Language Lives in Sparse Dimensions: Toward Interpretable and Efficient Multilingual Control for Large Language Models

Anonymous ACL submission

Abstract

Large language models exhibit strong multilingual capabilities despite limited exposure to non-English data. Prior studies show that English-centric large language models map multilingual content into English-aligned representations at intermediate layers and then project them back into target-language token spaces in the final layer. From this observation, we hypothesize that this cross-lingual transition is governed by a small and sparse set of dimensions, which occur at consistent indices across the intermediate to final layers. Building on this insight, we introduce a simple, training-free method to identify and manipulate these dimensions, requiring only as few as 50 sentences of either parallel or monolingual data. Experiments on a multilingual generation control task reveal the interpretability of these dimensions, demonstrating that the interventions in these dimensions can switch the output language while preserving semantic content, and that it surpasses the performance of prior neuron-based approaches at a substantially lower cost.

1 Introduction

Large language models (LLMs), though predominantly trained on English-centric corpora, have nevertheless demonstrated strong multilingual capabilities even with limited exposure to non-English data, enabling a wide range of multilingual applications. Recent studies (Wendler et al., 2024) have shown that English-centric LLMs map multilingual content into English-aligned representations at intermediate layers, before projecting them back into target-language token spaces in the later layers.

Various studies have sought to uncover the internal mechanisms of this transfer by investigating how specific components of LLMs manage language-dependent behavior such as vocabulary and grammar. Neuron-level studies (Sundar et al., 2025; Kojima et al., 2024; Tang et al., 2024) analyze the activation patterns of feed-forward network units across large-scale multilingual corpora, identifying language-specific neurons whose manipulation can steer generation toward a target language. Sparse-Autoencoder (SAE) based approaches (Deng et al., 2025) instead train massive sparse-autoencoders on large corpora to isolate high-dimensional features tied to language choice.

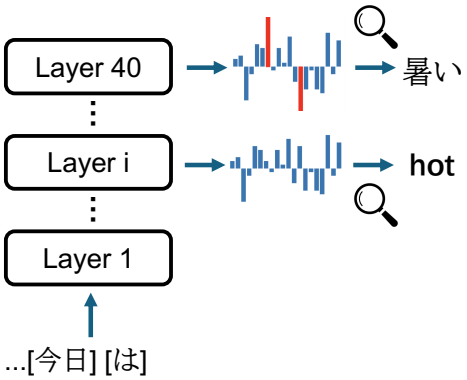


Figure 1: Hypothesis of language-specific dimensions. The logit lens shows that English-centric LLMs often surface the English answer in intermediate layers before producing the target-language answer. The prompt is "English:"Today is hot."-日本語:"今日は". We ask the model to predict the omitted word "hot" in Japanese. Dimensions in red are likely to be language-specific.

While these approaches examine which LLM components encode language identity, our perspective is to take a more intrinsic view of multilingual transfer through the lens of the trajectory of representations, tracing how representations transform across layers. As highlighted in red in Fig. 1, we examine the representation trajectories from intermediate to final layers, and reveal that the transition follows a distinct and identifiable pattern, with sparse distributions concentrated in a small set of dimensions. We put forward the hypothesis that the model leverages language-specific

dimensions to intervene in the transition from a shared, language-agnostic representation space to language-specific token spaces. While most semantic content remains stable in language-agnostic dimensions, these language-specific dimensions consistently control the projection into the target token space. Moreover, these language-specific dimensions are not arbitrary; they appear in similar dimensions across different layers of the model, with later layers amplifying their magnitude. These observations suggest that cross-lingual transfer relies on adjustments along a consistent set of dimensions rather than diffuse redistribution.

In this paper, we present an in-depth analysis of language-specific dimensions, and the insights lead us to develop a method for cross-lingual transfer that manipulates these dimensions. Our method identifies these dimensions with as few as ∼50 sentences, in contrast to neuron-based methods that require logging activations over millions of tokens or SAE-based approaches that train massive auxiliary autoencoders. The approach is training-free and can operate with either limited parallel data or even only monolingual data, leveraging the model's own cross-lingual mapping ability to perform low-resource translation.

Concretely, we compute the representation for each language over a small corpus of as few as ∼50 sentences, and select the top-K dimensions with the largest absolute differences between target language representation against English representation. In the inference stage, we overwrite these language-specific dimensions with the corresponding representation values, scaled by a coefficient, while leaving all other dimensions unchanged. By applying this single-layer intervention, our method can switch the output language to the desired target while preserving the underlying semantic content. We evaluate the approach across multiple models, including Llama2 (Touvron et al., 2023), Llama3.1 (Dubey et al., 2024), and Aya23 (Aryabumi et al., 2024), across five languages—French, German, Spanish, Chinese, and Japanese. The results confirm the existence of language-specific dimensions and show that our method outperforms neuron-based methods at steering the model to the desired target language.

In summary, our contribution is three-fold:

1. We provide a systematic analysis of language-specific dimensions in LLMs, showing that they play a central role in controlling the projection from a shared, language-agnostic space into language-specific token spaces.

2. We introduce a simple, training-free approach to identify and manipulate these dimensions using only a small amount of monolingual data (∼50 sentences), enabling efficient and explainable control of output language without large-scale logging or auxiliary training.

3. Through multilingual generation control experiments, we show that our method outperforms prior neuron-level approaches and could be further improved with parallel data.

2 Related Work

2.1 Multilingual LLMs

Multilingual processing has been a goal since the earliest transformer-based language models such as multilingual BERT (Devlin et al., 2019) and mBART (Liu et al., 2020). The most direct route to multilingual capability remains large-scale pretraining on mixed-language corpora, systems such as Llama2, 3.1 (Touvron et al., 2023; Dubey et al., 2024) and Aya23 (Aryabumi et al., 2024) can handle dozens of languages and achieve strong results on multilingual benchmarks. As various studies scale multilingual pretraining to boost performance, understanding how pretrained models carry out multilingual tasks, their internal representational factors, remains a key research direction.

2.2 Interpreting Multilingualism in LLMs

Prior work (Wendler et al., 2024) finds a layerwise progression in LLMs such as Llama2, where early layers encode general lexical and semantic features, middle layers representations tend to reside in an English-dominant space, with later layers shifting toward target language space, depending on prompt. In English-centric models such as Llama2, several studies (Wang et al., 2024; Mahmoud et al., 2025) train projection metrics that map representations from English space into target language spaces, by leveraging parallel corpora to obtain representations of the same semantic content across language spaces inside the model. Applying these mappings at inference can steer the output language and improve multilingual behavior.

Other methods, including FFN-neuron manipulation (Kojima et al., 2024; Tang et al., 2024; Sundar et al., 2025) and SAE-based interventions (Deng

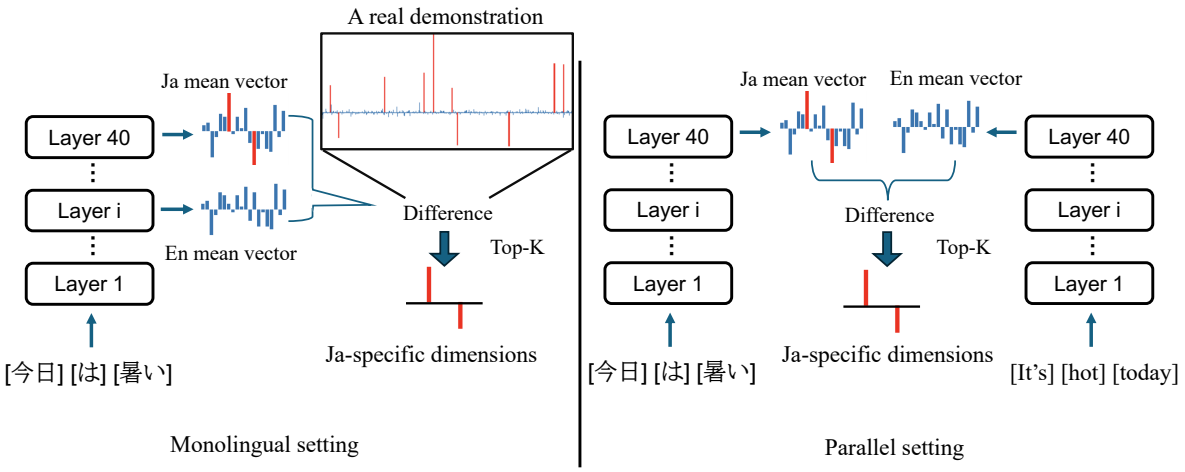


Figure 2: Two settings for identifying language-specific dimensions. Using Japanese as an example, we first average token representations within each sentence to obtain a sentence-level mean vector. Then, compare the Japanese and English means and keep the top-K dimensions with the largest absolute differences as Japanese-related dimensions. We demonstrate a real difference example of LLaMA2-13B on the left top.

et al., 2025), also enable language control. However, like the foregoing approaches, they generally rely on large datasets, frequently parallel data, which are difficult to obtain in many languages and domains, thereby limiting applicability. Thus, we seek a simple, training-free and highly data-efficient method that can achieve this objective.

3 Methodology

We first propose a method to identify language-specific dimensions. Given data availability constraints, we propose two scenarios to identify language-specific dimensions. (1) In Section 3.2.1 monolingual setting, we compare representations from an intermediate layer with those from the final layer to identify language-specific dimensions. (2) In Section 3.2.2 parallel setting, we compare final layer representations for English and the target language to identify language-specific dimensions. Then, we apply intervention to verify their effectiveness, with the setup described in Section 3.3. We modify only these dimensions at a chosen layer to steer the model toward the desired language while keeping language-agnostic semantics approximately stable. The procedure is described in Fig. 2.

3.1 Preliminary

As a preliminary step, we investigate whether the distribution of dimensions involved in the transition from intermediate to final layers exhibits systematic patterns related to language. Specifically, we analyze the Llama2-13B model, focusing on the transition from intermediate layer 22 to the final layer 40, as shown in Fig. 2. Following Wendler et al. (2024), we apply logit-lens and observe that layer 22 already begins to decode semantic content in English, indicating that the core semantics are already established at this stage. When comparing the representations of the intermediate layer with those of the final layer, the majority of dimensions exhibit only minor changes, whereas a small and sparse subset, as marked red in the actual difference demonstration Fig. 2, shows pronounced differences. Notably, we find that the same sparse set of dimensions also emerges when comparing later layers (>22) to the final layer, indicating that these spike indices are consistent across depth. This sparse "spike" distribution coincides with the logit-lens observation that the decoded representation shifts from English concepts to concepts in the target language. These preliminary findings motivate our hypothesis that language-specific dimensions govern the transition from a language-agnostic space to language-specific token spaces, and they form the foundation of our method for systematically identifying and manipulating such dimensions.

3.2 Identify Language-specific Dimensions

To identify language-specific dimensions, we compare target-language representations with a unified representation of the same underlying semantics, and localize the dimensions exhibiting sharp differences. In the monolingual setting, we contrast the final-layer representation with that of an inter-

mediate layer. In the parallel setting, given that the intermediate layer representation distribution aligns closely with English, we leverage parallel sentences to obtain paired representations in English and the target language.

3.2.1 Monolingual setting

First, we treat the LLM as an encoder. For each target-language sentence, the model produces token-level representations at every layer. We compute the mean of token-level representations at (i) a chosen intermediate layer and (ii) the final layer, yielding two sentence vectors per input. We then average these sentence vectors across all sentences to obtain corpus-level mean vectors for the intermediate and final layers, denoted by $\boldsymbol{\mu}_{\ell}^{(\text{lang})} \in \mathbb{R}^d$ and $\boldsymbol{\mu}_{L}^{\text{lang}} \in \mathbb{R}^d$, respectively. Finally, we compute the absolute difference per dimension

$$\boldsymbol{\delta}_{\text{diff}}^{(\text{lang})} = \left|\boldsymbol{\mu}_{L}^{(\text{lang})} - \boldsymbol{\mu}_{\ell}^{(\text{lang})}\right| \in \mathbb{R}^d, \quad (1)$$

and select the top-K indices

$$\mathcal{I}_{K}^{(\text{lang})} = \text{TopK}\big(\boldsymbol{\delta}_{\text{diff}}^{(\text{lang})}, K\big). \quad (2)$$

These dimensions are considered language-specific and their indices are recorded for the intervention step.

3.2.2 Parallel setting

With parallel data, we directly compare the final layer representations of parallel sentences across English and target language. For each parallel pair, we collect final layer sentence representations as in the monolingual setting. We then compute the corpus-level means at the final layer L for English and another for the target language, denoted by $\boldsymbol{\mu}_{L}^{(\text{en})} \in \mathbb{R}^d$ and $\boldsymbol{\mu}_{L}^{(\text{lang})} \in \mathbb{R}^d$, respectively. Then, we select the top-K dimensions with the largest absolute differences between the two language means use the same strategy as monolingual setting; these dimensions are taken as the language-specific dimensions for the target language, we record their index $\mathcal{I}_{K}^{(\text{lang})}$for the intervention step.

3.3 Inference-time Interventions

Given the identified language-specific dimensions, we perform inference-time interventions to validate the effectiveness of these dimensions. We overwrite the values of the identified dimensions to switch the output language to the desired target while preserving semantic content. These interventions can evaluate our central hypothesis that language-specific dimensions exist in the representations of the model.

We run inference on English inputs and intervene in a chosen intermediate layer j. Let $\mathbf{h}_j \in \mathbb{R}^d$ denote the representation at layer j, and let $\mathcal{I}_{K}^{(\text{lang})}$ be the top-K index set for target language $lang$, with corpus-level means $\boldsymbol{\mu}_{L}^{(lang)} \in \mathbb{R}^d$. Both $\mathcal{I}_{K}^{(\text{lang})}$ and $\boldsymbol{\mu}_{L}^{(lang)}$ are precomputed in Section 3.2. We overwrite the selected dimensions with a scaled corpus-level mean value and keep the rest unchanged:

$$\mathbf{h}'_j[i] = \begin{cases} \alpha * \boldsymbol{\mu}_{L}^{(lang)}[i], & \text{if } i \in \mathcal{I}_{K}^{(\text{lang})}, \\ \mathbf{h}_j[i], & \text{otherwise.} \end{cases} \quad (3)$$

Here α is a layer-dependent scaling coefficient controlling the intervention strength (later layers typically use larger α).

4 Experimental Settings

4.1 Models

We evaluate several models, including Llama2–7B, Llama2–13B, Llama3.1–8B, and Aya23–8B. These models come from different LLM families, are trained on different corpora, and vary in size. The Llama2 family models are predominantly English-centric, whereas the others include more multilingual data. Llama2–13B consists of 40 Transformer layers with a hidden size of 5,120. The other models use 32 layers with a hidden size of 4,096. All models are open-sourced on Hugging Face.

4.2 Data for Identification

We evaluate five English-to-X directions, including English to French, German, Spanish, Chinese and Japanese. We randomly sample 50 parallel sentence pairs from the development (dev) set of the FLORES-200 (team et al., 2022) for identifying language-specific dimensions.

4.3 Multilingual Generation Control

We follow Kojima et al. (2024) and adopt the multilingual generation control task for comparison. In this task, the model is prompted with a specially designed machine translation–like instruction: "Translate an English sentence into a target language. English: {source text} Target language:." We then apply interventions with our method to steer the output to a desired language. Because the target language is not explicitly specified, this setting tests whether the identified language-specific dimensions can control the output language. Also,

	zh	ja	fr	es	de
zh	400	**193**	105	100	118
ja		400	88	98	113
fr			400	**152**	143
es				400	140
de					400

Table 1: Overlap matrix of top-400 language-specific dimensions across languages for Llama2-7B in the monolingual setting. Bold numbers indicate language pairs that share more dimensions.

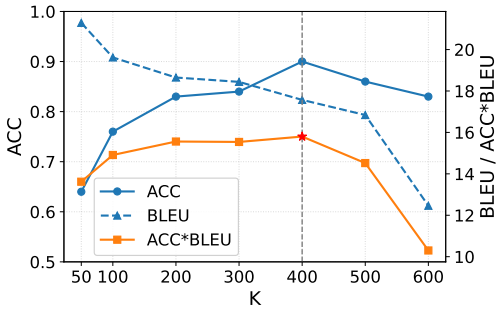


Figure 3: Selecting a top-K threshold for detecting language-specific dimensions. Ablation study conducted on Llama2-7B in the parallel setting.

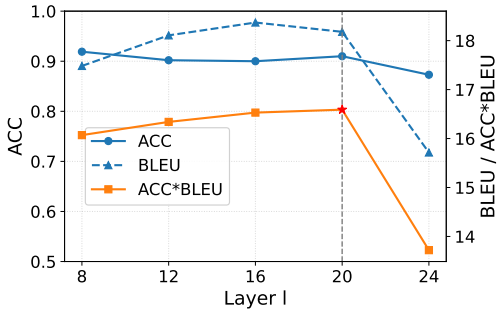


Figure 4: Choosing an anchor layer for monolingual setting. Ablation study conducted on Llama2-7B in the monolingual setting.

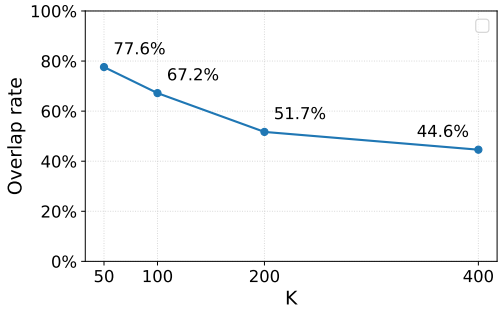


Figure 5: Overlap rate of language-specific dimensions identified by monolingual and parallel settings for different K. Experiments are conducted on Llama2-7B.

we use BLEU (Papineni et al., 2002) to assess whether the intervention preserves the source meaning while changing the output language.

For evaluation, we follow Kojima et al. (2024) and use the same 100 sampled test sentences per translation direction from each dataset of FLORES-200, IWSLT2017 (Cettolo et al., 2017), and WMT (Bojar et al., 2014, 2016, 2018) and compute the mean performance. If a dataset does not contain the given direction (e.g., IWSLT2017 lacks English to Spanish), it is omitted and the average is taken over the remaining.

We evaluate with a pipeline of three metrics:

- Accuracy (ACC): measures the probability that the intervention successfully induces generation in the target language.
- BLEU: evaluates translation quality for successful samples.
- ACC*BLEU: a composite score that reflects both success rate and translation quality.

For calculating accuracy, we use the fastText (Joulin et al., 2017) language identification classifier to detect the output language. A sample is counted as successful if the classifier's score exceeds a threshold of 0.5 (Kojima et al., 2024).

5 Results and Discussion

5.1 Pre-examination on language-specific dimension identification

We first search for an appropriate top-K threshold for selecting language-specific dimensions. Therefore, we run the multilingual generation control experiment on Llama2-7B, applying an intervention at layer 20 with $\alpha = 0.4$ in the parallel setting while varying K. As shown in Fig. 3, as K increases, ACC improves but BLEU drops. The composite ACC*BLEU improves from 13.62 at $K = 50$, to 14.91 at $K = 100$ (+1.29), and to 15.80 at $K = 400$ (+2.18). Considering this trade-off, the most essential language-specific dimensions appear to be captured at around $K = 100$, which yields a smaller loss in semantic fidelity while still improving accuracy. However, to maximize performance, we set $K = 400$ for the subsequent experiments, which is about 7.8% of the hidden size for Llama2-13B and

Model	Method	Fr		De		Zh		Ja		Es		Overall		
		ACC	BLEU	ACC	BLEU	ACC	BLEU	ACC	BLEU	ACC	BLEU	ACC	BLEU	A*B
Llama2-7b	**Neuron Kojima**	42.0	15.3	55.7	15.1	81.3	10.2	57.5	7.5	77.0	16.6	60.8	12.8	7.76
	Neuron Tang	83.7	**28.2**	72.7	**21.1**	86.7	**17.9**	20.5	**14.4**	98.0	**20.8**	72.3	**20.9**	15.14
	Ours (Monolingual)	98.3	23.4	97.3	19.4	84.7	16.4	**76.5**	9.9	98.7	17.30	91.1	18.3	**16.63**
	Ours (Parallel)	**99.1**	22.7	**98.8**	18.3	**88.3**	17.2	**76.5**	9.4	**99.3**	17.4	**92.6**	17.9	16.57
Llama2-13b	**Neuron Kojima**	15.7	7.3	30.3	9.5	96.0	15.9	64.5	8.6	14.0	4.3	47.4	9.9	4.71
	Neuron Tang	83.7	**32.5**	14.7	**23.5**	**99.3**	**19.1**	64.5	8.6	42.0	**25.8**	53.9	**21.4**	11.53
	Ours (Monolingual)	**96.2**	23.3	**99.2**	17.5	97.6	16.1	91.5	8.9	97.0	11.2	**96.9**	16.7	16.14
	Ours (Parallel)	93.1	26.0	98.6	17.9	99.1	17.6	**92.5**	**9.4**	**98.7**	13.4	96.3	18.1	**17.40**
Llama3.1-8b	**Neuron Kojima**	–	–	–	–	–	–	–	–	–	–	-	-	–
	Neuron Tang	–	–	–	–	–	–	–	–	–	–	-	-	–
	Ours (Mono)	**97.2**	**28.4**	**97.6**	**19.0**	**93.9**	**21.0**	**80.8**	**11.3**	**99.0**	**18.4**	**93.9**	**20.7**	**19.47**
	Ours (Parallel)	70.3	11.9	79.3	6.9	78.1	14.4	69.8	5.4	98.0	9.9	76.8	10.0	7.70
Aya23-8b	**Neuron Kojima**	–	–	–	–	–	–	–	–	–	–	-	-	–
	Neuron Tang	–	–	–	–	–	–	–	–	–	–	-	-	–
	Ours (Monolingual)	68.0	**21.7**	**69.4**	**14.7**	**22.8**	20.3	69.8	10.8	58.7	14.0	55.6	16.5	9.34
	Ours (Parallel)	**88.0**	20.1	58.8	14.4	**22.8**	**25.6**	**76.0**	**12.8**	**80.0**	**16.1**	**61.7**	**17.3**	**10.69**

Table 2: Result of multilingual generation control. The best performance for each metric is highlighted in bold. A*B stands for ACC*BLEU. Higher values indicate better performance. Results are averaged over three runs with distinct random seeds used to sample sentences for identifying language-specific dimensions.

9.8% for the other models.

In the monolingual setting, we need to choose an intermediate reference layer to compare with the final layer. We therefore set $K = 400$, apply an intervention at layer 20 with $\alpha = 0.4$, and evaluate Llama2-7B on the multilingual generation control task while varying the layer l from which dimensions are identified. As shown in Fig. 4, ACC and BLEU remain stable over a broad range of l, and the composite ACC*BLEU attains its best value near $l = 20$ but degrades at very deep layers. This indicates that the results are not highly sensitive to the choice of intermediate layer, and we use $l = 20$ in subsequent experiments.

After identifying language-specific dimensions, we examine whether different languages share them. Table 1 confirms clear cross-language overlap. Typologically closer pairs share more dimensions: Chinese and Japanese share 193 of 400 dimensions. Romance and Germanic languages also overlap more with one another. Specifically, French shares 152 and 143 dimensions with Spanish and German, respectively, substantially more than with Chinese (100) and Japanese (98). These results indicate that many language-specific dimensions are partially shared across languages rather than being unique to a single language.

Additionally, we assess the consistency of the language-specific dimensions identified by the monolingual and parallel settings. As shown in Fig. 5, when $K = 400$ the average agreement across languages is 44.6%. As K is reduced to 50, the agreement increases to 77.6%. This suggests that the two settings converge on the core language-specific dimensions.

For the experiments in the later sections, as established in earlier results, we fix top-$K = 400$ language-specific dimension identification for all models. In the monolingual setting, we compute differences between an intermediate layer and the final layer, using layer 20 for Llama2–7B, Llama3.1–8B, and Aya23–8B, and layer 22 for Llama2–13B. For the multilingual generation control task, we then apply single-layer inference-time

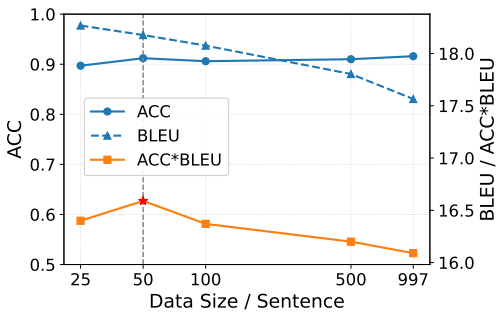
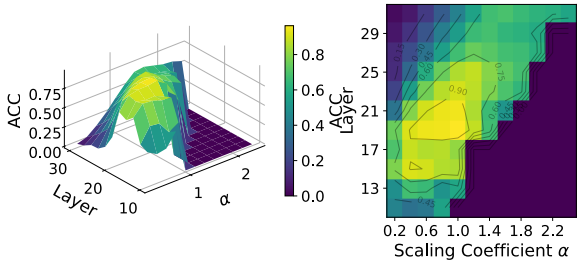
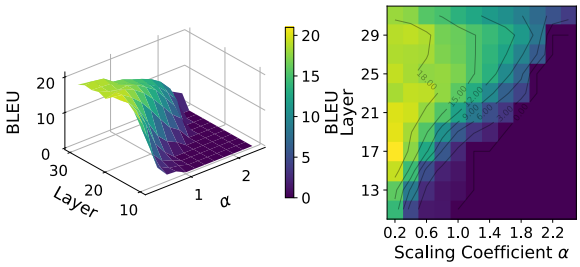


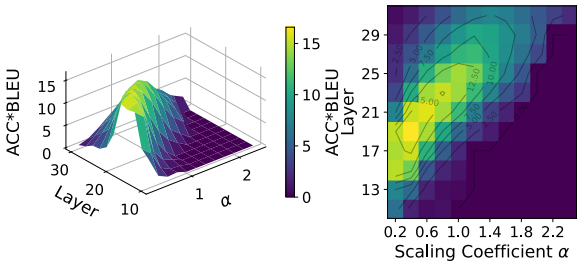
Figure 6: Identify Language-specific Dimensions. Ablation study on Llama2-7B in the monolingual setting.



(a) Overall ACC



(b) Overall BLEU



(c) Overall ACC*BLEU

Figure 7: Grid search results of Llama2-7B in monolingual setting. Purple regions denote unevaluated configurations; their values are set to zero for visualization.

interventions with scaling coefficient α. The intervention layer and α are selected via a grid search, hyper-parameter details are provided in Section 5.4 and Appendix A. Table 2 reports the results obtained under the best settings for each model.

5.2 Main Results of Controlling Multilingual Generation

We validate our method on the multilingual generation control task as described in section 4.3. In this task, we compare our method with two neuron-based intervention baselines. Neuron-Kojima (Kojima et al., 2024) ranks neurons by their ability to predict the target language from per sentence mean activations, then at inference replaces the selected neurons' activations to steer the output. Neuron-Tang (Tang et al., 2024) selects neurons with low cross language activation entropy after thresholding and controls generation by selectively turning these neurons on or off. We use the official code and settings for all experiments.

We show the results in Table 2, evaluated on four LLMs across five language directions. Both variants of our approach can steer the models to produce the desired target language. Considering ACC*BLEU, our monolingual setting achieves an overall score of 16.63 on Llama2-7B, surpassing Neuron-Kojima's 7.76 by more than two times and 1.49 points above Neuron-Tang's 15.14. On Llama2-13B, it reaches 16.14, outperforming Neuron-Kojima by 11.43 points, and Neuron-Tang by 4.61 points. In the parallel setting, our method attains 16.57 on Llama2-7B, roughly comparable with the monolingual setting. On Llama2-13B, the performance reaches 17.40, improving over the monolingual results by 1.26 points.

To assess generality, we evaluate on two latest models—Llama3.1–8B and Aya23–8B. On Llama3.1–8B in the monolingual setting, the overall ACC*BLEU reaches 18.70, the best among the settings we evaluated. On Aya23–8B in the parallel setting, it reaches 10.51. These results support our hypothesis about language-specific dimensions and indicate broad applicability across model families. We omit neuron-based baselines, as they are not compatible with these newer models.

5.3 Are 50 Sentences Sufficient for Identification

When identifying language-specific dimensions, we randomly sample 50 parallel sentence pairs from the FLORES-200. While using a small corpus reduces data and compute requirements, it can also introduce greater stochasticity. To assess whether larger data samples could reduce stochasticity and improve the accuracy of language-specific dimension identification, we conduct experiments with increased corpus sizes. The results are shown in Fig. 6. The experiments are conducted on Llama2-7B, with an intervention applied at layer 20 with

Input	-	Translate an English sentence into a target language.\n English: **Why might someone prefer to shop at a small, locally-owned business instead of a large chain store, even if the prices are higher?**\n Target language:
Output	$\alpha = 0.5$	为什么一个人可能会选择在小型本地商店购物，而不是大型连锁商店？尽管价格更高
	$\alpha = 0.7$	为什么一个人可能会选择在小型本地商店购物，而不是大型连锁商店，尽管价格更高？
	$\alpha = 0.9$	为什么一个人可能会选择在小型本地商店购物，而不是大型链式商店，尽管价格更高？

Table 3: Language steering examples with different intervention strengths. Examples are from Llama2–7B with an intervention at layer 27. They show how α affects generation. Characters in gray indicate content missing relative to the source sentence; characters in red indicate content that is mistranslated.

$\alpha = 0.4$. As corpus size increases, ACC*BLEU performance does not improve. This indicates that our method achieves strong performance with minimal data, obviating the need for large corpora.

5.4 Layerwise Analysis of Language-Specific Dimensions

To test our hypothesis that language-specific dimensions are consistent across layers, we apply an intervention to each intermediate layer while varying the scaling coefficient α. We visualize the grid search results of Llama2-7B with monolingual setting in Fig. 7. As shown in Fig. 7, for most intermediate layers, intervening on the same set of language-specific dimensions and choosing an appropriate α can yield a success rate exceeding 60% of steering the output to the target language, supporting our hypothesis that language-specific dimensions are aligned across layers.

From the figure, we observe that achieving higher accuracy requires larger scaling coefficient α at later layers. This is because the magnitude of the representation increases with depth, so stronger interventions are needed to dominate the signal. In contrast, translation quality, measured by BLEU, tends to decrease as α increases. A plausible explanation is that strong interventions rigidly steer the model toward the target language, reducing its freedom to plan and revise the generation. The model then produces more literal, word-by-word translations with lower fluency. An illustrative example with Llama2–7B is shown in Table 3. With the intervention applied at layer 27, at $\alpha = 0.5$, the model omits the clause "even if the prices are higher." At $\alpha = 0.9$, the word "chain" is rendered as "链式," whereas the idiomatic translation is "连锁商店" (chain store). At $\alpha = 0.7$, we get a good translation. This illustrates that an appropriate α should be chosen to let the model translate into the right target language while preserving some freedom to produce more refined generation. We also observe higher BLEU when the intervention is applied to earlier layers, indicating that early target-language signals allow later layers to devote more capacity to improving output quality.

6 Conclusion

In this work, we investigate how LLMs encode language identity by tracing the trajectory of representations across layers. We show that the transition from a English-aligned representation space to target-language token spaces is governed by a small and sparse set of language-specific dimensions that are consistent across layers. We introduc a simple, training-free method to identify and manipulate these dimensions, using as few as 50 sentences. Our inference-time interventions demonstrate that manipulating these dimensions enables controlled language switching while preserving semantic content, thereby revealing their interpretability and practical utility. Experimental results on multilingual generation control task confirm the effectiveness of our approach: it achieves strong performance in both monolingual and parallel settings, surpassing Neuron-Kojima by up to 12.69 points and Neuron-Tang by up to 5.87 points. Moreover, the parallel setting is generally stronger than the monolingual setting. These findings establish language-specific dimensions as a consistent and controllable mechanism underlying cross-lingual transfer in LLMs, opening new directions for both interpretability and efficient multilingual control.

7 Limitations

Most ablations, such as anchor layer, top-K, are conducted on Llama2–7B and then transferred to Llama3.1–8B and Aya23–8B. Although these models have similar depths, the optimal hyperparameters may differ by model. As shown in Table 2, in the parallel setting, the performance of Llama3.1–8B is markedly lower than in the monolingual setting, suggesting that the transferred hyperparameters are suboptimal for this model/setting. A systematic per-model search is left to future work.

Our empirical evaluation is limited to multilingual generation control with a specific prompt type. While our method surpasses neuron-based baselines in this setting, the extent to which these gains transfer to downstream applications such as question answering and summarization remains open. A more comprehensive assessment would combine task-specific metrics with human judgments and examine robustness across prompts.

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr F. Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, A. Ustun, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *ArXiv*, abs/2405.15032.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, and 2 others. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. 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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Omar Mahmoud, Buddhika Laknath Semage, Thommen George Karimpanal, and Santu Rana. 2025. Improving multilingual language models by aligning representations through steering. *ArXiv*, abs/2505.12584.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2024-07-28.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

Nllb team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cris tian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2024. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Annual Meeting of the Association for Computational Linguistics*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *ArXiv*, abs/2402.10588.

Chengzhi Zhong, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2025. What language do non-English-centric large language models think in? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26333–26346, Vienna, Austria. Association for Computational Linguistics.

A Details of Intervention Setting

The intervention configurations used to obtain the results in Table 2 are: Llama2–7B—layer 19, $\alpha = 0.4$ for monolingual and parallel; Llama2–13B—layer 19, $\alpha = 0.4$ for monolingual and $\alpha = 0.3$ for parallel; Llama3.1–8B—layer 21, $\alpha = 0.4$ for monolingual and layer 27, $\alpha = 0.8$ for parallel; Aya23–8B—layer 29, $\alpha = 1.2$ for monolingual and layer 27, $\alpha = 0.8$ for parallel. The intervention layer and α are selected via a grid search

A.1 Choosing an anchor layer for Llama2-13B

Ablation on Llama2–13B shows that using layer 22 as the intermediate reference and contrasting it with the final layer yields the most effective set of language-specific dimensions.

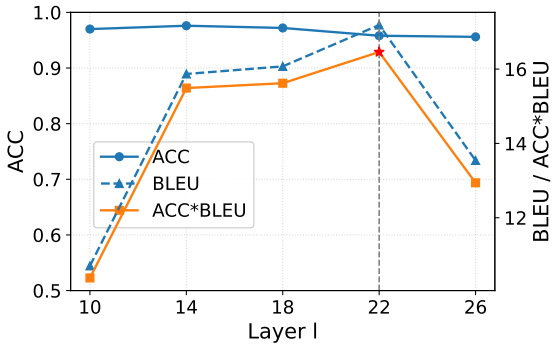


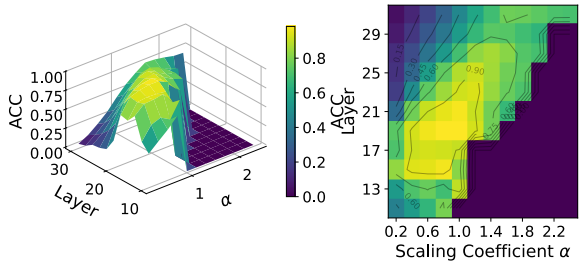
Figure 8: Choosing an anchor layer for monolingual setting. Ablation on Llama2–13B in the monolingual setting.

A.2 Extra Experiments

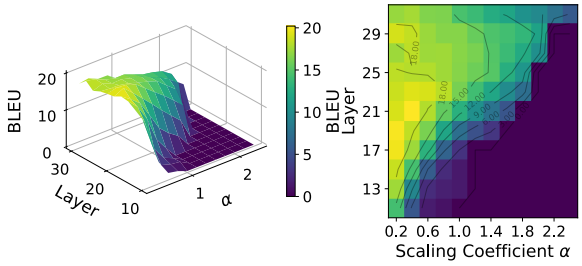
We present grid search results over the intervention layer and the scaling coefficient α for Llama2–7B in parallel setting, and Llama2–13B in monolingual and parallel settings.

Based on the grid search, the optimal intervention settings are as follows: for Llama2–7B under the parallel setting, layer 19 with $\alpha = 0.4$; for

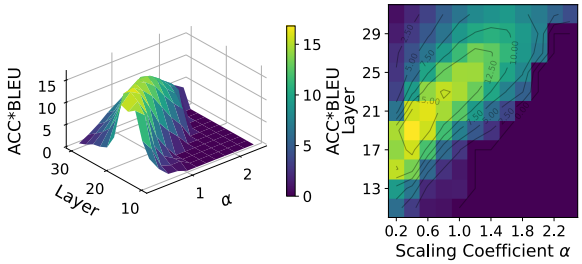
Llama2–13B under the monolingual setting, layer 19 with $\alpha = 0.4$; for Llama2–13B under the parallel setting, layer 19 with $\alpha = 0.3$.



(a) Overall ACC

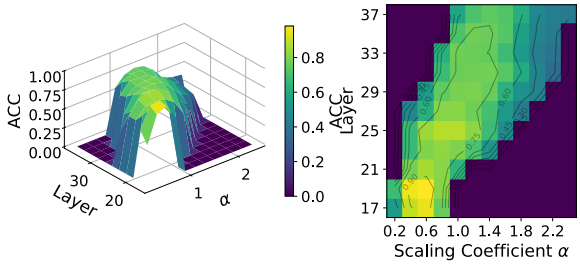


(b) Overall BLEU

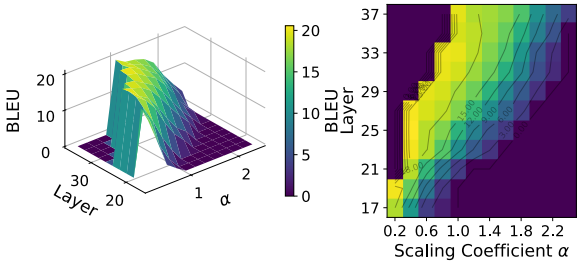


(c) Overall ACC*BLEU

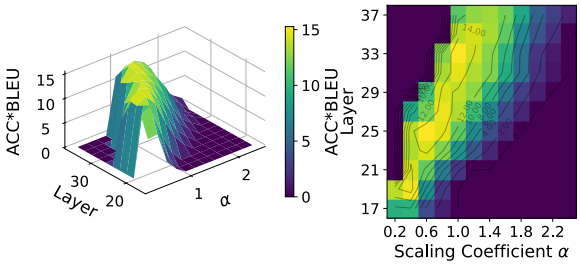
Figure 9: Grid search results of Llama2-7B in parallel setting. The results show that the most effective intervention should be applied to layer 19, with $\alpha = 0.4$.



(a) Overall ACC

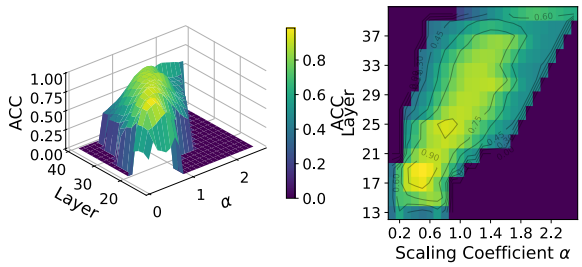


(b) Overall BLEU

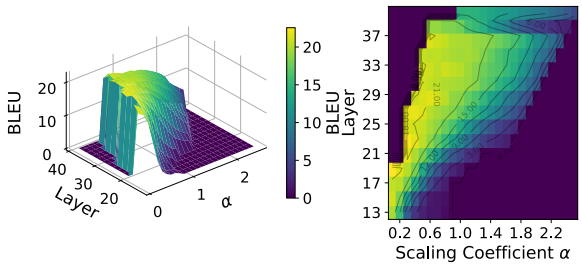


(c) Overall ACC*BLEU

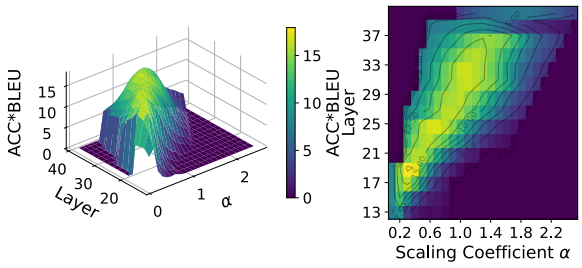
Figure 10: Grid search results of Llama2-13B in monolingual setting. The results show that the most effective intervention should be applied to layer 19, with $\alpha = 0.4$.



(a) Overall ACC



(b) Overall BLEU



(c) Overall ACC*BLEU

Figure 11: Grid search results of Llama2-13B in parallel setting. The results show that the most effective intervention should be applied to layer 19, with $\alpha = 0.3$.