

Localizing Cross-Lingual Generalization in Parameter Space

Yeoreum Lee¹, Jaeseong Lee², Seung-won Hwang^{1*}

Interdisciplinary Program in Artificial Intelligence, Seoul National University¹

Snowflake AI Research²,

leeyeoreum01@snu.ac.kr, jaeseong.lee@snowflake.com, seungwonh@snu.ac.kr

Abstract

Cross-lingual transfer is uneven across tasks, succeeding for natural language understanding but often failing for mathematical reasoning. We argue that this asymmetry is governed by the geometry of parameter space. Specifically, cross-lingual generalization is determined by whether English and target-language LoRA adapters share a low-error basin. To test this, we use linear mode connectivity (LMC) between English and target-language adapters, evaluating on three tasks. We find that NLI and QA model-language pairs are mostly linearly connected, whereas mathematical reasoning (Math) pairs are not, indicating that basin geometry is task-dependent. Such disconnection in Math is “local” rather than global. A small target-language task sample, applied only to the English adapter’s language encoding or semantic composition layers, is enough to repair it. Cross-lingual generalization is also factorizable into language-specific encoding and language-agnostic composition, each supplied separately from a cheaper signal. English task data routed to the composition layers and an unlabeled target-language corpus routed to the language encoding layers each alleviate it on their own for nearly every pair. Composition is the more influential of these, and it needs no target-language data. Together, these results suggest that cross-lingual generalization decomposes into language-specific and language-agnostic components in parameter space. This decomposition offers a practical route for low-resource cross-lingual adaptation.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of NLP tasks, including natural language understanding and reasoning tasks (OpenAI, 2023; Grattafiori et al., 2024; Gemma Team, 2024; Qwen Team,

2025; Gemini Team, 2023). However, these capabilities remain largely English-centric. This has driven growing interest in cross-lingual transfer, which leverages the strong capabilities of high-resource languages (e.g., English) to improve performance in other languages (Lauscher et al., 2020; Conneau et al., 2020; Hu et al., 2025; Bandarkar et al., 2025). Among these, a growing line of work performs cross-lingual transfer with LoRA for efficient adaptation to target languages (Whitehouse et al., 2024; Zhao et al., 2025).

But why is such cross-lingual transfer successful in some cross-lingual settings but not in others? Prior work approaches this question either externally, through properties of the language pair, or internally, through the mechanisms of an isolated pre-trained model. Specifically, the former predicts how well generalization will transfer between languages from properties of the language pair such as typological distance (Lin et al., 2019; Philippou et al., 2023). However, since these properties are tied to language pairs rather than models, they cannot account for variation across different models or tasks for the same pair. Alternatively, the latter probes the phenomenon internally, through neuron-level analyses (Tang et al., 2024; Kojima et al., 2024) and representation-level analyses that characterize how representations evolve along the depth of the model (Wendler et al., 2024; Wu et al., 2025; Zhong et al., 2025). Yet these analyses only examine the internal mechanisms of an isolated pre-trained model and do not explicitly compare it against an expert model specialized for the target language. We instead study cross-lingual generalization in parameter space, comparing the English adapter with a target-language expert. This shifts the question from external linguistic properties or internal pre-trained representations to the geometry of adaptation itself.

A natural lens for this geometry is *linear mode connectivity* (LMC), which evaluates whether two

* Corresponding author.

models fine-tuned from the same initialization share the same error basin (Frankle et al., 2020; Entezari et al., 2022). Applied to English and target-language adapters, LMC indicates whether they lie within a connected low-error region, providing a parameter-space probe for cross-lingual generalization. LMC has been studied for pre-trained language models (Qin et al., 2022) and multi-task LoRA adapters (Zhuang et al., 2025), but not yet for the cross-lingual case.

We cover cross-lingual generalization across three models and three tasks spanning natural language understanding (NLI and QA) and mathematical reasoning (Math). Our analysis surfaces three findings. First, basin geometry is task-dependent: NLI and QA model-language pairs are almost always linearly connected, while Math pairs largely fail to connect. Second, this deficit is local rather than global: adapting either the early or the middle layers alone restores connectivity in Math, while the final layers contribute little. Relative to the target-language adapter, this indicates that the English adapter falls short in input encoding and semantic composition, not in output decoding. Third, cross-lingual generalization is factorizable into language-specific input encoding and language-agnostic semantic composition. Each improves LMC on its own for nearly every pair, composition from English task data at the middle layers and language-specific processing from an unlabeled corpus at the early layers. Neither signal requires target-language task labels. Of the two, semantic composition is the more influential factor, and the language-specific signal adds little beyond it. Together, these findings reveal cross-lingual generalization as the parameter-space alignment of two components that can be addressed separately through layer-targeted training. In practice, English task supervision at the composition layers alone can extend an English adapter to a new target language, without any target-language data.

Our contributions are as follows:

- A parameter-space characterization of cross-lingual generalization through LMC, applied across three LLMs, three tasks, and four target languages.
- Evidence that cross-lingual generalization is task-dependent rather than monolithic, with NLU adapter pairs connecting cleanly while Math pairs largely fail to connect.

- A layer-level localization of the generalization deficit, showing that adapting only the adapter’s early or middle layers suffices to repair it.
- A parameter-space factorization of cross-lingual generalization into language-specific processing and language-agnostic composition. It is exploitable through layer-targeted training on English supervision alone, without any target-language data.

2 Related work

Cross-lingual transfer. A broad line of work has pursued cross-lingual transfer, improving its performance through various adaptation and merging strategies. Early work demonstrated that multilingual language models achieve zero-shot cross-lingual transfer (Pires et al., 2019; Conneau et al., 2020). Subsequent work has improved transfer performance through representation alignment (She et al., 2024), merging-based approaches that compose language and task modules in parameter space (Ansell et al., 2022; Zhao et al., 2025; Bandarkar et al., 2025; Chen et al., 2025), and leveraging internal English knowledge (Etxaniz et al., 2024; Zhu et al., 2024a). However, these studies build transfer by composing language and task modules. We instead use parameter-space geometry to characterize when cross-lingual generalization holds, rather than to produce it.

Localizing cross-lingual generalization. What drives cross-lingual generalization has been analyzed through external linguistic features (Lin et al., 2019; Ahuja et al., 2022; Philipp et al., 2023; Blaschke et al., 2025). Riemenschneider and Frank (2025) go further and argue that cross-lingual transfer is better treated as an evaluation method than as a capability in itself. These studies describe transfer from the outside and leave open where, if anywhere, it is realized inside the model. This motivates a line of work that asks where inside a pre-trained LLM cross-lingual generalization is realized, locating it along the depth of the network.

Neuron-level analyses suggest that this generalization is realized in the interior of the network. Activations sensitive to a specific language tend to concentrate near the outer portions rather than the interior (Tang et al., 2024; Kojima et al., 2024; Zhu et al., 2024b). Logit-lens and activation-patching studies localize generalization to the same interior

region from a representational standpoint (Wendler et al., 2024; Zhao et al., 2024; Schut et al., 2025; Wu et al., 2025). They show that the inner forward pass operates in a language-agnostic (or English-aligned) space, with the target language emerging only near the output. Some work argues that language-agnostic and language-specific representations can coexist along the same layers (Chang et al., 2022; Brinkmann et al., 2025; Dumas et al., 2025; Fierro et al., 2025). Other work finds that the dominant internal latent language is itself shaped by the pre-training mixture rather than defaulting to English (Zhong et al., 2025). However, all of these analyses operate within the internal representations and computations of the model. They leave unexamined the parameter-space geometry between an isolated pre-trained model and an expert model specialized for the target language.

Linear mode connectivity. A series of studies has established that models sharing the same initialization occupy the same error basin, giving rise to LMC as a tool for probing the structure of error landscapes (Frankle et al., 2020; Neyshabur et al., 2020; Entezari et al., 2022). Qin et al. (2022) provided the first systematic LMC analysis of pre-trained language models. Recent work has examined LMC of LoRA adapters in multi-task scenarios (Zhuang et al., 2025). However, all previous LMC studies in NLP compare endpoints trained on the same language data, leaving the geometry between English and target-language adapters outside their scope.

3 Evaluation design for research questions

Research questions. We investigate cross-lingual generalization through three questions, each established in a later section. First, is cross-lingual generalization task-dependent rather than uniform across tasks (§4)? Second, when it breaks on Math, is the deficit layer-localized rather than global (§5)? Third, does the deficit factorize into two separately suppliable components (§6)? All three questions rest on the same parameter-space probe, which we lay out in the remainder of this section.

Tasks and datasets. Our experiments cover three tasks that range from NLU to mathematical reasoning: NLI, QA, and Math, with English as the pivot language. Across all three tasks, we use datasets whose multilingual training and test sets are trans-

lations of a common English source. This ensures that differences in basin structure across adapters reflect language rather than data content.

Models and training. To ensure that our analysis is not tied to a single model, we use three instruction-tuned LLMs of similar scale: Llama 3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-8B (Qwen Team, 2025), and Gemma 2-9B-it (Gemma Team, 2024) as base models. Each model is fine-tuned independently on English or corresponding target language using LoRA adapters (Hu et al., 2022), sampling 5,000 examples per language. Target-language data is scarce in practice, so an adapter fine-tuned directly on such data is the ideal comparison endpoint for the English adapter. All adapters are trained for 3 epochs using LlamaFactory (Zheng et al., 2024). Further experimental details are provided in Appendix A.

Linear mode connectivity. To understand cross-lingual generalization through the lens of parameter space, we use linear mode connectivity. LMC tests whether two models trained from the same initialization occupy the same error basin (Frankle et al., 2020; Neyshabur et al., 2020; Entezari et al., 2022). We adapt it to the cross-lingual setting by measuring connectivity between an English adapter and each target-language adapter, both fine-tuned from the same base model. Given English adapter parameters θ_{en} and target-language adapter parameters θ_{tgt} , we evaluate the task metric $M(\theta_\alpha)$ along the linear path $\theta_\alpha = (1 - \alpha)\theta_{\text{en}} + \alpha\theta_{\text{tgt}}$ for $\alpha \in \{0, 0.1, \dots, 1.0\}$. If performance does not degrade along this path, no error barrier exists between the two adapters, indicating that they are linearly connected and lie in the same error basin. Specifically, we regard a adapter pair as linearly connected when no valley appears along the path and the English adapter matches the target-language adapter in performance. The barrier height (Entezari et al., 2022) quantifies the maximum performance drop along this path:

$$B = \sup_{\alpha} [(1 - \alpha)M(\theta_{\text{en}}) + \alpha M(\theta_{\text{tgt}}) - M(\theta_\alpha)]. \quad (1)$$

Since the target-language adapter is fine-tuned directly on target-language task data, it converges to a low-error basin on the target-language distribution. When the English adapter is linearly connected to it, the interpolation path remains within this basin, placing the English adapter in the same

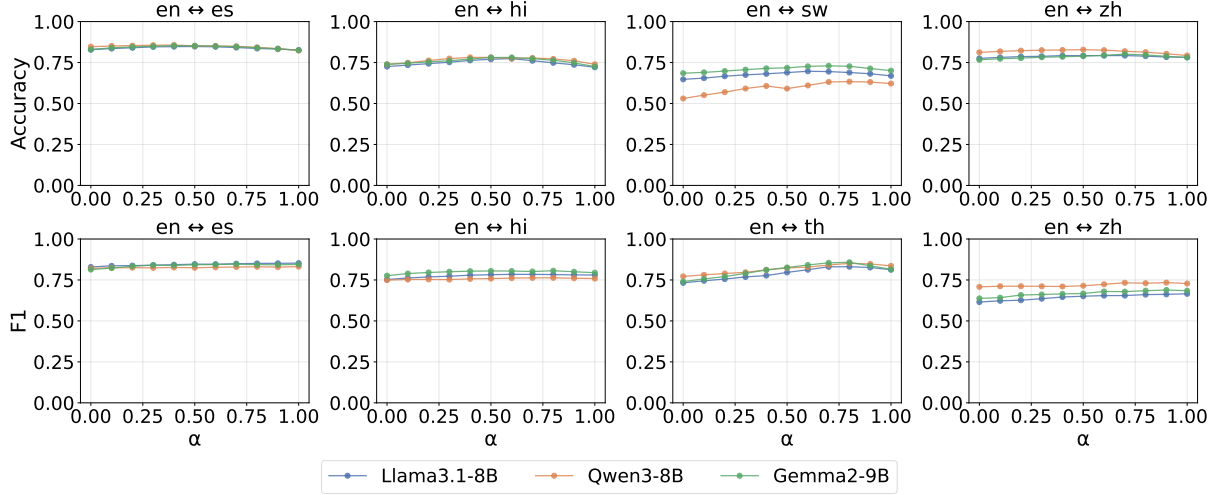


Figure 1: Linear interpolation paths between the English adapter ($\alpha = 0$) and the target-language adapter ($\alpha = 1$) for NLI and QA. The first row reports accuracy evaluated on XNLI, and the second row reports F1 evaluated on XQuAD. Each subplot reports task performance in the respective target language along the interpolation path.

target-language low-error region. We therefore interpret LMC between English and target-language adapters as parameter-space evidence of cross-lingual generalization. Specifically, the English adapter generalizes to the target-language distribution as effectively as the target-language oracle adapter.

Multilingual workflow. The internal behavior of multilingual LLMs has been characterized by the multilingual workflow (MWork) hypothesis, under which a multilingual LLM processes an input through three sequential stages (Zhao et al., 2024; Wendler et al., 2024). The early-layer stage encodes input tokens, written in the query language, into a concept-space representation that is largely language-agnostic (and empirically closer to English). The middle-layer stage operates on this representation to build a semantically correct next-token representation. The final-layer stage then decodes this representation back into the output space of the original input language. We utilize MWork to interpret LMC between an English adapter and a target-language adapter. When the two adapters are not linearly connected, we attribute the failure to the English adapter lacking sufficient ability at one or more MWork stages. Building on this view, we perform a layer-wise analysis to localize where such deficiency arises.

4 Cross-lingual generalization is mirrored by linear mode connectivity

Summary. Cross-lingual generalization manifests as LMC in parameter space between the En-

glish and target-language adapters. This connectivity is sharply task-dependent. It holds for NLI and QA but breaks for Math. Prior work has offered limited insight into why it varies in this way, treating cross-lingual generalization as a largely monolithic phenomenon.

4.1 NLI and QA

Setup. For NLI, we train on the machine-translated multilingual NLI dataset (Laurer et al., 2024) and evaluate on XNLI (Conneau et al., 2018), covering Arabic (ar), German (de), Spanish (es), French (fr), Hindi (hi), Russian (ru), Swahili (sw), Vietnamese (vi), and Chinese (zh). For QA, we train on the translated SQuAD data from XTREME (Hu et al., 2020) and evaluate on XQuAD (Artetxe et al., 2020), covering Arabic (ar), German (de), Spanish (es), Hindi (hi), Russian (ru), Thai (th), Vietnamese (vi), and Chinese (zh). These languages were chosen to cover commonly used low-, medium-, and high-resource languages. In the main text, we report Spanish (es), Hindi (hi), and Chinese (zh) for both tasks, as well as Swahili (sw) for NLI and Thai (th) for QA, with the remaining languages provided in Appendix B. Task metrics are accuracy for NLI and F1 for QA.

NLU exhibits no error barrier and satisfies cross-lingual generalization. The NLI and QA paths in Figure 1 are mostly flat with no error barrier. For these NLU tasks, the English and target-language adapters lie in the same low-error basin, indicating that the English adapter already generalizes to the target language. As edge cases, a single

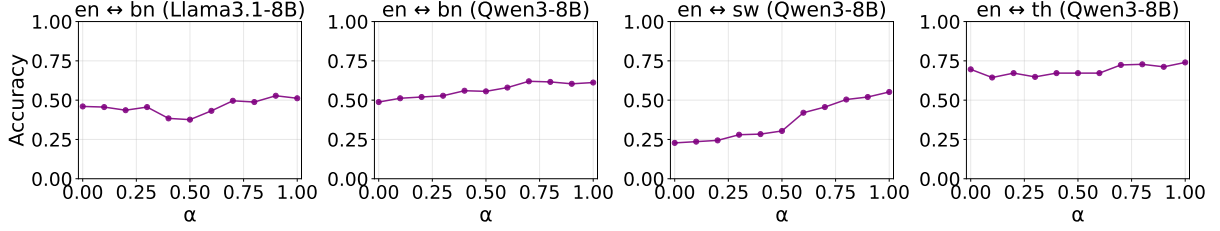


Figure 2: Linear interpolation path between English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) on Math. Each subplot shows accuracy evaluated on MGSM in the respective target language.

NLI pair (sw) and a single QA pair (th) show a slight rise in the middle of the interpolation path (third column in Figure 1). This mild fluctuation still keeps both adapters within a single shared basin on the target language.

4.2 Mathematical reasoning

Setup. For mathematical reasoning (Math), we train on MGSM8KInstruct (Chen et al., 2024) and evaluate on MGSM (Shi et al., 2023). MGSM covers Bengali (bn), German (de), Spanish (es), French (fr), Japanese (ja), Russian (ru), Swahili (sw), Thai (th), and Chinese (zh), from which we retain a model-specific subset (Table 1). We restrict our analysis to (model, language) pairs whose target-language adapter improves over the zero-shot instruction-tuned model (the base model in our setup) on the LMC evaluation set. This improvement ensures that the target-language adapter is the oracle that our analysis assumes, rather than a failed one. Such failures occur when the instruction-tuned model is already strong and small-scale fine-tuning degrades it below the zero-shot baseline. For NLI and QA, every language satisfies the criterion for every model, so all models share the same language set in each task. For Math, the screen is selective and produces model-specific sets, reported in Table 1. Task metrics are accuracy for Math.

Math adapters are not linearly connected and fail to satisfy cross-lingual generalization. Figure 2 shows that in Math, the English and target-language adapters consistently fail to share a basin on the target language, although the form of this failure varies across pairs. First, for Llama3.1-8B on bn and Qwen3-8B on th, an error barrier separates the two adapters, placing them in different basins even though their endpoint performance is comparable. This leaves it unclear whether the English adapter generalizes to the target language. The monotonically increasing cases also occur for Qwen3-8B on bn and sw, with sharper rises than

Lang	Llama3.1-8B	Qwen3-8B	Gemma2-9B
bn	+0.190	+0.296	-0.072
de	-0.140	-0.064	-0.068
es	-0.108	-0.056	-0.104
fr	-0.156	-0.108	-0.168
ja	-0.088	-0.152	-0.100
ru	-0.132	-0.044	-0.116
sw	-0.104	+0.080	-0.048
th	-0.112	+0.108	-0.060
zh	-0.132	-0.128	-0.052

Table 1: Language selection for Math (MGSM). Each entry is the accuracy improvement of the target-language adapter over the base model (i.e., instruction-tuned model) on the target-language test set. Entries with positive value (selected pairs) are shown in bold.

in the NLU setting, most pronounced for sw. This indicates that the English adapter lies even further outside the target-language basin and thus lacks, to an even greater extent than in NLU, the additional cross-lingual ability that Math requires. In summary, the contrast between NLU and Math shows that cross-lingual generalization is not monolithic. Its basin geometry is governed by task type, with most pairs linearly connected in NLU and LMC rarely observed in Math.

5 Localizing the abilities that improve cross-lingual generalization

Summary. When cross-lingual generalization breaks, the deficit is local rather than global, recoverable by adapting a single layer group and confined to the early and middle layers. Section 4 showed that this breakdown occurs for Math while NLU pairs stay connected, which leaves open whether the Math gap reflects a global mismatch or a localized one. Relative to the target-language oracle, we ask which ability the English adapter lacks on Math, and whether the deficit can be repaired by adapting only the corresponding layers. We pursue this question through MWork, which assigns distinct functional roles to the early, middle, and final layers. We localize the deficit by adapting

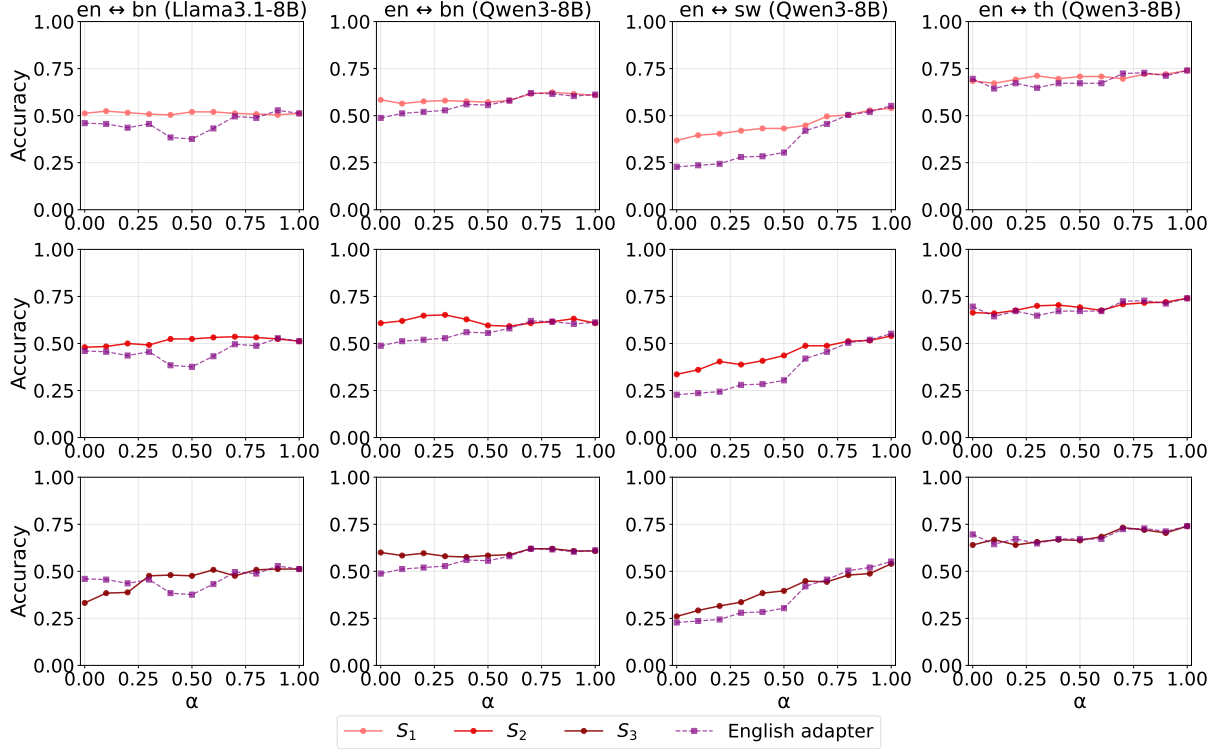


Figure 3: Linear interpolation path between ability-adapted English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) on Math. Rows correspond to early-layer (S_1), middle-layer (S_2), and final-layer (S_3) adaptation, respectively. Each column corresponds to a target language, and each subplot shows accuracy evaluated on MGSM in that language.

each group in isolation on the target-language task dataset. The group that restores connectivity then pinpoints the ability the English adapter lacks.

5.1 Setup

Layer-group adaptation. Recall that MWork attributes early layers to language-specific input encoding, middle layers to language-agnostic semantic composition, and final layers to language-specific output decoding. We isolate each of these three layer groups and continually fine-tune them to enhance the corresponding ability. We then examine how this targeted adaptation affects cross-lingual LMC and, in turn, cross-lingual generalization. Specifically, for each target language, we randomly sample 500 target-language training examples from the Math dataset (MGSM8KInstruct). Starting from the English adapter parameters θ_{en} , we continually fine-tune only a chosen layer group $S \subseteq \{1, \dots, L\}$ on these samples for 3 epochs. The remaining layers stay frozen at the English adapter values. We then evaluate the task metric $M(\theta_\alpha^S)$ along the linear path $\theta_\alpha^S = (1 - \alpha)\theta_{\text{en}}^S + \alpha\theta_{\text{tgt}}^S$ between the ability-adapted English adapter θ_{en}^S and the target-language adapter θ_{tgt}^S , with $\alpha \in \{0, 0.1, \dots, 1.0\}$. A previously disconnected

pair may become linearly connected once we strengthen a specific ability. We take this as evidence that the corresponding ability is what the English adapter lacks for cross-lingual generalization.

Analysis setup. We report results across all three layer groups: early ($S_1 = \{1, \dots, \lceil L/3 \rceil\}$), middle ($S_2 = \{\lceil L/3 \rceil + 1, \dots, \lceil 2L/3 \rceil\}$), and final layers ($S_3 = \{\lceil 2L/3 \rceil + 1, \dots, L\}$). We focus this analysis on Math, the task whose adapter pairs are linearly disconnected.

5.2 Results

One of the abilities the English adapter lacked is language-specific input encoding. Adapting only the early layers makes most adapter pairs linearly connected on Math (first row of Figure 3). Early-layer adaptation can relocate the English adapter into the target-language basin without altering the remaining layers. The only residual case is Qwen3-8B on sw, which does not become linearly connected, yet its steep monotonic increase before adaptation is clearly attenuated. Within MWork, this restoration indicates that the English adapter’s input encoding generalizes less well to target-language inputs than the target-language

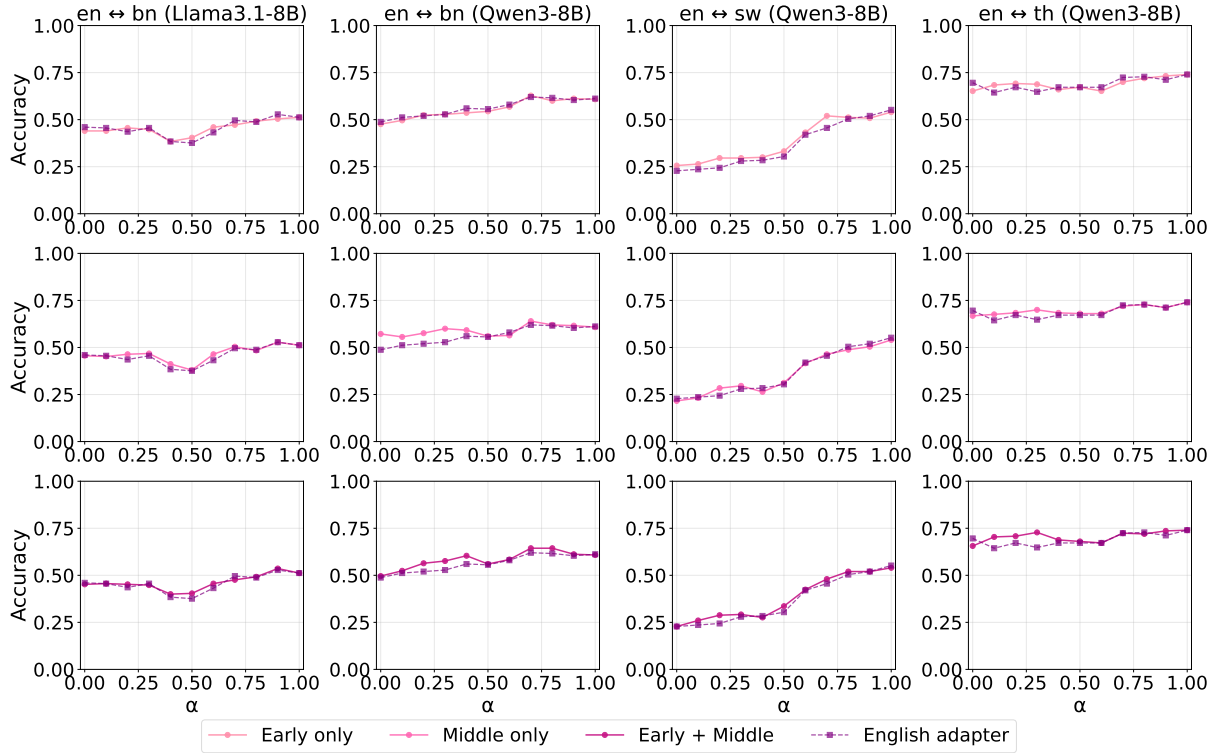


Figure 4: Linear interpolation path between routed-adaptation English adapter $\tilde{\theta}_{\text{en}}$ ($\alpha = 0$) and target-language adapter θ_{tgt} ($\alpha = 1$) on Math. Rows correspond to routing only the corpus signal to S_1 (Early only), routing only the English task signal to S_2 (Middle only), and routing both signals jointly (Early and Middle), respectively. Each column corresponds to a target language, and each subplot shows accuracy evaluated on MGSM in that language.

adapter. Early-layer adaptation thus identifies language-specific input encoding as one of the abilities the English adapter must improve for cross-lingual generalization.

Language-agnostic semantic composition is another ability in which the English adapter was deficient. Middle-layer adaptation also makes most adapter pairs linearly connected on Math, restoring the English adapter’s semantic composition to the level of the target-language adapter (second row of Figure 3). Qwen3-8B on sw still increases monotonically, but its barrier is mitigated relative to the naive English adapter. This suggests that the naive English adapter’s language-agnostic semantic composition does not fully generalize to target-language inputs as well as the target-language adapter does. Middle-layer adaptation thus identifies language-agnostic semantic composition as another ability the English adapter lacks relative to the target-language adapter, which can limit cross-lingual generalization.

Language-specific output decoding is less beneficial for achieving cross-lingual generalization. Among the three layer groups, final-layer adapta-

tion produces the weakest connectivity, although only some of its error barriers become milder after adaptation (third row of Figure 3). More pairs fail to become linearly connected than under early- or middle-layer adaptation, and these failures show increases, with Llama3.1-8B on bn as the most notable case. The final layers likely still operate on the original upstream English representation, so adapting their decoding alone may not relocate the adapter into the target-language basin. Final-layer adaptation thus identifies language-specific decoding as an ability that does not drive cross-lingual generalization, in contrast to the early and middle layers.

6 Factorizing cross-lingual generalization without target-language labels

Summary. Cross-lingual generalization factorizes in parameter space into two components, a language-specific input encoding and a language-agnostic semantic composition. Each can be supplied separately from a cheaper signal matched to its role under MWork, without any target-language task labels. The localization in Section 5 relied

on the full target-language task signal, which bundles language and task together. We establish factorization by decomposing this signal and routing each part to its matched layers. We route a target-language corpus to S_1 , since input encoding is language-specific. Semantic composition, by contrast, is language-agnostic, so English task data alone suffices at S_2 .

6.1 Setup

Routed adaptation. Following the per-layer adaptation of Section 5, we now route two distinct training signals to two disjoint layer groups. Specifically, for each target language we randomly sample 500 sentences from the target-language CulturaX (Nguyen et al., 2024) corpus and 500 English training examples from the Math dataset (MGSM8KInstruct). Starting from the English adapter parameters θ_{en} , we jointly fine-tune the language encoding layers S_1 on the corpus and the composition layers S_2 on the English task for 3 epochs. This yields a routed English adapter $\tilde{\theta}_{\text{en}}$. We then evaluate the task metric $M(\tilde{\theta}_{\alpha})$ along the linear path $\tilde{\theta}_{\alpha} = (1 - \alpha)\tilde{\theta}_{\text{en}} + \alpha\theta_{\text{tgt}}$ for $\alpha = \{0, 0.1, \dots, 1.0\}$.

Analysis setup. We compare three routing variants, namely the corpus signal alone to S_1 (Early only), the English task signal alone to S_2 (Middle only), and both signals jointly (Early and Middle). Throughout this section, we quantify the effect of routed adaptation as the reduction in barrier height B (Eq. 1) relative to the naive English adapter.

6.2 Results

Semantic composition is the most dominant factor. Each single-signal routing lowers the barrier on Math for nearly every pair, without any target-language task labels (second and third rows of Table 2, first and second rows of Figure 4). Composition routing (Middle only), the English task at S_2 , is the stronger factor. On average it lowers the barrier from 0.102 to 0.084 and is the only variant to raise accuracy above the naive adapter, from 0.468 to 0.478. Language routing (Early only), the corpus at S_1 , also lowers the barrier, to 0.086, but slightly reduces accuracy. Cross-lingual generalization thus factorizes into a language-specific input encoding and a language-agnostic semantic composition, with semantic composition the most important factor.

Composition alone accounts for much of the gain. Adding the language signal to composition does not improve on composition alone (fourth row of Table 2, third row of Figure 4). The joint routing (Early and Middle) matches composition routing on the average barrier (0.084) but reaches a lower accuracy of 0.458, below the latter’s 0.478. Composition routing therefore attains the highest accuracy and the lowest barrier of any variant. This result is also shown for Qwen3-8B bn, where composition routing reaches the lowest barrier along its interpolation path (Figure 4). The other routings lower the barrier less than composition alone. Composition thus emerges as the primary, label-free driver of cross-lingual generalization.

Baseline	Avg Acc. (Avg. B)
naive	0.468 (0.102)
E	0.456 (0.086)
M	0.478 (0.084)
E+M	0.458 (0.084)

Table 2: Average accuracy and barrier height B of the routing variants on Math (MGSM). Averages are computed over the four (model, language) pairs: Llama3.1-8B bn, Qwen3-8B bn, Qwen3-8B sw, and Qwen3-8B th. Rows denote the naive English adapter (naive) and three routing variants, namely Early only (E), Middle only (M), and Early and Middle (E+M). **Bold** denotes the highest accuracy and lowest barrier height.

Composition routing extends an English model to a new language without target-language data. This factorization has a concrete practical consequence. Because the composition layers are language-agnostic, English task supervision at S_2 alone improves cross-lingual connectivity for nearly every pair. This requires no target-language data, neither task labels nor an unlabeled corpus. The supervision also uses only a small number of English task samples, which keeps the recipe practical in low-resource settings. Target-language task annotations, by contrast, remain a primary bottleneck for cross-lingual adaptation. Composition routing thus offers a label-efficient path to extending an English-trained model to a new target language.

7 Conclusion

Cross-lingual generalization is governed by whether English and target-language adapters share a low-error basin in parameter space. Using linear mode connectivity (LMC) between English and

target-language LoRA adapters across three tasks, we surface three findings. First, the geometry is task-dependent, with NLI and QA model-language pairs mostly linearly connected while mathematical reasoning (Math) pairs largely fail to connect. Second, the Math deficit is layer-localized, since a small target-language task sample on only the language encoding or semantic composition layers can restore LMC. Third, the geometry is factorizable. An unlabeled target-language corpus routed to the language-specific layers and English task supervision routed to the composition layers each lower the barrier on their own for nearly every pair. Composition is the stronger of the two, as English supervision on it alone achieves this without any target-language data, offering a practical route for low-resource cross-lingual adaptation.

Limitations

Our study has several limitations. First, mathematical reasoning is the only non-NLU task in our experiments. Our experimental design requires parallel multilingual training data translated from the same English source, so that differences in basin structure can be attributed to language rather than data content. Few multilingual reasoning datasets satisfy this constraint. Second, MWork was originally formulated based on observations of pre-trained models, and its layer-wise functional decomposition may not transfer cleanly to fine-tuned adapters. Our interpretations under MWork should therefore be regarded as suggestive rather than conclusive. Third, our experiments are limited to 8B-parameter models, and whether larger or smaller models exhibit the same basin structure has not been tested. Fourth, due to a limited computational budget, our analysis is confined to a LoRA-only setting. Whether full fine-tuning yields the same connectivity patterns remains an open question. Fifth, our Math analysis covers only the small and model-specific subset of (model, language) pairs that pass the adapter-improvement criterion, leaving most languages and the entire Gemma model outside its scope. With broader multilingual reasoning datasets, future work could extend the analysis to the languages and models that this criterion currently excludes.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning &

Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2026-25522672, Development of Unified Reasoning Technology Mimicking Human Cognition for Hierarchical Understanding and Unbounded Problem Solving).

References

- Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. [Multi task learning for zero shot performance prediction of multilingual models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.
- Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. [Composable sparse fine-tuning for cross-lingual transfer](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. 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.
- Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2025. [Layer swapping for zero-shot cross-lingual transfer in large language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. 2025. [Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684, Vienna, Austria. Association for Computational Linguistics.
- Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. [Large language models share representations of latent grammatical concepts across typologically diverse languages](#). 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 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

- Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. [The geometry of multilingual language model representations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. [Breaking language barriers in multilingual mathematical reasoning: Insights and observations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.
- Zhipeng Chen, Kun Zhou, Liang Song, Wayne Xin Zhao, Bingning Wang, Weipeng Chen, and Ji-Rong Wen. 2025. [Extracting and combining abilities for building multi-lingual ability-enhanced large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17563–17580, Suzhou, China. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Alexis Conneau, Rutu Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- 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](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.
- Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. 2022. [The role of permutation invariance in linear mode connectivity of neural networks](#). In *International Conference on Learning Representations*.
- 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.
- Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. [How do multilingual language models remember facts?](#) In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16052–16106, Vienna, Austria. Association for Computational Linguistics.
- Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. 2020. [Linear mode connectivity and the lottery ticket hypothesis](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3259–3269. PMLR.
- Gemini Team. 2023. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Gemma Team. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. [Large language models are cross-lingual knowledge-free reasoners](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1525–1542, Albuquerque, New Mexico. Association for Computational Linguistics.
- Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. [On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971,

- Mexico City, Mexico. Association for Computational Linguistics.
- Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. [Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli](#). *Political Analysis*, 32(1):84–100.
- Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. [From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. [Choosing transfer languages for cross-lingual learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.
- Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. 2020. [What is being transferred in transfer learning?](#) In *Advances in Neural Information Processing Systems*, volume 33, pages 512–523. Curran Associates, Inc.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. [CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.
- OpenAI. 2023. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. [Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Yujia Qin, Cheng Qian, Jing Yi, Weize Chen, Yankai Lin, Xu Han, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2022. [Exploring mode connectivity for pre-trained language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6726–6746, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Frederick Riemenschneider and Anette Frank. 2025. [Cross-lingual generalization and compression: From language-specific to shared neurons](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13470–13491, Vienna, Austria. Association for Computational Linguistics.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. [Do multilingual llms think in english?](#) *Preprint*, arXiv:2502.15603.
- Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. [MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.
- 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. 2023. [Language models are multilingual chain-of-thought reasoners](#). In *The Eleventh International Conference on Learning Representations*.
- 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.
- 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.
- Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. [Low-rank adaptation for multilingual summarization: An empirical study](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1202–1228, Mexico City, Mexico. Association for Computational Linguistics.
- Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. [The semantic hub hypothesis: Language models share semantic representations across languages and modalities](#). In

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

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.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. [LlamaFactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

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

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024a. [Question translation training for better multilingual reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Sizhe Liu, Shujian Huang, Shuaijie She, Chris Wendler, and Jiajun Chen. 2024b. [Multilingual contrastive decoding via language-agnostic layers skipping](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8775–8782, Miami, Florida, USA. Association for Computational Linguistics.

Zhan Zhuang, Xiequn Wang, Wei Li, Yulong Zhang, Qiushi Huang, Shuhao Chen, Xuehao Wang, Yanbin Wei, Yuhe Nie, Kede Ma, Yu Zhang, and Ying Wei. 2025. [Come together, but not right now: A progressive strategy to boost low-rank adaptation](#). In *Forty-second International Conference on Machine Learning*.

A Experimental details

A.1 Datasets

The per-language sizes of our training and evaluation sets vary across tasks. For training, we draw on the machine-translated multilingual NLI set with

approximately 25,000 examples per language for NLI (Laurer et al., 2024), the translated SQuAD set with approximately 80,000 for QA (Hu et al., 2020), and MGSM8KInstruct (Chen et al., 2024) with approximately 7,500 for Math. From each set, we sample 5,000 examples per language so that the training size is fixed across tasks. For evaluation, we test NLI on the XNLI test set with 5,000 examples per language (Conneau et al., 2018), QA on the XQuAD test set with 1,190 (Artetxe et al., 2020), and Math on MGSM (Shi et al., 2023) with 250.

A.2 Training

All adapters use LoRA (Hu et al., 2022) with rank 16, alpha 32, and dropout 0.05, applied to the query, key, value, and output projections as well as the gate, up, and down projections. The maximum sequence length is set per task to match the input distributions, with 1,024 tokens for Math, 512 for NLI, and 1,536 for QA. These lengths reflect the long reasoning chains in MGSM8KInstruct, the short premise and hypothesis pairs in NLI, and the Wikipedia-style passages in QA. All adapters are trained by using 4 NVIDIA RTX A6000 GPUs.

We optimize with a cosine schedule, a warmup ratio of 0.15, and weight decay of 0.01. The learning rate is 10^{-4} for Math, 2×10^{-4} for NLI, and 10^{-4} for QA. We train all three tasks for 3 epochs. The effective batch size is held at 16 across tasks, with the per-device batch size and gradient accumulation steps set to 2 and 8 for Math and QA and to 4 and 4 for NLI. Training uses bfloat16 mixed precision, gradient clipping at 0.5, and a fixed seed of 42.

We consider the following adaptation regimes, all initialized from the English adapter:

- **Ability-adapted English adapter $\theta_{\text{en}}^{S_k}$:** From the English adapter θ_{en} , freeze all LoRA parameters outside layer group $S_k \in \{S_1, S_2, S_3\}$ and continue SFT on target-language task data. We use random-sampled 500 examples per language with seed 42.
- **Routed-adaptation English adapter $\tilde{\theta}_{\text{en}}$:** From the English adapter, freeze all LoRA parameters outside the assigned routing slices and continue training with two parallel SFT losses. We use random-sampled 500 examples per language with seed 42. The total loss

Baseline	Llama3.1-8B bn	Qwen3-8B bn	Qwen3-8B sw	Qwen3-8B th	Avg Acc. (Avg. B)
naive	0.460 (0.110)	0.488 (0.062)	0.228 (0.162)	0.696 (0.074)	0.468 (0.102)
E	0.440 (0.092)	0.476 (0.066)	0.256 (0.142)	0.652 (0.044)	0.456 (0.086)
M	0.456 (0.104)	0.572 (0.034)	0.216 (0.162)	0.668 (0.036)	0.478 (0.084)
E+M	0.452 (0.082)	0.496 (0.056)	0.228 (0.156)	0.656 (0.042)	0.458 (0.084)

Table 3: Per-pair accuracy and barrier height B of the routing variants on Math (MGSM). There are four baselines with three routing variants: naive English adapter (naive), namely Early only (E), Middle only (M), and Early and Middle (E+M). For each model-language pair, **bold** denotes the highest accuracy and lowest barrier height.

is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{lang}} \mathcal{L}_{\text{lang}} + \lambda_{\text{task}} \mathcal{L}_{\text{task}},$$

where $\mathcal{L}_{\text{lang}}$ is the standard causal-LM SFT cross-entropy computed on the target-language CulturaX (Nguyen et al., 2024) corpus (SFT-formatted), and $\mathcal{L}_{\text{task}}$ is computed using only English task data. We set $\lambda_{\text{lang}} = \lambda_{\text{task}} = 1$. Each loss term exclusively updates the parameters of its assigned layer slice (i.e., $\mathcal{L}_{\text{lang}}$ updates layers $k \in S_{\text{lang}}$, and $\mathcal{L}_{\text{task}}$ updates layers $k \in S_{\text{task}}$). Specifically, the Language routing (Early only) variant isolates updates to S_1 via $\mathcal{L}_{\text{lang}}$, leaving $\mathcal{L}_{\text{task}}$ inactive ($S_{\text{task}} = \emptyset$). The Composition routing (Middle only) configuration does the reverse, updating S_2 via $\mathcal{L}_{\text{task}}$ while $\mathcal{L}_{\text{lang}}$ is disabled. In Joint routing (Early + Middle), both losses are employed simultaneously, with $\mathcal{L}_{\text{lang}}$ modifying S_1 and $\mathcal{L}_{\text{task}}$ modifying S_2 .

A.3 Inference

At inference, decoding is greedy with temperature 0 and a batch size of 8. We set the maximum number of new tokens to 512 for Math, 8 for the NLI label, and 64 for the QA short answer. Inputs are formatted with the LlamaFactory template (Zheng et al., 2024). Every inference was conducted using 4 NVIDIA RTX A6000 GPUs.

A.4 Discussion of artifacts

All artifacts used in this research are publicly available under licenses permitting academic use. For the datasets, XNLI and MGSM are licensed under CC BY-NC 4.0, XQuAD under CC BY-SA 4.0, and MGSM8KInstruct under the MIT License, while the CulturaX corpus follows mC4 and OSCAR licenses. For the experimental codebase, we utilized LlamaFactory, which operates under the Apache License 2.0. Regarding the models, Qwen is distributed under the Apache License 2.0, whereas Llama and Gemma are used under their custom research licenses.

Lang	MGSM	MSVAMP
bn	+0.104	+0.022
de	+0.184	-0.032
es	+0.060	-0.073
fr	+0.068	-0.060
ja	+0.004	-0.023
ru	+0.116	-0.056
sw	+0.268	+0.372
th	+0.080	-0.040
zh	-0.104	-0.141

Table 4: Language selection for Math (MGSM and MSVAMP) on Qwen3-1.7B. Each entry is the accuracy gain of the target-language adapter over the base model (i.e., instruction-tuned model) on the target-language test set. **Bold** entries indicates pairs with positive gain (i.e., the selected pairs).

B Additional results

B.1 NLI and QA

The NLI and QA paths in Figures 5 and 6 are mostly flat with no error barrier. For NLU tasks, the English and target-language adapters lie in the same low-error basin, indicating that the English adapter already generalizes to the target language. As edge cases, three pairs rise monotonically from the English endpoint to the target-language endpoint: sw on NLI and ru and th on QA. The first two appear in the main text, while ru is additionally surfaced here. This indicates that the English adapter lies outside the target language’s low-error basin, suggesting that it lacks, relative to the target-language adapter, the ability required for cross-lingual generalization.

B.2 Sensitivity analysis of layer grouping

We examine layer-group adaptation under two additional splits to verify that the layer-localized deficit in Section 5 holds even if the layer split is different. Following prior work that treats the middle layer as performing language-agnostic composition (Zhao et al., 2024; Bandarkar et al., 2025), we split layers by expanding the middle layer group. Specifically, we set two split variants: a 15 : 70 : 15 and a

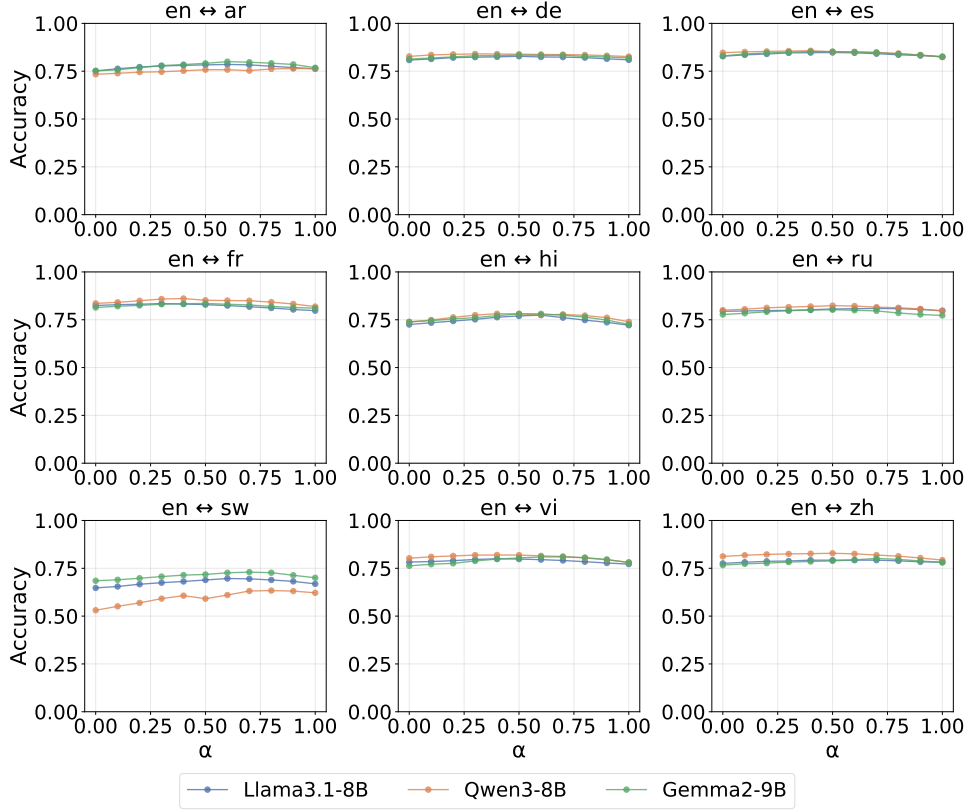


Figure 5: Linear interpolation path between English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) on NLI. Each subplot shows accuracy evaluated on XNLI in the respective target language.

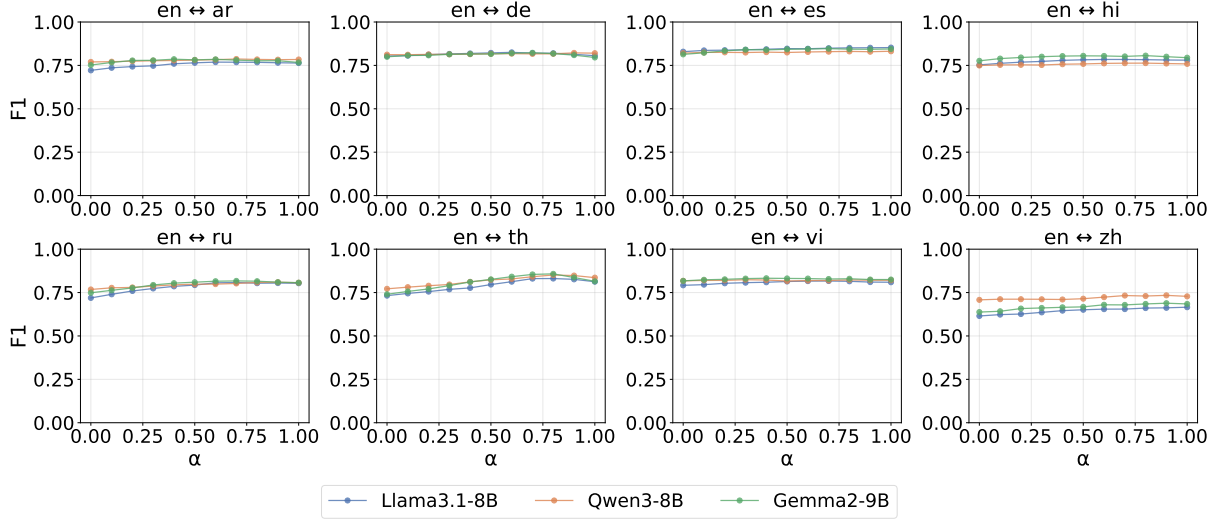


Figure 6: Linear interpolation path between English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) on QA. Each subplot shows F1 score evaluated on XQuAD in the respective target language.

20 : 60 : 20 split for the early (S_1), middle (S_2), and final (S_3) layer groups, respectively.

In Figure 7 and Figure 8, layer-group adaptation typically makes adapters linearly connected regardless of the split, consistent with the main text. Middle layer adaptation mitigates the barrier robustly across both splits, while early-layer adap-

tation becomes less effective as the middle-layer group grows larger. This suggests that within the early layer group, layers closer to the middle layers contribute more to cross-lingual generalization.

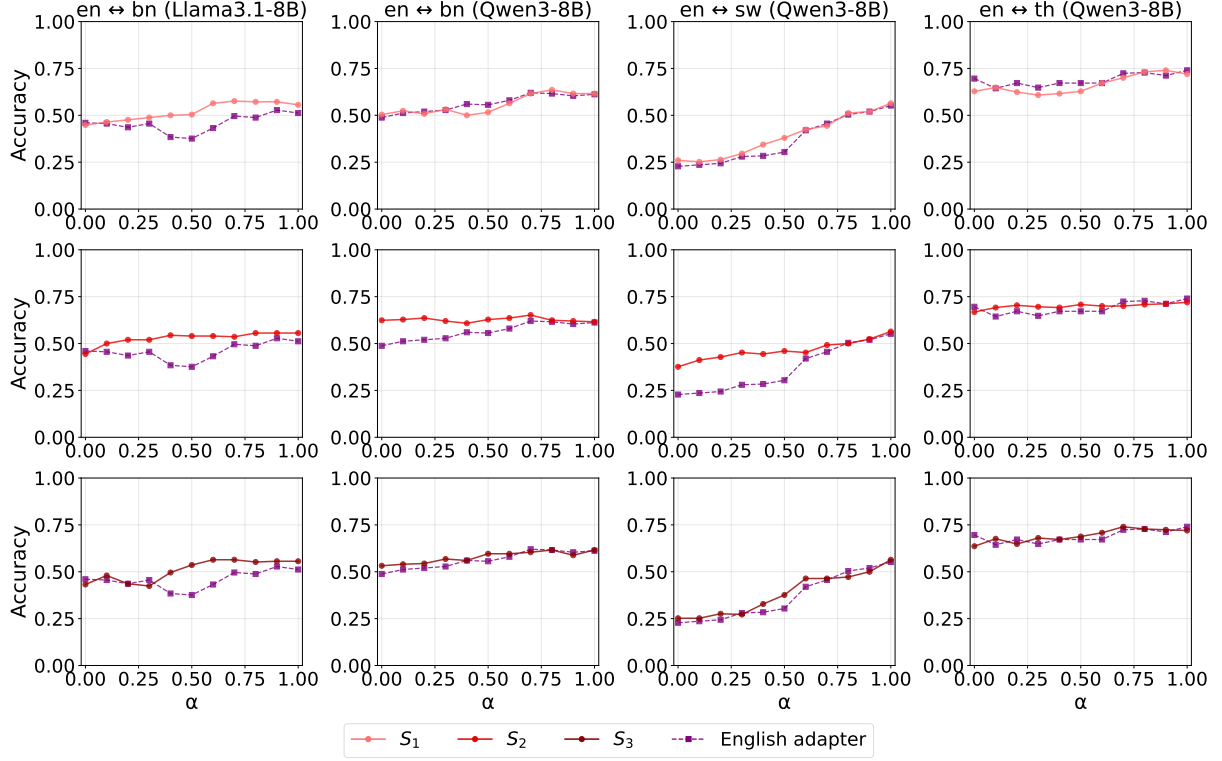


Figure 7: Linear interpolation path between ability-adapted English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) on Math. Rows correspond to early-layer (S_1), middle-layer (S_2), and final-layer (S_3) adaptation, respectively, where the layers are split in a 15:70:15 ratio. Each column corresponds to a target language, and each subplot shows accuracy evaluated on MGSM in that language.

Baseline	Avg Acc. (Avg. B)
naive	0.367 (0.029)
E	0.276 (0.014)
M	0.379 (0.023)
E+M	0.330 (0.028)

Table 5: Average accuracy and barrier height B of the routing variants of Qwen3-1.7B on Math (MGSM). Averages are computed over 8 languages. Rows denote the naive English adapter (naive) and three routing variants, namely Early only (E), Middle only (M), and Early and Middle (E+M). **Bold** denotes the highest accuracy and lowest barrier height.

B.3 Per-pair breakdown of routing variants

Table 3 reports the barrier and accuracy of every model-language pair, complementing the average results (Table 2) given in the main text. In most cases, every routing variant reduces the barrier compared to the English adapter. We note that composition routing (Middle only) lowers the barrier height to about half that of the naive English adapter on Qwen3-8B bn and Qwen3-8B th. Composition routing also yields a large accuracy improvement, from 0.488 to 0.572 on Qwen3-8B bn, and there is no destructive degradation on other pairs.

C Analysis on Qwen3-1.7B

To demonstrate that our findings generalize to smaller models and to more languages, we conduct additional analysis using Qwen3-1.7B. For NLI and QA, we use the same languages as in the main text. For Math, we use the 8 languages that satisfy the language-selection criteria for MGSM as reported in Table 4, covering Bengali (bn), German (de), Spanish (es), French (fr), Japanese (ja), Russian (ru), Swahili (sw), and Thai (th). To further verify that our findings hold on another Math dataset, we additionally use MSVAMP (Chen et al., 2024), for which we use the 2 languages (bn and sw) following the criterion. All other settings remain the same as in the main text.

Cross-lingual linear mode connectivity. We observe the same asymmetry in cross-lingual generalization across tasks. In Figure 9 and Figure 10, the English adapter and the target-language adapter are linearly connected across all languages on NLI and QA. In contrast, Figure 11 and Figure 12 show the opposite behavior, suggesting that this failure of linear connectivity in Math is not confined to a single dataset. These results indicate that the task-

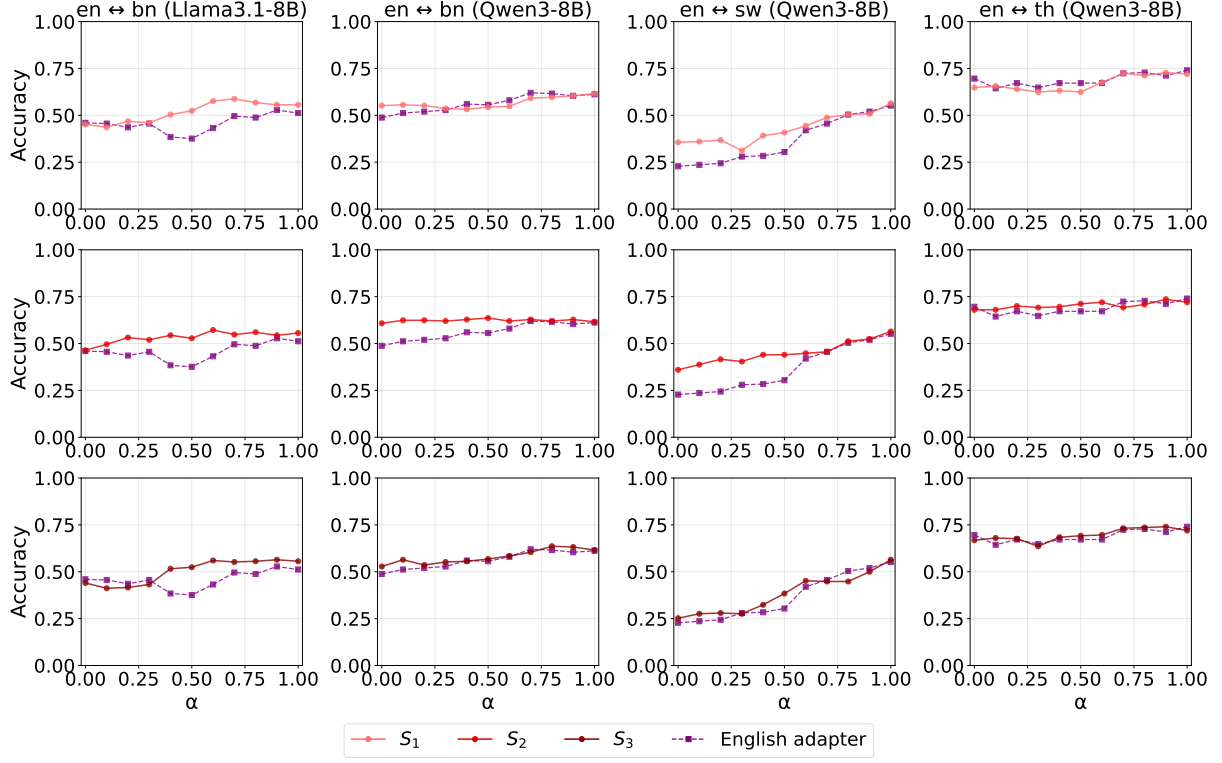


Figure 8: Linear interpolation path between ability-adapted English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) on Math. Rows correspond to early-layer (S_1), middle-layer (S_2), and final-layer (S_3) adaptation, respectively, where the layers are split in a 20:60:20 ratio. Each column corresponds to a target language, and each subplot shows accuracy evaluated on MGSM in that language.

dependent nature of cross-lingual generalization persists even as model size decreases.

Layer-group localization for cross-lingual generalization. We next investigate whether the deficit of cross-lingual generalization on Math is layer-localized in smaller and more diverse languages. Figure 13 shows that, as in Figure 3, adapting only a subset of layers is sufficient to recover LMC in Qwen3-1.7B as well. Moreover, we observe that adapting early or middle layers is more effective than adapting the final layers. These results confirm that the conclusion in Section 5, that cross-lingual generalization can be recovered by adapting early and middle layers in this setting.

Factorization without target-language labels. We further examine whether cross-lingual generalization factorizes in parameter space even for smaller models and a wide range of languages. Figure 14 and Table 5 show that language routing (Early only) and composition routing (Middle only) both lower the barrier. Composition routing, in particular, raises accuracy above the naive English adapter. This mirrors the results in Table 2 and strengthen the conclusion from Section 6 that se-

mantic composition is a key factor in cross-lingual generalization.

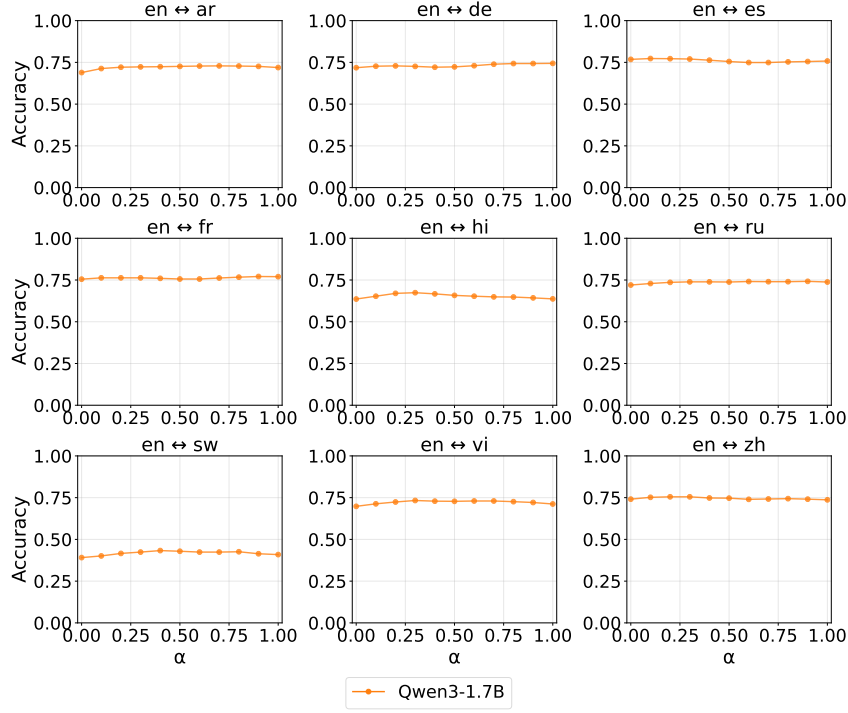


Figure 9: Linear interpolation path between English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) of Qwen3-1.7B on NLI. Each subplot shows accuracy evaluated on XNLI in the respective target language.

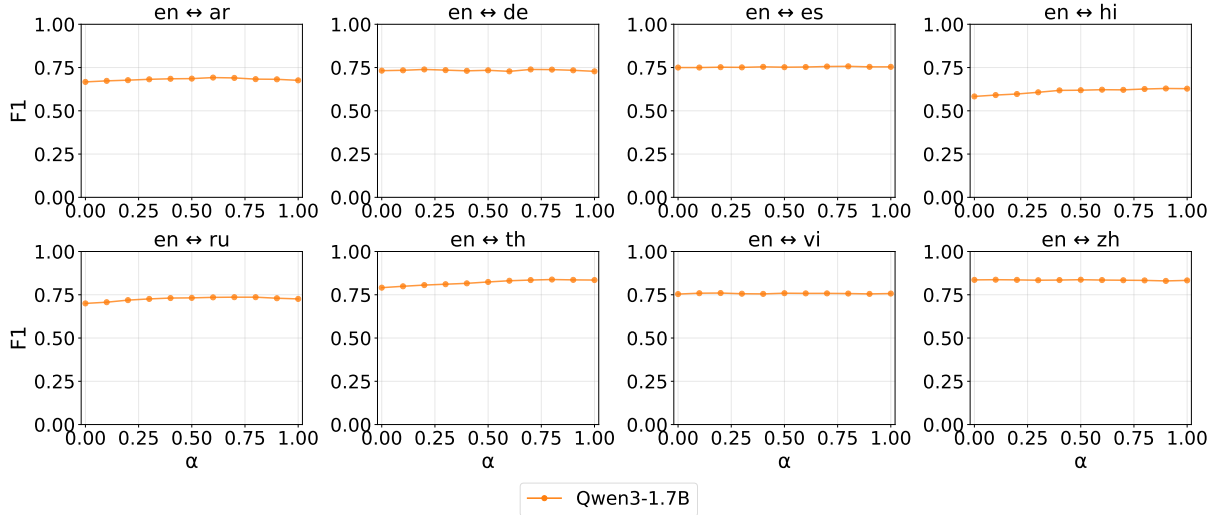


Figure 10: Linear interpolation path between English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) of Qwen3-1.7B on QA. Each subplot shows F1 score evaluated on XQuAD in the respective target language.

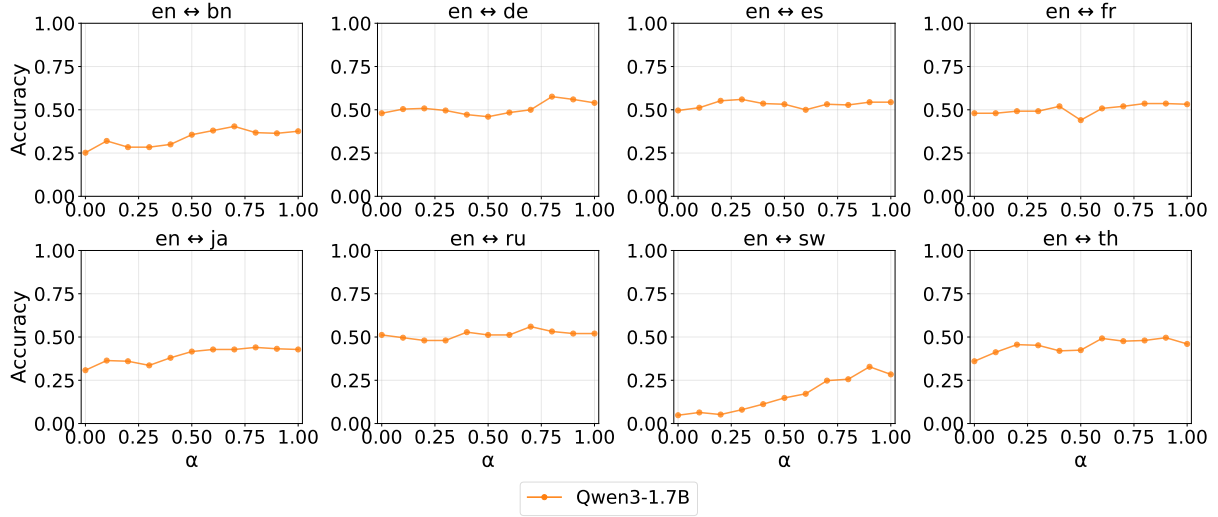


Figure 11: Linear interpolation path between English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) of Qwen3-1.7B on Math. Each subplot shows accuracy evaluated on MGSM in the respective target language.

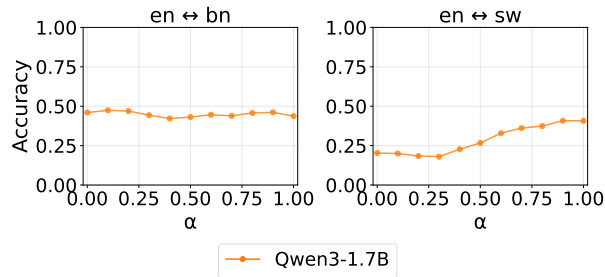


Figure 12: Linear interpolation path between English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) of Qwen3-1.7B on Math. Each subplot shows accuracy evaluated on MSVAMP in the respective target language.

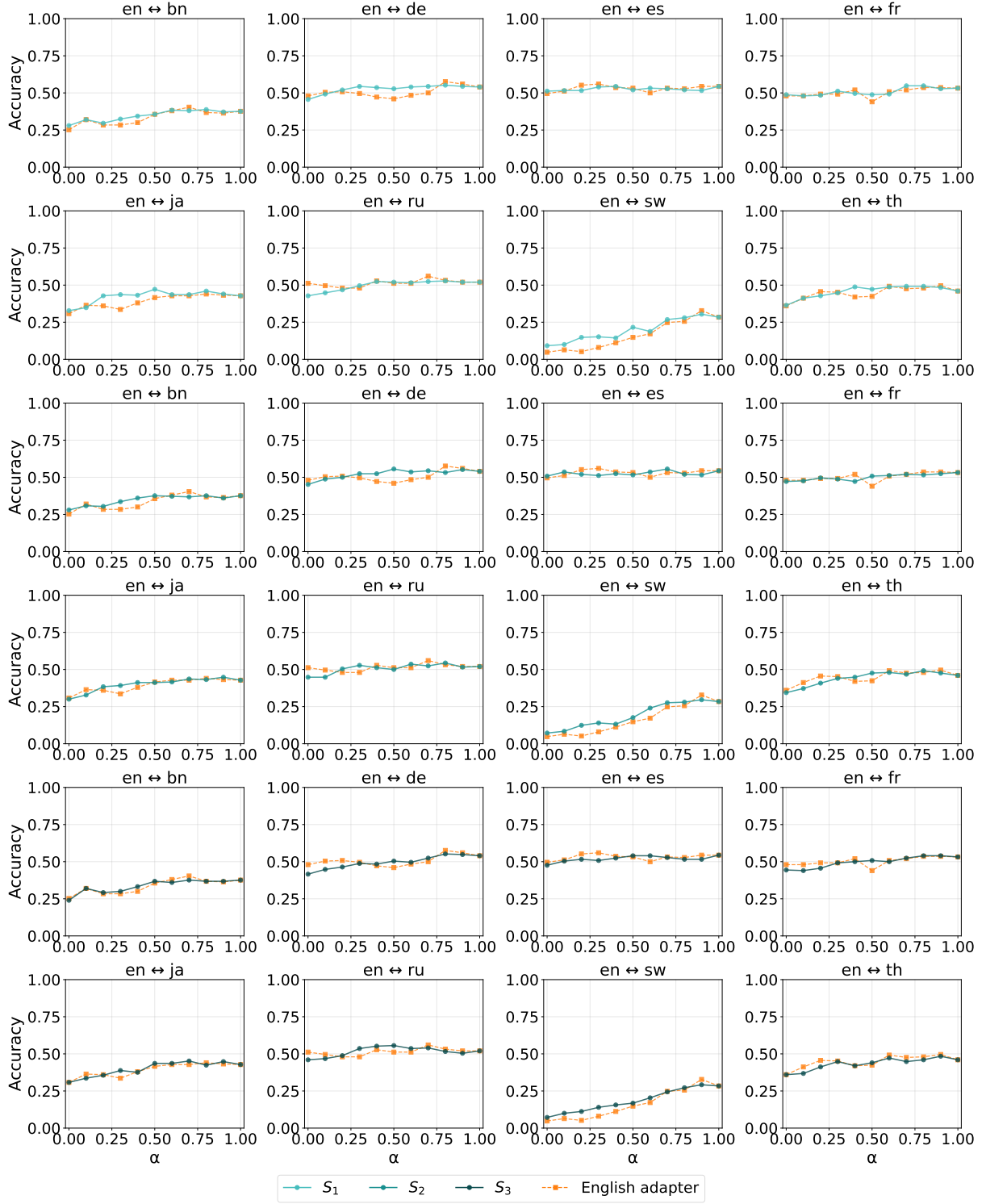


Figure 13: Linear interpolation path between ability-adapted English adapter ($\alpha = 0$) and target-language adapter ($\alpha = 1$) of Qwen3-1.7B on Math. The top (rows 1-2), middle (rows 3-4), and bottom (rows 5-6) row pairs report early (S_1), middle (S_2), and final (S_3) layer adaptation, respectively, covering all eight target languages. Each subplot shows accuracy evaluated on MGSM in that language.

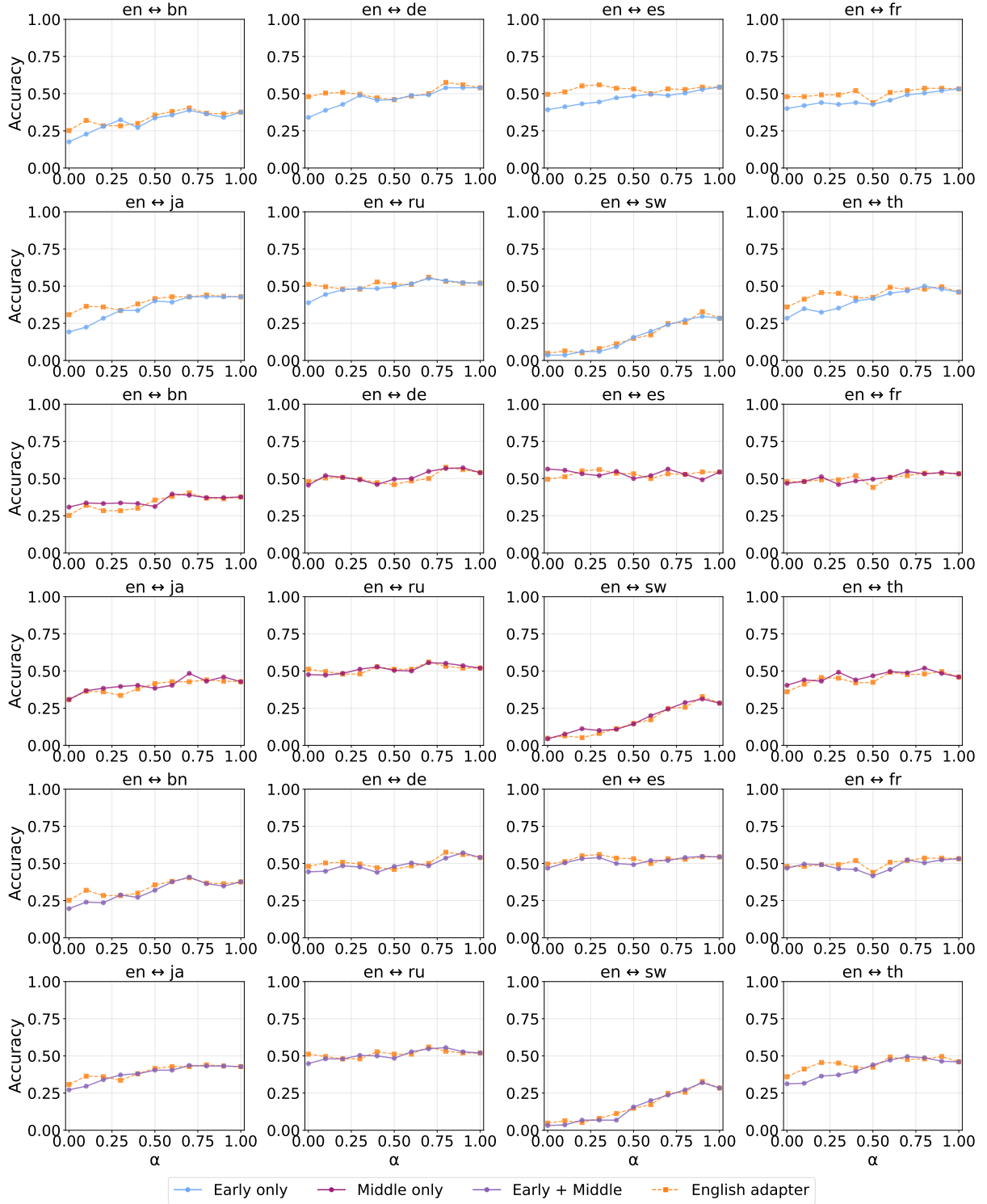


Figure 14: Linear interpolation path between routed-adaptation English adapter $\tilde{\theta}_{\text{en}}(\alpha = 0)$ and target-language adapter $\theta_{\text{tgt}}(\alpha = 1)$ on Math. The top (rows 1-2), middle (rows 3-4), and bottom (rows 5-6) row pairs report language (Early only), composition (Middle only), and joint (Early and Middle) routing, respectively, covering all eight target languages. Each subplot shows accuracy evaluated on MGSM in that language.