CoCo-CoLa: Evaluating and Improving Language Adherence in Multilingual LLMs

Elnaz Rahmati[*] **Alireza S. Ziabari**[*] **Morteza Dehghani**

University of Southern California
{erahmati, salkhord, mdehghan}@usc.edu

Abstract

Multilingual Large Language Models (LLMs) develop cross-lingual abilities despite being trained on limited parallel data. However, they often struggle to generate responses in the intended language, favoring high-resource languages such as English. In this work, we introduce *CoCo-CoLa* (Correct Concept - Correct Language), a novel metric to evaluate language adherence in multilingual LLMs. Using fine-tuning experiments on a closed-book QA task across seven languages, we analyze how training in one language affects others' performance. Our findings reveal that multilingual models share task knowledge across languages but exhibit biases in the selection of output language. We identify language-specific layers, showing that final layers play a crucial role in determining output language. Accordingly, we propose a partial training strategy that selectively fine-tunes key layers, improving language adherence while reducing computational cost. Our method achieves comparable or superior performance to full fine-tuning, particularly for low-resource languages, offering a more efficient multilingual adaptation.[1]

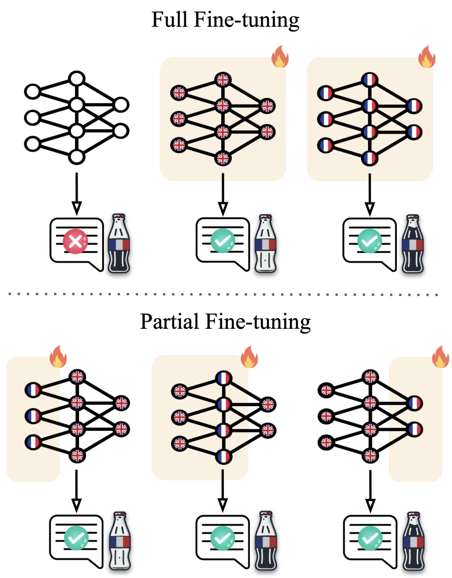

Figure 1: Evaluation of correctness and language adherence on French input. The soda level visualizes the CoCo-CoLa ratio, with higher levels indicating stronger adherence to the input language. Our results show that partially fine-tuning the final layers of an English-tuned model on French achieves language adherence and accuracy comparable to a model fully fine-tuned on French.

1 Introduction

Multilingual LLMs are pre-trained on raw text from multiple languages, typically consisting of separate corpora for each language. Remarkably, despite this lack of explicit parallel data to facilitate cross-lingual associations, these models develop an implicit understanding of inter-language relations and cross-lingual word associations (Wen-Yi and Mimno, 2023). Instruction tuning further enhances their ability to follow prompts, and models trained on multilingual data often exhibit zero-shot cross-lingual transfer of instruction-following capabilities (Chirkova and Nikoulina, 2024). However, this generalization is uneven: while high-resource languages in pretraining benefit significantly from instruction tuning, lower-resource or unseen languages often struggle to follow instructions reliably, frequently exhibiting degraded performance or defaulting to generating output in a preferred language (Nguyen et al., 2024; Chirkova and Nikoulina, 2024). To address these issues, we investigate how multilingual LLMs learn the same task across different languages.

A crucial step toward addressing the limitations of multilingual LLMs is understanding how they internally process and encode multilingual knowledge. Interpretability research has traditionally focused on monolingual models, leveraging techniques such as representation probing (Orgad et al.,

[*]Equal contribution.

[1]Our code is available at https://github.com/elnazrahmati/CoCo-CoLa/

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 62–77
November 8-9, 2025 ©2025 Association for Computational Linguistics

2024; Saphra and Lopez, 2019) and model patching (Ghandeharioun et al., 2024; García-Carrasco et al., 2024). These methods have been widely used to examine LLMs' performance across tasks such as mathematics (Nikankin et al., 2024; Zhou et al., 2024), and general knowledge (Jiang et al., 2024; Burns et al., 2022; Singh et al., 2024; Golgoon et al., 2024; Chowdhury and Allan, 2024; Rai et al., 2024). Studies on model internals suggest that Multi-Layer Perceptrons (MLPs) retrieve task-relevant information, while attention layers refine and promote the correct response (Geva et al., 2021; Meng et al., 2022). Furthermore, knowledge is often identified in earlier layers and reinforced in later layers (Fan et al., 2024).

However, these interpretability techniques have primarily been applied to monolingual models, which were initially dominant due to the early focus on English-language pertaining (Touvron et al., 2023; Jiang et al., 2023; Team et al., 2024; Abdin et al., 2024). The rise of multilingual LLMs trained on diverse languages (Gao et al., 2024; Shaham et al., 2024; Soykan and Şahin, 2024), necessitates extending interpretability research beyond English. Multilingual LLMs present additional challenges: representations of different languages are intertwined within a shared space; cross-lingual alignment varies across languages; and shared tokens between languages impact their process. These complexities make it difficult to isolate language-specific knowledge, benchmark cross-lingual generalization, and interpret how multilingual LLMs acquire and apply linguistic information. Given the prevalence of mid- and low-resource languages, understanding these mechanisms is crucial not only for improving cross-lingual transfer but also for mitigating the "curse of multilinguality" — the performance degradation observed as the number of supported languages increases.

Recent efforts have begun tackling these challenges by probing internal representations (Li et al., 2024), analyzing the emergence of cross-lingual transfer (Wang et al., 2024a), and studying token representation alignment on cross-lingual transfer (Gaschi et al., 2023). Furthermore, researchers attempt to separate the linguistic abilities from task abilities by developing language- and task-specific adapters (Pfeiffer et al., 2020; Parovic et al., 2023), subnetworks (Choenni et al., 2023), or layers (Bandarkar et al., 2024). However, despite this progress, most prior works treat multilinguality as a monolithic phenomenon, focusing on general cross-lingual transfer or aggregating all languages into a single block of linguistic knowledge. Less attention has been given to understanding how LLMs process individual languages at a more granular level, particularly within the context of task learning.

In this work, we focus on language adherence by first identifying both shared and distinct patterns in cross-lingual task acquisition, revealing how multilingual models internalize and apply linguistic knowledge (Section 3). We find that training on a task in one language improves performance in other languages. However, this benefit is not always directly observable due to an inherent model bias towards generating output in a preferred language, rather than strictly adhering to the input language (Section 4.1). To quantify this bias, we introduce *CoCo-CoLa* (Correct Concept, Correct Language), a novel metric designed to assess a model's ability to generate responses in the intended input language, particularly for languages not included in supervised finetuning (SFT). Furthermore, we propose a *partial training method* that selectively fine-tunes specific model layers which reveals the relation between language adherence and model layers (Section 4.2). This approach enables more efficient language adaptation, achieving comparable or even superior performance compared to full model retraining, especially for low-resource languages. Finally, we show that the issue of language adherence can be addressed by finetuning only the final layers of LLMs on a small balanced multilingual data (Section 4.3).

2 Related Work

This work builds on several active research areas that inform our study of multilingual task learning in LLMs. Specifically, we draw from (1) Multilingual interpretability, which helps us analyze how LLMs process different languages and how their internal structures influence multilingual task learning; (2) Representation alignment, which provides insights into token-level similarities across languages and how shared representations facilitate cross-lingual generalization; (3) Adapters, which separate language knowledge from task-specific knowledge, offering a structured framework for understanding their interactions; and (4) Subnetworks, which identify task- and language-specific parameters within existing models, offering an alternative to external adapters and directly informing our approach to efficient partial training.

Interpretability. Li et al. (2024) use probing techniques to analyze accuracy changes across layers in LLMs, showing that high-resource languages exhibit patterns similar to English, with accuracy increasing from lower to upper layers. However, this pattern is inconsistent for low-resource languages. Wang et al. (2024b) examine cross-lingual transfer by analyzing neuron overlap in different languages using checkpoints from BLOOM's pre-training (Le Scao et al., 2023). They find a strong correlation between neuron overlap and cross-lingual transfer, though neuron overlap does not increase monotonically during training, and patterns vary across model sizes. Similarly, Zhao et al. (2024a) investigate language-specific neurons and assess how these neurons affect both English and non-English language performance.

Representation alignment. Beyond studying multilingualism in LLMs, some research focuses on improving model performance across languages through representation alignment. Gaschi et al. (2023) align English and Arabic model representations using a bilingual dictionary before fine-tuning on a target task. Zhang et al. (2024) align English representations with other languages using question-translation data before instruction-tuning. Additionally, Salesky et al. (2023) introduce a pixel representation method to enhance alignment and improve translation quality.

Adapters. Another approach for cross-lingual transfer involves integrating adapters into the model. This technique is based on the assumption that task-solving knowledge can be separated from language knowledge. Pfeiffer et al. (2020) introduce MAD-X, a framework where language and task adapters are trained separately, with each block's representations passing through a language adapter before a task adapter. Building on this, later works aim to refine adapter creation and composition methods. For instance, Parović et al. (2022) propose BAD-X, which replaces monolingual adapters with bilingual adapters, improving performance for low-resource languages. Zhao et al. (2024b) introduce AdaMergeX, where adapters for language-task pairs are trained independently and later combined through linear operations (addition and subtraction) to generate adapters for new language-task pairs.

Subnetworks. To enhance cross-lingual transfer without adding new parameters, some methods focus on identifying existing task- and language-specific parameters within the model. Choenni et al. (2023) fine-tune models for specific languages or tasks, extract the most affected neurons, and use the resulting subnetworks to enable multilingual task adaptation. Bandarkar et al. (2024) take a layer-wise approach in multiple steps: they train separate language- and task-expert models, analyze parameter changes to identify key layers for language and task learning, and use layer-swapping techniques to create a math expert in a new language. Consistent with Zhao et al. (2024a), their findings suggest that initial and final layers primarily encode linguistic information, while middle layers are task-specific.

3 Preliminary Analysis

In the preliminary section of this paper, we first isolate language effects from task learning by choosing multi-lingual parallel QA data (Section 3.1), examine fine-tuning performance across multiple languages (Section 3.2), explore how well LLMs generalize knowledge across languages (Section 3.3), and which model components are most affected during training (Section 3.4). Then, in Section 4.1, we introduce **CoCo-CoLa** metric to measure language adherence in multilingual LLMs followed by an efficient partial training method to increase the model adherence (Section 4.2).

3.1 Setup

To investigate how multilingual LLMs learn a new task in a monolingual setting, we train four different models on a Closed-Book Question-Answering (CBQA) task. We include two sizes of the Llama-3.2 series (Dubey et al., 2024) to analyze the effect of model size on multilingual performance and behavior, given that these models are specifically optimized for multilingual dialogue. We also include Llama-3.1-8B as a point of comparison, as it, while not explicitly optimized for multilingualism, was trained on a small multilingual corpus. To test generalizability to multilinguialy balanced models, we include Gemma-3-4B (Team et al., 2025), which was trained with UniMax (Chung et al., 2023) for addressing language imbalances.

We select CBQA because it is language-dependent and demonstrates a model's ability to act as a knowledge base (Wang et al., 2021). To isolate the impact of language differences from the effects of learning a new task or acquiring new knowledge, we use the Mintaka CBQA dataset (Sen et al.,

	Llama-1B			Llama-3B			Llama-8B			Gemma-4B		
Language	PLM	SFT	Δ	PLM	SFT	Δ	PLM	SFT	Δ	PLM	SFT	Δ
English	13.27	38.44	25.17	32.85	53.09	20.24	12.92	50.98	38.06	30.69	53.67	22.98
French	11.30	40.27	28.97	22.90	43.80	20.90	18.53	50.85	32.32	21.67	48.43	26.76
German	7.16	40.34	33.18	23.79	48.10	24.31	11.04	44.35	33.31	23.76	45.79	22.03
Hindi	5.27	21.18	15.91	7.33	30.39	23.06	6.21	35.29	29.08	8.84	43.96	35.12
Italian	7.06	41.58	34.52	21.87	42.73	20.86	16.48	43.22	26.74	20.44	50.05	29.61
Portuguese	5.38	38.23	32.85	20.06	37.04	16.98	18.38	31.11	12.73	19.96	44.16	24.20
Spanish	6.13	41.71	35.58	22.01	45.69	23.68	16.60	45.46	28.86	26.33	48.13	21.80

Table 1: Performance of pre-trained models (PLM), fine-tuned models (SFT), and their difference (Δ = SFT - PLM) on CBQA data across languages.

2022). Mintaka provides identical question-answer pairs in nine languages, allowing us to keep the question content consistent and thus isolate the influence of language itself. The dataset was originally created in English and later translated into Arabic, French, German, Hindi, Italian, Japanese, Portuguese, and Spanish.

One challenge with Mintaka is that some answer types are not translated across languages. To keep question-answer pairs within the same language, we use Google Translate to convert these answers into the language of their respective questions and apply back-translation for accuracy checks. Additionally, since our goal is to study how models learn new tasks in languages they have been exposed to before, we exclude Arabic and Japanese.

3.2 SFT Performance

Our initial step is to assess the model's ability to learn the task in each individual language, effectively measuring how learning difficulty varies across languages. To do this, we perform SFT for all models on each language of the CBQA dataset for three epochs and generate answers for given questions. Next, we select the best model based on the validation loss. Further implementation details are provided in Appendix A.1.

Table 1 shows a comparison of accuracy between the pre-trained model and the best checkpoint of the language-specific SFT model across different languages. SFT significantly improves performance for all languages with relatively consistent accuracy levels, except for Hindi in all Llama model sizes and Portuguese for Llama-8B, which exhibit notably lower accuracy. This discrepancy is likely due to undertraining. Among the SFT models, English achieves the highest accuracy in all models, except Llama-1B that performs best in Spanish. The largest accuracy gains are observed in English (+38.06%) for Llama-8B, German (+24.31%) for Llama-3B, Spanish (+35.58%) for Llama-1B, and Hindi (+35.12%) for Gemma-4B, indicating that these languages benefited the most from fine-tuning. The comparable accuracy across languages indicates comparable knowledge acquisition.

However, two critical questions remain: (1) Do models share learned knowledge uniformly across languages, or do they correctly answer distinct subsets of questions depending on the language? (2) Are there specific parts of the model that are responsible for encoding language-specific information? To address these questions, we first analyze the overlap in correct answers across languages using the Jaccard Index, followed by an investigation of parameter updates to determine whether certain components of the model specialize in handling linguistic differences.

3.3 Cross-lingual Task Knowledge

To further investigate the extent of cross-lingual task knowledge transfer within the model, we analyze the overlap in correct answers across languages. Specifically, we measure how consistently the model arrives at the same correct answers in different languages, providing insight into whether knowledge is shared across languages. It is important to note that there is no overlap between the knowledge present in the training and evaluation data. This ensures that any correct answers during evaluation are derived from knowledge acquired during pretraining rather than memorization. Consequently, the model's ability to generate correct responses across languages indicates that it has internalized the underlying task knowledge from the training data, rather than relying solely on language-specific cues. Let L_A and L_B represent two languages, and let C_{L_A} denote the set of correct answers for L_A. To quantify the degree

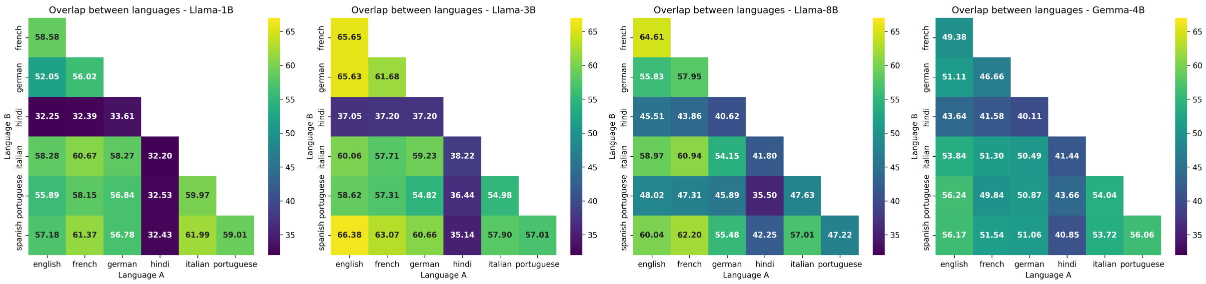

Figure 2: Jaccard similarity index between different languages, measuring the proportion of overlapping correctly answered questions between pairs of languages.

of shared task knowledge between languages, we compute the Jaccard Index, also known as Intersection over Union (IoU), between C_{L_A} and C_{L_B} (see Equation 1). The Jaccard Index is a natural choice for this analysis as it directly measures the proportion of overlapping correct answers relative to the total distinct answers across languages. This allows us to assess knowledge consistency and cross-lingual transfer within the model.

$$IoU(A, B) = \frac{|C_{L_A} \cap C_{L_B}|}{|C_{L_A} \cup C_{L_B}|} \tag{1}$$

The results, shown in Figure 2, indicate that on average approximately 60% of correctly answered questions are shared across languages for all models, suggesting a strong degree of shared knowledge among languages. However, Hindi exhibits significantly lower overlap with other languages in Llama-3.2 models, suggesting weaker generalization for this language. Interestingly, in Llama-8B, Hindi shows higher overlap compared to Llama-3.2 models, but Portuguese experiences a notable drop in overlap. Additionally, Llama-3B demonstrates a higher rate of shared knowledge compared to Llama-8B, despite both models achieving comparable accuracy across languages (see Table 1). This highlights the importance of multilingual optimization in enhancing cross-lingual transfer among languages. For Gemma-4B, despite comparable accuracy across languages, the overall overlap is lower than that observed in the Llama models, indicating less cross-lingual knowledge sharing.

3.4 Parameter Updates

To investigate language-specific encoding in LLMs, we analyze parameter updates during fine-tuning and compare them across languages to determine whether certain components of the model specialize in processing linguistic information. Meng et al. (2022) suggest that MLP modules primarily store knowledge, while attention modules control information retrieval and selection. SFT models correctly answer approximately 40% of evaluation questions in all languages. However, they require fine-tuning to improve their ability to select and output the correct information. As a result, we expect substantial modifications in the attention modules, particularly in the final layers, while changes in the MLP modules remain limited. Since these datasets differ only in language, not in task or knowledge, analyzing the model updates allows us to pinpoint which layers or components are most crucial for learning language-specific representations.

To compute parameter update, we follow Bandarkar et al. (2024) and calculate the average parameter modifications for each module in each layer. Denoting the pre-trained weight matrix as W_p and the fine-tuned weight matrix as W_f, the average magnitude of differences is given by:

$$\Delta W = \frac{1}{n} \sum_{i=1}^{n} |W_p^{(i)} - W_f^{(i)}| \tag{2}$$

The results for English and French are shown in Figure 3, with the remaining languages in Figure 6. As expected, significant modifications occur in the attention modules of the final six layers for Llama-1B and the final 14 layers for Llama-3B, Llama-8B, and Gemma-4B models across all languages. However, in Llama-3.2 models and Gemma-4B model, we observe substantial changes in the MLP modules in these layers for all languages except English, suggesting that these variations might be tied to language-specific processing rather than task-related learning. Surprisingly, for Llama-8B, even the model fine-tuned on English shows a high rate of change similar to other languages. Considering the unexpectedly low accuracy of the Llama-8B pre-trained model across all languages compared to Llama-3B, this larger modification could be related

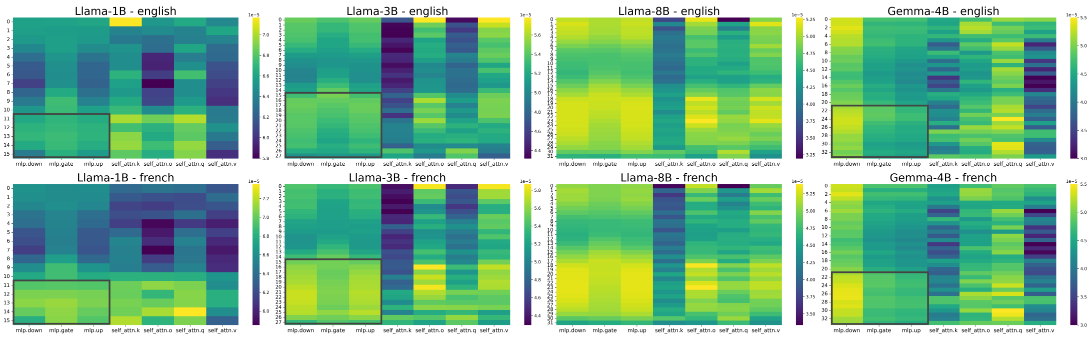

Figure 3: Heatmaps of parameter update magnitudes during monolingual fine-tuning on English (top) and French (bottom) across different LLMs. Gray boxes show MLP modules where parameter update differs between languages.

to learning the task or acquiring new knowledge rather than just language adaptation.

4 Approach

Our previous analysis suggests that while task knowledge is largely shared across languages, the *way* this knowledge is processed and accessed differs. Although a Jaccard Index analysis revealed substantial overlap in correct answers, our investigation of parameter updates showed that models trained on non-English languages required more substantial modifications in their MLP modules compared to English, even when achieving comparable accuracy. This raises an important question: Do these modifications reflect deviations in knowledge acquisition, or are they more related to language generation? In this section, we first introduce a metric to analyze linguistic bias in multilingual LLM outputs. Then, we propose a partial training strategy aimed at reducing this bias by selectively fine-tuning specific model components.

4.1 Correct Concept in Correct Language

According to Dubey et al. (2024), only 8% of the pre-training data used for Llama-3 models is multilingual, while the rest is dominated by English general knowledge, mathematics, and code. This suggests a strong bias toward English. Given this imbalance, we hypothesize that the observed MLP module changes in non-English languages may not indicate new knowledge acquisition but rather adjustment in language selection during response generation. Supporting this, Chirkova and Nikoulina (2024) found that when Llama-2-13B is instruction-tuned on English and tested in other languages, it generates responses in a different language from input language in over 30% of cases, with this behavior influenced by training hyperparameters.

To investigate this further, we introduce **CoCo-CoLa** (**Co**rrect **Co**ncept - **Co**rrect **La**nguage), a metric designed to measure how well the model adheres to the input language while generating correct responses. Let L_i denote the input language, $C_{L_i \to L_o}$ the set of correct output in language L_o when passing language L_i as input. We define the CoCo-CoLa ratio as follows:

$$\text{CoCo-CoLa}(L_i) = \frac{|C_{L_i \to L_i} - \bigcup_{L_o \neq L_i} C_{L_i \to L_o}|}{|C_{L_i \to L_i} \,\Delta\, \bigcup_{L_o \neq L_i} C_{L_i \to L_o}|} \tag{3}$$

The denominator uses the symmetric difference between $C_{L_i \to L_i}$ and correct answers in other languages because many answers involve named entities, such as well-known places, books, and individuals. Since most of the languages in this work use similar scripts, named entities often appear in identical forms across multiple languages. This redundancy leads to overlap between $C_{L_i \to L_i}$ and $\bigcup_{L_o \neq L_i} C_{L_i \to L_o}$, which the symmetric difference helps mitigate by ensuring that shared named entities do not artificially inflate the metric.

Given that these models are primarily trained on English, when the input is in L_i the output is usually either L_i or English. Thus, $\bigcup_{L_o \neq L_i} C_{L_i \to L_o}$ is largely dominated by $C_{L_i \to en}$, meaning that most language switching occurs between the input language and English rather than other languages.

To further simplify the calculation, we filter the data to include only questions where the correct answers in L_i and English are different. Under this condition, $C_{L_i \to L_i} \cap C_{L_i \to en} = \emptyset$, allowing the CoCo-CoLa ratio to reduce to:

$$\text{CoCo-CoLa}(L_i) = \frac{|C_{L_i \to L_i}|}{|C_{L_i \to L_i}| + |C_{L_i \to en}|} \tag{4}$$

Table 2: CoCo-CoLa ratio (Ratio) and cumulative accuracy (Acc) of pretrained model (PLM), English-tuned model ($\rightarrow en$), and L_i-tuned model ($\rightarrow L_i$) across languages for Llama-1B, Llama-3B, Llama-8B, and Gemma-4B.

Language	Metric	Llama-1B			Llama-3B			Llama-8B			Gemma-4B		
		PLM	$\rightarrow en$	$\rightarrow L_i$	PLM	$\rightarrow en$	$\rightarrow L_i$	PLM	$\rightarrow en$	$\rightarrow L_i$	PLM	$\rightarrow en$	$\rightarrow L_i$
French	Acc	12.07	52.66	55.73	20.57	62.55	52.97	12.89	58.64	66.01	18.67	65.16	63.23
	Ratio	49.42	13.47	88.58	52.51	14.73	89.45	58.11	12.32	87.54	50.77	19.22	90.14
German	Acc	8.05	51.97	50.92	16.99	49.30	57.01	10.43	59.95	52.27	15.26	64.59	60.24
	Ratio	53.87	10.50	91.02	56.53	19.64	89.26	57.49	11.03	87.21	42.82	15.23	92.64
Hindi	Acc	8.65	29.34	27.42	15.77	38.26	39.67	9.79	37.29	39.21	12.58	47.81	49.66
	Ratio	43.16	13.28	90.79	31.93	10.04	77.47	43.67	10.74	90.68	40.86	9.39	97.19
Italian	Acc	7.76	51.35	62.39	16.63	53.17	46.02	11.77	61.88	58.55	14.62	62.99	67.98
	Ratio	51.32	10.00	93.60	56.68	16.29	87.91	52.11	10.90	91.35	48.76	14.84	91.08
Portuguese	Acc	10.22	54.85	57.57	17.60	55.52	50.64	16.23	60.75	42.90	17.11	63.81	61.16
	Ratio	56.40	12.73	91.07	63.37	15.99	85.10	51.41	11.49	90.73	51.89	14.98	90.69
Spanish	Acc	9.75	57.52	59.02	19.17	57.55	60.38	14.13	58.34	54.27	17.69	65.88	60.65
	Ratio	52.28	12.01	91.24	61.68	15.84	89.18	61.98	9.40	91.35	51.15	14.70	91.36

To evaluate language adherence and accuracy, we pass the input in L_i to pre-trained, *en-tuned*, and L_i*-tuned* models. We then compute the CoCo-CoLa ratio and the cumulative accuracy, defined as the proportion of correct answers either in L_i or English. The results, presented in Table 2, show that while the *en-tuned* models and the L_i*-tuned* models achieve comparable cumulative accuracy on L_i input, the CoCo-CoLa ratio is significantly lower for the *en-tuned* model. This suggests that although the *en-tuned* model can correctly process the question in L_i and retrieve the correct answer at the same rate as the L_i*-tuned* model, it frequently generates the answer in English instead of L_i. Furthermore, analyzing the CoCo-CoLa ratio of the pre-trained model reveals that the model already exhibits a bias toward generating English responses, though this bias is less pronounced than in the *en-tuned* model. These findings support our hypothesis that the varying rate of parameter updates across languages is related to output language preference. Since the model is already inherently biased toward English, *en-tuned* results in the least MLP change compared to other languages.

4.2 Partial Training for Language Adaptation

In this section, we aim to disentangle task learning from output generation in language L_i. Our previous results reveal two key observations. First, as shown in Section 4.1, both the *en-tuned* model and the L_i*-tuned* model achieve comparable cumulative accuracy on L_i, indicating that they learn the task equally well. The only difference is their CoCo-CoLa score, meaning that while both models understand the task to the same degree, they generate outputs in different languages. Second, from Section 3.4, we observed that the *en-tuned* and L_i*-tuned* models undergo different parameter updates. Some of these updates are necessary for learning the task itself, while others may specifically steer the model toward producing responses in the intended language.

Based on these observations, we hypothesize that fine-tuning specific layers of an *en-tuned* model on L_i can enable it to generate responses in L_i without requiring full model updates. Specifically, these layers correspond to the parameters that were updated in the L_i*-tuned* model but not in the *en-tuned* model. To test this hypothesis, we first identify the layers that undergo language-specific updates. We then fine-tune only these layers in the *en-tuned* model and compare the results to fine-tuning other layers. This comparison allows us to isolate the parameters responsible for output language.

Identifying language layers. We select layers for partial training based on the variation in parameter update rates observed in Section 3.4. For the Llama-1B model, we train three variants by unfreezing different sets of layers: (1) layers 11-16, (2) layers 1-5 (chosen to match the parameter count of the final six layers), and layers 1-10 (including all parameters except the final six). We expect the first variant to be the most language-related and to result in the largest improvement in the CoCo-CoLa ratio, while the other two should have a smaller effect. For Llama-3B and Gemma-4B, we similarly train two variants each: unfreezing layers 15–28 and 1–14 for Llama-3B, and layers 21–34 and 1–20 for Gemma-4B. Again, we

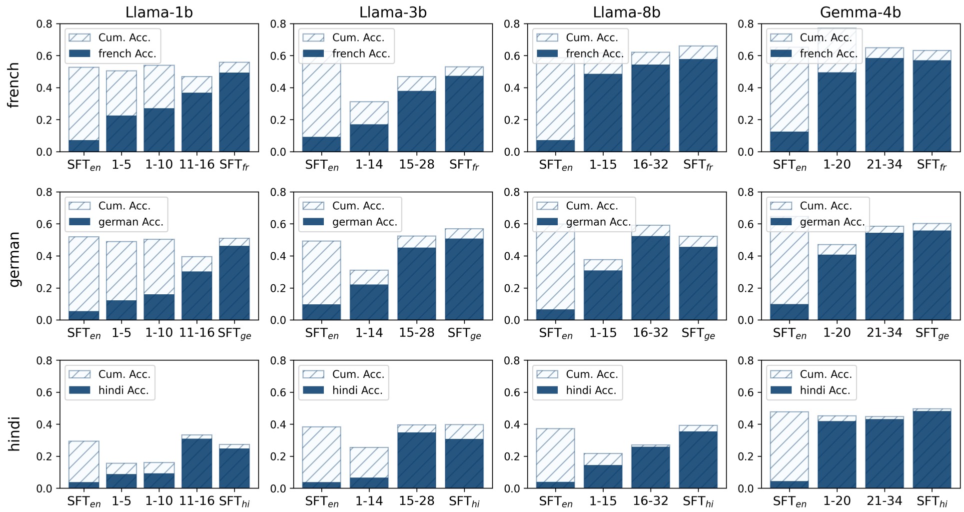

Figure 4: Cumulative accuracy and L_i accuracy on *en-tuned* (SFT_{en}) and L_i-tuned models (SFT_{L_i}), along with partially trained models, across all Llama model sizes.

expect the final-layer variants to have a stronger relationship to language generation. For Llama-8B, which does not show clear variations in update rates across languages (as noted in Section 3.4), we instead select layers based on the most updated MLP modules. Specifically, we choose layers 16–32 and layers 1–15 for partial training to determine which part of the model is more responsible for language generation. Through this analysis, we aim to verify whether the final layers play a greater role in controlling the output language

Partial training evaluation. To evaluate the effectiveness of partial training, we compare all partially trained models to both their fully *en-tuned* and fully L_i*-tuned* models. Figure 4 presents cumulative accuracy and L_i accuracy across three languages, while results for the remaining three languages are included in Figure 7. In addition, CoCo-CoLa ratios for partially trained models are also available in Appendix A.4, providing further insight into the extent to which partial fine-tuning improves output language consistency.

As shown in Figure 4, among the partially trained models, unfreezing the final layers results in the highest accuracy and CoCo-CoLa ratio for all models, highlighting the crucial role these layers play in determining the output language. Notably, the accuracy of this partially trained configuration closely approaches that of the fully L_i*-tuned* model, suggesting that the earlier layers already encode sufficient information for question answering, even without direct exposure to L_i during training. Interestingly, Hindi-which initially exhibited lower performance than other languages-benefits significantly from cross-lingual transfer, achieving better results with partial training than with full training in both Llama-3.2 models. Llama-3B demonstrates even stronger cross-lingual transfer, with improved accuracy for Italian and Portuguese as well. For Llama-8B and Gemma-4B, training the second half of the model yields the highest CoCo-CoLa ratio; however, the differences in L_i accuracy across partial training configurations are less pronounced than in the Llama-3.2 models. These models also show improved accuracy with partial training compared to full training for German, Italian, and Portuguese in Llama-8B, and for French, Portuguese, and Spanish in Gemma-4B. For low-resource languages, partially training only the final layers of an *en-tuned* model can achieve similar or even better accuracy compared to full fine-tuning in the target language. Beyond its effectiveness, partial training is significantly more efficient, reducing training time to half and memory usage to 65% of full training. Furthermore, the model achieves higher accuracy in fewer training steps, requiring less than one epoch, meaning it is trained on fewer data points.

These findings confirm the hypothesis that the final layers are linked to output language selection, whereas initial and middle layers have less effect on the output language. Our results are aligned with concurrent work suggesting that LLMs process input in three stages: understanding the input, reasoning and knowledge retrieval in a shared space among languages, and generating output

	French		German		Hindi		Italian		Portuguese		Spanish		Average	
Model	Ratio	Acc	Ratio	Acc	Ratio	Acc	Ratio	Acc	Ratio	Acc	Ratio	Acc	Ratio	Acc
Llama-1B	82.15	47.53	64.47	39.02	90.75	28.00	83.58	50.39	73.89	41.51	72.72	41.65	77.93	41.35
Llama-3B	78.37	42.19	79.55	36.14	85.34	33.54	80.11	41.61	74.27	49.04	78.92	44.22	79.43	41.12
Llama-8B	75.95	67.62	85.29	49.75	96.89	32.00	87.83	59.77	88.63	64.55	86.94	38.01	86.92	51.95
Gemma-4B	88.35	57.38	88.04	55.88	83.71	37.32	90.22	64.21	87.00	60.69	87.34	61.81	87.45	56.22

Table 3: CoCo-CoLa ratio (Ratio) and cumulative accuracy (Acc) of models partially trained on balanced multilingual data, with averages across all languages.

(Wendler et al., 2024; Dumas et al., 2025; Schut et al., 2025). Although it remains debated whether this shared knowledge space is language agnostic (Dumas et al., 2025) or whether the model simply thinks in English (Wendler et al., 2024; Schut et al., 2025), these works, alongside ours, all suggest that the process happening in middle layers is not dependent on the input language. However, what previous works overlooks is that the final stage is defective and cannot generate the response in the correct language. We believe this phenomenon has led to misleading evaluations and the belief that multilingual LLMs think better in English (Etxaniz et al., 2024). Our work emphasizes the importance of considering both correctness and language adherence, as relying on output accuracy against the ground truth does not provide a complete picture of a model's ability to reason and operate in non-dominant languages.

4.3 Improving Language Adherence in Multilingual LLMs

As demonstrated in Section 4.1, multilingual LLMs exhibit a strong linguistic bias toward English, the most prevalent language in their training data. In Section 4.2, we further established that this bias is closely linked to the model's final layers. To investigate whether this bias can be mitigated and to enable the model to better adhere to the input language, we take the *en-tuned* model and, rather than adapting it to a single target language, we partially fine-tune the language-related layers using a balanced multilingual dataset, where all languages appear with equal frequency in the training data.

As shown in Table 3, the average CoCo-CoLa ratio for multilingually fine-tuned Gemma-4B and Llama-8B reaches 87.45% and 86.92%, respectively, while Llama-1B and Llama-3B achieve slightly lower ratios of 77.93% and 79.43%. These results are similar to the monolingual models partially trained for each language (Appendix A.4). These findings indicate that, even when starting from a model pretrained on biased data, fine-tuning only the final layers on a balanced multilingual dataset substantially improves language adherence across all languages. Notably, for Llama-8B and Gemma-4B, the accuracy of the resulting multilingual model is competitive with models fully fine-tuned for each individual language, despite using only 200 datapoints per language during training.

5 Conclusion

In this work, we first analyzed shared knowledge across seven languages and identified key differences in the parameters most affected when training models for each language. Building on these insights, we proposed the CoCo-CoLa ratio, a metric for evaluating language adherence in multilingual LLMs, and used it to evaluate both pre-trained and fine-tuned LLMs. Our findings show that pre-trained models tend to generate English outputs regardless of the input language and that fine-tuning on English further amplifies this bias.

To address this problem, we leveraged insights from parameter updates and CoCo-CoLa results to develop a partial training method that improves language adherence in English-trained models. Our analysis demonstrated a more efficient alternative to full fine-tuning, achieving comparable or even superior performance while significantly reducing the number of updated parameters. Additionally, we showed that partial training on balanced multilingual data achieves similar language adherence to monolingual training. Given the widespread availability of instruction-tuned and task-specific English models, partial training of final layers presents a fast and efficient approach for improving language adherence and adapting LLMs to new languages.

Limitations

We acknowledge that training hyperparameters can influence the linguistic bias of fine-tuned models, as highlighted by Chirkova and Nikoulina (2024). For instance, while smaller learning rates may reduce bias, they can also lead to degraded task performance. Due to resource constraints, we used a single set of hyperparameters optimized for task performance. Additionally, we applied the same hyperparameter settings across all languages and model sizes, though fine-tuning them individually for each model-language pair could potentially yield better results.

Moreover, linguistic bias in pre-trained models and the observed trends in parameter updates across languages are influenced by factors such as model architecture, training procedures, data proportions, and even the order in which the model encounters training data. As a result, the specific layers we identified for each model size may differ when tested on other LLMs. Additionally, our observations suggest that certain languages are undertrained in Llama models. However, due to the lack of publicly available information on training data and procedures, we cannot make definitive claims regarding language-specific training discrepancies.

Another limitation is that our study focuses on languages that mostly come from the same language family, and are relatively close to each other. As a result these languages exhibit significant token overlap, facilitating cross-lingual transfer. The models we evaluated were also trained on a limited set of languages with similar characteristics. The studied languages mainly fall into the mid- or high-resource category, meaning our findings may not generalize to massively multilingual models trained on a more diverse set of languages.

Ethical Statement

This research investigates language adherence in multilingual large language models and proposes partial training methods for efficient adaptation. Our work aims to enhance linguistic fairness and accessibility by mitigating biases that favor high-resource languages. We acknowledge that training data composition and fine-tuning decisions can introduce unintended biases, which may disproportionately affect underrepresented languages. While our findings contribute to more equitable multilingual model adaptation, they are limited to languages present in the model's pretraining data and may not generalize to unseen languages. We encourage further work to assess our method's applicability to a broader set of languages, particularly low-resource and non-Indo-European languages.

This study does not involve human subjects, personal data, or user interactions, and we adhere to ethical guidelines for computational research. Our experiments were conducted using publicly available models and datasets, ensuring transparency and reproducibility.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2024. Layer swapping for zero-shot cross-lingual transfer in large language models. *arXiv preprint arXiv:2410.01335*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 695–708, Tokyo, Japan. Association for Computational Linguistics.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, pages 613–641.

Tanya Chowdhury and James Allan. 2024. Probing ranking llms: Mechanistic interpretability in information retrieval. *arXiv preprint arXiv:2410.18527*.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *Preprint*, arXiv:2304.09151.

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

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. *Preprint*, arXiv:2411.08745.

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.

Siqi Fan, Xin Jiang, Xiang Li, Xuying Meng, Peng Han, Shuo Shang, Aixin Sun, Yequan Wang, and Zhongyuan Wang. 2024. Not all layers of llms are necessary during inference. *arXiv preprint arXiv:2403.02181*.

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.

Jorge García-Carrasco, Alejandro Maté, and Juan Trujillo. 2024. Extracting interpretable task-specific circuits from large language models for faster inference. *arXiv preprint arXiv:2412.15750*.

Felix Gaschi, Patricio Cerda, Parisa Rastin, and Yannick Toussaint. 2023. Exploring the relationship between alignment and cross-lingual transfer in multilingual transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3020–3042, Toronto, Canada. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscope: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*.

Ashkan Golgoon, Khashayar Filom, and Arjun Ravi Kannan. 2024. Mechanistic interpretability of large language models with applications to the financial services industry. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 660–668.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024. On large language models' hallucination with regard to known facts. 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 1041–1053, Mexico City, Mexico. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Daoyang Li, Mingyu Jin, Qingcheng Zeng, Haiyan Zhao, and Mengnan Du. 2024. Exploring multilingual probing in large language models: A cross-language analysis. *arXiv preprint arXiv:2409.14459*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2024. Arithmetic without algorithms: Language models solve math with a bag of heuristics. *arXiv preprint arXiv:2410.21272*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Marinela Parovic, Alan Ansell, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual transfer with target language-ready task adapters. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 176–193, Toronto, Canada. Association for Computational Linguistics.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 7654–7673, Online. Association for Computational Linguistics.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Elizabeth Salesky, Neha Verma, Philipp Koehn, and Matt Post. 2023. Multilingual pixel representations for translation and effective cross-lingual transfer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13845–13861, Singapore. Association for Computational Linguistics.

Naomi Saphra and Adam Lopez. 2019. Understanding learning dynamics of language models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *Preprint*, arXiv:2502.15603.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Gürkan Soykan and Gözde Gül Şahin. 2024. Linguistically-informed multilingual instruction tuning: Is there an optimal set of languages to tune? *arXiv preprint arXiv:2410.07809*.

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

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Cunxiang Wang, Pai Liu, and Yue Zhang. 2021. Can generative pre-trained language models serve as knowledge bases for closed-book QA? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3241–3251, Online. Association for Computational Linguistics.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024a. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Hetong Wang, Pasquale Minervini, and Edoardo M Ponti. 2024b. Probing the emergence of cross-lingual alignment during llm training. *arXiv preprint arXiv:2406.13229*.

Andrea W Wen-Yi and David Mimno. 2023. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1131, Singapore. 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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024. Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8037–8051, Miami, Florida, USA. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024a. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2024b. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. *arXiv preprint arXiv:2402.18913*.

Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. 2024. Pre-trained large language models use fourier features to compute addition. *arXiv preprint arXiv:2406.03445*.

A Appendix

A.1 Implementation details

We experimented with dropout rates of 0.1 and 0.05, and learning rates of 5e-5, 1e-5, 5e-6, 1e-6, 5e-7, and 1e-7 for training on the English CBQA task. The best setting (dropout = 0.1, learning rate = 5e-6) was selected based on the minimum validation loss. These hyperparameters were used consistently across all languages and models throughout the paper.

For all training runs in our experiments, we used the hyperparameters listed in Table 4. All experiments were conducted with a fixed random seed of 42. We implemented our models using Transformers 4.46.3 and Torch 2.5.1, with Accelerate 1.1.0 and DeepSpeed 0.16.1 for multi-GPU training. All experiments were run on NVIDIA RTX A6000 GPUs, with all experiments taking approximately 48 hours on eight GPUs.

Parameter	**value**
num_epochs	3
save_steps	100
eval_steps	100
logging_steps	100
batch_size	64
gradient_accumulation	1
weight_decay	0.01
bf16	True

Table 4: Training hyperparameters

A.2 Language specific knowledge

Beyond measuring similarities between languages using the Jaccard Index, we also analyze differences by identifying answers that are known in language A but unknown in language B. This allows us to examine the distribution of languages within the 40% of answers that are not correctly predicted by both languages. The results, presented in Figure 5, reveal an almost symmetrical distribution of known and unknown answers across most language pairs. However, notable deviations occur for languages with significantly lower overall accuracy. Specifically, Hindi shows a greater disparity in the Llama-3.2 models, while both Hindi and Portuguese exhibit this trend in the Llama-8B model.

A.3 Parameter update

Due to space constraints, the main text presents results for only four languages. However, the analysis of model updates for Italian, Spanish, and Portuguese follows similar trends and can be found in Figure 6. These additional results confirm the patterns observed in other languages, reinforcing our findings on language-specific parameter updates.

A.4 Partial Training

Due to space limitations, the results of partial training on Italian, Portuguese, and Spanish are provided in Figure 7. Additionally, the CoCo-CoLa ratios for both partially trained and fully trained models are shown in Table 5 for Llama-1B, Table 6 for Llama-3B, and Table 7 for Llama-8B. These comparisons highlight the consistently superior CoCo-CoLa ratio in the partial training of final layers.

Language	SFT_{en}	1-5	1-10	11-16	SFT_{L_i}
French	13.47	44.63	50.22	78.72	88.58
German	10.50	25.12	31.77	76.66	91.02
Hindi	13.28	56.82	58.49	92.73	90.79
Italian	10.00	32.12	65.17	86.18	93.60
Portuguese	12.73	45.18	56.33	75.43	91.07
Spanish	12.01	34.61	34.41	81.66	91.24

Table 5: CoCo-CoLa Ratios (%) for different languages across finetuned Llama-3.2-1B models.

Language	SFT_{en}	1-14	14-27	SFT_{L_i}
French	14.73	54.64	81.18	89.45
German	19.64	71.40	86.04	89.26
Hindi	10.04	26.40	88.41	77.47
Italian	16.29	65.45	86.91	87.91
Portuguese	15.99	61.76	84.45	85.10
Spanish	15.84	72.38	85.50	89.18

Table 6: CoCo-CoLa Ratios (%) for different languages across finetuned Llama-3.2-3B models.

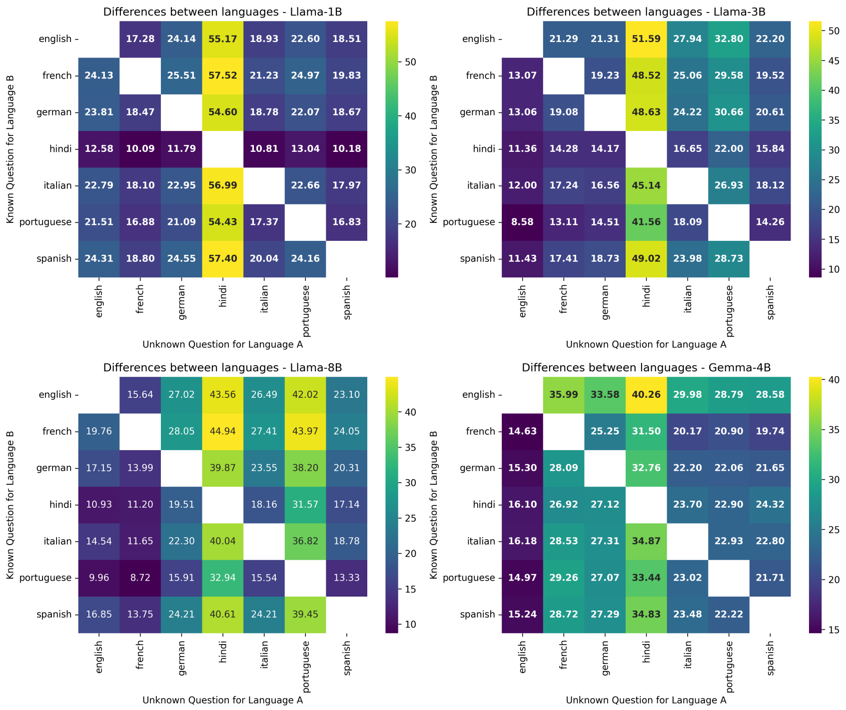

Figure 5: Difference in known knowledge between each pair of languages across different model sizes.

Language	SFT_{en}	1-15	16-31	SFT_{L_i}
French	12.32	78.93	87.77	87.54
German	11.03	81.91	88.69	87.21
Hindi	10.74	67.08	96.06	90.68
Italian	10.90	78.92	90.28	91.35
Portuguese	11.49	74.68	90.11	90.73
Spanish	9.40	75.82	93.55	91.35

Table 7: CoCo-CoLa Ratios (%) for different languages across finetuned Llama-3.1-8B models.

Language	SFT_{en}	1-20	21-34	SFT_{L_i}
French	19.22	64.26	89.99	90.14
German	15.23	86.70	93.03	92.64
Hindi	9.39	92.74	96.30	97.19
Italian	14.84	85.03	91.20	91.08
Portuguese	14.98	70.93	88.40	90.69
Spanish	14.70	68.14	90.19	91.36

Table 8: CoCo-CoLa Ratios (%) for different languages across finetuned Gemma-3-4B models.

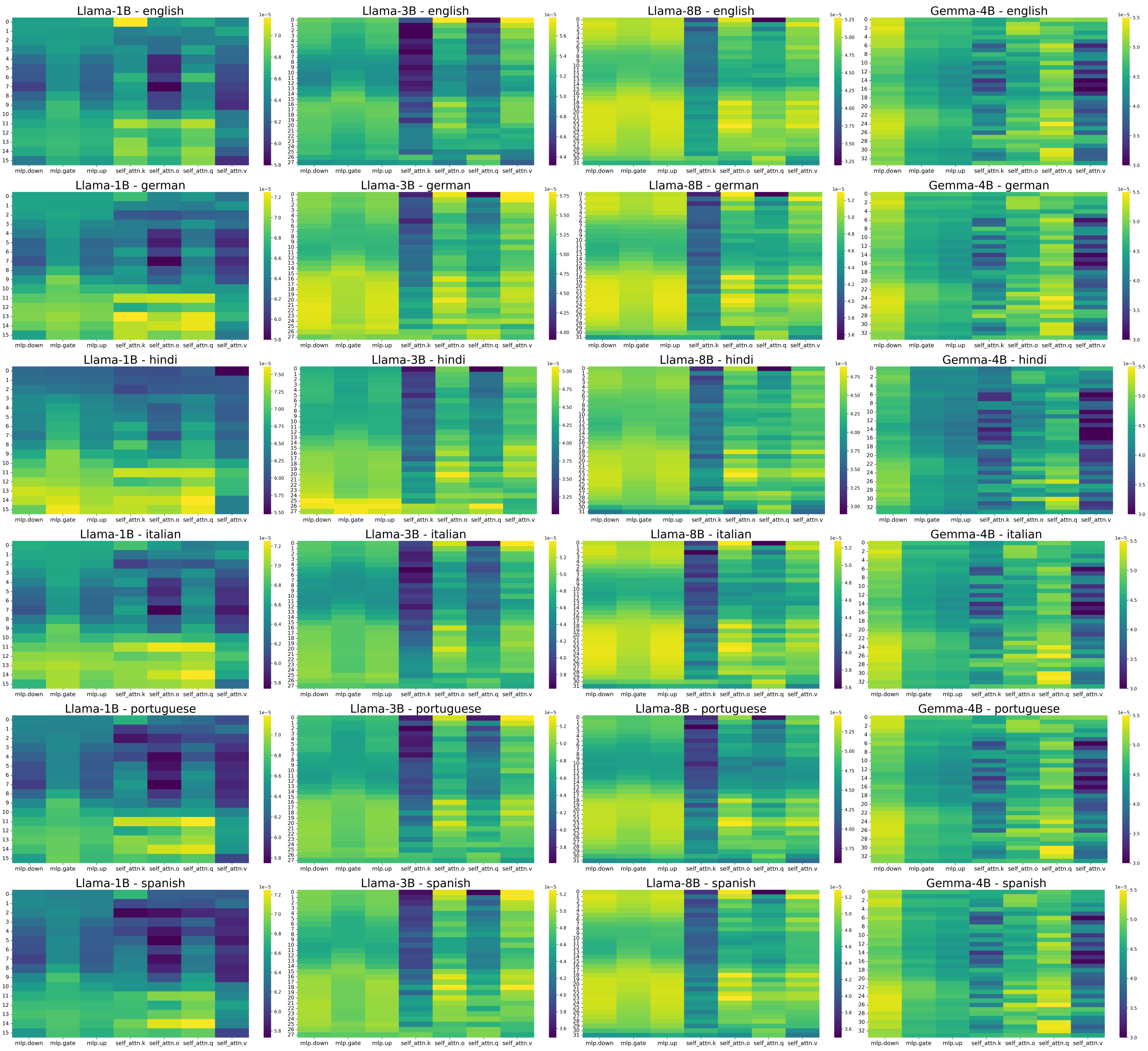

Figure 6: Average magnitude of difference between pretrained and monolingually fine-tuned models for Llama-1B, Llama-3B, and Llama-8B.

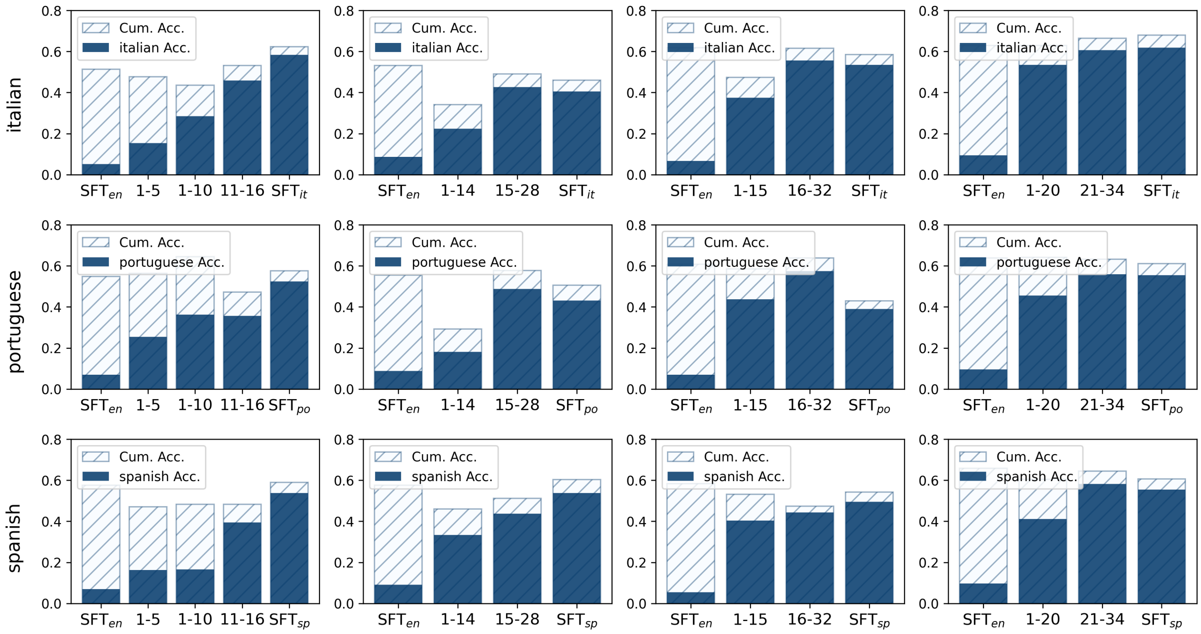

Figure 7: Cumulative accuracy and L_i accuracy on *en-tuned* (SFT_{en}) and L_i-tuned models (SFT_{L_i}), along with partially trained models, across all Llama model sizes.