Cross-Lingual Activation Steering for Multilingual Language Models

Rhitabrat Pokharel[1], Ameeta Agrawal[1], Tanay Nagar[2]

[1]Department of Computer Science, Portland State University, USA
{pokharel,ameeta}@pdx.edu

[2]Carnegie Mellon University
tnagar@andrew.cmu.edu

Abstract

Large language models exhibit strong multilingual capabilities, yet significant performance gaps persist between dominant and non-dominant languages. Prior work attributes this gap to imbalances between shared and language-specific neurons in multilingual representations. We propose Cross-Lingual Activation Steering (CLAS), a training-free inference-time intervention that selectively modulates neuron activations. We evaluate CLAS on several multilingual downstream tasks and achieve improvements while maintaining high-resource language performance. We discover that effective transfer operates through functional divergence rather than strict alignment, and performance gains correlate with increased language cluster separation. Our results demonstrate that targeted activation steering can unlock latent multilingual capacity in existing models without modification to model weights.

1 Introduction

Large Language Models (LLMs) perform well on many multilingual tasks, but a substantial performance gap remains between dominant (high-resource) and non-dominant (low-resource) languages. This gap is largely attributed to the heavy skew of pre-training corpora toward dominant languages, which enables models to develop richer representations in those languages (Zhang et al., 2024; Xu et al., 2025a).

To address this problem, prior work has explored data- and training-based solutions such as multilingual instruction tuning (Chen et al., 2024b; Shaham et al., 2024), supervised fine-tuning (Chen et al., 2024a), and model alignment using smaller multilingual datasets (She et al., 2024; Gao et al., 2024; Pokharel et al., 2025). Although effective, these methods depend on annotated datasets or additional training. A complementary line of work studies multilingual behavior at the neuron level as a lightweight alternative. Researchers have identified shared, partially shared, and language-specific neurons (Wang et al., 2024; Tang et al., 2024), and shown that shared neurons and overlapping subspaces in middle and upper layers play a central role in cross-lingual transfer (Tezuka and Inoue, 2025; Xu et al., 2025b). At the same time, direct manipulation of language-specific neurons has produced mixed results (Mondal et al., 2025). Other work suggests that multilingual models often reason through an English-like latent space before generating outputs in the target language (Etxaniz et al., 2024; Zhao et al., 2024a).

Building on these insights, we introduce Cross-Lingual Activation Steering (CLAS), a test-time neuron-level intervention that rebalances shared and language-specific representations during inference. CLAS gently suppresses the dominant shared neurons that otherwise overwhelm finer linguistic signals, while selectively amplifying or attenuating language-specific neurons to steer the model toward the desired target language. This targeted reweighting improves performance on non-dominant languages without any additional data or parameter updates. Figure 1 provides an overview of our proposed approach.

While Mondal et al. (2025) also study test-time neuron interventions, their method has an important limitation: it replaces selected neuron values with fixed constants, ignoring the input and removing useful variation. In contrast, our approach keeps the model's original signal and adjusts it proportionally, so the result still depends on the input. Similarly, INCLINE (Wang et al., 2025) requires tuning a separate hyperparameter for each language and task, which is time- and resource-intensive, while our method only needs tuning once per task and modifies a small number of layers.

Our main contributions are as follows:

- We propose CLAS, a training-free activation

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 815–827
October 17-21, 2026 ©2026 Association for Computational Linguistics

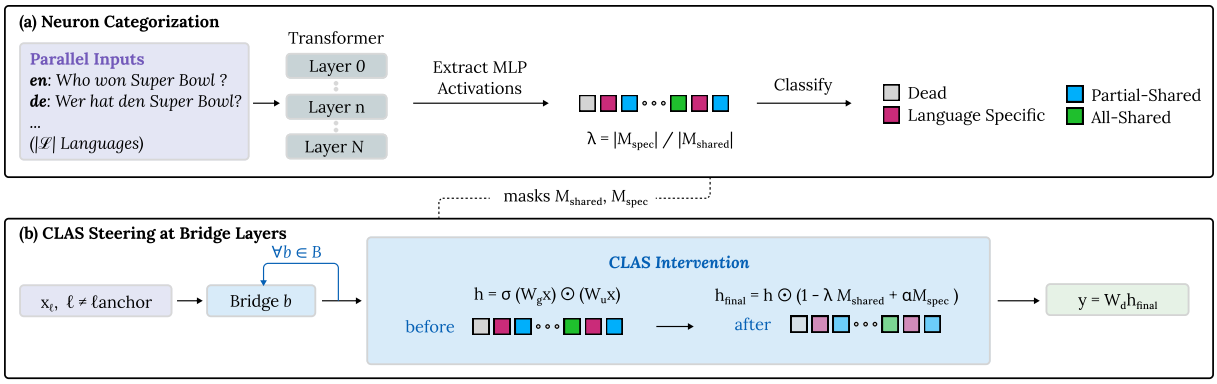


Figure 1: Overview of CLAS, with two high level steps (1) Neuron Categorization and (2) Activation Steering.

steering mechanism that selectively modulates neurons to enhance cross-lingual transfer during inference[1].

- We demonstrate CLAS effectiveness across diverse tasks and models, achieving improvements on both classification and generation benchmarks while maintaining anchor language stability.
- We conduct comprehensive analysis of cross-lingual representations and find that effective transfer is driven by functional divergence rather than proximity to the anchor language.

2 Cross-Lingual Activation Steering (CLAS)

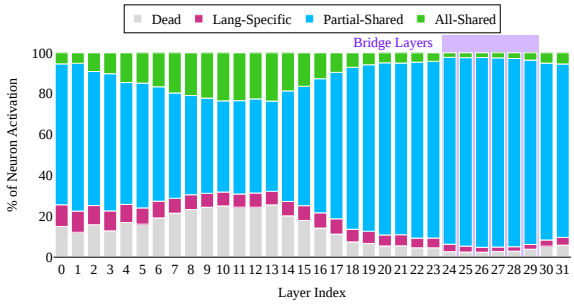
CLAS consists of three stages: (i) construct parallel inputs across languages, (ii) summarize neuron behavior using simple activation statistics and group neurons into coarse categories, and (iii) apply a lightweight steering rule that modulates activations at inference time.

2.1 Preliminaries

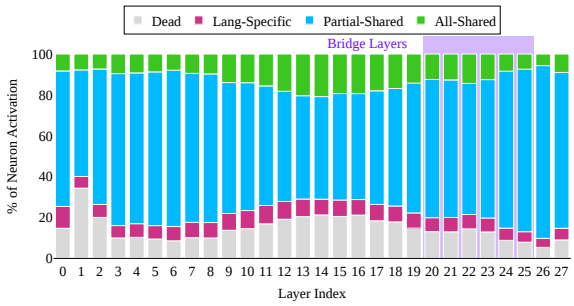
We define a configuration $\mathcal{C} = \{\mathcal{L}, \ell_{\text{anchor}}, \mathcal{B}, \alpha\}$, where $\mathcal{L}$ is the set of languages used for analysis, ℓ_{anchor} is an anchor language (typically the model's strongest language), $\mathcal{B}$ is the set of "bridge layers" selected for intervention, and α is the hyperparameter that controls the strength of boosting or suppressing selected activations during steering.

2.2 Parallel Input Construction

To analyze neuron behavior across languages, we require aligned multilingual inputs that express the same underlying content. For each sample index i, we construct a set of parallel inputs $\mathbf{x}^{(i)} = \{x_\ell^{(i)} \mid \ell \in \mathcal{L}\}$, where each $x_\ell^{(i)}$ is the same text expressed in language ℓ. These parallel inputs are used only for measuring neuron activations and do not update the model in any way.



(a) Llama



(b) Qwen

Figure 2: Distribution of types of neuron per layer across models. Llama has a total of 32 layers and Qwen has 28 layers. The bridge layers (purple shade) are selected near the final layers where the partial-shared language activations are the highest.

2.3 Neuron Statistics and Categorization

To understand the model's multilingual structure and identify where CLAS should intervene, we first analyze neuron activations across languages and layers. Following (Wang et al., 2024), we group neurons into four mutually exclusive categories: dead (never active), language-specific (active for one language), partial-shared (active for some lan-

[1]The code will be made publicly available at https://github.com/PortNLP/CLAS.

guages), and all-shared (active for all languages). However, we differ from prior work in how these categories are computed. Wang et al. (2024) assign categories at the instance level: a neuron is considered all-shared if it activates for all languages on a single parallel example. As a result, neuron categories can vary across inputs and tasks. In contrast, we adopt a dataset-level view. We compute the mean activation of each neuron using 100 parallel samples from the XQuAD dataset across 12 languages and two models (Llama 3.1 8B and Qwen 2.5 7B) and use this aggregate statistic to assign each neuron to a single category. This makes each neuron's category a stable property of the model and language set rather than something that changes across examples. We observe a similar phenomenon with XNLI dataset.

Layer-wise Activation Patterns For each layer, we compute the percentage of neurons in each category. Figure 2 summarizes the resulting distributions. We observe several consistent patterns:

Early layers. Early layers contain a high proportion of partial-shared neurons, suggesting that these layers are involved in mapping language-specific or language-related features into shared representations (Tang et al., 2024; Zhang et al., 2025).

Middle layers. In the middle layers, we observe an increase in all-shared neurons alongside a rise in dead neurons and a reduction in partial-shared neurons. This is consistent with prior findings that mid-level representations become more language-agnostic and semantically focused (Zhang et al., 2025; Xu et al., 2025b).

Final layers. In the upper layers, partial-shared neurons become more prevalent again, and dead neurons decrease, suggesting that language-specific processing re-emerges near the output as the model prepares to generate text in the target language (Tang et al., 2024; Zhang et al., 2025).

These observations are descriptive and do not imply that these categories are sharply separated or functionally pure; rather, they provide a coarse summary of how multilingual structure is distributed across layers.

Selecting Bridge Layers Based on these statistics, we select a small set of *bridge layers* for intervention. In both LLaMA and Qwen, layers 24-29 and 20-25, respectively, exhibit relatively high proportions of partial-shared neurons together with relatively low proportions of dead and language-specific neurons. We therefore hypothesize that intervening in these layers provides a useful balance: representations are shared enough to support cross-lingual steering, while not yet so specialized that small perturbations directly disrupt surface generation. We exclude the final two layers from intervention, as these layers are known to be strongly specialized for producing language-specific outputs and are particularly sensitive to perturbation (Zhao et al., 2024b). All layer selections are fixed prior to evaluation and are not tuned on test data.

2.4 Activation Steering via CLAS

Given an input x_ℓ in language ℓ, the model computes an intermediate MLP activation using the standard SwiGLU transformation:

$$\mathbf{h} = \sigma(W_g\mathbf{x}) \odot (W_u\mathbf{x}),$$

where W_g and W_u are the gating and up-projection matrices, and σ is the nonlinear activation function.

For non-anchor languages $\ell \neq \ell_{\text{anchor}}$, CLAS applies a lightweight, deterministic modification to this activation using masks derived from neuron categories to adjust the relative contribution of shared and language-specific neurons.

Adaptive Masking and Heuristic To effectively steer the model's intermediate representations, we must account for the structural imbalance between neuron categories. Let M_{shared} denote the combined mask for all shared neurons (i.e., the union of partial and all-shared neurons), and let M_{spec} denote the mask for language-specific neurons. In the target layers (e.g., the bottom layers), shared neurons inherently dominate the representation space ($|M_{\text{shared}}| \gg |M_{\text{spec}}|$).

To balance our intervention and prevent the abundant shared neurons from overpowering the scarce language-specific signals, we introduce an adaptive scaling factor λ, defined as the ratio of active specific neurons to shared neurons:

$$\lambda = \frac{|M_{\text{spec}}|}{|M_{\text{shared}}|},$$

where $|\cdot|$ represents the count of active neurons within a given mask. Because shared neurons form the majority of the active dimensions, λ strictly remains a fraction ($\lambda \ll 1$). This formulation ensures that the intervention strength is not static, but naturally grounded in the precise architectural makeup of each individual layer.

Inference-Time Steering Intervention We apply a single-step multiplicative intervention designed to reweight the neuron contributions:

$$\mathbf{h}_{\text{final}} = \mathbf{h} \odot \left(\mathbf{1} - \lambda M_{\text{shared}} + \alpha M_{\text{spec}}\right),$$

where α is a configurable hyperparameter that dictates the amplification or suppression of language-specific neurons.

Subtracting $\lambda M_{\text{shared}}$ gently suppresses the shared neurons, preventing them from drowning out finer language-specific signals. Adding αM_{spec} then explicitly amplifies or suppresses the language-specific neurons depending on the desired steering direction: a positive α pushes the representation toward the target language, while a negative α pulls it away. The adjusted activation is then passed through the down-projection:

$$\mathbf{y} = W_d \mathbf{h}_{\text{final}}.$$

2.5 Anchor Language Handling

For the anchor language ℓ_{anchor} (i.e. English), no modification is applied: $\mathbf{h}_{\text{final}} = \mathbf{h}$. We keep the anchor activations untouched because the anchor language serves as a stable reference point for cross-lingual alignment. Modifying it would risk introducing unnecessary distortion into a representation that is already well supported by the model. All adjustments are therefore applied only to non-anchor languages, allowing them to shift relative to a fixed reference.

3 Experimental Setting

To assess CLAS's performance, we detail our setup using 2 models and 5 datasets.

3.1 Models

We use LLaMA 3.1 8B Instruct (Grattafiori et al., 2024) and Qwen 2.5 7B Instruct (Team, 2024). Both models are kept frozen; CLAS is applied only at inference time and does not modify model parameters.

3.2 Datasets and Languages

We evaluate on the following five benchmarks, covering a total of 21 languages. More details are presented in Appendix A.

XNLI (Conneau et al., 2018) is a multilingual natural language inference dataset in 15 languages. We prompt the model with a premise and hypothesis, reporting accuracy based on the highest-probability target token: Entailment, Neutral, or Contradiction.

XStoryCloze (Lin et al., 2022) is a multilingual commonsense reasoning dataset based on StoryCloze. Given a four-sentence context and two candidate endings, the model selects the correct ending. We evaluate on 11 languages and report accuracy.

PAWS-X (Yang et al., 2019) is a cross-lingual paraphrase identification dataset with adversarial sentence pairs that have high lexical overlap but different meanings. We evaluate on 4 languages and report accuracy.

HarryPotter Quiz (Chua et al., 2025) is a cross-lingual factual knowledge benchmark that tests whether models can access the same knowledge across languages. We evaluate on 5 languages and report accuracy.

XQuAD (Artetxe et al., 2020) is a multilingual question answering dataset in 12 languages. Each instance requires the model to predict the correct answer span within the provided text. The results are reported using the token-level F1 score.

3.3 Baselines

We consider three baselines: **(1)** unmodified base model, **(2)** Int μ (Mondal et al., 2025), which substitutes the model's intermediate representations with mean activations during intervention; and **(3)** INCLINE (Wang et al., 2025) which learns alignment matrices applied during inference to project the intermediate representations of low-performing languages toward the representation space of a high-performing anchor language. INCLINE uses language-specific hyperparameter and fully reproducing this setup would require tuning a separate parameter for each language, which is beyond the scope of our experiments. Instead, we fix $\alpha = 1$ across all settings as this was reported to perform well for most languages (Wang et al., 2025).

4 Results and Discussion

In this section, we discuss the results of the experiments with CLAS and how it improves multilingual performance.

4.1 Main Results

Table 1 summarizes the results on five benchmarks, comparing CLAS against the baselines. Across all the benchmarks, CLAS consistently improves accuracy over the baselines. For Llama, CLAS yields substantial gains on XNLI (+2.61) and XStoryCloze (+2.20), while moderate enhancements

	Llama				Qwen			
	Base	**Int μ**	**INCLINE**	**CLAS (Ours)**	**Base**	**Int μ**	**INCLINE**	**CLAS (Ours)**
XNLI	42.42	42.19	39.19	**45.03** (+2.61)	59.08	59.05	41.34	**59.29** (+0.21)
XStoryCloze	83.38	82.97	79.05	**85.58** (+2.20)	78.68	78.96	61.08	**79.70** (+1.02)
PAWS-X	65.70	62.87	59.58	**66.00** (+0.30)	78.58	79.03	68.30	**79.07** (+0.49)
HarryPotter	51.59	47.84	50.75	**52.59** (+1.00)	53.67	49.58	28.42	**55.50** (+1.83)
XQuAD	26.85	26.98	**33.50**	32.10 (+5.25)	32.73	21.78	18.38	**33.94** (+1.21)

Table 1: The performance on different benchmarks across Llama and Qwen. The best performing model is in **bold**. The improvements of CLAS over base are shown within parentheses.

	en	ar	es	eu	hi	id	my	ru	sw	te	zh	**Avg.**
Llama												
Base	92.26	88.09	91.07	70.02	90.07	87.49	74.26	91.40	71.01	79.68	90.73	83.38
Int μ	92.39	89.01	91.33	71.74	86.30	84.98	74.39	91.79	71.08	78.89	90.14	82.97
INCLINE	92.39	90.21	95.10	67.77	84.58	88.88	52.75	92.52	71.14	52.81	94.77	79.05
CLAS	92.26	85.44	**94.77**	**74.92**	89.61	**92.12**	73.13	**94.04**	**78.09**	78.95	**94.71**	**85.58**
Qwen												
Base	97.68	91.46	95.10	52.95	87.03	93.58	60.69	94.97	54.73	61.02	95.30	78.68
Int μ	97.68	91.59	95.10	52.42	87.09	93.91	61.68	94.37	55.79	62.41	95.23	78.96
INCLINE	97.68	90.54	52.88	52.68	51.62	53.67	52.81	93.71	52.81	57.31	52.81	61.08
CLAS	97.68	**91.53**	**95.96**	**53.34**	**87.62**	**93.85**	**61.68**	**95.70**	**56.52**	**65.32**	**95.50**	**79.70**

Table 2: Performance on XStoryCloze (accuracy) across multiple languages and models. For each language, **bold** represents the cases where CLAS beats the base model.

	en	ar	de	el	es	hi	ro	ru	th	tr	vi	zh	**Avg.**
Llama													
Base	23.47	23.20	38.54	25.18	34.23	29.47	30.01	27.32	19.23	23.74	34.91	12.84	26.85
Int μ	23.47	21.25	33.08	26.68	35.31	29.39	36.98	27.25	15.83	21.64	40.00	12.90	26.98
INCLINE	34.96	20.39	45.79	35.68	44.88	34.02	34.13	21.87	26.87	44.18	36.01	23.19	33.50
CLAS	23.47	**27.03**	**50.86**	23.71	**50.57**	28.32	**41.84**	**34.87**	**21.90**	**30.38**	**37.49**	**14.77**	**32.10**
Qwen													
Base	48.88	28.73	34.77	37.06	38.53	33.12	34.48	27.58	28.51	29.19	31.67	20.24	32.73
Int μ	14.53	23.11	11.99	48.95	35.43	17.83	31.31	17.39	12.63	21.25	15.75	11.20	21.78
INCLINE	29.91	25.10	3.40	7.61	13.37	0.14	20.87	22.47	39.56	20.47	33.94	3.66	18.38
CLAS	48.88	28.29	36.03	**38.91**	**40.53**	25.98	**38.13**	**27.93**	**29.49**	**30.28**	**35.86**	**26.92**	**33.94**

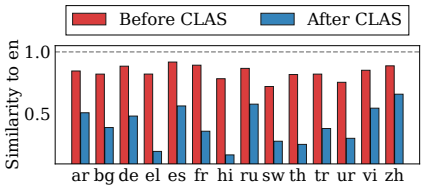
Table 3: The performance (F1 scores) on the XQuAD dataset across multiple languages and models. For each language, **bold** represents the cases where CLAS beats the base model.

on HarryPotter (+1.00) and PAWS-X (+0.30). For Qwen, the improvements are generally smaller but remain positive across all tasks, peaking on HarryPotter (+1.83) and XQuAD (+1.21). This narrower margin of improvement reflects Qwen's inherently stronger baseline multilingual capabilities, where CLAS acts more conservatively to provide stable, consistent enhancements without disrupting existing knowledge. Crucially, CLAS demonstrates significantly greater reliability than existing intervention baselines. While Int μ exhibits mixed results generally degrading Llama's performance and yielding only marginal or negative shifts on Qwen, INCLINE suffers from performance degradation across both models. This regression is particularly drastic on Qwen, where INCLINE causes accuracy to drop steeply (e.g., from 53.67 to 28.42 on HarryPotter). In contrast, CLAS achieves positive gains across all tasks and both model families.

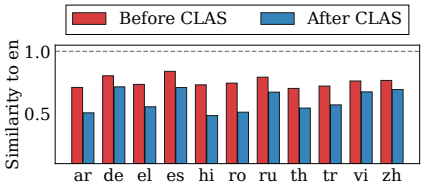
4.2 Language-Level Performance

To better understand where the overall gains come from, we analyze performance by language on XStoryCloze (Table 2) and XQuAD (Table 3). As intended by design, CLAS preserves the anchor language English performance which remains essentially unchanged across both models and tasks.

On XStoryCloze (discriminative multiple-choice), CLAS improves many non-English languages. For Llama, gains are large and broadly distributed, e.g., Swahili (+7.08), Basque (+4.90), Indonesian (+4.63), and Chinese (+3.98), with



(a) XNLI



(b) XQuAD

Figure 3: Cosine similarity with English across languages on each task using Llama model.

small regressions in a few cases (e.g., Arabic and Burmese). For the stronger Qwen model, CLAS acts as a consistent refinement, yielding targeted improvements (notably Telugu +4.30 and Swahili +1.79) without regressions. This stability contrasts with INCLINE, which collapses on Qwen for XStoryCloze, dropping performance to roughly the low-50s across many languages. Results for the remaining datasets are in Appendix B.

On XQuAD (generative QA), effects are more uneven across languages, consistent with generation being more sensitive to representation shifts. CLAS again preserves the anchor, while INCLINE is inconsistent, boosting English on Llama (23.47 $\rightarrow$ 34.96) but degrading it on Qwen (48.88 $\rightarrow$ 29.91). For Llama, CLAS delivers a strong +5.25 average F1 gain, driven by large improvements in Spanish (+16.34), German (+12.32), and Romanian (+11.83). For Qwen, CLAS gives a modest +1.21 average F1 improvement, combining clear wins in some languages (e.g., Chinese and Vietnamese) with regressions in others (notably Hindi), highlighting that test-time interventions are more stable on discriminative tasks than on generation.

4.3 Cross-Lingual Alignment Analysis

In this subsection, we analyze the impact of our activation steering mechanism on representation alignment and examine how these geometric shifts relate to cross-lingual performance gains on XNLI and XQuAD. Figure 3 shows cosine similarity between each target language and English in the bridge layers before and after CLAS. Across tasks and models, CLAS generally reduces similarity to English rather than increasing it. This effect is strongest for Llama, especially on XQuAD, where several languages move noticeably farther from English, while Qwen shows smaller and more uniform changes.

This pattern matches our neuron analysis presented earlier in Section 2.3. CLAS rebalances partial-shared and language-specific neurons, reducing over-reliance on English-centric features and allowing more target-language structure to influence predictions. As a result, CLAS improves performance not by pulling languages closer to English, but by relaxing excessive alignment. The effect is larger on XQuAD than on XNLI, likely because span extraction is more sensitive to surface form. Reducing English alignment can help match target-language expressions, though it can also increase variance and occasional regressions.

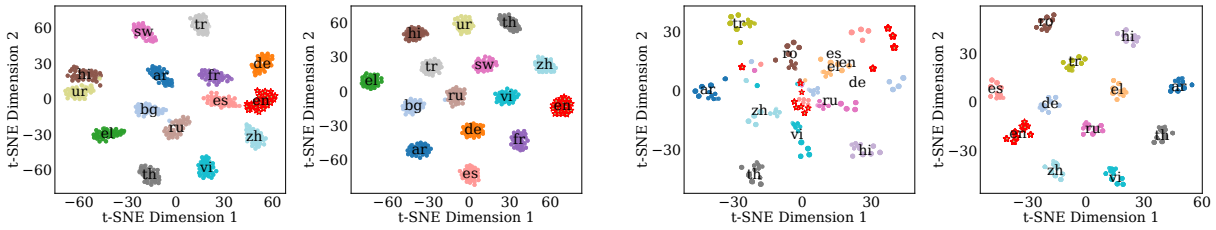
4.4 Representation Space Visualization

We use t-SNE to project the penultimate layer's hidden states for the English anchor and all other languages into a shared 2D space. Figure 4 shows the representations and centroids before and after CLAS, revealing three patterns.

CLAS reorganizes representations without collapsing them into English. After CLAS, language clusters remain distinct and are often more clearly separated, suggesting gains come from restructuring language-specific spaces rather than forcing assimilation to English.

Moderate reorganization correlates with larger gains. On XNLI, increased cluster structure is broadly helpful. On XQuAD, the largest gains occur when CLAS resolves overlap for languages that start in mixed regions (e.g., *ar*, *de*, *es*); already well-separated languages (e.g., *hi*, *vi*, *zh*) change less and typically gain less. This suggests that CLAS is most helpful when it resolves representational overlap, rather than when representations are already well-formed.

Distance to English is not predictive. Neither initial nor final proximity to the English cluster reliably explains performance changes. Some languages far from English (e.g., *ur*, *hi*) improve largely as seen in (Lim et al., 2025), while some closer languages improve little or regress. After CLAS, high-performing languages are scattered rather than concentrated near English. This suggests that CLAS effectiveness depends on how representations are reorganized, not simply on how close they are to the anchor language.



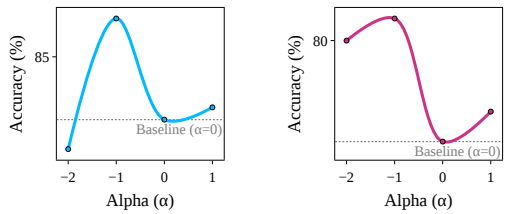
(a) XNLI Llama (Before vs After)

(b) XQuAD Llama (Before vs After)

Figure 4: t-SNE visualization of multilingual representations before (left) and after (right) CLAS intervention. Each point represents a sentence embedding, with colors denoting languages. English, marked with (red) stars, is the anchor language.

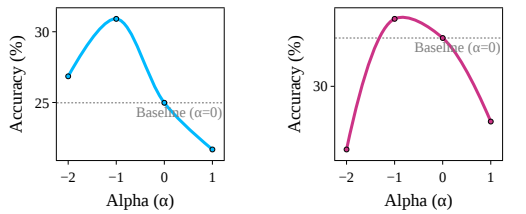
Question	Llama (base)	CLAS
Was benannte die Generalkonferenz für Maß und Gewicht 1960 nach Tesla? (What did the General Conference on Weights and Measures name after Tesla in 1960?)	SI-Maßeinheit der magnetischen Flussdichte. stop. Die SI-Maßeinheit der magnetischen Flussdichte (SI unit of magnetic flux density. stop. The SI unit of magnetic flux density)	SI-Maßeinheit der magnetischen Flussdichte. (SI unit of magnetic flux density.)
Die Lutherbibel beeinflusste wessen englische Übersetzung der Bibel? (The Luther Bible influenced whose English translation of the Bible?)	William Tyndale Die Lutherbibel beeinflusste die englische Bibel von William Tyndale. Tyndale war ein eng (William Tyndale's English Bible was influenced by the Luther Bible. Tyndale was a close)	William Tyndale. (William Tyndale.)

Table 4: Comparison of outputs for German XQuAD samples. English translations provided via Google Translate.



(a) XStoryCloze (Llama)

(b) XStoryCloze (Qwen)



(c) XQuAD (Llama)

(d) XQuAD (Qwen)

Figure 5: Impact of α across XStoryCloze and XQuAD. The left column shows the performance for the Llama (blue) and the right column for the Qwen (pink).

4.5 Optimal Steering Intensity and Direction

As shown in Figure 5, Llama and Qwen both achieve peak accuracy on XStoryCloze and XQuAD when $\alpha = -1$. Under the CLAS formulation, negative α to suppresses language-specific neurons. Since $-\lambda$ already regulates dominant shared neurons, this trend indicates that these tasks benefit most from masking language-specific signals and relying entirely on shared semantic structures. However, optimal steering varies. Appendix C shows the settings that benefit from positive steering, where selectively amplifying language-specific neurons captures necessary target-language nuances. This aligns with prior findings that optimal cross-lingual transfer remains inherently task-dependent (Tang et al., 2024; Wang et al., 2024).

4.6 Importance of Bridge Layer Selection

To study whether intervening on other layers produces comparable results, we evaluate the impact of CLAS on the XStoryCloze with Llama. Specifically, we compare our selected bridge layers (24–30) against two distinct representational regions: the early layers (0–6), which contain a higher density of language-specific neurons, and the middle layers (6–16), where all-shared neuron activity is the highest. As illustrated in Figure 6, applying the intervention to the early or middle layers yields heavily suboptimal cross-lingual performance, achieving average scores of only 72.57 and 75.43, respectively. Intervening too early likely disrupts low-level lexical processing (Tang et al., 2024; Kojima et al., 2024), while suppressing the shared neurons in the middle layers risks degrading the model's language-agnostic representations (Xu et al., 2025b; Zhang et al., 2025). In contrast, applying CLAS to the bridge layers (24–30) results in a performance spike, reaching an average

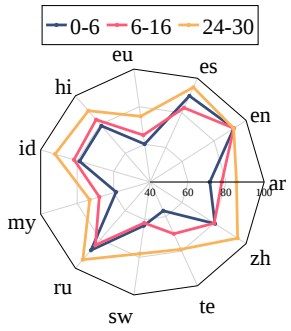


Figure 6: Multilingual performance comparison across different Llama layers (early layers 0-6, middle layers 6-16, and bottom layers 24-30).

of 85.58. This demonstrates that identifying the correct intervention site is critical; the bridge layers provide the optimal semantic space where shared representations can be safely dampened and allow language-specific features to steer the final output distribution.

4.7 Qualitative Analysis

Table 4 shows the qualitative impact of CLAS on German generation. In the examples, the baseline model suffers from severe repetition loops and verbosity leakage. In contrast, CLAS successfully suppresses these behaviors. This demonstrates CLAS effectively generates concise response while maintaining accuracy.

5 Related Work

Cross-lingual Transfer Cross-lingual transfer in multilingual LLMs has been widely studied, with prior work showing that transfer quality varies across tasks, languages, and models. For example, Hu et al. (2025) analyze factors that influence cross-lingual performance on reasoning tasks. Many approaches improve transfer through additional training such as adding more data during instruction tuning (Shaham et al., 2024), combining supervised fine-tuning with preference alignment (Lai et al., 2024), and constructing new multilingual pre-training datasets (He et al., 2025). Other work uses translation-based fine-tuning (Lee et al., 2025), layer-wise fine-tuning (Bandarkar et al., 2025), or language-specific adapters (Zhao et al., 2025). Continued pre-training has also been shown to improve transfer for some language pairs (Wu et al., 2025). An alternative line of work uses prompts rather than parameter updates. For example, Tanwar et al. (2023) use multilingual in-context examples, and Yoo et al. (2025) study in-context learning in code-switching settings.

Neuron Behavior and Cross-lingual Transfer Several studies examine cross-lingual transfer at the neuron level. Huang et al. (2025) show that activation similarity across languages is associated with better transfer. Other work identifies language-specific and shared neurons (Tang et al., 2024; Zhang et al., 2025; Tezuka and Inoue, 2025), and shows that shared neurons often concentrate in middle and upper layers (Xu et al., 2025b). Results on directly intervening on neurons are mixed. Wang et al. (2024) show that neuron roles vary by task and model while Mondal et al. (2025) find that manipulating language-specific neurons yields limited gains. Our work differs in that we apply test-time steering that blends rather than overwrites activations, preserving representational structure. We show that the direction and magnitude of steering matter, and that carefully controlled neuron-level interventions can improve cross-lingual transfer. Sundar et al. (2025) target language-specific neurons to push multilingual embedding into a shared aligned space whereas we show that effective cross-lingual transfer actually relies on maintaining distinct separation between language clusters. Lastly, INCLINE (Wang et al., 2025) learns per-layer alignment matrices at inference time but requires per-language tuning. On the other hand, CLAS tunes a single parameter per task.

6 Conclusion

We presented CLAS, a training-free activation steering method that improves cross-lingual transfer by rebalancing shared and language-specific neurons at inference time. CLAS improves performance across both discriminative and generative tasks, and our analysis shows that gains come from functional divergence rather than forcing representations to align closely with the anchor language. Languages benefit most by shifting from ambiguous regions into distinct clusters, whereas initial anchor proximity does not reliably predict success. These patterns are task-dependent, highlighting the need for task-aware cross-lingual methods. Future work could explore adaptive steering strategies that adjust intervention strength based on representational structure, better understand saturation effects in already well-separated clusters, and extend CLAS to other modalities and training settings.

Limitations

We evaluate CLAS on two multilingual instruction-tuned LLMs (Qwen and Llama) and across multiple multilingual benchmarks. However, our analysis and suggested intervention are anchored to English i.e. we quantify alignment shifts via cosine similarity to an English anchor. This may not reflect behavior under alternative references/anchors or truly language pair-specific settings. Additionally, it is important to note that CLAS is not uniformly beneficial: while average gains for languages are positive, we also observe language-specific regressions (e.g, Hindi, Greek). This indicates that test-time steering can be brittle for certain model/language combinations. Finally, our mechanistic analysis does not isolate the role of attention heads or other circuit components, limiting the granularity of causal attribution.

Ethical Considerations

CLAS is a test-time activation intervention and can change model behavior in ways that are not always predictable across languages, including occasional performance degradations. As with other steering approaches, such interventions can be re-purposed in undesirable ways (e.g, modulating output without transparency). Thus, we recommend caution and task-specific testing and validation before real-world deployment. Our experiments use publicly-available benchmarks and do not involve human subjects or personal data.

Acknowledgments

We appreciate the valuable feedback from the anonymous reviewers and the members of the PortNLP Lab. We also gratefully acknowledge the support of the David E. Wedge Awards for Excellence from Portland State University's Maseeh College of Engineering and Computer Science.

References

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

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2025. Layer swapping for zero-shot cross-lingual transfer in large language models. In *The Thirteenth International Conference on Learning Representations*.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024a. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024b. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2025. Crosslingual capabilities and knowledge barriers in multilingual large language models. *Preprint*, arXiv:2406.16135.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6101–6117, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kaiyu He, Tong Zhou, Yubo Chen, Delai Qiu, Shengping Liu, Kang Liu, and Jun Zhao. 2025. Semantic pivots enable cross-lingual transfer in large language models. *arXiv preprint arXiv:2505.16385*.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian

Huang. 2025. Large language models are cross-lingual knowledge-free reasoners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1525–1542, Albuquerque, New Mexico. Association for Computational Linguistics.

Chongxuan Huang, Yongshi Ye, Biao Fu, Qifeng Su, and Xiaodong Shi. 2025. From neurons to semantics: Evaluating cross-linguistic alignment capabilities of large language models via neurons alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28956–28974, Vienna, Austria. Association for Computational Linguistics.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.

Jungseob Lee, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. 2025. Cross-lingual optimization for language transfer in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15100–15119, Vienna, Austria. Association for Computational Linguistics.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. 2025. Language-specific latent process hinders cross-lingual performance. *arXiv preprint arXiv:2505.13141*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and 1 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9019–9052.

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. Language-specific neurons do not facilitate cross-lingual transfer. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 46–62, Albuquerque, New Mexico. Association for Computational Linguistics.

Rhitabrat Pokharel, Yufei Tao, and Ameeta Agrawal. 2025. Capo: Confidence aware preference optimization learning for multilingual preferences. *arXiv preprint arXiv:2511.07691*.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. 2025. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2375–2401, Vienna, Austria. Association for Computational Linguistics.

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.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hinata Tezuka and Naoya Inoue. 2025. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31730–31780, Suzhou, China. Association for Computational Linguistics.

Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. 2024. Sharing matters: Analysing neurons across languages and tasks in llms. *arXiv preprint arXiv:2406.09265*.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025. Bridging the language gaps

in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria. Association for Computational Linguistics.

Linjuan Wu, Hao-Ran Wei, Huan Lin, Tianhao Li, Baosong Yang, Fei Huang, and Weiming Lu. 2025. Enhancing LLM language adaption through cross-lingual in-context pre-training. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27140–27154, Suzhou, China. Association for Computational Linguistics.

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

Yuemei Xu, Kexin Xu, Jian Zhou, Ling Hu, and Lin Gui. 2025b. Linguistic neuron overlap patterns to facilitate cross-lingual transfer on low-resource languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27646–27661, Suzhou, China. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *Preprint*, arXiv:1908.11828.

Haneul Yoo, Jiho Jin, Kyunghyun Cho, and Alice Oh. 2025. Code-switching in-context learning for cross-lingual transfer of large language models. *arXiv preprint arXiv:2510.05678*.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. 2025. How does alignment enhance llms' multilingual capabilities? a language neurons perspective. *arXiv preprint arXiv:2505.21505*.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204, Bangkok, Thailand. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024a. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319.

Dataset	$\|\mathcal{L}\|$	#Samp.	Metric
XNLI (natural language inference)	15	5010	Acc.
XStoryCloze (commonsense reasoning)	11	1511	Acc.
PAWS-X (paraphrase identification)	4	2000	Acc.
HarryPotter (MCQA)	5	300	Acc.
XQuAD (question answering)	12	1190	F1

Table 5: Summary of evaluation datasets. #Samp. represents the number of samples per language.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2025. AdaMergeX: Cross-lingual transfer with large language models via adaptive adapter merging. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9785–9800, Albuquerque, New Mexico. Association for Computational Linguistics.

A Dataset Summary

Table 5 presents an overview of the datasets used in this study.

B Language-Level Performances

Tables 6, 7, and 8 show the language level results on XNLI, PAWS-X, and HarryPotter Quiz across Llama and Qwen models.

C Effect of α

Figure 7 shows the effect of α on XNLI, PAWS-X and HarryPotter Quiz datasets.

	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	**Avg.**
Llama																
Llama	52.63	38.88	42.28	43.65	40.58	46.95	44.53	41.78	43.75	41.10	40.68	44.17	36.09	44.81	44.63	42.42
Int μ	52.63	38.12	40.68	45.77	41.3	46.59	43.11	41.10	43.79	41.42	44.39	43.81	36.37	42.85	41.38	42.19
INCLINE	52.63	39.24	38.74	35.77	38.08	38.62	35.75	44.25	33.51	35.51	42.87	39.60	33.35	46.89	46.43	39.19
CLAS	52.63	39.10	41.68	49.42	44.85	50.38	47.96	41.74	45.27	38.36	40.98	49.36	44.25	46.41	50.72	**45.03**
Qwen																
Qwen	71.74	60.45	60.75	66.22	58.02	65.56	64.47	55.56	61.87	40.31	57.8	59.18	51.07	62.11	63.81	59.08
Int μ	71.78	60.54	60.74	66.33	58.06	65.51	64.33	55.51	61.70	39.96	57.78	59.10	51.16	62.18	63.81	59.05
INCLINE	71.74	54.51	44.71	58.14	33.37	33.33	56.59	33.33	60.82	34.07	33.33	33.33	33.33	36.55	33.33	41.34
CLAS	71.74	60.64	60.94	66.87	57.92	65.99	65.21	56.21	61.86	39.90	57.09	59.40	50.98	61.68	65.35	**59.29**

Table 6: The performance on XNLI across multiple languages and models.

	en	de	es	fr	**Avg.**
Llama					
Base	68.20	63.45	64.45	69.20	65.70
Int μ	68.20	62.20	54.70	71.70	62.87
INCLINE	69.00	66.45	54.55	57.75	59.58
CLAS	68.20	66.60	65.70	65.70	**66.00**
Qwen					
Base	84.20	78.85	78.75	78.15	78.58
Int μ	84.20	79.00	79.85	78.25	79.03
INCLINE	84.50	79.95	78.70	46.25	68.30
CLAS	84.20	79.55	79.25	78.40	**79.07**

Table 7: Performance on the PAWS-X across multiple languages and models.

	en	de	fr	it	es	**Avg.**
Llama						
Base	65.00	50.67	50.67	48.33	56.67	51.59
Int μ	65.33	53.67	44.33	46.67	46.67	47.84
INCLINE	61.00	52.00	49.00	48.67	53.33	50.75
CLAS	65.00	53.67	52.00	47.67	57.00	**52.59**
Qwen						
Base	57.00	55.00	51.33	48.33	60.00	53.67
Int μ	53.67	51.33	50.67	41.00	55.33	49.58
INCLINE	58.67	24.33	40.67	24.33	24.33	28.42
CLAS	57.00	59.00	54.33	47.67	61.00	**55.50**

Table 8: Accuracy on the multilingual HarryPotter Quiz across multiple languages and models.

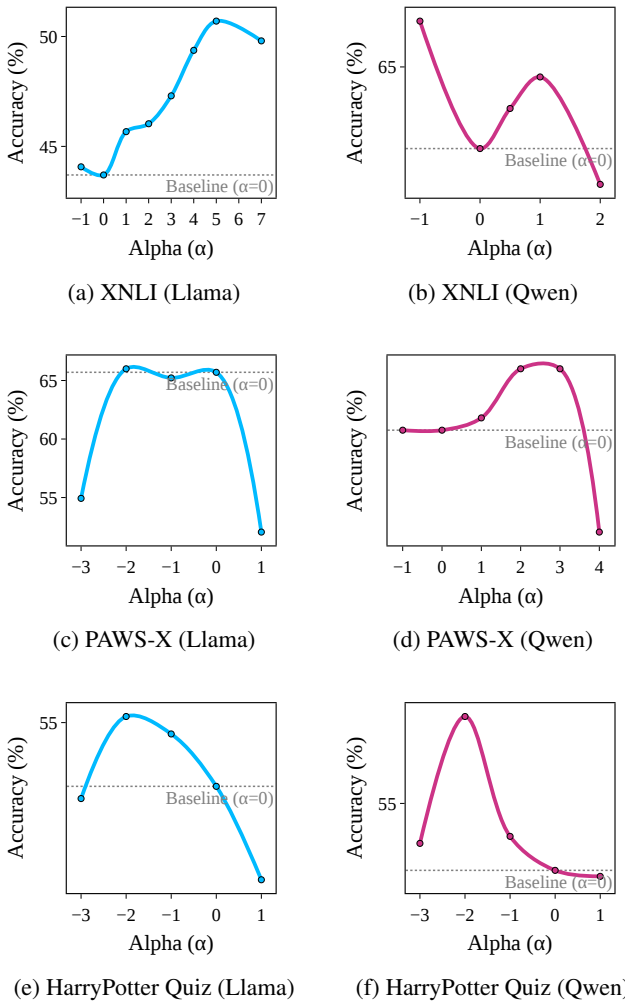


Figure 7: Impact of the hyperparameter α on performance across datasets. The left column shows the performance for the Llama model, while the right column shows the performance for the Qwen model.