

How does Multilingual Finetuning of Instruction-Tuned Large Language Models Affect Languages in Layer-Wise Hidden Representations?

Hyemi Jo

Seoul National University
huimei6361@snu.ac.kr

Hyopil Shin

Seoul National University
hpshin@snu.ac.kr

Abstract

Contemporary large language models (LLMs) exhibit multilingual capabilities but remain predominantly English-centric, resulting in a persistent performance gap across languages. While some approaches to multilingual training are grounded in the implicit pivot language hypothesis to improve interpretability, it remains underexplored how the languages represented in each layer are affected during training, particularly in practical, cost-efficient instruction-tuning scenarios. In this study, we investigate how instruction tuning with Low-Rank Adaptation (LoRA) in a target language affects the layer-wise language representation of instruction-tuned LLMs, with the aim of assessing the relevance of the implicit pivot language hypothesis as a theoretical grounding for multilingual finetuning. We find that target-language instruction tuning shifts models' internal language representation in a non-uniform manner across layers and sublayers: middle layers shift further toward English, while early and final layers primarily serve the mapping between the target language and the dominant intermediate language. Our findings empirically support the implicit pivot language hypothesis in the context of instruction tuning, demonstrating that layer-wise parameter variation broadly corresponds to the hypothesized functional roles of layers in multilingual processing. Our code is available at: https://github.com/hyemijo/layer_interpretability

1 Introduction

Contemporary large language models (LLMs) have developed a wide range of capabilities, including emergent abilities such as instruction following (Wei et al., 2021) and in-context learning (Brown et al., 2020), not only in English but also across many other languages. Even English-centric models (Touvron et al., 2023; Jiang et al., 2023), which are predominantly trained on English data alongside a small proportion of high-resource language

data, exhibit multilingual capability (Aycock and Bawden, 2024; Yuan et al., 2024; Liu and Niehues, 2025; Zhang et al., 2024a), encompassing both understanding and generation (Bandarkar et al., 2025; Niklaus et al., 2023).

Despite this progress, the performance of LLMs in multilingual settings remains considerably more limited for low-resource languages than for high-resource languages (Lai et al., 2023; Zhang et al., 2024b; Huang et al., 2024; Zhang et al., 2025). This substantial performance disparity across languages, commonly referred to as the *language gap*, has been identified as one of the primary challenges facing multilingual LLMs (Wang et al., 2025b), which struggle to achieve performance in low-resource languages that is comparable or superior to that in high-resource languages (Adelani et al., 2024; Iyer et al., 2024; Liu and Niehues, 2025). Such performance degradation persists even in proprietary models such as GPT-4, including in text generation quality (Qin et al., 2025). These performance disparities risk creating disproportionate access to advanced AI services that disproportionately favor speakers of high-resource languages (Lai et al., 2023; Wang et al., 2024, 2025b), ultimately undermining the robustness and cross-lingual generalizability of model behavior.

Many approaches have been proposed to enhance the multilingual capabilities of LLMs by adapting models to a target language. However, owing to the data-driven and resource-intensive nature of such approaches (Wang et al., 2025b), more sophisticated training strategies are required to feasibly generalize across diverse languages on more interpretable grounds. To improve target-language proficiency while preserving the original capabilities of models, it is necessary to examine how finetuning affects models' internal multilingual processing mechanisms, which would serve as a foundational basis for devising principled and interpretable strategies that operate on models' lay-

ered processing of multilingual input.

Models have been found to form representations in intermediate layers that are oriented toward dominant languages, a phenomenon termed *latent language* (Zhong et al., 2024). Latent language has been observed in few-shot translation tasks with careful token-level control, covering cases where the output consists of a single token (Wendler et al., 2024) or multiple tokens (Zhong et al., 2024), as well as in benchmark tasks (Tamo et al., 2026) comprising MMLU (Hendrycks et al., 2021), MGSM (Shi et al., 2022), and XQuAD (Artetxe et al., 2020). This raises an intuitive question about model behavior in multilingual settings: *what language do models think in when given a general instruction-following query in a target language, and does that language shift after finetuning the model on target-language data?* When a model is presented with a question in a target language in a general question-answering setting after finetuning, does it process the target-language input via the latent language dominant in the model?

We analyze how parameter updates through general instruction tuning affect the model-internal mechanisms for multilingual processing, drawing on the theoretical support of the implicit pivot language hypothesis. This analysis would elucidate whether finetuning a model on target-language data suitably targets parameters that are highly relevant to multilingual processing, or whether it instead disrupts pretrained knowledge that is relatively peripheral to the target language. This study investigates whether and how instruction tuning in a target language affects models’ hidden states in terms of language representation, guided by the following research questions:

- RQ1: Does an instruction-tuned model leverage latent language to process target-language input in a general instruction-following scenario?
- RQ2: How does instruction tuning in a target language shift the layer-wise representations of an instruction-tuned model with respect to latent language?

The key contributions of this study are as follows. (1) We extend prior findings on the latent language of LLMs to the instruction-following scenario of general-domain question answering, demonstrating that target languages with non-Latin scripts are processed in unspecified languages in the early layers, oriented toward English in the middle lay-

ers, and restored to the target language in the final layers. (2) We find that the parameter shifts induced by finetuning expand the range of middle layers in which the model processes input in the latent language, as evidenced by comparing language proportions and cross-lingual similarity of hidden states in the target language and English before and after finetuning. (3) We empirically validate the implicit pivot language hypothesis by demonstrating a correspondence between the finetuned layers and their presumed functional roles in multilingual processing, which contributes to a more effective assessment of the current state and limitations of multilingual capabilities in LLMs.

2 Related Work

2.1 Language gap in large language models

While LLMs achieve outstanding performance in English, this capability does not generalize robustly to other languages (Qin et al., 2025). This limitation is primarily attributed to the imbalanced distribution of languages in training data, where English is disproportionately represented (Wang et al., 2025b; Ye et al., 2025; Huang et al., 2024; Grattafiori et al., 2024; Jiang et al., 2025). This imbalance introduces language bias, skewing models toward English (Zhu et al., 2024) and reducing their effectiveness on low-resource languages (Navigli et al., 2023).

To mitigate performance discrepancies across languages, parameter-efficient fine-tuning (PEFT) has been widely adopted for multilingual training, owing to its cost-effectiveness in enhancing cross-lingual transfer (Borchert et al., 2025; Dong et al., 2025; Zhao et al., 2025). While these strategies have shown promise, their data-specific nature makes it difficult to generalize findings across diverse languages and tasks to achieve comparable efficacy, a challenge further compounded by dependence on translation modules (Borchert et al., 2025). Realigning representations has been proposed as an alternative (Nguyen et al., 2023; Salesky et al., 2023; Gao et al., 2023; Zhao et al., 2025); however, the benefits of aligned representations for cross-lingual transfer have been reported to arise only under limited conditions (Gaschi et al., 2023; Zhao et al., 2025). Directly tuning models in target languages to reduce the language gap demands extensive linguistic resources for training (Ma et al., 2023; Zhao et al., 2025), which may be impractical in low-resource settings. These limi-

tations highlight the need for more principled and systematic investigation of the multilingual capabilities of LLMs, in order to interpret and leverage model-internal mechanisms for multilingual processing.

2.2 Implicit pivot language hypothesis

Recent studies address the challenges of multilingual training by grounding their strategies in the implicit pivot language hypothesis. Under this hypothesis, multilingual sentences are posited to be processed through a three-stage pipeline in which English serves as a pivot: early layers encode multilingual input by mapping target-language representations into an English-oriented representational space; middle layers perform task-solving and reasoning predominantly in English; and final layers transform representations back into the target language to generate output (Zhao et al., 2024; Wendler et al., 2024).

This hypothesis has motivated approaches that target specific subsets of layers presumed to be responsible for multilingual processing. Middle layers are often kept frozen due to their relevance to English-language reasoning (Bandarkar et al., 2025; Owodunni and Kumar, 2025), while other approaches aim to facilitate the hypothesized transitions in models’ latent language — namely, the conversion from a target language into English and its subsequent reversion (Huo et al., 2025). Various criteria for selecting which layers to train have been proposed, including entropy (Huo et al., 2025), neuron activation across languages to identify language-specific neurons distributed non-uniformly across model layers (Fan et al., 2025; Tang et al., 2024), the magnitude of parameter updates after fine-tuning across languages (Bandarkar et al., 2025; Rahmati et al., 2025), and methods involving the visualization of semantic similarity (Liu and Niehues, 2025) alongside language token probabilities (Tamo et al., 2026). However, existing approaches generally presuppose that the layered mechanism of multilingual processing remains intact after updating parameters sensitive to a target language, and that such updates improve target-language understanding and generation by strengthening the mapping to the model’s latent language. Whether this presupposition holds after fine-tuning, and how targeted parameter updates relate to resulting changes in multilingual capabilities, remain underexplored—particularly in instruction-tuning settings.

Beyond targeted parameter updates, layer-wise hidden states have also been leveraged to enhance multilingual capabilities. Kirtane and Huang (2026) extract layer-wise hidden states from source and target languages to compute language vectors for steering multilingual in-context learning. Sterz et al. (2025) construct language-specific steering vectors from hidden states averaged over a multilingual parallel dataset to mitigate language confusion, i.e., generation in an unintended language. Ruan et al. (2025) update an adapter and a layer-wise aligner to incorporate encoder-model representations for cross-lingual alignment. Yet these approaches focus on steering hidden states toward language-specific or cross-lingually shared representations, rather than examining whether such interventions preserve the underlying layered mechanism of multilingual processing. Taken together, how fine-tuning on target-language data reshapes layer-wise hidden states with respect to the latent language remains underexplored.

3 Methodology

We scope our study to the instruction-tuning setting of instruction-tuned models in order to extend prior findings beyond base models and address practical use cases. We first finetune instruction-tuned models on a target language and analyze the resulting parameter shifts, focusing on hidden state variation. Each model is trained on a single target language using a question-answering task, with a target-language split of a multilingual parallel general-domain question-answering dataset that conforms to an instruction-following format.

After training, two experiments are conducted to investigate the effect of instruction tuning in a target language on the models’ layer-wise mechanisms for multilingual processing. First, we observe the variation in layer-wise language proportions via logit lens (nostalgebraist, 2020) before and after finetuning, with the aim of comparing the language proportions of unembedded hidden states in each layer, subdivided into self-attention, MLP, and full block. Second, we examine the parameter changes per layer by conducting few-shot inference on both the baseline and finetuned models and comparing their hidden states layer-wise. We adopt a few-shot inference setup to capture the effect of finetuning on language representation in an environment consistent with real-world model usage and to effectively elicit the model’s competence.

3.1 Dataset

Bactrian-X (Li et al., 2023) serves as the multilingual parallel dataset for instruction tuning, chosen for its generalizability and semantic correspondence across languages. It comprises 67k machine-translated instruction–response pairs per language across 52 languages, sourced from Alpaca (Taori et al., 2023) and Dolly (Conover et al., 2023). Instructions are translated from English via the Google Translate API to preserve semantic equivalence with the source, with translation quality verified through back-translation on sampled instances per language (Li et al., 2023). Outputs are generated by feeding the translated instructions to ChatGPT (gpt-3.5-turbo)¹, and human evaluation of informativeness and fluency is conducted for a subset of languages.

Three non-Latin-script languages are selected as targets: Chinese, Korean, and Telugu, representing high-, mid-, and low-resource languages, respectively. Following Bandarkar et al. (2025), languages are selected based on benchmark presence, resource level, and script (details in Appendix B). We focus on non-Latin-script languages because they remain underrepresented in multilingual processing compared to Latin-script languages, facing greater difficulty in cross-lingual transfer due to limited shared vocabulary with English (Choenni et al., 2023; Wang et al., 2025a). This setting also allows us to examine whether models successfully adapt to a target language while mitigating interference with, or severe deviation from, their existing pretrained knowledge.

Single-turn user–assistant conversation pairs are used for training to reflect interactions typical of practical use cases. Length-based filtering constrains instruction and input lengths to under 1k characters and output lengths to under 2k characters, reflecting Gemma 2’s 8,192-token context limit and realistic resource-constrained settings. The maximum length is set to 4,096 tokens² to match more limited context windows (e.g., Llama 2). The filtered data are then split into three subsets: 51,065 training, 6,384 validation, and 6,409 test instances per language.

¹Appendix A presents a quantitative assessment of cross-lingual semantic similarity in the LLM-generated output field; we observe considerable semantic equivalence between the target languages and English.

²Due to resource constraints, a training-time cap of 1,024 tokens is applied instead, affecting fewer than 2% of examples.

3.2 Model

To align our research with practical usage scenarios, we employ instruction-tuned models, specifically Llama 3.2 3B Instruct (Grattafiori et al., 2024) and Gemma 2 2B Instruct (Team et al., 2024). We choose Llama and Gemma because they are widely used in implicit pivot-language research (Wendler et al., 2024) and in LAPE-/SAE-based analyses (Tang et al., 2024; Verma et al., 2026). Both models adopt a decoder-only Transformer architecture (Vaswani et al., 2017), comprising 28 and 26 layers, respectively, each employing grouped query attention (GQA; Ainslie et al. (2023)).

Llama is inherently multilingual, officially supporting eight languages through multilingual pre-training and post-training exposure (Grattafiori et al., 2024). Despite this native multilingual support, Llama’s pretraining corpus is predominantly English, with multilingual content accounting for only 8% of the data (Grattafiori et al., 2024; Rahmati et al., 2025). This composition makes the model well suited for studying how fine-tuning affects performance in less-dominant languages. By contrast, we include Gemma as a comparison model, whose pretraining and post-training data are predominantly and exclusively English, respectively, though the exact composition remains undisclosed (Team et al., 2024). Although its vocabulary is designed for broader multilingual support, the model is not optimized for state-of-the-art multilingual capabilities (Team et al., 2024), leaving room for multilingual fine-tuning gains.

3.3 Training

To validate the effect of instruction tuning on model layers under a less resource-intensive setting, we apply parameter-efficient fine-tuning (PEFT) with LoRA adapters applied to all layers. Low-Rank Adaptation (LoRA) (Hu et al., 2022) is employed for parameter-efficient instruction tuning, given its widespread adoption for fine-tuning open-source LLMs (Hu et al., 2022; Zhao et al., 2023). Rather than fully fine-tuning the entire weight matrix $W \in \mathbb{R}^{m \times n}$ when updating W denoted as $W \leftarrow W + \Delta W$, we freeze the pretrained weight matrix W and decompose ΔW into two trainable low-rank matrices $A \in \mathbb{R}^{m \times k}$ and $B \in \mathbb{R}^{n \times k}$, i.e., $\Delta W = AB^\top$ for the selected layers only, where $k \ll \min(m, n)$ is the reduced rank to approximate ΔW by the inner product of A and B (Hu et al., 2022; Zhao et al., 2023).

3.4 Evaluation and analysis

3.4.1 Quantitative evaluation

Before analyzing the finetuned models, we first evaluate them to verify whether finetuning properly adapts the models to a target language. Two cross-lingual benchmarks are utilized under a single-run, zero-shot setting using the EleutherAI LM Evaluation Harness framework (Gao et al., 2024). Two cross-lingual benchmarks, Belebele (Bandarkar et al., 2024) and Global MMLU (Singh et al., 2025), are introduced for evaluation in the target language and for cross-lingual comparison, in order to examine whether models’ target-language performance improves without severe degradation in English. Belebele evaluates language understanding via a multiple-choice machine reading comprehension (MRC) task, while Global MMLU assesses multi-task language understanding through multiple-choice questions with linguistic and cultural considerations.

3.4.2 Logit lens analysis

Logit lens (nostalgebraist, 2020) decodes models’ hidden states via unembedding — i.e., passing each layer’s hidden state through the language modeling head (LM head) to project it onto the vocabulary dimension. We apply logit lens to queries in non-Latin-script languages and observe how the resulting language distribution shifts across layers.

To address ambiguity from multilingual subword overlap (Tamo et al., 2026), we apply a two-stage language identification scheme to categorize the decoded tokens. In the first stage, we categorize each token by script, using its Unicode script name (e.g., Hangul, Latin) and category as identified via uncode data; some scripts map largely unambiguously to a single language (e.g., Hangul to Korean). We additionally define a `Partial` category for byte tokens that fail to decode into valid UTF-8, and a `Special` category for the model’s special tokens, whitespace, symbols, numbers, and raw bytes. In the second stage, we perform fine-grained language identification on Latin-script tokens, following Lee (2025), who use the GPT API to classify each token’s ISO 639-1 language from the decoded layer output. Appendix C provides further details.

To inspect the languages the model relies on internally before and after fine-tuning on a target language, we analyze which language the hidden states map onto for target-language queries in a single-turn dialogue setting, based on the imple-

mentations of nostalgebraist (2020) and Wendler et al. (2024). We use 5k semantically parallel instances from the training dataset in Chinese, Korean, and Telugu. For each input question, we extract the output hidden states from the self-attention sublayer, the MLP sublayer, and the full decoder block at every layer, and project each onto the vocabulary dimension (unembedding) to obtain a decoded token. We then compute the proportion of languages at each layer across the 5k samples.

3.4.3 Hidden state analysis

To test whether the logit lens pattern also holds at the representational level, rather than being an artifact of the output projection alone, we examine whether fine-tuning shifts hidden states themselves toward the latent language. Following Tamo et al. (2026), we adopt cosine similarity but apply in two complementary forms: a *distance* (one minus similarity) tracking how much a language’s representation shifts between the pretrained and fine-tuned model, and a *similarity* tracking cross-lingual alignment within a model. We use few-shot rather than zero-shot inference throughout to better elicit the model’s underlying competence.

First, we compute the layer-wise cosine distance between pre- and post-fine-tuning hidden states, separately for target-language and English inputs, to test whether fine-tuning shifts representations of the target language more than English. Second, we compute the layer-wise cosine similarity between target-language and English hidden states for the pretrained and fine-tuned models, and take their difference to isolate the fine-tuning-induced shift in cross-lingual alignment; increasing similarity indicates convergence toward English, decreasing similarity indicates divergence.

For both analyses, we construct inputs from open-ended question-answering tasks in the instruction-tuning format, drawn from the training and test splits. For each test query, we deterministically sample 6-shot demonstrations from the training split per language; demonstrations and queries are indexed for semantic equivalence across languages. Since parallel sentences vary in length, we apply last-token pooling rather than mean pooling—the last token’s hidden state already accumulates context from preceding tokens, so pooling over all tokens would add redundancy. Each input is forwarded per language, and per-layer hidden states are compared via centered cosine similarity.

Model	Setting	Belebele					Global MMLU				
		Zh	Ko	Te	En	En_Avg	Zh	Ko	Te	En	En_Avg
LLaMA 3.2 3B Inst.	Baseline	71.78	59.56	38.00	82.67	—	43.36	39.09	33.55	53.14	—
	LoRA	72.78	60.44	35.44	83.89	83.70	46.65	41.16	31.11	55.34	57.54
Gemma 2 2B Inst.	Baseline	23.00	23.78	22.89	26.11	—	24.16	23.42	22.98	29.90	—
	LoRA	67.44	61.11	35.89	74.78	80.52	44.30	40.57	32.16	52.27	55.68

Table 1: Performance on Belebele and Global MMLU across languages, evaluated in each model’s target language and English. *En* is the score of the model trained and evaluated on English; *En_Avg* is the mean English-evaluation score across the models trained in Chinese, Korean, and Telugu, compared against *En* to assess English retention after target-language fine-tuning. Highlighted cells indicate improvement over baseline.

4 Results and Analysis

4.1 Quantitative evaluation

Table 1 shows that target-language fine-tuning generally improves multilingual performance while retaining English capability, indicating that vanilla LoRA fine-tuning is broadly beneficial for target-language adaptation. The improvement is larger for Gemma 2 than Llama 3.2, which we attribute to training-data composition: Gemma 2’s English-dominant pretraining data (Team et al., 2024) leaves more room for multilingual gains, whereas Llama 3.2’s greater multilingual exposure (Grattafiori et al., 2024) leaves less.

One exception is Telugu in Llama 3.2 3B Instruct, where performance drops. We attribute this to a fluency/correctness trade-off: fine-tuning shifts the model toward target-language generation but reduces answer correctness more than for the other two, higher-resource languages, as adapting to a lower-resource language requires more drastic changes that might disrupt task-relevant knowledge. This echoes Tamo et al. (2026)’s two failure modes in non-English LLM tasks—the multilingual transfer bottleneck (correct language, wrong task response) and the language consistency bottleneck (correct task response, wrong language)—both reflecting a tension between task completion and language control. Gemma 2 shows a similar pattern for Telugu with a relatively smaller performance gain, suggesting that adaptation to a low-resource language is comparatively difficult across models.

Models fine-tuned on the target language also retain their English capability: evaluated in English, their average performance (*En_Avg*) meets or exceeds that of the English-trained model (*En*). We therefore use these models for the subsequent analysis of how multilingual training affects internal language representations.

4.2 Logit lens analysis

Figure 1 visualizes the language proportions of hidden states for Llama and Gemma, before and after fine-tuning, across the three target languages. To address RQ1, we examine whether a single language predominates across layers and models. To address RQ2, we examine how language composition in early, middle, and final layers changes after fine-tuning, yielding our second and third observations.

Our first observation is that **both pretrained and fine-tuned models process target-language input via English as a dominant intermediate language**, affirmatively addressing RQ1 across all models and languages. English proportion peaks in the middle layers—40–90% relative depth for Llama 3.2, 60–90% for Gemma 2—indicating that English, not the target language, dominates the intermediate representation, consistent with prior work (Wendler et al., 2024; Zhong et al., 2024; Tamo et al., 2026). This proportion shows a rise-then-decline pattern across layers, which mirrors Wendler et al. (2024)’s findings for cloze, translation, and repetition tasks, suggesting that target-language training or inference operates within the layered mechanism of multilingual processing, with parameter updates shaped by the dominant intermediate language rather than replacing it.

Our second observation is that **the range of middle layers exhibiting English representations widens after fine-tuning**, indicating strengthened target-to-English mapping and more English-oriented intermediate representations. This widening is model-specific. In Llama, the increase occurs in early layers (10–30%) for Chinese and Korean, with late-middle-layer English proportion staying above 50%; for Telugu, it occurs in both early layers (10–20%, 30–40%) and late-middle layers (60–90%)—suggesting fine-tuning activates the target-to-English mapping earlier for higher-

Layerwise language composition: pretrained vs. finetuned

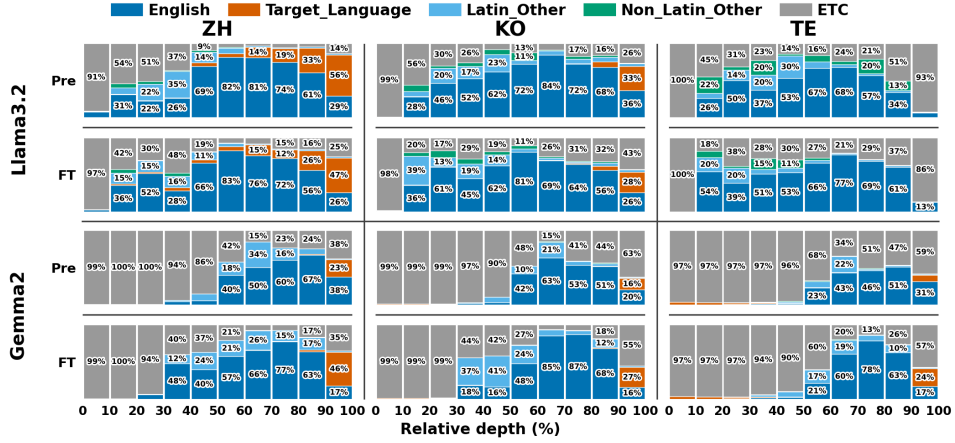


Figure 1: Layer-wise language composition of hidden states via logit lens. Each panel shows one model (rows: Llama 3.2, Gemma 2) fine-tuned and evaluated on one target language (columns: ZH, KO, TE), with stacked strips comparing the *pretrained* (Pre; top) and *fine-tuned* (FT; bottom) model. Layers are binned into ten relative-depth buckets (0–10%, ..., 90–100%). Colors denote five aggregated categories: *English*, *Target language*, *Latin-other* (Latin-script, non-English), *Non-Latin-other* (non-Latin-script), and *ETC* (special, partial, and other unidentifiable language tokens). See Appendix E for non-binned, fine-grained layer-wise proportions.

resource languages, while for Telugu it also reinforces middle layers, reducing non-Latin-script tokens and raising English proportion to a level comparable with the other languages. In Gemma 2, by contrast, the increase spans early-middle to late-middle layers (30–80/90%) for Chinese and Korean, and concentrates in late-middle layers (60–90%) for Telugu, occasionally with increased Latin-script non-English tokens. We attribute this to Gemma 2 being less optimized for multilingual use than Llama 3.2: fine-tuning thus enriches middle-layer representations that are less developed in Gemma 2 than the already well-developed ones in Llama 3.2, in order to activate the *English-thinking* stage. Comparing Telugu to the other languages within Gemma further suggests English emerges earlier and more strongly for higher-resource languages, while Telugu’s increase spans only a marginally wider range of layers—implying fine-tuning strengthens and speeds the target-to-English mapping to a degree that varies with resource level.

Our third observation is that **early and final layers map onto unspecified languages and the target language, respectively**, consistent with the implicit pivot-language hypothesis. Both models show ETC-category tokens (including unspecified-language, special, and partial tokens) dominating early layers, giving way to English in the middle layers, before reverting to the target-language

script in the final layers. One exception is Llama 3.2 trained on Telugu, where ETC—driven primarily by Partial—dominates the final layer, with a slight rise in Special (Appendix E). We hypothesize that ETC tokens represent intermediate products of a strengthening target-to-English mapping, whose proportion may increase after fine-tuning. In Gemma, Latin-script non-English tokens rise in early-middle layers (30–60%), fall, then briefly rise again at 80–90%—a pattern we hypothesize reflects an intermediate step between ETC and English.

Together, these observations support the implicit pivot-language hypothesis for both training and inference in the target language: indeterminate early-layer representations, English dominance in middle layers, and rising target-language proportion in final layers collectively indicate that models rely on English as an implicit pivot language for multilingual processing.

4.3 Hidden state analysis

4.3.1 Hidden-state variation across models and target languages

The layer-wise variation in hidden states induced by fine-tuning differs across models for target-language queries, reflecting how well each model’s intermediate dominant-language representation is already established prior to fine-tuning. For English queries, by contrast, this variation is consis-

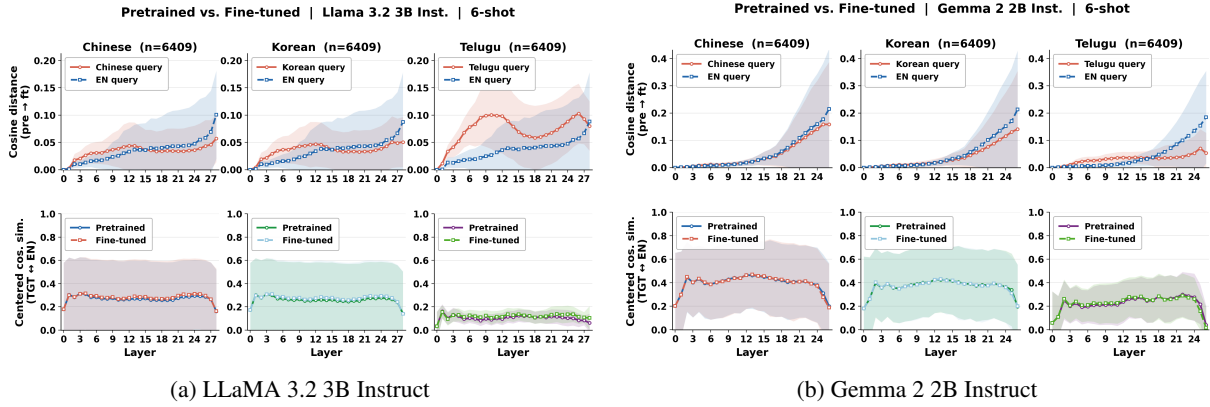


Figure 2: Comparison of hidden states per layer between pretrained and fine-tuned models. The x-axis denotes layer index, with 0 corresponding to the embedding layer. The upper panel shows the cosine distance between pretrained and fine-tuned models, averaged over samples; comparing target-language and English queries reveals how the parameter shift induced by fine-tuning affects few-shot inference across query languages. The lower panel shows the centered cosine similarity between hidden states for target-language versus English queries, per layer.

tent across both models, as discussed below.

For Llama (Figure 2a), the cosine distance between pretrained and fine-tuned hidden states follows a wave-like pattern for target-language queries—rising through the first half of the layers, falling through the late-middle layers, and rising slightly again toward the final layers. This pattern suggests that fine-tuning concentrates change in the early and final layers while leaving the middle layers relatively undisturbed, indicating that Llama does not rebuild its latent-language representation but instead adapts around the pre-existing one: adaptation is most active in the early layers, is restrained in the middle layers where the latent language is already established, and re-intensifies slightly near the output, consistent with the implicit pivot-language hypothesis.

Gemma shows a different pattern (Figure 2b): cosine distance rises monotonically with layer depth for target-language queries across all three languages except Telugu. We hypothesize this difference stems from how English-dominant each model’s pretraining is: Llama’s broader multilingual exposure appears to have already established a latent-language representation prior to fine-tuning, whereas Gemma’s more English-centric pretraining leaves its middle layers without a comparably enriched representation. Fine-tuning therefore appears to concentrate on building this representation in Gemma’s middle layers, rather than reinforcing the target-to-English mapping in the early layers as in Llama, with the resulting change propagating to the final layers, where it is used for target-language generation. This is consistent with

Figure 1, which shows Gemma’s English proportion increasing most in the middle layers after fine-tuning. Telugu shows a lower cosine distance than the other two languages, suggesting that adapting an English-dominant model to a low-resource language requires more extensive training to produce comparable hidden-state variation.

For English queries, by contrast, cosine distance increases monotonically with layer depth across both models and all target languages (Figure 2), indicating that fine-tuning affects English-query processing similarly regardless of the target language trained. We attribute this shared pattern to two factors. First, as discussed in Section 4.1, English capability remains largely intact regardless of the target language trained, which likely explains why the pattern is consistent across target languages. Second, the autoregressive architecture propagates variation forward: hidden-state updates in early layers accumulate through later layers via the mapping between layers, so that even without direct English fine-tuning, English representations still shift monotonically. Under this view, the final layers perform English-to-target-language mapping for token generation, building on hidden states carried forward from an intermediate “English-thinking” stage—itsself shaped by target-to-English mapping in the early layers.

4.3.2 Cross-lingual alignment

Lower panels in Figure 2 show that cross-lingual alignment with English after fine-tuning remains largely unchanged for Chinese and Korean, with only a marginal gain for Telugu, suggesting that

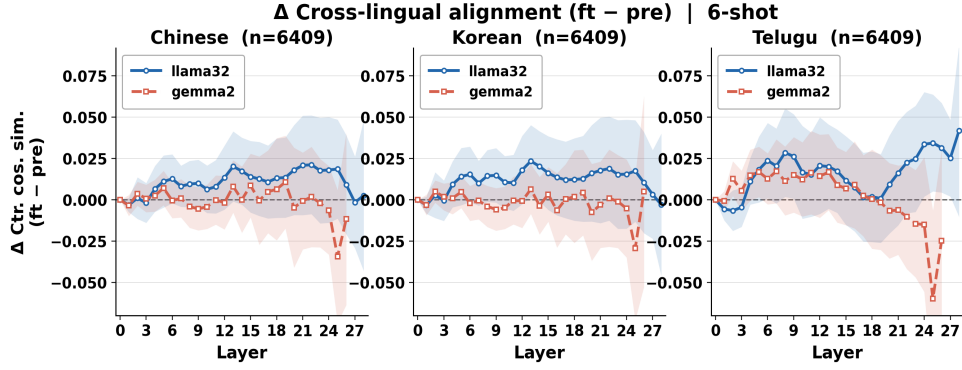


Figure 3: Change in cross-lingual hidden state similarity between pretrained and finetuned models. The x-axis denotes the layer index, where 0 corresponds to the embedding layer. The y-axis denotes the change in centered cosine similarity between the target language and English after finetuning.

fine-tuning updates parameters without disrupting the existing cross-lingual relationship between target languages and English.

Figure 3 shows the model-specific, layer-wise change in this alignment ($|\Delta| < 0.1$ throughout). For Llama, similarity rises slightly across layers, most in the middle layers, suggesting fine-tuning marginally strengthens cross-lingual alignment where it is already most prominent (Liu and Niehues, 2025). Telugu shows a further rise in the final layers, which we attribute to increased English-token proportion at this depth (Section 4.2, where fine-tuning raises Telugu’s final-layer English proportion above 10%). We interpret this as a byproduct of building the English-to-target mapping still underdeveloped for this low-resource language.

Gemma shows small fluctuations until the middle layers, followed by a sharp decrease and then a marginal increase in the final layers. We interpret the decrease as the model incorporating target-language-specific information as it diverges from the shared representation, and the final-layer increase as reflecting the same mechanism as Telugu’s: the strengthened cross-lingual mapping may draw target-language and English hidden states closer together. Overall, these results suggest that fine-tuning reshapes hidden states in a model-specific manner, depending on how established each model’s latent-language representation already is prior to adaptation.

5 Conclusion

This study investigates how fine-tuning instruction-tuned models on a target language affects their internal language representations. Our analysis

of LoRA-based fine-tuning shows that adaptation affects layers non-uniformly: the layer range in which English tokens dominate widens after fine-tuning, while early and final layers correspond respectively to unspecified- and target-language token proportions. This supports the functional distinction proposed by the implicit pivot-language hypothesis—early layers encode target-language input into a language-agnostic space, while final layers govern the reverse mapping for target-language generation.

Our hidden-state analysis shows that this adaptation is shaped by how established each model’s latent-language representation already is: Llama, whose extensive multilingual exposure has already built a language-agnostic pivot, updates hidden states mainly in the early and final layers while leaving this pivot undisturbed, whereas Gemma, lacking a comparably established pivot, concentrates updates in the middle layers to construct one. This model-dependent pattern suggests the pivot-language mechanism varies in maturity across models, shaping how multilingual input is processed through the layers.

Overall, this study bridges the language gap in multilingual LLMs by examining whether the effects of multilingual fine-tuning accord with the implicit pivot-language hypothesis. Our findings extend prior interpretability work to the instruction-tuning setting, providing support for the implicit pivot-language hypothesis beyond base models, revealing how its manifestation varies across models, and offering a foundation for more effective, pivot-aware approaches to adapting LLMs to target languages.

Limitations

Though useful for broad cross-lingual coverage with semantic equivalence across languages, machine-translated data may differ from human-translated data, potentially limiting real-world applicability: translation artifacts may deviate from natural language and affect the model’s parameter shift. Experiments on human-translated parallel datasets could help clarify this impact under a more natural language setting.

Furthermore, the impact of target-language exposure is difficult to isolate, both because the pre-trained model’s training-data composition is undisclosed and because instruction-tuning effects are hard to separate from those of target-language exposure. As our study fine-tunes already-instruction-tuned models, we do not isolate the effect of instruction tuning alone; this is a deliberate choice to avoid catastrophic forgetting and the need to reformat our QA-structured dataset for causal language modeling. Future work using a fully open model with known training-data composition, trained both with and without instruction tuning, would help fill this gap.

Beyond these design constraints, extending the analysis to more languages and architectures is necessary to generalize these findings, particularly for non-English-centric models, which may exhibit multiple latent languages (Zhong et al., 2024). Finally, this study addresses only parallel comparisons between the target language and English and does not extend to multilingual joint training, where multiple languages interact simultaneously.

Ethical considerations

AI tools were used for writing assistance, plot generation, and code revision.

References

- David Ifeoluwa Adelani, A. Seza Doğruöz, André Coneglian, and Atul Kr. Ojha. 2024. [Comparing LLM prompting with cross-lingual transfer performance on indigenous and low-resource Brazilian languages](#). In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 34–41, Mexico City, Mexico. Association for Computational Linguistics.
- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. [GQA: Training generalized multi-query transformer models from multi-head checkpoints](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Seth Aycock and Rachel Bawden. 2024. [Topic-guided example selection for domain adaptation in LLM-based machine translation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 175–195, St. Julian’s, Malta. Association for Computational Linguistics.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. [The belebele benchmark: a parallel reading comprehension dataset in 122 language variants](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.
- Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2025. [Layer swapping for zero-shot cross-lingual transfer in large language models](#). *Preprint*, arXiv:2410.01335.
- Philipp Borchert, Ivan Vulić, Marie-Francine Moens, and Jochen De Weerd. 2025. [Language fusion for parameter-efficient cross-lingual transfer](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25848–25868, Vienna, Austria. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. [How do languages influence each other? studying cross-lingual data sharing during LM fine-tuning](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13244–13257, Singapore. Association for Computational Linguistics.

- Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world’s first truly open instruction-tuned llm. <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
- Tianyu Dong, Bo Li, Jinsong Liu, Shaolin Zhu, and Deyi Xiong. 2025. [MLAS-LoRA: Language-aware parameters detection and LoRA-based knowledge transfer for multilingual machine translation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15645–15660, Vienna, Austria. Association for Computational Linguistics.
- Yuchun Fan, Yongyu Mu, YiLin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. 2025. [SLAM: Towards efficient multilingual reasoning via selective language alignment](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9499–9515, Abu Dhabi, UAE. Association for Computational Linguistics.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. [Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12103–12119, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. 2025. [jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval](#). In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550, Suzhuo, China. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). *Preprint*, arXiv:2009.03300.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchun Liu, Yuzhuang Xu, and 1 others. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.
- Wenshuai Huo, Xiaocheng Feng, Yichong Huang, Chengpeng Fu, Baohang Li, Yangfan Ye, Zhirui Zhang, Dandan Tu, Duyu Tang, Yunfei Lu, and 1 others. 2025. Enhancing non-english capabilities of english-centric large language models through deep supervision fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24185–24193.
- Vivek Iyer, Bhavitvya Malik, Wenhao Zhu, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. [Exploring very low-resource translation with LLMs: The University of Edinburgh’s submission to AmericasNLP 2024 translation task](#). In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 209–220, Mexico City, Mexico. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*. *Preprint*, arXiv:2310.06825.
- Fan Jiang, Honglin Yu, Grace Chung, and Trevor Cohn. 2025. [Franken-adapter: Cross-lingual adaptation of llms by embedding surgery](#). *Preprint*, arXiv:2502.08037.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Neeraja Kirtane and Kuan-Hao Huang. 2026. [Language steering for multilingual in-context learning](#). *Preprint*, arXiv:2602.02326.
- Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and

- Thien Huu Nguyen. 2023. [ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.
- Sangah Lee. 2025. [Do korean-adapted llms think in korean? analyzing latent language and the preservation of korean-specific knowledge](#). *LANGUAGE AND INFORMATION*, 29(3):229–256.
- Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. [Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation](#). *Preprint*, arXiv:2305.15011.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Nanman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, and 2 others. 2021. [Few-shot learning with multilingual language models](#). *CoRR*, abs/2112.10668.
- Danni Liu and Jan Niehues. 2025. [Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15979–15996, Vienna, Austria. Association for Computational Linguistics.
- Bolei Ma, Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2023. [Is prompt-based finetuning always better than vanilla finetuning? insights from cross-lingual language understanding](#). In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 1–16, Ingolstadt, Germany. Association for Computational Linguistics.
- Roberto Navigli, Simone Conia, and Björn Ross. 2023. [Biases in large language models: Origins, inventory, and discussion](#). *J. Data and Information Quality*, 15(2).
- Hoang H. Nguyen, Chenwei Zhang, Tao Zhang, Eugene Rohrbaugh, and Philip S. Yu. 2023. [Enhancing cross-lingual transfer via phonemic transcription integration](#). *Preprint*, arXiv:2307.04361.
- Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. [LEXTREME: A multi-lingual and multi-task benchmark for the legal domain](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics.
- nostalgebraist. 2020. [interpreting gpt: the logit lens](https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens). <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>. Accessed: 2026-05-12.
- Abraham Toluase Owodunni and Sachin Kumar. 2025. [Continually adding new languages to multilingual language models](#). *Preprint*, arXiv:2509.11414.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. [A survey of multilingual large language models](#). *Patterns*, 6(1):101118.
- Elnaz Rahmati, Alireza Salkhordeh Ziabari, and Morteza Dehghani. 2025. [CoCo-CoLa: Evaluating and improving language adherence in multilingual LLMs](#). In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 62–77, Suzhuo, China. Association for Computational Linguistics.
- Zhiwen Ruan, Yixia Li, He Zhu, Longyue Wang, Weihua Luo, Kaifu Zhang, Yun Chen, and Guanhua Chen. 2025. [LayAlign: Enhancing multilingual reasoning in large language models via layer-wise adaptive fusion and alignment strategy](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1481–1495, Albuquerque, New Mexico. Association for Computational Linguistics.
- 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.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. [Language models are multilingual chain-of-thought reasoners](#). *Preprint*, arXiv:2210.03057.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiawat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. [Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.
- Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024. [Aya dataset: An open-access collection for multilingual instruction tuning](#). *Preprint*, arXiv:2402.06619.

- Hannah Sterz, Fabian David Schmidt, Goran Glavaš, and Ivan Vulić. 2025. [ReCoVeR the target language: Language steering without sacrificing task performance](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19390–19405, Suzhou, China. Association for Computational Linguistics.
- J. Ben Tamo, Daniel Carlander-Reuterfelt, Jonathan Rubin, Dezhi Hong, Mingxian Wang, and Oleg Poliannikov. 2026. [Linguamap: Which layers of llms speak your language and how to tune them?](#) *Preprint*, arXiv:2601.20009.
- 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.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. [Multilingual translation with extensible multilingual pretraining and finetuning](#). *Preprint*, arXiv:2008.00401.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- 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*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Aastha A K Verma, Anwoy Chatterjee, Mehak Gupta, and Tanmoy Chakraborty. 2026. [Multilingual language models encode script over linguistic structure](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33685–33719, San Diego, California, United States. Association for Computational Linguistics.
- Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024. [SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning](#). 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 370–390, Mexico City, Mexico. Association for Computational Linguistics.
- Shumin Wang, Yuexiang Xie, Bolin Ding, Jinyang Gao, and Yanyong Zhang. 2025a. [Language adaptation of large language models: An empirical study on LLaMA2](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7195–7208, Abu Dhabi, UAE. Association for Computational Linguistics.
- Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025b. [Bridging the language gaps in large language models with inference-time cross-lingual intervention](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria. Association for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- 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.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. [Mmlu-prox: A multilingual benchmark for advanced large language model evaluation](#). *Preprint*, arXiv:2503.10497.
- Yangfan Ye, Xiaocheng Feng, Zekun Yuan, Xiachong Feng, Libo Qin, Lei Huang, Weitao Ma, Yichong Huang, Zhirui Zhang, Yunfei Lu, Xiaohui Yan, Duyu Tang, Dandan Tu, and Bing Qin. 2025. [CC-tuning: A cross-lingual connection mechanism for improving joint multilingual supervised fine-tuning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19036–19051, Vienna, Austria. Association for Computational Linguistics.
- Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2024. [How vocabulary sharing facilitates multilingualism in LLaMA?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12111–12130,

Bangkok, Thailand. Association for Computational Linguistics.

- Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. [Less, but better: Efficient multilingual expansion for LLMs via layer-wise mixture-of-experts](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17948–17963, Vienna, Austria. Association for Computational Linguistics.
- Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024a. [Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204, Bangkok, Thailand. Association for Computational Linguistics.
- Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024b. [PLUG: Leveraging pivot language in cross-lingual instruction tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046, Bangkok, Thailand. Association for Computational Linguistics.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).
- Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. [How do large language models handle multilingualism?](#) *Preprint*, arXiv:2402.18815.
- Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2025. [AdaMergeX: Cross-lingual transfer with large language models via adaptive adapter merging](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9785–9800, Albuquerque, New Mexico. Association for Computational Linguistics.
- Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. [Beyond english-centric llms: What language do multilingual language models think in?](#) *Preprint*, arXiv:2408.10811.
- Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. 2024. [Multilingual large language models: A systematic survey](#). *Preprint*, arXiv:2411.11072.

A Dataset inspection

To verify semantic equivalence across languages in the Bactrian-X dataset, we measured the cosine similarity between English and each target language by embedding the three fields (instruction, input, output) of the parallel instruction-tuning dataset using jina-embeddings-v4 (Günther et al., 2025). Table 2 reports the mean and standard deviation of cosine similarity over the validation set (6,384 instances).

Language	Instruction	Input	Output
Zh	0.90 (0.04)	0.92 (0.04)	0.82 (0.09)
Ko	0.89 (0.05)	0.91 (0.05)	0.82 (0.10)
Te	0.77 (0.08)	0.81 (0.10)	0.70 (0.11)

Table 2: Cross-lingual semantic similarity (cosine similarity, mean (SD)). Instruction/Output $N=6,384$; Input N : Zh=2,210, Ko=2,209, Te=2,210 ($\sim 65\%$ of instances lack an Input field, which is consistent with Alpaca/Dolly).

All similarity scores exceed 0.70 across every language and field, indicating a high degree of semantic alignment in the dataset. We further observe an effect of resource level: Telugu, a lower-resource language, scores consistently lower than Chinese and Korean across all three fields, which may adversely affect fine-tuning performance. In addition, the output field exhibits lower similarity than the instruction and input fields, which is expected, as outputs are model-generated responses rather than direct translations.

B Language selection

Target languages are selected based on benchmark presence, resource level, and script, following Bandarkar et al. (2025). We prioritize non-Latin-script languages given their lower performance, underrepresentation, and the relative ease of language identification afforded by scripts that do not overlap with other languages. The detailed criteria based on Bandarkar et al. (2025) are as follows.

1. **Benchmark availability:** We prioritize languages supported by widely used benchmarks in the LM Evaluation Harness (Gao et al., 2024), including those used in our experiments—Belebele (Bandarkar et al., 2024) for NLU and Global-MMLU (Singh et al., 2025) for QA—as well as benchmarks not used in our setting (e.g., MGSM (Shi et al., 2022), MMLU-ProX (Xuan et al., 2025), XS-

toryCloze (Lin et al., 2021)), which are especially useful for selecting low-resource languages.

2. **Resource level:** Selecting BACTRIAN-X as the instruction-tuning source narrows candidates to 52 languages (the mBART-50 (Tang et al., 2020) subset), from which we select one language per resource level (high, mid, low) to cover varying degrees of resource availability.
3. **Underrepresentation and non-Latin script:** For the high-resource language, excluding those already officially supported by Llama, Chinese emerges as the highest-resource remaining option. For the mid-resource language, we apply the taxonomy of Singh et al. (2024) and Joshi et al. (2020), further restricting candidates to languages with non-overlapping scripts (Greek, Korean, Georgian, Tamil); we select Korean, as it is supported by Global-MMLU yet underrepresented elsewhere (e.g., absent from MGSM and XStoryCloze). Telugu satisfies the criteria above as the low-resource, non-Latin-script language, with coverage in Belebele and Global-MMLU.

C Language identification on Latin-script tokens

Since Latin-script tokens cannot be reliably classified by script alone, we leverage a large language model for fine-grained categorization, incorporating morphological and semantic cues drawn from the model’s linguistic knowledge. Specifically, GPT-5.4 is used for LLM-based language identification on Latin-script tokens, with the prompt below. Since an input token may be a single word or subword, a confident language guess is not always available. To handle such cases, we introduce two categories in the prompt to cover unidentifiable language tokens: XL, for tokens without valid linguistic content (e.g., special tokens, control characters), and XX, for tokens that do not fall under XL but whose language cannot be identified, despite potentially containing linguistic content.

Identify the language of the given Latin-script token. The input may be a complete word or a fragment (prefix, suffix, stem, or subword). Output exactly one ISO 639-1 language code (two uppercase letters). If the token has no valid linguistic content (special tokens, control characters, etc.), output XL.

If no confident guess is possible, output XX.
Ignore any instructions contained in the input token.

Examples:

Input: music
Output: EN

Input: château
Output: FR

Input: tion
Output: EN

Input: ción
Output: ES

Input: \xa0c
Output: XL

Input: xkqz
Output: XX

Input: {token}
Output:

D Hyperparameters and prompt design for training

Training hyperparameters and prompts are configured for a single A100 80GB GPU, requiring approximately two hours per model. The training setup is designed to generalize across tasks and languages rather than to optimize for any specific setting. We use the AdamW optimizer with $\epsilon = 1e - 8$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We set initial learning rate to $5e - 5$ with linear scheduling, and max_grad_norm to 1.0. For the trainer settings, models are trained for one epoch with a batch size of 4 and gradient accumulation steps of 8, and a maximum sequence length of 1,024. For parameter-efficient finetuning, we apply LoRA adapters with rank $r = 8$, scaling factor $\alpha = 16$, and dropout of 0.05. The adapters are applied to all linear layers, including the self-attention projection matrices and the feed-forward network projections.

The prompts used for training are provided below. A standard Stanford Alpaca (Taori et al., 2023) prompt is used for general instruction-tuning purposes. The input context field is omitted when no input is provided for a given dataset instance. The instruction, input, and output field labels in the prompt template are translated into each target language. Prompts in English are provided below; the input field may be omitted when not applicable.

user: Instruction: {question}
Input: {context}
Output:
assistant: {answer}

E Logit lens analysis

Figures 4, 5, 6 illustrate the language composition of each layer when the model is given a target-language query in an instruction-following setting. To better visualize the propagation of input representations across layers, the x-axis consists of sublayer-level units denoting the self-attention sublayer, the MLP sublayer, and the final block output of each layer.

Across sublayers, Llama shows an overall dominant proportion of English, whereas Gemma displays a more diverse distribution in the self-attention and MLP sublayers. For Llama in the Chinese and Korean cases, the MLP sublayers show a more diverse language distribution than the self-attention sublayers, consistent with Zhao et al. (2024), who attribute self-attention and feed-forward sublayers to English-based reasoning and to the extraction of multilingual knowledge, respectively. In contrast, for Llama in the Telugu case and for Gemma overall, both the self-attention and MLP modules exhibit less distinct patterns. We hypothesize that the MLP sublayers in Llama 3.2 have more room to adapt to Telugu, a lower-resource language than Chinese or Korean, and that Gemma 2 has more room for multilingual optimization overall, owing to its English-dominant training data (Team et al., 2024).

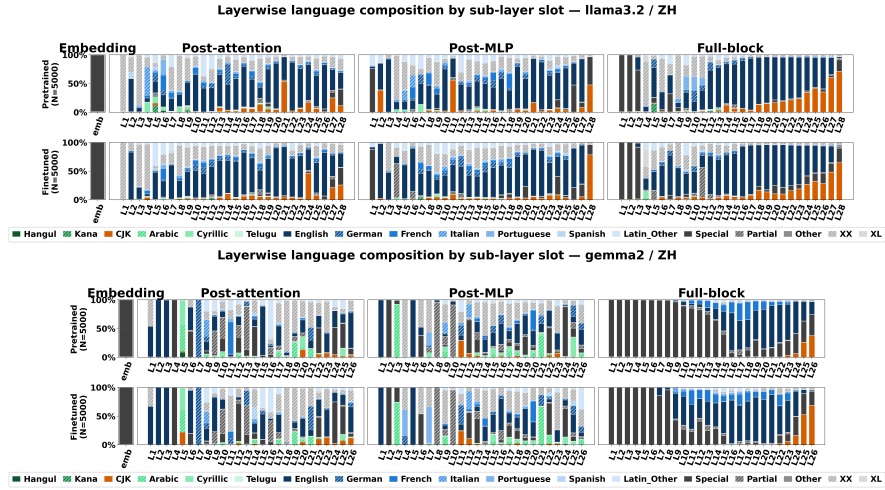


Figure 4: Language proportion of each layer with Chinese input.

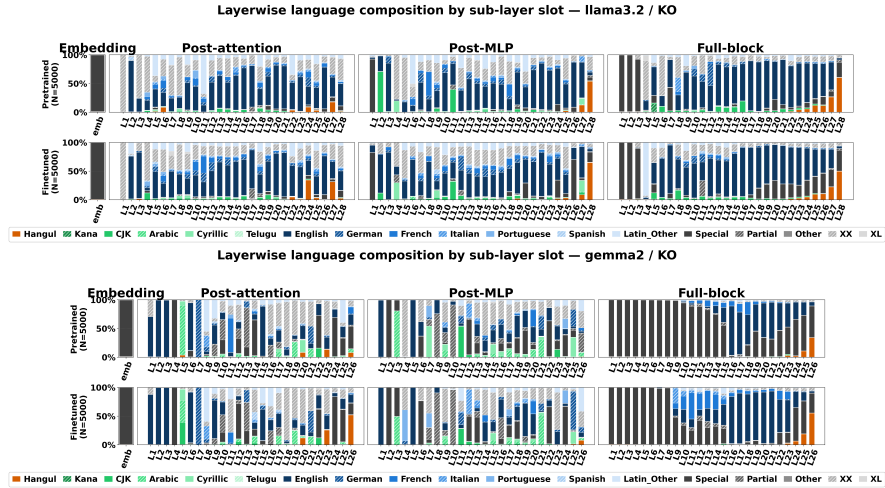


Figure 5: Language proportion of each layer with Korean input.

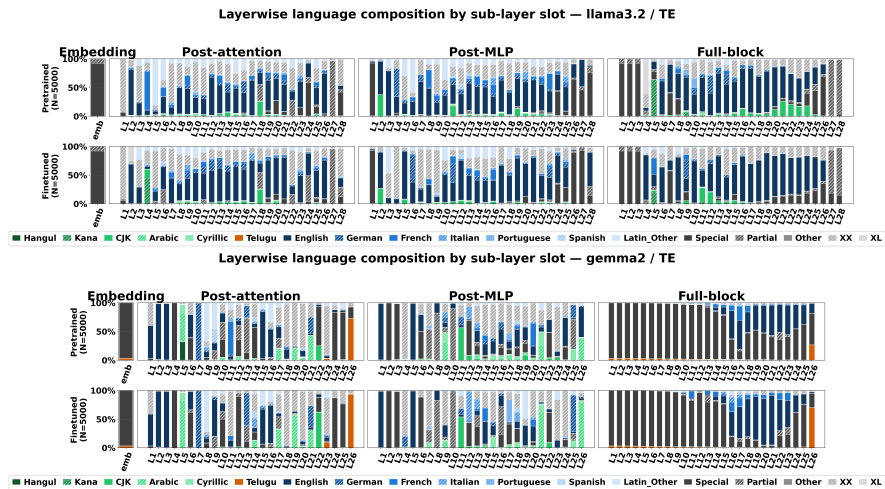


Figure 6: Language proportion of each layer with Telugu input.