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

**Anonymous ACL submission**

## 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 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. We further find that steering vectors built from high-resource languages have large magnitudes that overwhelm the compact representations of low-resource languages, causing "linguistic collapse" in the generated text. Multilingual aggregation mitigates this by producing smaller, more balanced vectors, at the cost of weaker control in specialized languages. Our findings establish lexicons as a practical signal for sentiment control in low-resource languages and offer a mechanistic account, grounded in vector magnitude, for understanding when monolingual or multilingual steering will succeed.[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., 2024; 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 steering, and use this framework to investigate *when and why* sentiment control transfers across languages. Our contributions are as follows:

- **A lexicon-based framework for cross-lingual sentiment steering.** We introduce multilingual sentiment lexicons as a scalable, resource-efficient signal for activation steering, removing the dependence on sentence-level supervision that limits existing methods. We show that aggregated multilingual lexicons outperform English-derived steering vectors for most mid- and low-resource languages across three model families, while monolingual lexicons remain stronger in narrower

[1] Code will be provided in camera-ready.

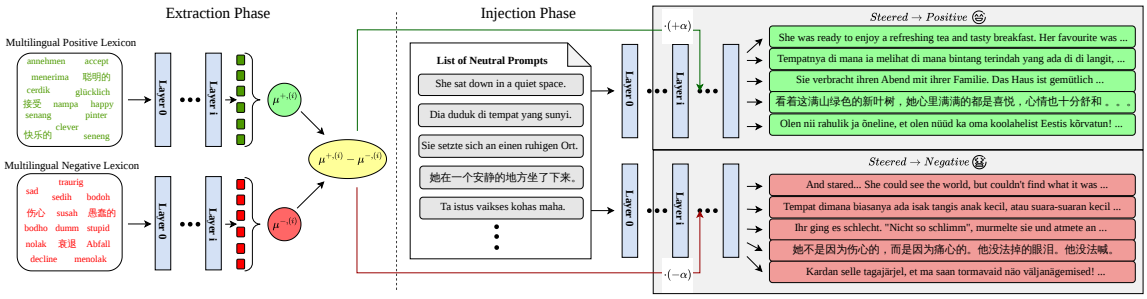



Figure 1: **Workflow of the Cross-Lingual Sentiment Control using Multilingual Lexicon.** The methodology is divided into two stages: the Extraction Phase, where activation differences between multilingual positive and negative lexicons are calculated at a specific layer (Layer $i$) to derive the steering vector $\mu^{+,(i)} - \mu^{-,(i)}$. In the Injection Phase, this vector is applied to neutral prompts using steering coefficients ($+\alpha$ or $-\alpha$) to influence the model's latent representations. The resulting "Steered" outputs demonstrate the model's ability to shift the emotional tone of neutral text into highly positive or negative sentiments across multiple languages

cases where a model is heavily specialized in a particular language.

- **An empirical comparison of monolingual and multilingual lexicon steering across 11 languages and three model families.** We provide the systematic comparison of these two steering configurations on a typologically diverse set of languages spanning high-, mid-, and low-resource tiers.
- **A mechanistic explanation of when steering succeeds or fails based on vector magnitude.** We show that sentiment vectors extracted from low-resource languages have much smaller magnitudes than those from high-resource languages. Steering vectors built from English lexicons have large magnitudes that overwhelm low-resource representations, breaking the generated text. Multilingual lexicon aggregation produces smaller, more balanced vectors that avoid this mismatch, explaining why it works better for low-resource languages than English-only steering.

## 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., 2024), 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 $\{(x_i, y_i)\}_{i=1}^{n}$:

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

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. 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.

**Cross-Lingual Alignment and Universality.** While the Linear Representation Hypothesis suggests stable sentiment directions within a single language, the cross-lingual universality of these directions 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.

However, recent studies indicate that pre-training on unbalanced corpora often results in an "Anglocentric" internal structure, where high-resource languages align closely with English subspaces, while mid-to-low resource languages suffer from weaker alignment or representation degeneration (Ahuja et al., 2023; Lauscher et al., 2020). This representational gap poses a challenge for cross-lingual steering: if the "sentiment axis" in a low-resource language is orthogonal to or disjoint from the English axis, monolingual steering will fail. In this context, we propose that aggregating 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 poor alignment, such aggregation may fail to overcome the lack of semantic grounding.

## 3 Methodology

### 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 $t_i \in \mathcal{V}$ 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.

**Layer-wise Centroids.** We compute the centroid for the positive and negative groups by averaging the phrase 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, $|\mathcal{L}|$ represents the number of lexicon entries in each set.

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

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

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

### 3.2 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 $h^{(i)}(t_{\textsf{last}})$ denote the original hidden state activation at the final token position $t$ for layer $i$. The steering direction $\mathbf{v}^{(i)}$, which can be either $\mathbf{v}_{\textsf{pos}}^{(i)}$ or $\mathbf{v}_{\textsf{neg}}^{(i)}$, is injected into this activation before it is passed to the subsequent layer $i+1$. We modify this activation by adding the scaled steering direction:

$$\tilde{h}^{(i)}(t_{\textsf{last}}) \leftarrow h^{(i)}(t_{\textsf{last}}) + \alpha \cdot \mathbf{v}^{(i)} \,,$$

where $\tilde{\mathbf{h}}_t^{(i)}$ is the steered hidden state passed to the next layer; and $\alpha$ is the steering coefficient (hyperparameter) controlling the intervention strength.

### 3.3 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. This approach aims to analyze layers with high sentiment density and subsequently isolate a sparse subset of neurons that contribute most significantly to the sentiment representation. The probing and attribution processes are conducted using the vector representations of lexicon entries $\phi^{(i)}(l)$ as defined in Section 3.1.

**Layer-wise Probing.** We hypothesize that sentiment information is non-uniformly distributed across the model's depth. To identify the most informative layers, we conduct linear probing on the representations $\phi^{(i)}(l)$ for each layer $i$. We train a linear probe 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 ranked according to their `macro-F1` scores on a held-out set of lexicon representations. Since the optimal subset of layers for intervention may vary across different architectures and linguistic contexts, we evaluate three layer configurations (all layers, the top 50%, and the top 5). We select the configuration that yields the highest **Success Rate (SR)** for each specific combination of model, target language, and steering direction (Monolingual vs. Multilingual), as detailed in Tables 1 and 2.

**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. In this stage, the individual neurons of the lexicon representation from $\phi^{(i)}(l)$, denoted by $\{n_{i,1}, n_{i,2}, n_{i,3}, \ldots, n_{i,j}\}$, serve as the input features.

We calculate the *gain* for each neuron $n_{i,j}$, denoted by $G(n_{i,j})$, which measures the average reduction in training log-loss contributed by that specific neuron across all split points in the ensemble:

$$G(n_{i,j}) = \frac{1}{T} \sum_{t=1}^{T} \Delta\mathcal{L}(n_{i,j,t}) \,,$$

where $T$ denotes the total number of trees; and $\Delta\mathcal{L}(n_{i,j,t})$ represents the improvement in the objective function attributed to neuron $n_{i,j}$ in tree $t$. We define the active steering set $\mathcal{N}_i$ by selecting only neurons with a non-zero gain ($G(n_{i,j}) > 0$). By only intervening on 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). 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

| Lang | Res. | Gemma-3 SR (%) | Gemma-3 Layers | Qwen-3 SR (%) | Qwen-3 Layers | Llama-3.2 SR (%) | Llama-3.2 Layers |
|---|---|---|---|---|---|---|---|
| English (en) | High | 94.3 | all | 45.1 | all | 65.8 | all |
| German (de) | High | 90.5 | all | 46.1 | top50% | 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 | top50% |
| Estonian (et) | Mid | 77.8 | all | 23.5 | top50% | 25.4 | top50% |
| Basque (eu) | Mid | 60.7 | all | 12.2 | all | 20.4 | top50% |
| Hausa (ha) | Low | 58.9 | top50% | 2.1 | top5 | 9.6 | top5 |
| Javanese (jv) | Low | 37.9 | top50% | 9.8 | top50% | 4.2 | top50% |
| Sundanese (su) | Low | 27.2 | all | 5.7 | top50% | 5.9 | top50% |

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).

| Lang | Res. | Gemma-3 SR (%) | Gemma-3 Layers | Qwen-3 SR (%) | Qwen-3 Layers | Llama-3.2 SR (%) | Llama-3.2 Layers |
|---|---|---|---|---|---|---|---|
| English (en) | High | 95.0 | all | 44.5 | all | 66.4 | all |
| German (de) | High | 90.7 | all | 45.4 | top50% | 53.8 | all |
| Chinese (zh) | High | 86.8 | all | 53.4 | all | 34.3 | all |
| Japanese (ja) | High | 80.2 | all | 23.9 | top50% | 20.8 | all |
| Indonesian (id) | Mid | 82.3 | all | 42.9 | top50% | 38.6 | all |
| Icelandic (is) | Mid | 62.3 | top50% | 12.1 | top50% | 20.0 | all |
| Estonian (et) | Mid | 78.0 | top50% | 19.2 | top50% | 22.1 | top50% |
| Basque (eu) | Mid | 54.7 | top50% | 9.2 | top50% | 19.6 | all |
| Hausa (ha) | Low | 54.6 | top50% | 1.8 | top5 | 9.3 | top50% |
| Javanese (jv) | Low | 37.8 | top50% | 10.2 | top50% | 4.7 | all |
| Sundanese (su) | Low | 25.6 | top50% | 4.7 | top50% | 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).

$\alpha$ 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 our multilingual 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. However, this advantage is not universal. 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**, 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

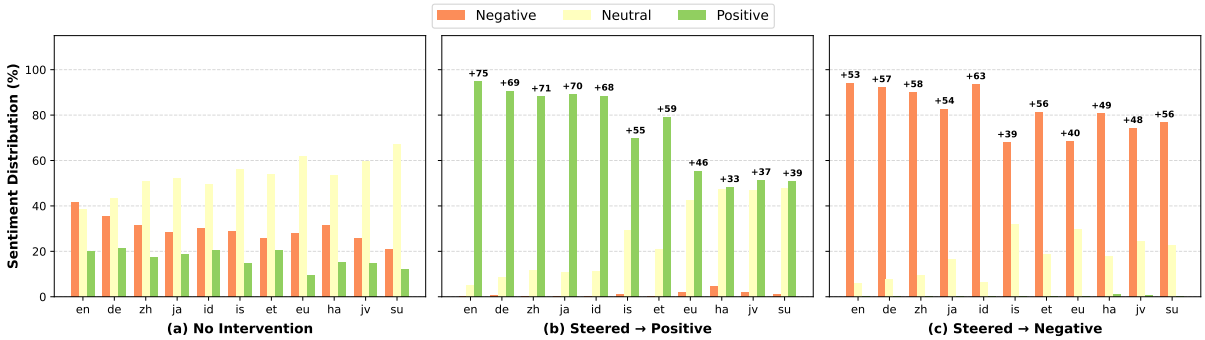



Figure 2: **Impact of Steering on Sentiment Distribution (Gemma-3).** The distribution of generated sentiments across 11 languages under three conditions: (a) No Intervention (as baseline), (b) Positive Steering, and (c) Negative Steering. The proposed method successfully shifts the predominantly neutral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

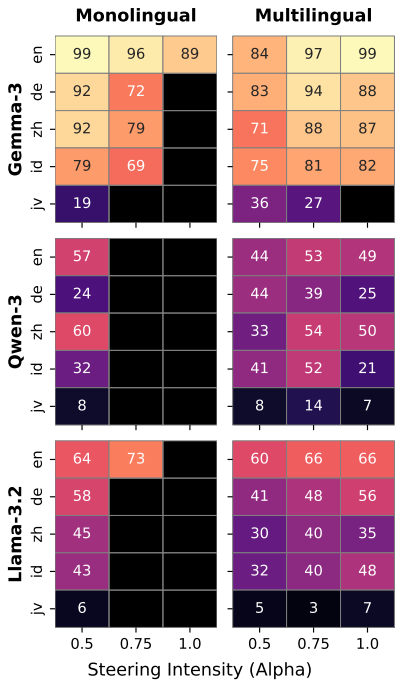



Figure 3: **Impact of Steering Intensity on Success Rate and Language Integrity.** Heatmaps show performance as the steering coefficient $\alpha$ increases from 0.5 to 1.0, for steering vectors derived from monolingual (English) and multilingual lexicons. Values represent Success Rate (%); black cells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

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 $\alpha \rightarrow 1.0$, 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. 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 F, `Gemma-3` operates in a high-variance regime ($L_2 >$

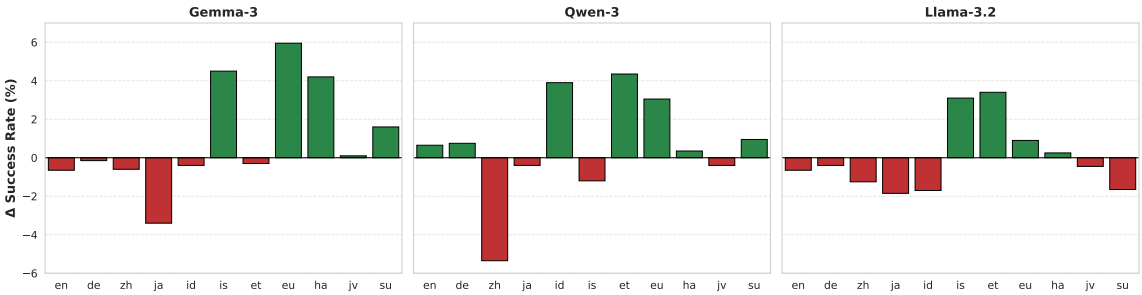



Figure 4: **Heterogeneous Impact of Multilingual Aggregation Across Model Families.** This figure highlights the divergent impact of the Multilingual steering direction across three distinct model families. Positive values (green) indicate that the Multilingual direction outperforms the Monolingual direction.

$10,000$) compared to the compressed spaces of Llama-3.2 ($\sim 25$) 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 7). 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:

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

where $\mathbf{SR}_m$ denotes the success rate for configuration $m \in \{\text{Multi}, \text{Mono}\}$. Note that $\Delta$ is computed across 11 languages. The patterns across the three models highlight how model architecture and pretraining priors dictate the optimal steering strategy.

**Architectural Drivers of Steering Efficacy.** The patterns in Figure 4 reveal that a model's pretraining 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 G), 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 F).

## 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 verified 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.

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.

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.

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.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. 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.

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.

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.

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.

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.

## 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 linguistice 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. We evaluated the performance of our **`GPT-4.1`** sentiment classifier by auditing 50 samples from each predicted class (Positive, Negative, Neutral). The results demonstrate high reliability. **Positive:** 100% (50/50) accuracy. **Negative:** 94% (47/50) accuracy; 3 samples were found to be neutral. **Neutral:** 96% (48/50) accuracy; 1 sample was negative and 1 was positive.

| 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. Probes were implemented with `StandardScaler` and a `liblinear` solver using L1 regularization ($C = 1.0$).

**XGBoost Refinement.** The neuron selection process utilized an **`XGBoost`** classifier with `n_estimators=50` and a `binary:logistic` objective. We only retained neurons that yielded a non-zero importance gain to ensure a compact and effective steering direction.

### A.4 Generation and Hyperparameters

For generation, we set a constraint of **at most 128 new tokens**. 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 $\alpha$ 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`**.

# 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 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:**

## D Additional Sentiment Distribution Plots

In this section, we provide the detailed sentiment distribution results for the other two models evaluated in our study: **Qwen-3** (Figure 5) and **Llama-3.2** (Figure 6). These results complement the analysis provided in Section 5.1 regarding **Gemma-3**.

**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.

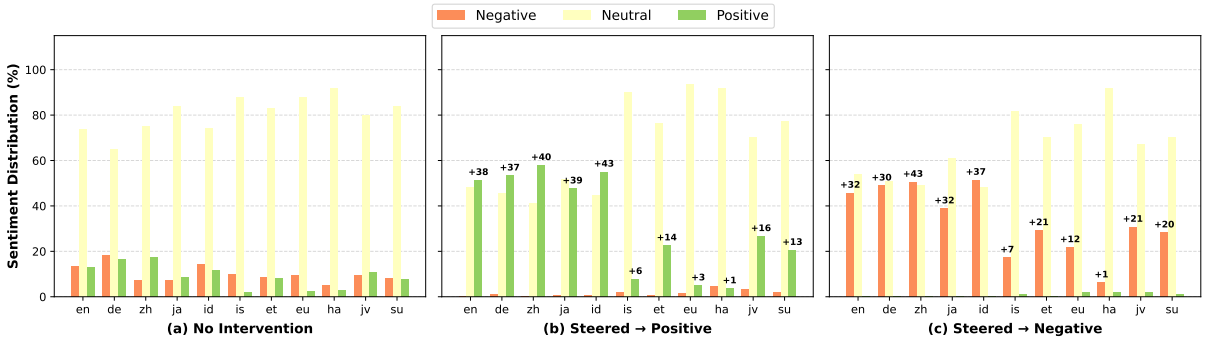



Figure 5: **Impact of Steering on Sentiment Distribution (Qwen-3).** Steering induces a noticeable shift toward the target sentiment, confirming the method's cross-model applicability. However, Qwen-3 exhibits a stronger tendency to preserve neutral tonality compared to other models, with neutral outputs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

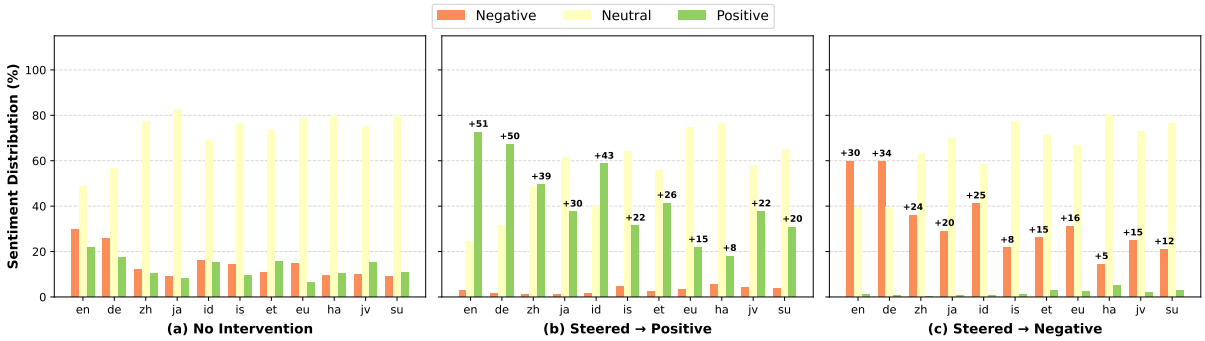



Figure 6: **Impact of Steering on Sentiment Distribution (Llama-3.2).** While the steering vectors successfully increase the probability of the target sentiment across all languages, Llama-3.2 retains a higher proportion of neutral outputs compared to Gemma-3, particularly in low-resource languages (e.g.,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

## E Complete Steering Results with SA and LIDR

Tables 4 and 5 illustrate the comparative strengths of Gemma-3, Qwen-3, and Llama-3. 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 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 7, the models exhibit vastly different latent space scales and layer-wise dynamics.

### F.1 Scale and Specialization Ceiling

The absolute magnitude of steering vectors varies by three orders of magnitude between architectures:

- **`Gemma-3`** operates in a high-variance regime, with norms peaking at over $10,000$ in its middle layers.
- **`Qwen-3`** exhibits a moderate scale, peaking at approximately $1,150$.
- **`Llama-3.2`** utilizes a highly compressed or normalized space, with norms remaining below $25$.

| Lang | Res. | Gemma-3 | | | | Qwen-3 | | | | Llama-3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SR | SA | LIDR | Cfg | SR | SA | LIDR | Cfg | 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 | top50% | 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 | top50% | 20.8 | 38.1 | 58.2 | all |
| Indonesian (id) | Mid | 82.3 | 91.0 | 89.8 | all | 42.9 | 47.9 | 85.1 | top50% | 38.6 | 51.3 | 69.1 | all |
| Icelandic (is) | Mid | 62.3 | 63.8 | 97.6 | top50% | 12.1 | 13.2 | 95.5 | top50% | 20.0 | 24.1 | 89.6 | all |
| Estonian (et) | Mid | 78.0 | 80.1 | 97.5 | top50% | 19.2 | 22.3 | 87.1 | top50% | 22.1 | 29.5 | 80.2 | top50% |
| Basque (eu) | Mid | 54.7 | 56.0 | 96.0 | top50% | 9.2 | 10.2 | 92.0 | top50% | 19.6 | 26.6 | 79.0 | all |
| Hausa (ha) | Low | 54.6 | 60.8 | 90.0 | top50% | 1.8 | 5.1 | 61.8 | top5 | 9.3 | 19.8 | 74.0 | top50% |
| Javanese (jv) | Low | 37.8 | 62.5 | 66.0 | top50% | 10.2 | 30.9 | 49.1 | top50% | 4.7 | 34.0 | 29.9 | all |
| Sundanese (su) | Low | 25.6 | 56.8 | 50.3 | top50% | 4.7 | 24.8 | 44.8 | top50% | 7.5 | 29.1 | 46.6 | all |

Table 4: **Best Results (Mono).** SR: Success Rate, SA: Sentiment Acc, LIDR: Language ID Rate. All metrics in %.

| Lang | Res. | Gemma-3 | | | | Qwen-3 | | | | Llama-3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SR | SA | LIDR | Cfg | SR | SA | LIDR | Cfg | SR | SA | LIDR | 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 | top50% | 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 | top50% |
| Estonian (et) | Mid | 77.8 | 80.2 | 96.8 | all | 23.5 | 25.9 | 90.8 | top50% | 25.4 | 33.7 | 78.8 | top50% |
| Basque (eu) | Mid | 60.7 | 61.9 | 95.2 | all | 12.2 | 13.3 | 91.3 | all | 20.4 | 26.5 | 84.7 | top50% |
| Hausa (ha) | Low | 58.9 | 64.4 | 90.1 | top50% | 2.1 | 4.9 | 60.9 | top5 | 9.6 | 16.2 | 75.7 | top5 |
| Javanese (jv) | Low | 37.9 | 62.7 | 64.5 | top50% | 9.8 | 28.6 | 52.3 | top50% | 4.2 | 31.2 | 33.5 | top50% |
| Sundanese (su) | Low | 27.2 | 63.7 | 46.1 | all | 5.7 | 24.3 | 48.0 | top50% | 5.9 | 25.9 | 47.9 | top50% |

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

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** ($\sim 24$), 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 7, 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.

### F.2 The Compactness of Low-Resource Representations

Conversely, 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.

### F.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

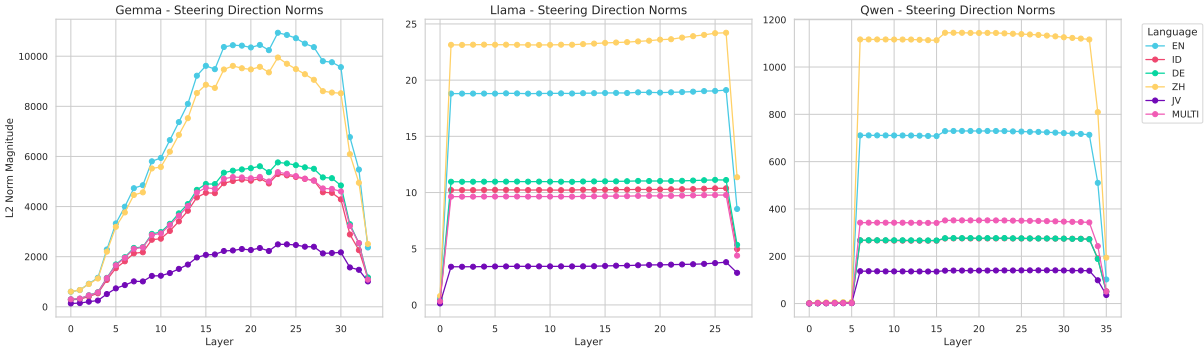
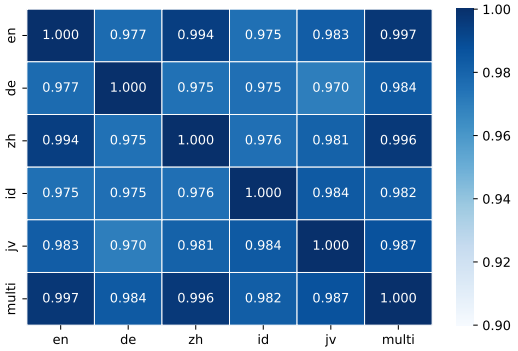



Figure 7: Comparison of $L_2$ norm magnitudes for sentiment steering directions across **Gemma-3**, **Llama-3.2**, and **Qwen-3**.

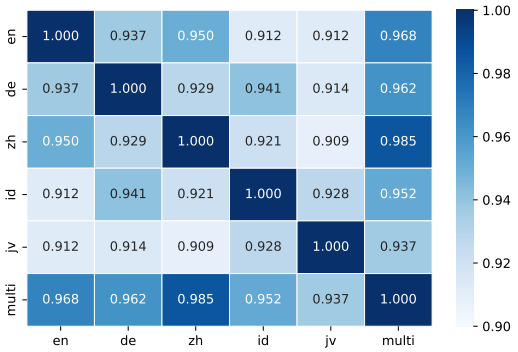
6, remaining static until the final layers. This implies sentiment is treated as a global, invariant attribute.

- 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.

# G Cross-Lingual Vector Alignment




(a) Gemma-3




(b) Qwen-3

(c) Llama-3.2

Figure 8: Cosine similarity heatmaps between monolingual and multilingual steering directions across three architectures.

## G.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.

## G.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.