximize semantic contrast and ensure robust steering vectors. We compare two configurations: **Monolingual** (English only) and **Multilingual**, which aggregates and deduplicates lexicons across 11 typologically diverse languages spanning three resource tiers: high-resource (English, German, Chinese, Japanese), mid-resource (Indonesian, Icelandic, Estonian, Basque), and low-resource (Hausa, Javanese, Sundanese). Detailed preprocessing and model configurations are provided in Appendix A.

Task, Dataset Construction, and Evaluations. We evaluate cross-lingual steering on an open-ended generation task: given a neutral prompt, the model generates a continuation, and we measure whether steering shifts the sentiment of the generated text toward the target polarity while preserving the language of the prompt. To support this evaluation, we constructed a large-scale suite of 11,000 neutral prompts across 11 languages. We first generated 1,000 unique neutral English prompts using a hierarchical structure (10 categories $\times$ 10 topics $\times$ 10 prompts per topic), then translated them into 10 target languages using **`Gemini`**. For low-resource languages, we employed style-guided prompts (e.g., specifying Javanese Ngoko Lugu and Sundanese Basa Loma) to strictly maintain a natural, conversational style. To ensure baseline reliability, we manually audited a 10% random sample (100 prompts) of the unique English set (see Appendix A.2).

To assess steering, models generate continuations from these neutral points with a constraint of **at most** 128 **new tokens**. This setup ensures sufficient context for sentiment analysis without forcing a fixed generation length. A separate tuning split of 50 prompts was used for hyperparameter selection (see Appendix B for topic lists and prompts). Continuations are evaluated using **`GPT-4.1`** (OpenAI et al., 2024) as a sentiment judge and **`GlotLID`** (Kargaran et al., 2023) for language identification. The use of large language models for automated text annotation has been shown to outperform crowdworkers across various classification tasks (Gilardi et al., 2023), with specific validation demonstrating high human-alignment in sentiment and emotion analysis (Calderon et al., 2025; Ruder et al., 2025). To confirm this reliability extends to our specific low-resource settings, we conducted native-speaker validation audits (detailed in Appendix A.5). Based on these automated evaluations, we report three metrics: (1) *Sentiment Accuracy* (**`SA`**) and (2) *LID Rate* (**`LIDR`**) as constituent components, and (3) *Success Rate* (**`SR`**), which serves as our primary metric for analysis. The *Success Rate* represents a strict intersection, requiring a generation to satisfy **both** correct sentiment and correct language simultaneously. Steering intensity α was optimized via grid search on the tuning split (details in Appendix A).

5 Results and Analysis

5.1 Effectiveness of Cross-Lingual Steering

We first evaluate the impact of our steering directions on the model's generation. Figure 2 visualizes the sentiment distribution shifts for the **`Gemma-3`** model across 11 languages using the steering direction derived from the multilingual lexicon (see Figure 7 for the corresponding distributions utilizing

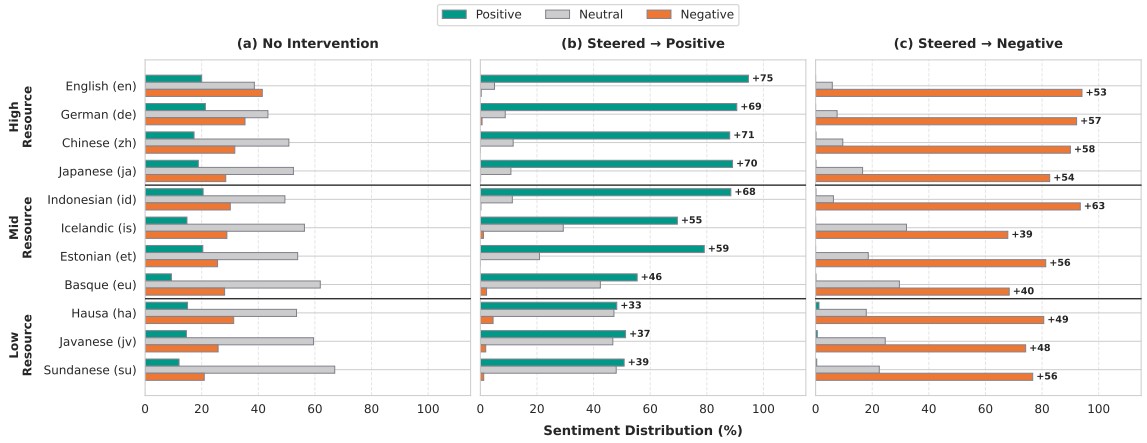


Figure 2: **Impact of Steering on Sentiment Distribution (Gemma-3).** The distribution of generated sentiments across 11 languages under three conditions: (a) No Intervention (as baseline), (b) Positive Steering, and (c) Negative Steering. The proposed method successfully shifts the predominantly neutral baseline toward the target sentiment. *Corresponding distributions for* **`Qwen-3`** *and* **`Llama-3.2`** *are provided in Appendix D.*

		Gemma-3		Qwen-3		Llama-3.2	
Lang	**Res.**	**SR (%)**	**Layers**	**SR (%)**	**Layers**	**SR (%)**	**Layers**
English (en)	High	94.3	all	45.1	all	65.8	all
German (de)	High	90.5	all	46.1	top-50%	53.4	all
Chinese (zh)	High	86.2	all	48.0	all	33.1	all
Japanese (ja)	High	76.8	all	23.5	all	18.9	all
Indonesian (id)	Mid	82.0	all	46.8	all	37.0	all
Icelandic (is)	Mid	66.8	all	10.9	all	23.1	top-50%
Estonian (et)	Mid	77.8	all	23.5	top-50%	25.4	top-50%
Basque (eu)	Mid	60.7	all	12.2	all	20.4	top-50%
Hausa (ha)	Low	58.9	top-50%	2.1	top-5	9.6	top-5
Javanese (jv)	Low	37.9	top-50%	9.8	top-50%	4.2	top-50%
Sundanese (su)	Low	27.2	all	5.7	top-50%	5.9	top-50%

Table 1: **Cross-Lingual Steering Performance with Multilingual Lexicons.** Success Rate (SR, %) across three models and 11 languages. Each cell reports SR at the best-performing layer configuration (all, top-50%, or top-5).

		Gemma-3		Qwen-3		Llama-3.2	
Lang	**Res.**	**SR (%)**	**Layers**	**SR (%)**	**Layers**	**SR (%)**	**Layers**
English (en)	High	95.0	all	44.5	all	66.4	all
German (de)	High	90.7	all	45.4	top-50%	53.8	all
Chinese (zh)	High	86.8	all	53.4	all	34.3	all
Japanese (ja)	High	80.2	all	23.9	top-50%	20.8	all
Indonesian (id)	Mid	82.3	all	42.9	top-50%	38.6	all
Icelandic (is)	Mid	62.3	top-50%	12.1	top-50%	20.0	all
Estonian (et)	Mid	78.0	top-50%	19.2	top-50%	22.1	top-50%
Basque (eu)	Mid	54.7	top-50%	9.2	top-50%	19.6	all
Hausa (ha)	Low	54.6	top-50%	1.8	top-5	9.3	top-50%
Javanese (jv)	Low	37.8	top-50%	10.2	top-50%	4.7	all
Sundanese (su)	Low	25.6	top-50%	4.7	top-50%	7.5	all

Table 2: **Cross-Lingual Steering Performance with Monolingual English Lexicons.** Success Rate (SR, %) across three models and 11 languages. Each cell reports SR at the best-performing layer configuration (all, top-50%, or top-5).

monolingual lexicon). With no intervention (Figure 2a), the model generates continuations based on its intrinsic probability distribution. Since the prompts are semantically neutral (e.g., 'He turned on the computer'), the resulting sentiments are naturally diverse or predominantly neutral. In contrast, applying the steering direction results in a dramatic distributional shift: Positive Steering (Figure 2b) successfully aligns the majority of generations with positive sentiment, while Negative Steering (Figure 2c) induces negative sentiment. These shifts indicate that lexicon-derived directions provide a viable basis for modulating sentiment across diverse languages. However, this responsiveness is not uniform across architectures. While the sentiment shift is nearly total in **`Gemma-3`**, we observe varying degrees of sensitivity in **`Qwen-3`** and **`Llama-3.2`** (see Appendix D). In these models, although the intervention significantly increases the target sentiment, they maintain a higher proportion of neutral generations, particularly in lower-resource languages such as Hausa (`ha`) and Sundanese (`su`). This suggests that model-specific biases and language-specific representation density play a role in the effectiveness of lexicon-derived steering.

To quantify this effectiveness, Tables 1 and 2 summarize the performance of Multilingual and Monolingual (english) steering directions. First, both strategies significantly alter the model's behavior to align the output with the target sentiment compared to the uncontrolled baseline. Second, the Multilingual Lexicon aggregation emerges as a critical enabler for mid- and low-resource languages. This is most evident in **`Gemma-3`**, where the Multilingual vector consistently outperforms the Monolingual counterpart in lower-resource settings. For instance, the Multilingual approach improves the success rate for Icelandic (`is`) from 62.3% to

66.8% (+4.5%) and for Basque (eu) from 54.7% to 60.7% (+6.0%). Even for Hausa (ha), we see an improvement of over 4 percentage points (54.6% to 58.9%).

Interestingly, in **Qwen-3**, we observe a notable boost in Indonesian (id), rising from 42.9% to 46.8%. This demonstrates that aggregating diverse lexicons creates a more robust, language-agnostic sentiment direction that helps bridge the representation gap in under-resourced languages. To ensure the robustness of these observed gains against stochastic decoding variance, we conducted rigorous statistical significance testing (Two-Sample Proportion Z-Tests) across all resource tiers. We find that the multilingual steering advantage is highly significant in Mid- and capable Low-resource regimes (see Appendix F). While the Multilingual direction is generally superior for low-resource languages across all models, the Monolingual vector remains more effective for specific high-resource pairs, most notably **Qwen-3** in Chinese (zh), where performance drops from 53.4% to 48.0% when using the Multilingual lexicon. In **Llama-3.2**, the differences between Mono and Multi are often marginal in high-resource settings (e.g., 66.4% vs 65.8% for English). These variations suggest that the optimal steering strategy is heavily influenced by the model's pre-training dynamics, which we analyze further in Section 5.3.

5.2 Sensitivity and Robustness Analysis

To determine the optimal steering intensity and evaluate the stability of our vectors, we conducted a hyperparameter tuning sweep on the coefficient $\alpha \in \{0.50, 0.75, 1.00\}$. We focused this analysis on a representative subset of languages ranging from high- to low-resource: English (en), German (de), Chinese (zh), Indonesian (id), and Javanese (jv). Figure 3 visualizes the results.

Fragility of High-Norm Monolingual Steering Directions. The left column of Figure 3 reveals that Monolingual steering is highly unstable at higher intensities. As $\alpha \to 1.0$, we observe widespread *linguistic collapse* (black cells), particularly in **Qwen-3** and **Llama-3.2**, where the model output becomes incoherent. We hypothesize that this fragility stems from the high magnitude (norm) of the Monolingual steering direction. Since English sentiment representations are highly separable, the resulting steering direction encodes a "sharp" and distinct direction with a large norm.

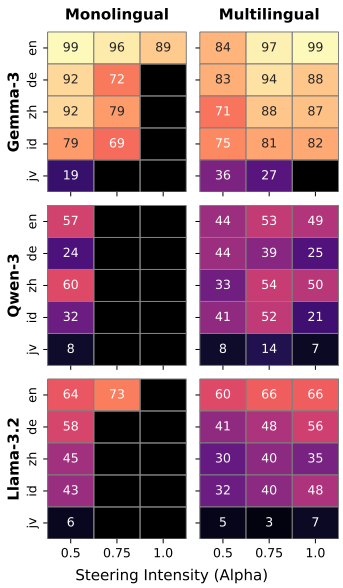


Figure 3: **Impact of Steering Intensity on Success Rate and Language Integrity.** Heatmaps show performance under the all-layers configuration as the steering coefficient α increases from 0.5 to 1.0, for steering vectors derived from monolingual (English) and multilingual lexicons. Values represent Success Rate (%); black cells indicate linguistic collapse, where the generated text was incoherent and thus could not be evaluated for sentiment.

When this high-norm vector is injected into the latent space of other languages, especially low-resource ones like Javanese where representations are less distinct, it pushes the activations far out of the natural distribution, causing the generation to break down.

Robustness and Norm Dampening in Multilingual Steering Directions. In contrast, the Multilingual steering directions (right column) generally demonstrate superior stability. By aggregating lexicons from diverse languages, the resulting vector represents a "consensus" direction. We posit that this aggregation process often creates a dampening effect on the vector norm, particularly when the component languages have varying alignment. This strictly moderated norm allows for stronger steering intensities without shattering the model's linguistic integrity.

Power vs. Stability in Gemma-3. **Gemma-3** demonstrates a unique power-stability trade-off: while achieving superior steering shifts, it is prone to linguistic collapse at $\alpha = 1.0$ (Figure 3). This fragility stems from a massive latent scale disparity: as measured across layers in Appendix G, **Gemma-3** operates in a high-variance regime ($L_2 >$

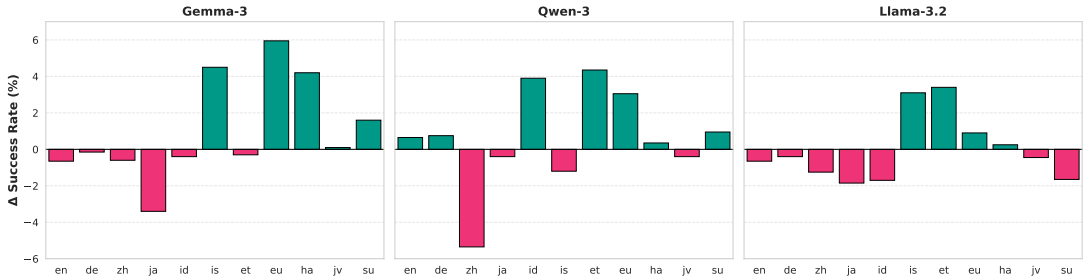


Figure 4: **Heterogeneous Impact of Multilingual Aggregation Across Model Families.** This figure highlights the divergent impact of the Multilingual steering direction across three distinct model families. Positive values (teal) indicate that the Multilingual direction outperforms the Monolingual direction.

$10,000$) compared to the compressed spaces of **`Llama-3.2`** (~ 25) and **`Qwen-3`** ($\sim 1,150$). Within **`Gemma-3`**, high-resource anchors (EN, ZH) define a "norm ceiling," while low-resource languages like Javanese occupy a compact subspace (see Figure 8). Consequently, the multilingual steering signal ($\sim 5,000$) is $2.5\times$ larger than the native Javanese norm ($\sim 2,000$). At maximum intensity, this disproportionate force shatters the linguistic integrity of lower-resource representations while remaining effective for high-norm languages. This suggests that universal steering requires intensity scaling relative to a language's native representational density.

5.3 Cross-Lingual Dynamics and Model Architecture

While Section 5.2 analyzed the stability of the steering vectors, this section examines their relative effectiveness. Figure 4 illustrates the performance difference, defined as:

$$\Delta = \textbf{SR}_{\text{Multi}} - \textbf{SR}_{\text{Mono}},$$

where $\textbf{SR}_m$ denotes the success rate for configuration $m \in \{\text{Multi}, \text{Mono}\}$ under its optimal layer configuration for that specific model and target language. Note that Δ is computed across 11 languages. The patterns across the three models highlight how model architecture and pre-training priors dictate the optimal steering strategy.

Architectural Drivers of Steering Efficacy. The patterns in Figure 4 reveal that a model's pre-training priors dictate its responsiveness to multilingual aggregation. **`Gemma-3`** exhibits the most consistent cross-lingual synergy, where multilingual vectors provide a coherent reinforcement signal for low-resource languages (e.g., `ha`, `su`) with minimal performance drops in high-resource languages: a benefit likely rooted in its optimized sub-word vocabulary. In contrast, **`Qwen-3`** incurs a "Specialization Penalty"; its bilingual-centric optimization for Chinese creates highly distinct representations that resist multilingual "noise," leading to a $> 5\%$ performance drop. For **`Llama-3.2`**, multilingual aggregation serves as a critical bridge for mid-resource languages (e.g., `is`, `et`) where the English vector is semantically distant, yet it remains insufficient to overcome the baseline "neutral anchor" in the lowest-resource settings (`jv`, `su`). This resistance occurs despite high cross-lingual vector alignment (see Appendix H), a paradox likely driven by the model's compressed latent scale and static layer-wise dynamics which create an internal numerical rigidity against external perturbations (see Appendix G).

6 Conclusion

We introduced multilingual sentiment lexicons as a scalable signal for cross-lingual activation steering, providing a practical alternative to sentence-level supervision where such data is unavailable. Across three model families and 11 languages, aggregated multilingual lexicons outperform English-derived steering vectors for mid- and low-resource languages, while monolingual lexicons remain preferable when a model is heavily specialized in a particular language. Our analysis further shows that successful intervention depends on internal model properties beyond surface-level similarity between sentiment directions: architectural specialization can decouple vector similarity from steering performance, and the scale mismatch between high-resource directions and compact low-resource representations can trigger linguistic collapse. These findings motivate a shift toward adaptive steering that calibrates intervention magnitude and lexicon configuration to each target language's representational scale.

Limitations

Task and Model Scope. Our experiments were conducted exclusively on **open-ended sentence completion** tasks using pre-trained base models. We did not evaluate the efficacy of these steering vectors on instruction-following tasks (e.g., QA, summarization) or on instruction-tuned checkpoints (Chat models). Consequently, it remains unverified whether the observed steering dynamics persist when the model is constrained by strict instruction formatting or safety alignment filters. Furthermore, our analysis is limited to **efficient-scale models** in the 3B–4B parameter range. It is unclear if the trade-offs between stability and potency scale linearly to larger models (e.g., 70B+ parameters), where the latent space typically exhibits greater robustness and separability.

Generalization to Unseen Languages. The Multilingual directions in this study were aggregated using lexicons from a specific set of 11 languages. We evaluated the steering performance only on these in-domain languages and did not assess zero-shot transferability to unseen languages (languages whose lexicons were excluded from the aggregation). Evaluating on unseen languages would be an interesting extension and would further test the universality of the resulting directions. That said, the practical impact of this limitation is modest: sentiment lexicons for new languages can be obtained reliably through word-level translation, which is substantially more accurate and less error-prone than translating full sentences. Extending the multilingual lexicon set to cover additional languages is therefore straightforward in practice, even if the present results do not directly demonstrate zero-shot transfer.

Evaluation Metrics and Reliability. Our evaluation relied primarily on automated classifier-based metrics to measure sentiment shift and language consistency. First, we did not perform a deep quantitative analysis of textual fluency (e.g., perplexity) or topical relevance, meaning high sentiment scores do not guarantee that the generated text maintains semantic coherence or "naturalness." Second, we acknowledge that the reliability of LLM-based evaluators (e.g., `GPT-4`) degrades in low-resource settings. These models may harbor English-centric biases or lack the nuance to accurately adjudge sentiment in languages like Sundanese or Javanese. Consequently, the reported success rates in these languages reflect the evaluator's probability distribution but may not perfectly correspond to native-level perceptions of sentiment.

Ethics Statement

Cultural Homogenization and Translation Artifacts. Our methodology relies on aggregating sentiment lexicons (`NRC-VAD`) which are constructed via automated translation from English without recalibrating valence scores for the target cultures. Consequently, the resulting "universal" vectors inherently propagate an **Anglocentric definition of sentiment**. By projecting English sentiment values onto other languages, we risk enforcing a homogenized alignment where concepts deemed "positive" in Western norms override local cultural nuances (e.g., definitions of politeness or positivity in Javanese that differ from English). Users of this method should be mindful that the "universality" achieved here is mathematical rather than cultural; the method prioritizes cross-lingual robustness over cultural precision, potentially risking cultural erasure in low-resource languages by aligning them to a foreign sentiment standard.

Dual-Use and Safety. While this study aims to improve cross-lingual control, activation steering is a dual-use technology. The same mechanisms used to enforce positive sentiment can be inverted to amplify toxicity or hate speech. Furthermore, as noted in our limitations, our current evaluation does not guarantee semantic coherence. Deploying such steering methods in production without rigorous fluency and safety checks could lead to the generation of superficially "positive" but nonsensical or hallucinated content, potentially degrading user trust.

Artifact and License Compliance. All artifacts used in this study are restricted to their intended research purposes. The code libraries are utilized under standard open-source licenses (e.g., MIT, Apache 2.0), and all datasets and pre-trained models are accessed in strict accordance with their original creators' terms of use.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

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

Nora Belrose. 2023. Diff-in-means concept editing is worst-case optimal. *EleutherAI Blog*.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. LLaMAs have feelings too: Unveiling sentiment and emotion representations in LLaMA models through probing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6124–6142, Vienna, Austria. Association for Computational Linguistics.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Christopher SG Khoo and Sathik Basha Johnkhan. 2018. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802, St. Julian's, Malta. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

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

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey. *Preprint*, arXiv:2408.12599.

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

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*, 1 edition. Synthesis Lectures on Human Language Technologies. Springer Cham.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024a. In-context vectors: making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024b. Fantastic semantics and where to find them: Investigating which layers of generative LLMs reflect lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14551–14558, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

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

Deniss Ruder, Andero Uusberg, and Kairit Sirts. 2025. Assessing the reliability and validity of GPT-4 in annotating emotion appraisal ratings. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 1–11, Albuquerque, New Mexico. Association for Computational Linguistics.

Sonal Sannigrahi and Jesse Read. 2022. Isomorphic cross-lingual embeddings for low-resource languages. *Preprint*, arXiv:2203.14632.

Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498, Miami, Florida, US. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. 2025. Analyzing the generalization and reliability of steering vectors. *Preprint*, arXiv:2407.12404.

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

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2562–2578, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Experimental Details

A.1 Model and Implementation

All models (**`Llama-3.2-3B`**, **`Gemma-3-4B-PT`**, and **`Qwen3-4B-Base`**) were implemented using the `transformers` library (Wolf et al., 2020) in `bfloat16` precision. We specifically utilize the **base (pre-trained)** versions to investigate sentiment representations in the foundational latent space, thereby avoiding biases from instruction-tuning or RLHF. Because all evaluated models are in the $\sim$3B to 4B parameter range, they require approximately 16 GB of VRAM for inference and activation steering. We distributed the experiments across several computing environments based on resource availability: publicly available cloud notebooks (dual NVIDIA T4, 2×16 GB), commercial cloud GPU instances (NVIDIA RTX A4500, 20 GB), and a shared academic compute cluster (NVIDIA DGX-A100).

A.2 Data and Lexicon Preprocessing

Lexicon Settings. As discussed in Section 3.1, we specifically utilize the multilingual **`NRC-VAD`** lexicon (Mohammad, 2018) to obtain the sentiment direction. The **`NRC-VAD`** lexicon provides human ratings for valence, arousal, and dominance, with scores ranging from 0 to 1. In our experiments, we investigate the efficacy of cross-lingual steering by comparing two distinct lexicon configurations: *monolingual* and *multilingual*.

For the *monolingual* configuration, we rely exclusively on English lexicon entries. The preprocessing involves removing invalid entries (**`NaN`** values) and filtering the remaining entries based on their valence scores. Entries with a valence score > 0.8 are assigned to the positive set $\mathcal{L}^+$, while entries with a valence score < 0.2 are assigned to the negative set $\mathcal{L}^-$. We deliberately employ these strict thresholds to maximize the semantic contrast between the opposing sets, ensuring that the resulting steering direction is robust and not diluted by neutral or weakly polarized terms.

For the *multilingual* configuration, we aggregate lexicon entries from 11 languages, selected to cover a spectrum from high-resource to low-resource languages: English (`en`), German (`de`), Chinese (`zh`), Japanese (`ja`), Indonesian (`id`), Icelandic (`is`), Estonian (`et`), Basque (`eu`), Hausa (`ha`), Javanese (`jv`), and Sundanese (`su`). The preprocessing for the aggregated set involves a rigorous deduplication process. First, we identify and remove ambiguous duplicates: entries appearing in multiple languages that exhibit conflicting valence scores (i.e., one instance having a score < 0.2 and another > 0.8). Second, for duplicates with consistent sentiment polarity (both < 0.2 or both > 0.8), we retain only the first occurrence. From this refined set of unique multilingual entries, we construct $\mathcal{L}^+$ and $\mathcal{L}^-$ using the same valence thresholds (> 0.8 and < 0.2) as the *monolingual* configuration.

Dataset Construction and Validation We constructed a multilingual evaluation suite consisting of 1000 neutral prompts across 11 languages. We utilized **`Gemini`** to generate 10 categories and 100 sub-topics, which served as the foundation for creating 1,000 neutral evaluation sentences. The full list of categories and prompts is provided in Appendix B. These were subsequently translated into 10 target languages, with specialized prompts for low-resource languages (Hausa, Javanese, Sundanese) to ensure linguistic naturalness. To ensure the neutrality of the baseline, we manually audited a random sample of 100 prompts. While the majority were neutral, we observed that approximately 3% contained implicit sentiment (e.g.,"The departure board showed the flight was 'delayed"' or "A client reported an issue with the software"). These instances highlight that while **`Gemini`** generates objectively factual statements, certain real-world contexts carry inherent negative or positive connotations.

Lang.	ISO	Llama-3.2	Qwen3-4B	Gemma3-4B
English	en	**8.47**	8.53	8.54
German	de	12.39	12.42	**10.80**
Chinese	zh	9.75	**7.62**	8.03
Japanese	ja	13.41	12.43	**9.84**
Indonesian	id	12.51	12.61	**9.41**
Icelandic	is	17.42	17.62	**15.14**
Estonian	et	13.33	13.44	**11.78**
Basque	eu	15.44	15.61	**13.53**
Hausa	ha	15.51	15.22	**13.68**
Javanese	jv	14.35	14.49	**12.64**
Sundanese	su	15.02	15.16	**12.54**

Table 3: Average tokens per neutral prompt across target languages. Lower values indicate higher tokenization efficiency and a more optimized vocabulary for that language.

A.3 Steering Methodology Details

Linear Probing. For layer selection, we split the activation dataset into training (70%) and validation (30%) sets using stratified sampling

with a fixed random state (1331). Probes were implemented as Logistic Regression models via `scikit-learn`, utilizing a `StandardScaler` and a `liblinear` solver with L1 regularization (C=1.0) initialized with the same fixed seed. Importantly, L1 regularization in this layer-probing stage is employed solely to mitigate overfitting in the high-dimensional activation space and improve probe generalization. The actual isolation of sparse neurons is handled exclusively by the non-linear **`XGBoost`** attribution.

Neuron Selection. For each layer, neuron-level attribution uses an **`XGBoost`** model (Chen and Guestrin, 2016) with a `binary:logistic` objective, trained on all lexicon entries of $\mathcal{D}^{(i)}$ for 100 boosting rounds with learning rate $\eta = 0.1$, maximum tree depth 3, and random seed 1331; all other hyperparameters are left at their library defaults (including $\lambda = 1$ and $\gamma = 0$). Neuron importance is computed with **`XGBoost`**'s `gain` importance, i.e., the average split gain over all splits that use the neuron. We retain only the neurons with non-zero gain, which yields a sparse set of neurons for steering.

A.4 Generation and Hyperparameters

For generation, we set a constraint of **at most** 128 **new tokens**. To ensure highly stable and reproducible sampling trajectories despite the stochastic nature of the decoding process, all generation runs were executed with a fixed global random seed (1331) initialized across `Python`, `NumPy`, and `PyTorch`. The specific sampling parameters were:

- **Llama:** Temperature $0.6, p = 0.9$.
- **Qwen:** Temperature $0.6, k = 20, p = 0.95$.
- **Gemma:** Default settings ($k = 64, p = 0.95$).

The steering intensity α was tuned via grid search on a 50-prompt tuning split. Based on the highest **Success Rate** (our primary metric), we selected $\alpha = 0.5$ for monolingual steering across all models. For multilingual steering, we used $\alpha = 0.75$ for **`Gemma`** and **`Qwen`**, and $\alpha = 1.0$ for **`Llama`**.

A.5 Validation of LLM-as-a-Judge for Multilingual Sentiment

Evaluating sentiment shifts in open-ended generative text across typologically diverse languages requires a metric capable of capturing complex, context-dependent semantics. While traditional lexicon-based scoring approaches offer a highly interpretable alternative (Taboada et al., 2011; Liu, 2012), they present severe structural limitations for our specific experimental setup. Our evaluation suite spans 11 languages, including highly agglutinative systems (e.g., Basque, Japanese, Indonesian, Javanese, Sundanese) and morphologically fusional languages (e.g., Estonian, Icelandic). In these typologies, sentiment-bearing roots undergo extensive affixation and internal stem mutations, causing static dictionary engines to suffer from catastrophic surface-form miss rates without flawless, language-specific lemmatizers.

Furthermore, lexicon approaches are inherently compositionally blind; they aggregate static token weights and struggle to resolve contextual polysemy or implicit sentiment distributed across narrative structures (Khoo and Johnkhan, 2018). To overcome these limitations, we utilize an LLM-based evaluator (**`GPT-4.1`**), which natively resolves morphological variants via sub-word tokenization and captures deep contextual dependencies.

To empirically assess the reliability of our evaluation metric and address broader cross-lingual limitations noted in recent literature (Xu et al., 2025), we manually audited the **`GPT-4.1`** judge across both high-resource (English) and low-resource (Javanese) settings. In both cases, 150 generated instances were sampled (50 per predicted sentiment class), fully randomized, and evaluated with the human annotator blinded to the model's predictions. The English baseline was audited by one of the authors, whereas the Javanese evaluation was conducted by a volunteer native speaker recruited from the authors' network to ensure a rigorous assessment of low-resource linguistic nuances. To ensure a fair comparison, the human annotator evaluated the text using criteria identical to the LLM judge (detailed in Appendix C). The annotator provided verbal consent for their labels to be used for research validation; no personal or demographic data was collected.

The detailed confusion matrices (Table 4) provide deeper insight into the model's error distribution across resource tiers. While there is an expected performance drop in the low-resource Javanese setting (achieving a macro-average precision of 78%), the matrix reveals that errors are predominantly boundary confusions between polar and neutral classes rather than severe polar inversions. For instance, the judge only predicted a

English (High-Resource)				
Predicted	**Human Annotation (Gold) Negative**	**Positive**	**Neutral**	**Precision**
Negative	**47**	0	3	94%
Positive	0	**50**	0	100%
Neutral	1	1	**48**	96%
			Macro-Average	**96.7%**
Javanese (Low-Resource)				
Predicted	**Human Annotation (Gold) Negative**	**Positive**	**Neutral**	**Precision**
Negative	**46**	1	3	92%
Positive	7	**36**	7	72%
Neutral	9	6	**35**	70%
			Macro-Average	**78.0%**

Table 4: Confusion matrices for the `GPT-4.1` sentiment judge across 150 randomized samples per language.

positive sentiment for a true negative sample in 7 out of 150 instances, maintaining excellent directional precision for negative sentiment shifts (92%). These empirical findings confirm that despite the structural complexities of low-resource languages, the contextual LLM judge provides a highly reliable metric for capturing directional sentiment shifts.

B Prompts and Dataset Topics

B.1 Prompts for Data Generation

We utilized specific prompts to generate the categories, topics, and neutral sentences. For low-resource languages (Hausa, Javanese, Sundanese), a specialized translation prompt was used to ensure stylistic accuracy.

1. Category and Topic Generation Prompt
I would like to conduct an experiment to examine whether a Large Language Model (LLM) can be steered to generate positive or negative continuations from neutral prompts. Please suggest 10 suitable topic categories, each containing 10 broad subcategories that can serve as a basis for constructing neutral prompts. I plan to create 10 prompts for each subcategory; therefore, the subcategories should remain general and widely applicable rather than overly specific. For reference, I prefer the subcategories to be phrased in a style similar to the examples below:

- **Preferred style (broad and illustrative):**
  1. Morning routine (waking up, breakfast, commuting)
  2. House chores (cleaning, cooking, laundry)
  3. Daily commute (bus, train, traffic)
  4. Leisure activities (reading, walking, hobbies)
  5. Communication (texting, calling, chatting)
  6. Time management (schedules, delays, appointments)
  7. Weather and environment (rain, heat, wind)
  8. Personal interactions (neighbors, family, friends)
  9. Weekend activities (shopping, rest, recreation)
  10. Unexpected daily events (forgotten items, late arrivals)

- **Less preferred style (too narrow or fragmented):**
  1. Morning routines
  2. Commuting (bus, train, car)
  3. Household chores (e.g., laundry, dishes)
  4. Running errands
  5. Grocery shopping
  6. The weather forecast
  7. Receiving mail or a package
  8. Walking in the neighborhood
  9. Preparing a simple meal
  10. The end of the day

2. Neutral Sentence Generation Prompt
I would like to conduct an experiment to examine whether a Large Language Model (LLM) can be steered to generate positive or negative continuations from neutral prompts.

To prepare the dataset, I need your help in creating these neutral prompts. I already have a selection of topics from the [category] category, consisting of the following topics: [topics]

Please generate [amount] neutral prompts for each topic. First, write each prompt in English, then translate it into [languages].

The output should be formatted in JSONL with a unique prompt_id, category, topic, and language codes, as shown below. The prompt_id should start from 1 and increment sequentially.

```
{“prompt_id”: 1, “category”: “morning routine”, “topic”: “waking up”, “en”: “The alarm clock went off at 6 AM.”, “id”: “Jam weker berbunyi pukul 6 pagi.”, “de”: “Der Wecker klingelte um 6 Uhr morgens.”, “zh”: “早上6点，闹钟响了。”}
```

Note: The topic details (i.e., explanations in parentheses) do not need to be included in the JSONL output.

3. Low-Resource Translation Prompt
I am building a parallel dataset for low-resource languages. Below is a list of JSON objects containing English sentences (“en").

Please update each JSON object by adding translations for the following keys:

1. “ha": Hausa (Standard).
2. “jv": Javanese (Style: Ngoko Lugu/Casual. Strictly NO Krama/Formal language).
3. “su": Sundanese (Style: Basa Loma/Familiar. Use daily conversation style, comparable to Javanese Ngoko).

Constraints:

- Output ONLY the final valid JSONL format.
- Do not add any conversational text before or after the data.
- Please enclose the final output in a single code block.

Input Data:
[JSONL dataset containing English neutral sentences]

B.2 List of Categories and Topics

Below is the complete list of 100 topics categorized by domain. Topics marked with an asterisk (*) were also selected for the preliminary tuning dataset (50 sentences).

1. Work & Professional Life
Starting the workday*, Team collaboration*, Managing tasks*, Workplace communication, Taking a break, Meetings, Skill development, Performance feedback, Problem-solving*, Ending the workday*.

2. Home & Household
Morning routine*, Household chores*, Meal preparation*, Home maintenance, Running errands, Family interactions, Managing finances*, Pet care, Receiving mail, Evening routine*.

3. Social & Relationships
Making plans*, Social gatherings*, Casual conversation*, Community involvement, Meeting new people, Family events*, Online social media, Maintaining contact, Discussing topics, Public interactions*.

4. Travel & Transport
Daily commute*, Planning a trip*, Air travel, Using public transport, Accommodation, Navigating a new place*, Eating out*, Sightseeing, Unexpected travel events*, Returning home.

5. Health & Wellness
Physical exercise*, Diet and nutrition*, Sleep patterns*, Medical appointments*, Mental well-being*, Personal hygiene, Feeling unwell, Taking medication, Rest and recovery, Health goals.

6. Education & Learning
Attending a class*, Studying*, Group projects, Taking an exam*, Doing homework, Interacting with instructors*, Using resources, Learning a new skill*, Class registration, School environment.

7. Technology & Digital Life
Using a smartphone, Working on a computer*, Email management, Software updates, Troubleshooting*, Online shopping, Digital entertainment*, Data management, Digital security*, Video calls*.

8. Hobbies & Leisure
Creative activities*, Outdoor pursuits*, Consuming media*, Games*, Attending events, Cooking and baking, Crafts, Following sports, Simple relaxation, Weekend plans*.

9. Commerce & Finance
Grocery shopping*, Retail shopping, Customer service*, Receiving a service, Public services, Comparing products*, Making a large purchase, Subscription services*, Banking tasks, Following a budget*.

10. Environment & Surroundings
The weather*, Changing seasons, Environmental noise*, Public spaces*, Recycling and waste*, Weather forecast, Natural events, Indoor environment, Local wildlife, The physical landscape*.

C Sentiment Evaluation Guidelines

To ensure a rigorous and fair comparison between the automated metric and human evaluation, both the **GPT-4.1** judge and the human annotator evaluated the generated texts using identical task descriptions and sentiment definitions.

C.1 LLM Judge Prompt

We used **GPT-4.1** to evaluate the sentiment of the generated texts. The model was provided with a system instruction to act as a precise classifier and a user prompt containing detailed definitions and constraints. The full prompt structure is provided below:

Sentiment Evaluation Prompt Template
System Message:
`You are a precise sentiment analysis classifier.`

User Message:
`You are an expert linguist acting as a sentiment judge.`

`The user will provide a text that may be in {language} or a mix of languages.`

Task:
`Analyze the text and classify its dominant sentiment based on the definitions below.`

`If the text contains multiple parts (e.g., instructions, questions, and narrative), focus on the sentiment conveyed by the` **`main narrative content`** `or the` **`generated statement`**`, ignoring purely instructional phrasing (like "Answer:", "Choose A/B", etc.).`

Definitions:
`- "Positive": The text expresses a positive evaluation, such as indicating that the situation, action, or object is good, helpful, safe, beneficial, desirable, comfortable, satisfying, or otherwise favorable. The evaluation may be explicit (e.g., "I am happy") or implicit (e.g., "the process is efficient").`

`- "Negative": The text expresses a negative evaluation, such as indicating that the situation, action, or object is bad, difficult, harmful, inconvenient, frustrating, disappointing, risky, undesirable, or otherwise unfavorable. The evaluation may be explicit (e.g., "I am angry") or implicit (e.g., "the system is confusing").`

`- "Neutral": The text is purely factual, informative, descriptive, or consists of neutral questions/instructions without emotional valence.`

Constraint:
`The answer must be exactly one word from this list: [Positive, Negative, Neutral]`

Text to analyze:
`{sentence}`

Answer:

C.2 Human Annotator Instructions

The native Javanese human annotator was provided with the exact same task description and definitions as the LLM judge. The only modifications were the removal of system framing (e.g., "You are a precise sentiment analysis classifier") and the substitution of the output constraint with spreadsheet-specific data entry instructions. The provided instructions are detailed below:

Human Annotator Instructions

Task:
Analyze the text and classify its dominant sentiment based on the definitions below.

If the text contains multiple parts (e.g., instructions, questions, and narrative), focus on the sentiment conveyed by the **main narrative content** or the **generated statement**, ignoring purely instructional phrasing (like "Answer:", "Choose A/B", etc.).

Definitions:
- "Positive": The text expresses a positive evaluation, such as indicating that the situation, action, or object is good, helpful, safe, beneficial, desirable, comfortable, satisfying, or otherwise favorable. The evaluation may be explicit (e.g., "I am happy") or implicit (e.g., "the process is efficient").
- "Negative": The text expresses a negative evaluation, such as indicating that the situation, action, or object is bad, difficult, harmful, inconvenient, frustrating, disappointing, risky, undesirable, or otherwise unfavorable. The evaluation may be explicit (e.g., "I am angry") or implicit (e.g., "the system is confusing").
- "Neutral": The text is purely factual, informative, descriptive, or consists of neutral questions/instructions without emotional valence.

Note:
Select the dropdown on the "Judgement" column (rightside of the "Generated Text" column) to put your judgment.

Thank you.

D Additional Sentiment Distribution Plots

In this section, we provide detailed sentiment distribution plots that complement the analysis in Section 5.1. First, we present the multilingual steering results for **Qwen-3** and **Llama-3.2**. Second, we provide a comparison against a monolingual (English-only) steering baseline for **Gemma-3**.

D.1 Multilingual Steering Impact on Sentiment Distribution Across Models

Our results demonstrate that multilingual steering directions consistently influence the sentiment distribution across all three models, shifting the probability mass toward the target sentiment compared to the baseline. However, the magnitude of this effect varies significantly. **Gemma-3** exhibits the highest influence, where the intervention almost completely overrides the default neutral behavior (see Figure 2). In contrast, **Llama-3.2** and **Qwen-3** show a more moderate response, often retaining a significant portion of neutral outputs even after steering.

D.2 Monolingual (English) Steering Impact on Sentiment Distribution

To better understand the distinct contribution of language-specific lexicons, Figure 7 visualizes the sentiment distribution shifts for **Gemma-3** using steering directions derived solely from English lexicons. The monolingual directions prove highly effective for high-resource and some mid-resource languages (e.g., German, Indonesian), occasionally matching or even slightly exceeding the multilingual approach, suggesting a strong inherent cross-lingual alignment within **Gemma-3**. However, in specific mid- to low-resource settings (such as Icelandic, Basque, and Sundanese), the English vectors exhibit steeper performance drops and struggle more to overcome the baseline neutral anchor. While the multilingual intervention (Figure 2) does not completely eliminate this neutral tendency, it demonstrably mitigates the deterioration, providing a more robust and resilient steering signal for the most resource-constrained languages.

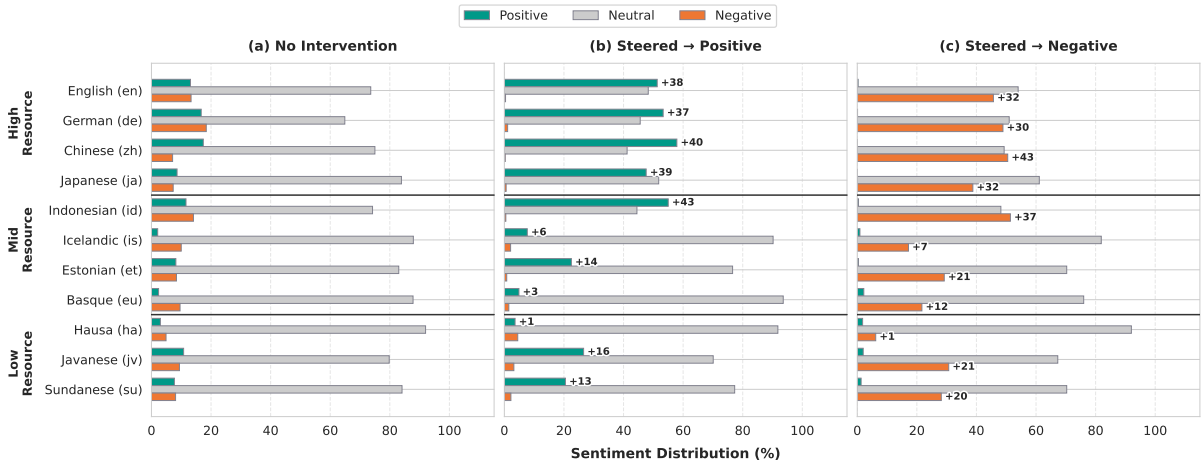


Figure 5: **Impact of Steering on Sentiment Distribution (Qwen-3).** Steering induces a noticeable shift toward the target sentiment, confirming the method's cross-model applicability. However, Qwen-3 exhibits a stronger tendency to preserve neutral tonality compared to other models, with neutral outputs remaining the dominant class in mid to low resource languages even after intervention.

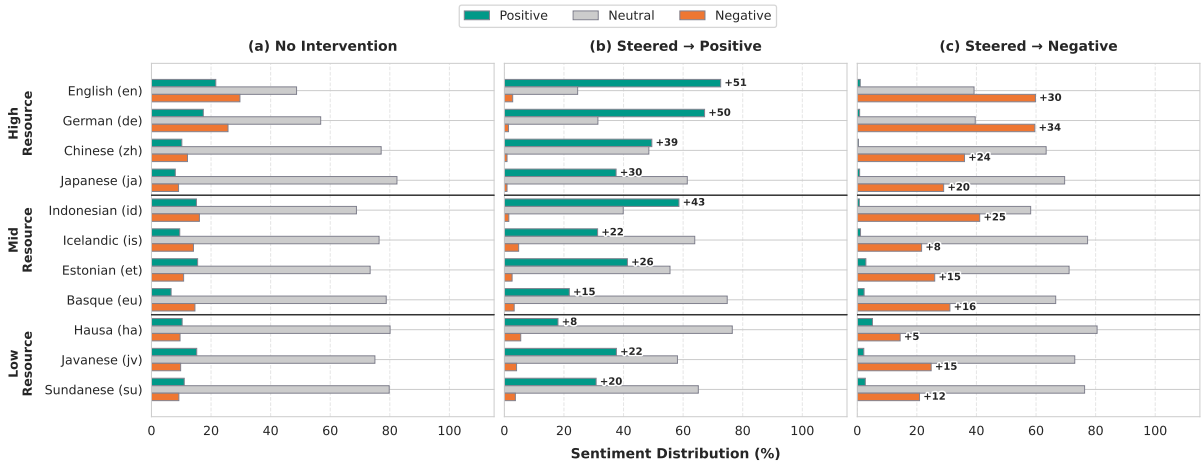


Figure 6: **Impact of Steering on Sentiment Distribution (Llama-3.2).** While the steering vectors successfully increase the probability of the target sentiment across all languages, Llama-3.2 retains a higher proportion of neutral outputs compared to Gemma-3, particularly in low-resource languages (e.g., Hausa, Javanese, Sundanese).

E Complete Steering Results with SA and LIDR

Tables 5 and 6 illustrate the comparative strengths of **`Gemma-3`**, **`Qwen-3`**, and **`Llama-3.2`**. Notably, **`Gemma-3`** maintains significantly higher Success Rates (SR) and Sentiment Accuracy (SA) in low-resource categories (e.g., Hausa, Javanese) compared to its counterparts in both monolingual and multilingual environments.

F Statistical Significance and Decoding Variance

In this section, we provide a rigorous statistical analysis of our evaluation results to confirm that the observed performance deltas between multilingual and monolingual steering directions are statistically significant and not artifacts of stochastic decoding variance. We also provide a granular analysis of model behavior in low-resource regimes.

F.1 Stochastic Decoding and Corpus-Level Stability

To evaluate the impact of our steering interventions, our experimental pipeline utilizes a massive sample footprint. For each language condition, we

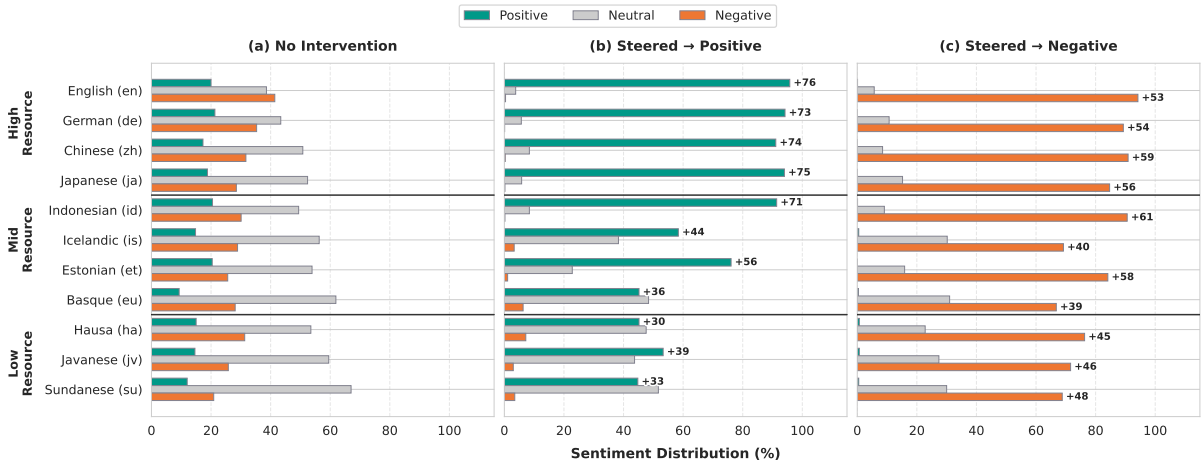


Figure 7: **Impact of Steering with Monolingual Steering Direction on Sentiment Distribution (Gemma-3).** The distribution of generated sentiments across 11 languages under three conditions: (a) No Intervention (as baseline), (b) Positive Steering, and (c) Negative Steering.

		Gemma-3				Qwen-3				Llama-3			
Lang	**Res.**	**SR**	**SA**	**LIDR**	**Config**	**SR**	**SA**	**LIDR**	**Config**	**SR**	**SA**	**LIDR**	**Config**
English (en)	High	95.0	95.0	99.8	all	44.5	54.6	84.1	all	66.4	67.0	98.6	all
German (de)	High	90.7	91.8	98.4	all	45.4	52.8	72.1	top-50%	53.8	64.2	83.0	all
Chinese (zh)	High	86.8	91.0	93.5	all	53.4	62.0	82.5	all	34.3	44.5	67.5	all
Japanese (ja)	High	80.2	89.3	88.6	all	23.9	31.3	57.5	top-50%	20.8	38.1	58.2	all
Indonesian (id)	Mid	82.3	91.0	89.8	all	42.9	47.9	85.1	top-50%	38.6	51.3	69.1	all
Icelandic (is)	Mid	62.3	63.8	97.6	top-50%	12.1	13.2	95.5	top-50%	20.0	24.1	89.6	all
Estonian (et)	Mid	78.0	80.1	97.5	top-50%	19.2	22.3	87.1	top-50%	22.1	29.5	80.2	top-50%
Basque (eu)	Mid	54.7	56.0	96.0	top-50%	9.2	10.2	92.0	top-50%	19.6	26.6	79.0	all
Hausa (ha)	Low	54.6	60.8	90.0	top-50%	1.8	5.1	61.8	top-5	9.3	19.8	74.0	top-50%
Javanese (jv)	Low	37.8	62.5	66.0	top-50%	10.2	30.9	49.1	top-50%	4.7	34.0	29.9	all
Sundanese (su)	Low	25.6	56.8	50.3	top-50%	4.7	24.8	44.8	top-50%	7.5	29.1	46.6	all

Table 5: **Best Results (Mono).** SR: Success Rate, SA: Sentiment Acc, LIDR: Language ID Rate. All metrics in %.

conduct $N = 2,000$ independent generation trials (derived from $1,000$ distinct prompts $\times$ 2 target sentiments).

While decoding parameters remain stochastic, evaluating across thousands of distinct, independent generation episodes leverages the Law of Large Numbers. Local, token-level variance effectively averages out at the corpus level. This stabilizes the observed success rates (SR), allowing them to closely converge toward the true expected performance of the steered model and minimizing the impact of seed-to-seed variance.

F.2 Macro- and Micro-Level Statistical Significance

To verify both the global and localized validity of our approach, we conducted Two-Sample Proportion Z-Tests to ensure that the observed performance deltas between multilingual and monolingual steering directions are mathematically meaningful, rather than products of stochastic decoding variance.

Macro-Level Aggregation. First, we aggregated the steering success rates by language resource level across all target models. Because each language tier aggregates multiple balanced conditions, the sample sizes provide immense statistical power (n = 8,000 trials for High/Mid tiers; n = 6,000 trials for the Low tier). Table 7 contrasts the Multilingual steering direction against the Monolingual (English) baseline, reporting the success rate for each condition, their difference (Δ SR, in percentage points), and the corresponding two-tailed test statistics.

The aggregate data demonstrate a distinct performance pattern: as language resource tiers drop from High to Mid/Low, the multilingual lexicon yields regular and mathematically meaningful ad-

Lang	Res.	Gemma-3 SR	Gemma-3 SA	Gemma-3 LIDR	Gemma-3 Config	Qwen-3 SR	Qwen-3 SA	Qwen-3 LIDR	Qwen-3 Config	Llama-3 SR	Llama-3 SA	Llama-3 LIDR	Llama-3 Config
English (en)	High	94.3	94.4	99.7	all	45.1	48.5	91.6	all	65.8	66.2	98.9	all
German (de)	High	90.5	91.4	98.5	all	46.1	51.1	78.8	top-50%	53.4	63.4	82.0	all
Chinese (zh)	High	86.2	89.0	95.0	all	48.0	54.2	85.8	all	33.1	42.8	71.5	all
Japanese (ja)	High	76.8	85.9	87.8	all	23.5	43.2	43.3	all	18.9	33.2	60.1	all
Indonesian (id)	Mid	82.0	91.0	88.9	all	46.8	53.2	83.2	all	37.0	49.9	69.8	all
Icelandic (is)	Mid	66.8	68.8	96.6	all	10.9	12.4	94.3	all	23.1	26.5	91.6	top-50%
Estonian (et)	Mid	77.8	80.2	96.8	all	23.5	25.9	90.8	top-50%	25.4	33.7	78.8	top-50%
Basque (eu)	Mid	60.7	61.9	95.2	all	12.2	13.3	91.3	all	20.4	26.5	84.7	top-50%
Hausa (ha)	Low	58.9	64.4	90.1	top-50%	2.1	4.9	60.9	top-5	9.6	16.2	75.7	top-5
Javanese (jv)	Low	37.9	62.7	64.5	top-50%	9.8	28.6	52.3	top-50%	4.2	31.2	33.5	top-50%
Sundanese (su)	Low	27.2	63.7	46.1	all	5.7	24.3	48.0	top-50%	5.9	25.9	47.9	top-50%

Table 6: **Best Results (Multi).** SR: Success Rate, SA: Sentiment Acc, LIDR: Language ID Rate. All metrics in %.

Model	Multi SR	Mono SR	Δ SR	Z-score	p-value
High-resource (4 langs, $n = 8{,}000$)					
Gemma-3	86.95	88.17	-1.22	-2.34	**0.0194**
Qwen-3	40.67	41.80	-1.13	-1.45	0.1466
Llama-3.2	42.80	43.83	-1.03	-1.31	0.1886
Mid-resource (4 langs, $n = 8{,}000$)					
Gemma-3	71.83	69.33	$+2.50$	$+3.47$	**0.0005**
Qwen-3	23.35	20.85	$+2.50$	$+3.81$	**0.0001**
Llama-3.2	26.48	25.08	$+1.40$	$+2.02$	**0.0429**
Low-resource (3 langs, $n = 6{,}000$)					
Gemma-3	41.33	39.33	$+2.00$	$+2.23$	**0.0255**
Qwen-3	5.87	5.57	$+0.30$	$+0.71$	0.4791
Llama-3.2	6.57	7.17	-0.60	-1.30	0.1938

Table 7: Two-sample proportion Z-tests comparing multilingual vs. monolingual (English) steering success rates (SR, %) by language resource tier. Bold p-values indicate rejection of the null hypothesis at $\alpha = 0.05$. Positive Δ SR favors the multilingual direction; negative favors the monolingual baseline.

vantages. We reject the null hypothesis across the board for the Mid-resource tier, and for **Gemma-3** in the Low-resource tier, where the base model possesses sufficient capacity to respond to cross-lingual steering. Notably, the sole significant result favoring the monolingual baseline occurs for **Gemma-3** in the High-resource tier (Δ SR $= -1.22$ pp, $p = 0.0194$). We interpret this reversal as consistent with our broader findings on architectural dynamics: where monolingual representations are already highly optimized, the multilingual lexicon direction introduces slight multilingual "noise" rather than a reinforcement signal, resulting in a marginal performance penalty.

Micro-Level Verification (Case Study: Gemma-3 Hausa). To confirm that the gains observed in viable low-resource models are mathematically sound at the individual language level, we performed a localized two-proportion z-test specifically on the **Gemma-3** Hausa results ($N = 2,000$ trials per condition):

- **Multilingual Baseline Success:** $1,178/2,000$ (58.9%)
- **Monolingual Baseline Success:** $1,092/2,000$ (54.6%)

This +4.3 pp margin yields a critical value of $z = 2.7447$, corresponding to a two-tailed p-value of 0.0060. This establishes that when baseline capacity thresholds are met, the multilingual lexicon's advantage is substantial and robust, clearing the $p < 0.05$ threshold by a wide margin, proving that the observed low-resource gains are not merely artifacts of stochastic decoding.

F.3 The Low-Resource Performance Spectrum in Latent Space

While the statistical significance holds firmly where models exhibit baseline capacity, individual language conditions in extreme low-resource regimes occasionally exhibit marginal absolute performance differences (e.g., $+0.1$ to $+0.3$ percentage points). Rather than indicating methodological failure, our analysis shows these variations reflect a clear architectural spectrum of base model capacity:

- **The Floor Effect (e.g., Qwen-3 Hausa):** At a success rate of just 1.8%, **Qwen-3** experiences a foundational downstream alignment failure in Hausa. Crucially, its baseline Language Identification Rate remains high (LIDR $> 60\%$), indicating the model recognizes the

linguistic context. However, the underlying semantic features are not sufficiently robust within the latent space to support targeted linear interventions. In this extreme regime, the steering vector lacks a mature geometric representation to act upon, resulting in statistically insignificant variance (e.g., a marginal $+0.3$ pp delta).

- **Neutral Transfer Equivalence (e.g., `Gemma-3` Javanese):** In contrast to the floor effect, **`Gemma-3`** possesses functional baseline capacity in Javanese, maintaining a stable success rate of roughly 37.8%. The marginal $+0.1$ pp delta represents a regime of *representation equivalence*. In this condition, both the cross-lingual multilingual vector and the monolingual vector achieve matching directional utility in the residual stream, resulting in stable, net-neutral transfer.

- **Positive Cross-Lingual Transfer (e.g., `Gemma-3` Hausa):** When a base model crosses the necessary capacity threshold to respond robustly to interventions in a low-resource language, the architectural advantage of the multilingual lexicon becomes highly visible, delivering the statistically significant boost of $+4.3$ percentage points verified in the previous section.

G Latent Space Geometry and Steering Norms

To understand the physical "intensity" of steering interventions, we calculate the L_2 norm of the steering direction vectors at each layer. As shown in Figure 8 and 9, the models exhibit vastly different latent space scales and layer-wise dynamics.

G.1 Scale and Specialization Ceiling

The absolute magnitude of steering vectors varies by three orders of magnitude between architectures, despite their comparable hidden dimensions (2560 neurons for **`Gemma-3`** and **`Qwen-3`**; 3072 for **`Llama-3.2`**). Specifically, **`Gemma-3`** operates in a high-variance regime with norms peaking over $10,000$ in its middle layers, whereas **`Qwen-3`** exhibits a moderate scale (peaking near $1,150$) and **`Llama-3.2`** utilizes a highly compressed space with norms remaining below 25. In all models, the norm magnitude is positively correlated with language resource levels and model specialization. The Chinese (`zh`) steering direction consistently defines the "norm ceiling" in **`Qwen-3`** ($\sim 1,150$) and **`Llama-3.2`** (~ 24), followed closely by English (`en`). This suggests that the model's most specialized features are encoded with the highest numerical weights.

The significant disparity in steering effectiveness between models can be partially attributed to this internal scaling of their latent spaces. As illustrated in Figure 8, the absolute magnitude of steering vectors in **`Gemma-3`** is several orders of magnitude larger than those in **`Llama-3.2`** and **`Qwen-3`**. This suggests a "stretched" or high-variance latent space where semantic features are represented with massive numerical weights. Such high-intensity representations may make specific features easier for the model's attention mechanisms to prioritize during token generation, explaining its responsiveness to sentiment steering.

Isolating Effective Intervention Magnitude. While Figure 8 illustrates the total latent scale, evaluating the non-zero gain vectors (Figure 9) reveals the concentrated "active" magnitude of the steering interventions. Filtering for non-zero gains reduces the L_2 norm magnitudes drastically across all architectures: **`Gemma-3`** drops from $> 10,000$ to peak values around 500. Similarly, **`Qwen-3`** reduces from $\sim 1,150$ to < 50, and **`Llama-3.2`** drops from ~ 25 to < 2.5. Despite this massive reduction in absolute magnitude, the hierarchical relation-

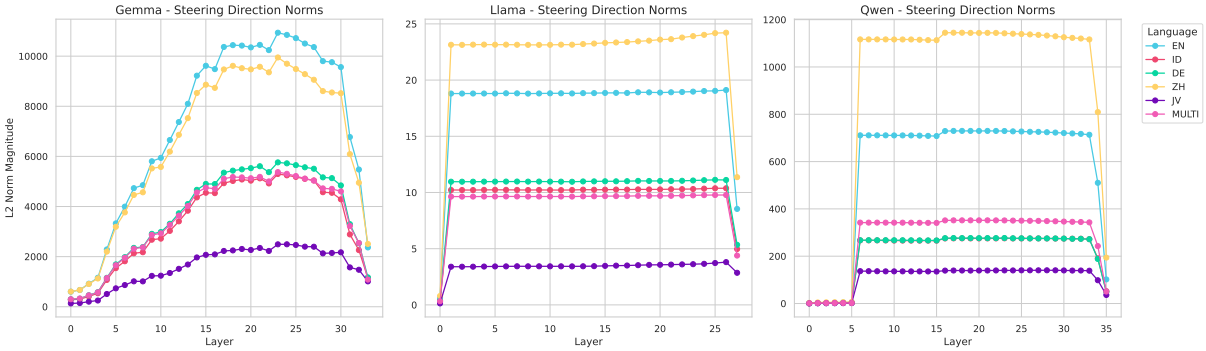


Figure 8: Comparison of L_2 norm magnitudes for sentiment steering directions (all neurons) across **Gemma-3**, **Llama-3.2**, and **Qwen-3**. Monolingual directions for German, Chinese, Indonesian, and Javanese are extracted and plotted solely to analyze the latent geometry and are not used for injection.

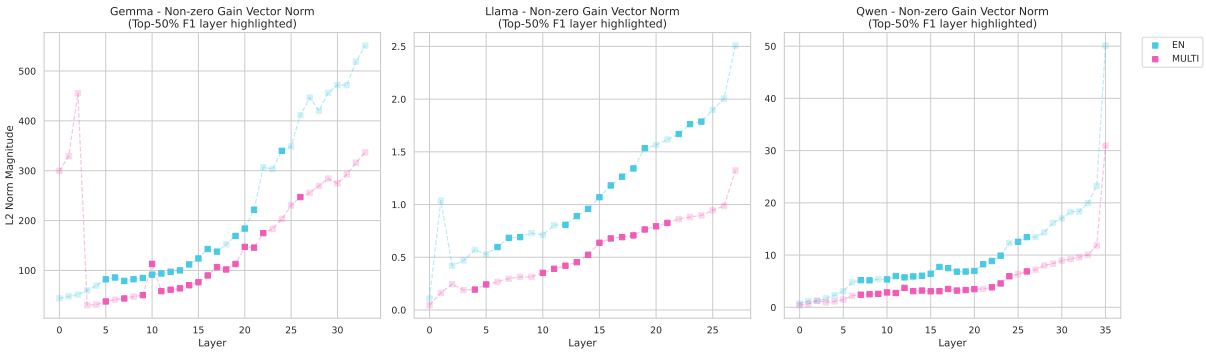


Figure 9: L_2 norm magnitude of the nonzero-gain steering vectors across layers for **Gemma-3**, **Llama-3.2**, and **Qwen-3**, comparing English and multilingual steering directions. Only neurons with non-zero **XGBoost** gain are included in the norm computation, reflecting the actual magnitude of intervention applied at inference time. Opaque markers denote layers in the top 50% by F1 score of the corresponding linear probe; faded markers denote layers outside this top half.

ship between the architectures (**Gemma** > **Qwen** > **Llama**) remains strictly preserved, confirming that **Gemma-3**'s high-variance regime is a fundamental architectural property rather than an artifact of inactive neurons.

G.2 The Compactness of Low-Resource Representations

In stark contrast to the high norm ceilings of English and Chinese, the Javanese (jv) direction possesses the smallest norm magnitude across all architectures: below 5 in **Llama-3.2** and below 200 in **Qwen-3**. This suggests that low-resource languages occupy a physically more compact region of the latent space. We hypothesize that the "linguistic collapse" observed at high steering intensities for Javanese occurs because external vectors (often scaled to high-resource norms like English) effectively "over-steer" these compact clusters, pushing activations far beyond their natural boundaries.

G.3 Layer-wise Dynamics

The models demonstrate distinct temporal profiles for sentiment representation:

- Static Encoding (**Llama** & **Qwen**): These models exhibit a "step-function" behavior. In **Llama-3.2**, the sentiment direction is fully established by Layer 1, and in **Qwen-3** by Layer 6, remaining static until the final layers. This implies sentiment is treated as a global, invariant attribute.
- Dynamic Progression (**Gemma**): **Gemma-3** shows a "bell-curve" progression where the norm builds through the initial layers, peaks between Layers 15 and 25, and tapers off toward the output. This indicates a more dynamic semantic refinement process.

While dense vectors exhibit static "step-functions" (**Llama**, **Qwen**) or "bell-curves" (**Gemma**) as shown in Figure 8, isolating the non-zero gain

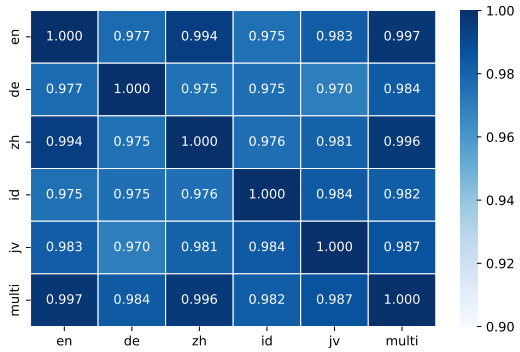
vectors reveals a different temporal dynamic. As illustrated in Figure 9, the non-zero gain magnitude steadily increases in the deeper layers across all three models. This indicates that while the global sentiment direction may be established early in the network, the concentrated, manipulable weights required to effectively steer that sentiment accumulate and scale up significantly toward the later layers.

G.4 Optimal Layer Configuration Asymmetry Between Steering Directions.

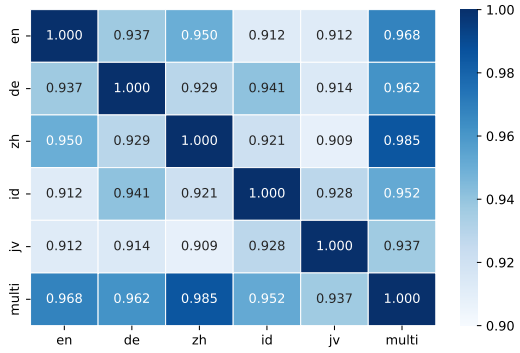
Another consequence of this norm disparity is visible in the optimal layer configurations reported in Tables 5 and 6. Across nearly all models and languages, the best-performing monolingual (English) steering direction requires fewer layers (e.g., top-5 or top-50%) to reach its optimal success rate, whereas multilingual directions more often require intervention across all layers. This phenomenon is visually supported by both our dense and sparse norm analyses. As shown in Figure 8, the English steering direction consistently exhibits a substantially larger L_2 norm than the multilingual direction at comparable layers. This disparity is further corroborated by the non-zero gain vectors (Figure 9), which explicitly map the optimal intervention zones. Within these optimal middle-to-late layers, the concentrated, non-zero magnitude of the English direction (blue) remains distinctly higher than that of the multilingual direction (pink) across all three architectures. Because the magnitude of the injected monolingual perturbation is definitively larger, fewer layers of intervention are needed to displace the model's activations sufficiently to shift sentiment. Conversely, the multilingual direction, being comparatively dampened in norm, requires broader layer-wise application to achieve its optimal steering efficacy.

Interestingly, this norm-layer relationship does not hold uniformly in **`Llama-3.2`** for several languages (Icelandic, Basque, Hausa, Javanese, Sundanese), where the optimal monolingual configuration in fact requires more layers than the multilingual counterpart. This anomaly may be attributed to the relatively high cross-lingual vector alignment observed in **`Llama-3.2`** between English and low-resource language (see Section H). For instance, the gap between English–Javanese similarity and Multilingual–Javanese similarity is markedly smaller in **`Llama-3.2`** ($0.970 - 0.968 = 0.002$) than in **`Gemma-3`** ($0.987 - 0.983 = 0.004$) or **`Qwen-3`** ($0.937 - 0.912 = 0.025$). Because the English steering direction in **`Llama-3.2`** is already nearly as aligned with low-resource languages like Javanese as the multilingual consensus direction is, monolingual injection is comparatively less destructive than in the other architectures. Consequently, it can tolerate broader layer-wise application without triggering the linguistic collapse observed elsewhere.

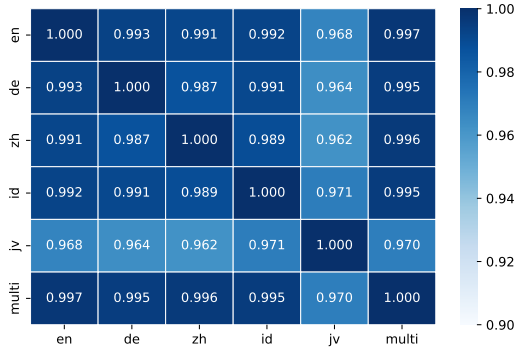
H Cross-Lingual Vector Alignment



(a) Gemma-3



(b) Qwen-3



(c) Llama-3.2

Figure 10: Cosine similarity heatmaps between monolingual and multilingual steering directions across three architectures. The displayed values represent the average cosine similarity across all layers.

H.1 The Universal Sentiment Axis

All three models exhibit high cross-lingual alignment, suggesting the existence of a "universal" sentiment axis within their latent spaces.

- **`Gemma-3`** maintains a high "consensus" direction, where the multilingual vector's similarity to all anchor languages ranges between 0.982 and 0.997.
- **`Qwen-3`** shows slightly more internal divergence, particularly regarding Chinese (`zh`), which maintains a 0.985 similarity to the multilingual vector, while the alignment for Javanese drops to 0.937.
- **`Llama-3.2`** demonstrates near-perfect uniformity, with the Multilingual (`multi`) vector achieving a cosine similarity of > 0.99 with English, German, Chinese, and Indonesian. Even the low-resource Javanese direction remains highly aligned at 0.970.

H.2 Multilingual Dampening Effect

The similarity data supports our "consensus" hypothesis regarding the robustness of multilingual steering. By aggregating diverse lexicons, the resulting vector moves toward a shared centroid, effectively dampening the high-norm "sharpness" characteristic of monolingual directions. This effect is most pronounced in **`Qwen-3`**/Chinese, where the specialized, high-intensity monolingual signal is diluted by the lower-norm profiles of the broader lexicon. While this dampening enhances stability across the "long tail" of languages, it accounts for the observed Specialization Penalty in high-resource anchors, where the reduced vector magnitude fails to meet the model's native representational threshold.

I Establishing the Causal Relevance of Targeted Intervention Sites

I.1 Experimental Setup

As briefly introduced in Section 3.2, we performed a causal intervention experiment to verify that our isolated layers and neurons actively govern model generation rather than inducing arbitrary changes or gibberish output. We conducted this experiment using the multilingual direction across two low-resource setups: **`Gemma-Hausa`** and **`Qwen-Javanese`**. We randomly sampled 50 prompts from our broader 1,000-prompt evaluation set and steered them toward both positive and negative targets, yielding 100 paired evaluation instances per model. To isolate the causal effect of our **`XGBoost`**-selected neurons, we established our proposed method against two rigorous null-hypothesis baselines:

- **Proposed Method (H_1):** Intervene strictly on the top-50% probe-ranked layers, using only the specific sparse neurons isolated by **`XGBoost`**.
- **Fully Random Baseline ($H_0^{(1)}$):** Intervene on a randomly sampled 50% of the layers across the architecture, using a completely random subset of neuron indices of identical size ($|\mathcal{N}_i|$).
- **Layer-Controlled Baseline ($H_0^{(2)}$):** Intervene on the exact same top-50% probe-ranked layers, but replace the **`XGBoost`**-selected coordinates with a random subset of neuron coordinates of identical size ($|\mathcal{N}_i|$).

I.2 Metrics and Statistical Significance

We evaluated causal efficacy using two metrics: **Sentiment Accuracy** (whether the generated sentiment matches the intended target) and **Strict Success Rate (SR)** (whether the text matches the target sentiment *and* retains the correct Language ID, ensuring no language-switching or gibberish occurred). Because the experimental outcomes are paired and binary, we applied McNemar's test to evaluate statistical significance against a standard threshold of $p < 0.05$.

I.3 Results and Discussion

Table 8 and Table 9 summarize the outcomes against the fully random and layer-controlled baselines, respectively. For both Gemma-Hausa and Qwen-Javanese, our proposed method (H_1) significantly outperforms both the fully random ($H_0^{(1)}$) and layer-controlled ($H_0^{(2)}$) baselines in Sentiment Accuracy ($p < 0.05$). By demonstrating that steering via these specific neurons changes model behavior in a predictable, task-relevant manner compared to random controls, we confirm that these representations are causally relevant, addressing concerns regarding correlative probing (Ravichander et al., 2021).

In Gemma-Hausa, our method yields a highly significant improvement in Strict Success Rate

Model & Language	Metric	H_1 (Ours)	$H_0^{(1)}$	McNemar p-value	Significance ($\alpha = 0.05$)
Gemma-Hausa	Sentiment Acc.	**66%**	3%	3.09×10^{-17}	**Significant**
	Strict SR	**56%**	0%	2.78×10^{-17}	**Significant**
Qwen-Javanese	Sentiment Acc.	**27%**	7%	0.0003	**Significant**
	Strict SR	**9%**	4%	0.2266	Not Significant

Table 8: Proposed Framework (H_1) vs. Fully Random Baseline ($H_0^{(1)}$). This baseline evaluates the effect of applying the steering vector to a random subset of layers (50% of all layers) and randomly sampled neuron indices, preserving the exact intervention size ($|N_i|$) of the original targeted configuration.

Model & Language	Metric	H_1 (Ours)	$H_0^{(2)}$	McNemar p-value	Significance ($\alpha = 0.05$)
Gemma-Hausa	Sentiment Acc.	**66%**	17%	5.91×10^{-12}	**Significant**
	Strict SR	**56%**	16%	1.51×10^{-9}	**Significant**
Qwen-Javanese	Sentiment Acc.	**27%**	12%	0.0107	**Significant**
	Strict SR	**9%**	6%	0.6072	Not Significant

Table 9: Proposed Framework (H_1) vs. Layer-Controlled Baseline ($H_0^{(2)}$). This baseline controls for layer location by intervening on the exact same top 50% probe-ranked layers but replacing our **XGBoost**-selected coordinates with random neuron indices.

($p < 10^{-8}$). Randomly perturbing neurons or layers frequently broke the Language ID or induced gibberish, whereas targeting our probe-selected neurons successfully flipped the sentiment while keeping the language generation intact.

For Qwen-Javanese, the strict success rate difference was not statistically significant. This aligns directly with our findings in Table 6, which demonstrates that Qwen has an overall Javanese baseline Success Rate of only 9.8%. Because the base model inherently struggles to maintain Javanese text generation, any representation steering is bottlenecked by the model's underlying language generation capabilities. Crucially, however, the Sentiment Accuracy remains statistically significant ($p = 0.0107$ and $p = 0.0003$), demonstrating that the targeted features do encode sentiment even when the base model's language constraints limit full generation success.

J Isolating the Multilingual Advantage from Vector Norm and Scaling Effects

Given the high cosine similarity between the monolingual and multilingual vectors as discussed in Section H, it is critical to determine whether the observed "multilingual advantage" is driven by true cross-lingual semantic directions or if it is an artifact of vector norm differences and alpha scaling (α). To isolate the geometric impact of the vector direction from these scaling factors, we present controlled experiments executed on **Gemma-3** for Hausa (ha). Hausa serves as a representative low-resource setting where the base model demonstrates robust performance, ensuring a reliable baseline environment to analyze steering dynamics.

J.1 Structural Divergence in Steering Pipelines

Our two-stage feature attribution mechanism independently derives the optimal layer configurations and their corresponding active neuron masks for both setups. Consequently, the selected subset of target layers inherently differs. As illustrated in Figure 11, the top-50% layer sets for the monolingual and multilingual directions exhibit structural divergence across all evaluated models. Specifically, the layer-wise Jaccard similarity is 0.70 (14 of 20 shared active layers) for **Gemma-3**, 0.40 (8 of 20 shared) for **Llama-3.2**, and 0.80 (16 of 20 shared) for **Qwen-3**.

Furthermore, even when both setups select the exact same layers, the specific active neurons targeted within those layers diverge substantially. To isolate this neuron-level divergence, we analyze the middle network layer for each model (using 0-based indexing): Layer 16 for **Gemma-3**, Layer 13

for **Llama-3.2**, and Layer 17 for **Qwen-3** (Figure 12). We chose these specific middle layers because they are independently identified as active target layers by both the monolingual and multilingual setups. By examining layers where the selection perfectly aligns, we provide an objective baseline to demonstrate that layer-level agreement does not guarantee neuron-level agreement. Within these shared layers, the active neuron masks exhibit minimal overlap (all $J \leq 0.34$). This proves that the two setups steer fundamentally distinct internal processing pathways, rather than the multilingual vector acting as a merely scaled version of the English circuit.

J.2 Norm-Controlled Baselines

To control for differences in scale, we first rescaled the English vector to match the Euclidean norm of the multilingual vector,

$$v_{\text{scaled}} = \frac{\|v_{\text{multi}}\|_2}{\|v_{\text{mono}}\|_2} \cdot v_{\text{mono}}.$$

Using this vector, we formalize the validation of the multilingual advantage through the following comparative configurations.

- **Original Multilingual Setup (H_1):** Intervene using the multilingual direction vector v_{multi} within the independently optimized multilingual layer/neuron configuration at $\alpha_{\text{multi}} = 0.75$.
- **Norm/α-Controlled Baseline ($H_0^{(A)}$):** Intervene using the rescaled monolingual English vector v_{scaled} within the native monolingual layer/neuron configuration at α_{multi}.
- **Configuration-Swapped Baseline ($H_0^{(B)}$):** Inject the rescaled monolingual English vector v_{scaled} directly into the multilingual layer/neuron configuration at α_{multi}.

J.3 Results and Statistical Significance

We evaluated these baselines against the evaluation dataset consisting of 1,000 unique base prompts. Steering each prompt toward two distinct sentiment targets yields a total of $N = 2000$ paired text evaluations per condition (Table 10).

We tested whether the multilingual advantage persists after controlling for vector scale and structural placement. Specifically, we used McNemar's test for paired nominal data, comparing binary outcomes (correct vs. incorrect sentiment generation) across identical prompt instances. We set the threshold for statistical significance at $p < 0.05$.

Setup	Vector	L/N Config	Scale	SA	SR
Mono	v_{mono}	Mono	Tuned	60.8	54.6
$H_0^{(A)}$	v_{scaled}	Mono	α_{multi}	61.3	56.1
$H_0^{(B)}$	v_{scaled}	Multi	α_{multi}	61.4	56.5
H_1 **(Multi)**	v_{multi}	**Multi**	**Tuned**	**64.4**	**58.9**

Table 10: Performance of norm-matched and configuration-swapped baselines on **Gemma-3** (Hausa). Metric scores (SA, SR) are in percentages (%). Statistical significance ($p < 0.05$) across identical prompt instances was achieved for Sentiment Accuracy (SA) against both baselines, but not for Steering Rate (SR).

- H_1 **vs.** $H_0^{(A)}$**:** This evaluates whether the multilingual direction yields an advantage over a perfectly norm- and scale-matched English vector operating in its native architecture zone. McNemar's test yields $p = 0.0285$. Since $p < 0.05$, we reject $H_0^{(A)}$, confirming that the multilingual advantage is statistically significant over a norm-corrected monolingual setup.
- H_1 **vs.** $H_0^{(B)}$**:** This evaluates if the advantage holds when the rescaled monolingual vector is forced into the identical multilingual configuration. McNemar's test yields $p = 0.0374$. Since $p < 0.05$, we reject $H_0^{(B)}$. Even within the cross-lingually optimized pathway, the rescaled English direction underperforms the true multilingual direction.

These results demonstrate that while the norm effect is present (rescaling the English vector yields a minor gain in sentiment accuracy ($60.8\% \rightarrow 61.3\%$)), it cannot fully close the performance gap. In high-dimensional LLM activation spaces, near-collinear cosine similarities can mask orthogonal directional components. Minor directional shifts in this high-dimensional space carry a unique cross-lingual utility that cannot be replicated by scalar adjustments to a monolingual vector.

K Empirical Analysis of Incoherent Outputs

During our investigations, severe linguistic collapse (manifesting as non-words, punctuation loops, or degenerate repetitive sequences) emerged under two distinct experimental contexts. To clarify the

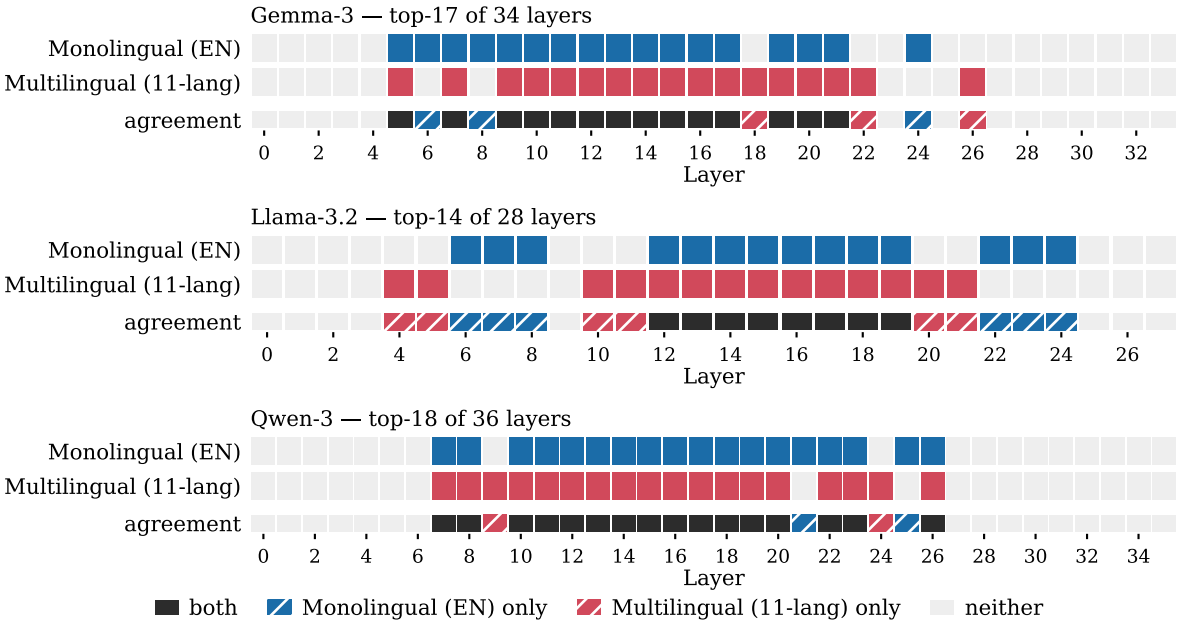


Figure 11: Layer selection divergence across **Gemma-3** (top-17 of 34), **Llama-3.2** (top-14 of 28), and **Qwen-3** (top-18 of 36). Blue and red striped regions indicate layers active exclusively in the Monolingual (EN) and Multilingual (11-lang) setups, respectively, while black indicates agreement.

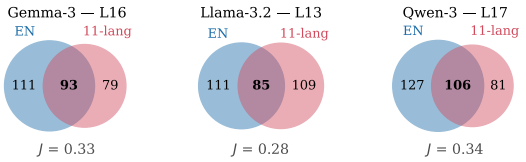


Figure 12: Neuron activation divergence between the Monolingual (EN) and Multilingual (11-lang) directions within the representative middle layer of each model. The minimal overlap across all three architectures confirms that the setups utilize distinct active neuron subnetworks.

failure modes discussed in the main paper, we decouple our preliminary baseline observations from the formal hyperparameter sweep.

1. **Ablation of Sparse Selection (Dense Intervention):** In Section 3.2 (the Sparse Neuron Selection via Non-linear Attribution paragraph), we argue that dense intervention (steering all neurons indiscriminately) injects excessive noise into foundational syntax circuits. As shown in the top half of Table 11, preliminary baselines applying vectors across all neurons consistently resulted in syntax breakdown. Because this dense configuration universally failed to produce evaluable text even at low intensities ($\alpha \leq 0.5$), it was excluded from the main evaluation sweep.

2. **Magnitude Oversaturation (Figure 3 Black Cells):** Even when successfully isolating features via sparse selection (e.g., non-zero gain neurons), excessive steering multipliers force the model outside its operational distribution. In Figure 3, these autoregressive collapses are visually encoded as black cells. We observe that the Monolingual vector shows a higher tendency to trigger collapse at $\alpha \geq 0.75$ across many non-English languages, whereas the Multilingual vector remains comparatively robust.

Table 11 compiles representative outputs illustrating both failure regimes. To clearly showcase these degradation patterns, all examples in this table utilize the Monolingual (en) steering direction. Driven by its larger vector norm (see Figure 8), it is particularly susceptible to both dense-induced collapse and magnitude oversaturation.

Model	Layer Config	Neuron Selection	α	Lang	Translated Prompt	Degenerate Output
			Dense Intervention: Disabling Sparse Selection			
Llama3.2-3B	top-5	all	0.5	en	He turned on his computer.	`\nTheLG-HeCJHG-4HEL-200D-J Question-Question-`
Llama3.2-3B	top-5	all	0.3	zh	He put on his running shoes.	`-- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -.. - - - - - - - - - - - - -........ - ........................ .......`
Qwen3-4B	top-5	all	0.2	id	The alarm went off at 6:30 AM.	`„„„ ordered of,bbbdon of ordered of of of of, of monn,bonitnakne bononononnebit of, amongonnebnebnebvitit, of are amongnebpitneby bpbnenebb ofbneonbonbbbhnepnenebhon itnepnebpitnedithitonnebf uccitititbbbb atvnenenelenelepitbononit hitneb`
Qwen3-4B	top-5	all	0.2	de	He looked at his phone screen.	`\n \n \n \n\n \n\n\n \n \n\n \n \n \n \n\n \n \n\n \n \n\n\n \n \n \n \n \n \n \n \n \n......... \n \n \n \n\n \n \n \n \n \n \n \n...... \nd\n\nd\n... \n \n \n... \n \n \n \n \n \n \n.........`
			Magnitude Oversaturation: Figure 3 Black Cells			
Gemma3-4B	all	non-zero gain	1.0	jv	She opened the spreadsheet to track expenses.	`Nalokir sing bisaukak d-r-s-s2 - ing - er or u\nAnus\n`
Qwen3-4B	all	non-zero gain	1.0	jv	The building stood next to the road.	`Nak Dak Dakakakak, Dak Dak Dak D, D D, D is the, D D, D In, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D, D; D, D, D; D, D, D; D, D; D, D, D; D, D`
Llama3.2-3B	all	non-zero gain	1.0	jv	I sent a text message about meeting up.	`Itu makan. Bukan. I. But. I. Not. I. I. Not. I. Not. I. Not. Not. I. Not. I. Not. Not. I. Not. Not. I. Not. Not. Not. I. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not. Not.`

Table 11: Representative degenerate outputs illustrating the two failure modes discussed in the paper. All examples utilize the Monolingual steering vector. Top: Linguistic collapse caused by dense intervention (steering all neurons), excluded from the main sweep. Bottom: Collapse stemming from magnitude oversaturation on sparse features ($\alpha = 1.0$), which corresponds directly to the unevaluable black cells in Figure 3.