Rethinking the Multilingual Reasoning Gap with Layer Swap

Maxence Lasbordes[1,2] **Amélie Chatelain**[1] **Djamé Seddah**[2]
[1]LightOn, Paris [2]Inria, Paris
{maxence.lasbordes, amelie}@lighton.ai
djame.seddah@inria.fr

Abstract

Recent reasoning Large Language Models produce a chain-of-thought (CoT) predominantly in English, even when prompted in non-English languages. Prior work suggests that forcing the CoT to remain in the input language (*native reasoning*) substantially degrades performance relative to allowing the model to reason in English before answering in the input language (*English-pivoted reasoning*). However, most studies of this native reasoning gap rely on inference-time interventions or limited native-language training data. We revisit this comparison at a larger scale and under comparable supervision. We construct long multilingual reasoning datasets across six languages (English, French, German, Spanish, Chinese and Swahili); fine-tune specialists in both native and English-pivoted regimes on top of `Qwen/Qwen3-8B-Base`, and evaluate across mathematics, science, general knowledge, and code. In this setting, the average native reasoning gap shrinks to 1.9–3.5% across the five non-English languages, considerably smaller than previously reported. Weight-space analysis of the native specialists reveals aligned fine-tuning updates in the middle layers and divergence in the outer layers. This points to a largely language-agnostic reasoning core surrounded by language-specific layers. Exploiting this structure, we introduce a Layer Swap: transferring the English specialist's stronger reasoning mid-layers into each native specialist, closing most of the native reasoning gap across the five non-English languages while preserving CoT in the target language. We release all models and datasets.[1]

1 Introduction

Reasoning models such as OpenAI o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), and Qwen3 (Yang et al., 2025) rely on long chain-of-thought (CoT) to tackle complex tasks in mathematics, code, and science. In current open reasoning models, the CoT is overwhelmingly produced in English, including on non-English inputs: the model switches back to the input language only for the final response (Saji et al., 2026; Park et al., 2025), a regime we refer to as *English-pivoted reasoning* (Figure 1). This default has practical costs. English-only reasoning limits interpretability for non-English users, reduces the linguistic and cultural nuance captured by native reasoning traces, and accumulates translation-style errors that compound with task complexity (Saji et al., 2026). Constraining the CoT to remain in the input language, *native reasoning* (Figure 1), is therefore desirable, but most attempts substantially degrade accuracy on key benchmarks (Barua et al., 2025; Zhang et al., 2025). We call models trained for native reasoning in a specific non-English language *native specialists*. However, these measurements are typically obtained either through prompting alone (Saji et al., 2026; Kang et al., 2025) or through fine-tuning on small amounts of native reasoning data (Barua et al., 2025; Qi et al., 2025), using post-training budgets far below those required for meaningful native reasoning supervision. Whether the reported gap persists once native post-training is brought to a scale comparable to English-pivoted post-training remains, to our knowledge, understudied.

In this work, we revisit the comparison under strictly comparable training conditions. We construct a large multilingual reasoning dataset spanning six languages (French, German, Spanish, Swahili, Chinese, and English), with approximately 500k samples per language up to 32k tokens, and perform supervised fine-tuning (SFT) on `Qwen/Qwen3-8B-Base` with roughly 10B tokens per language in both regimes (native and English-pivoted). Across general knowledge, mathematics, code and science benchmarks, the native reason-

[1] https://huggingface.co/collections/lightonai/multilingual-reasoning

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 50265–50280
October 24-29, 2026 ©2026 Association for Computational Linguistics

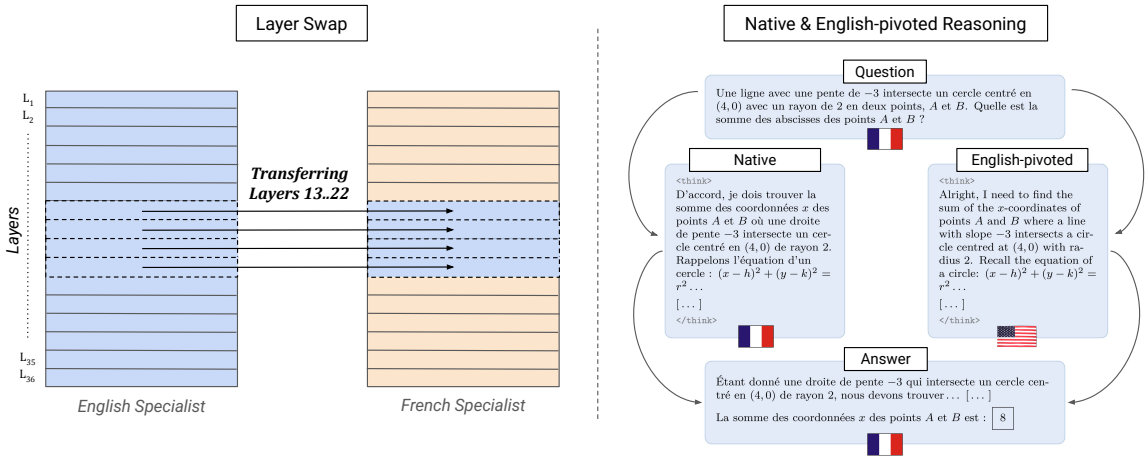

Figure 1: (*left*) **Layer Swap**: transferring a mid-stack window of the English specialist into the native specialist keeps the CoT in the input language while reducing the remaining native reasoning gap. (*right*) The two baselines compared in this work: **Native Reasoning** (CoT in the input language, here French) and **English-pivoted Reasoning** (CoT in English regardless of input), the default of open multilingual reasoning models.

ing gap shrinks to 1.9–3.5% on average across the five non-English languages (Figure 2), substantially smaller than previous reports suggest (Barua et al., 2025; Qi et al., 2025); the gap concentrates on complex reasoning benchmarks, with much smaller gaps on the other benchmarks. Weight-space analysis of the per-language SFT updates further shows that cross-language updates align tightly in the middle layers but diverge at the edges. We exploit this structure with a *Layer Swap* (Bandarkar et al., 2024) (Figure 1), a training-free method that transfers the English specialist's middle layers into each native specialist. To our knowledge, this technique has not previously been applied to long-CoT reasoning models, nor to pairs of experts that share the same reasoning skills but differ in their training language, the setting we study here. Layer Swap closes 83–89% of the gap on French and German, 60% on Swahili, 27% on Chinese, and matches the English-pivoted ceiling on Spanish, all while keeping the CoT in the target language (Figure 2).

Our contributions are: **(i)** a publicly released long-CoT reasoning corpus across six languages at 32k context, with CoTs in the target language, covering both European and non-European languages, including Swahili and Chinese; **(ii)** a large-scale, strictly controlled measurement of the native-vs-English-pivoted reasoning gap under matched SFT token budgets; **(iii)** a weight-space analysis revealing a largely language-agnostic reasoning core in the middle layers, and a training-free Layer Swap that exploits this structure to close most of the remaining gap while preserving the target-language CoT; and **(iv)** an input-language ablation showing that input understanding remains a primary bottleneck under matched native SFT.

2 Related Work

Native multilingual reasoning training Recent work has approached the native reasoning gap through both data and post-training objectives with various degrees of success. Publicly available native long-CoT corpora remain scarce (Ghosh et al., 2025): existing releases provide only a few hundred to a few thousand samples per language (Barua et al., 2025; Qi et al., 2025), while broader multilingual corpora either retain English CoT or do not target long-form reasoning. Within this limited regime, Barua et al. (2025) show that translated long-CoT supervision can effectively train non-English reasoners, motivating our translated-data pipeline at substantially larger scale across six languages.

Post-training approaches similarly reveal a substantial native reasoning gap. Under pure SFT on `Qwen/Qwen3-8B-Base`, Barua et al. (2025) find that target-language CoT underperforms English-pivoted reasoning, with AIME 24/25 gaps averaging ∼19% across nine languages (∼17% for French). Son et al. (2025) report a similar gap in Korean and mitigate it by inserting English anchor segments into the reasoning trace. Reinforcement learning-based (RL) methods offer mixed evidence: Huang et al. (2025) show that RL on non-English data can improve cross-lingual transfer, whereas *Cross-lingual Collapse* (Park et al., 2025) identifies a recurrent failure mode in which CoT drifts back to English as accuracy improves under GRPO (Shao

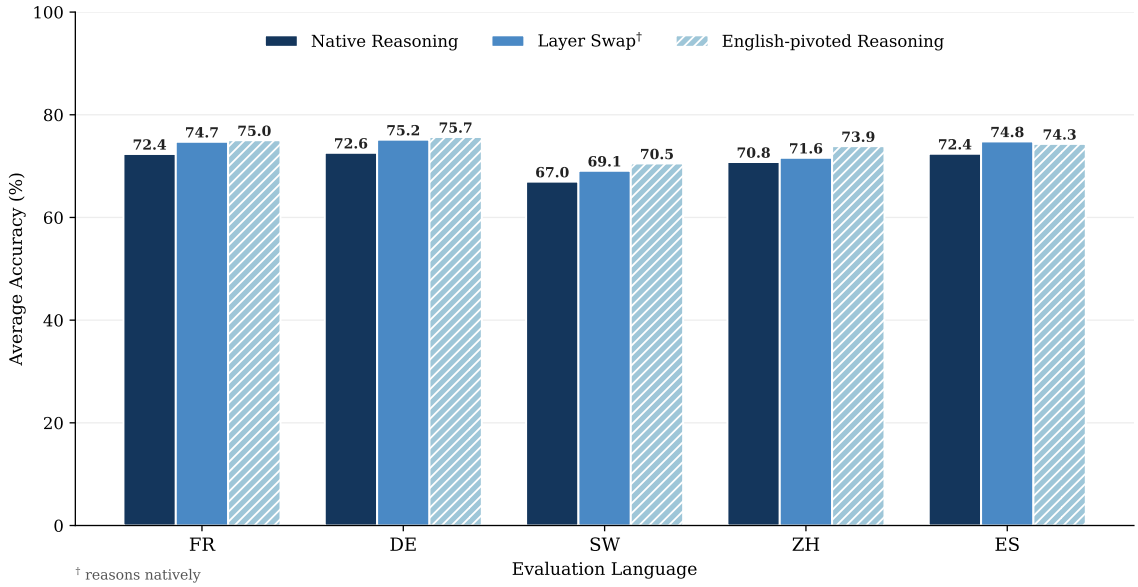

Figure 2: Average accuracy across MGSM-Rev2, Global-MMLU-Lite, GPQA-Diamond, AIME 24/25, and HumanEvalPlus in the target language, for XX ∈ {FR, DE, ES, ZH, SW}. Three settings are compared per language: **Native Reasoning** (`Qwen3-8B-xx`, CoT in XX), **Layer Swap** (`Qwen3-8B-xx-Swap`, a mid-stack window of the English specialist transferred into the native specialist, CoT in XX; see §5), and **English-Pivoted Reasoning** (`Qwen3-8B-xx-Pivot-EN`, CoT in English).

et al., 2024). Hybrid approaches combine both stages: *Think Natively* (Zhang et al., 2025) applies SFT followed by GRPO with language-consistency and cross-lingual alignment rewards, while concurrent independent work, *ReasonXL* (Gurgurov et al., 2026), combines SFT and RLVR on SmolLM3-3B across five European languages at 16k context. Like several prior works, *ReasonXL* evaluates trained native specialists against the original base model, which can make it challenging to isolate the effect of native reasoning from the effect of specialization itself. In this work, we instead compare native and English-pivoted specialists trained under identical conditions, differing only in the language used for CoT reasoning. Our experimental design enables this comparison at broader scope on an 8B model: six languages, including Chinese and Swahili; an extended 32k training context; and, for each language, matched native and English-pivoted specialists trained on similar Q&A data.

English as a latent reasoning language Another line of work argues that multilingual LLMs internally route reasoning through English-aligned representations. Using logit-lens probing on Llama-2, Wendler et al. (2024) show that intermediate states traverse an English-aligned region before resolving to the target language; Schut et al. (2025) confirm this with activation steering across several languages, finding stronger transfer from English-derived steering vectors than from native-language ones. The same bias appears behaviourally: Etxaniz et al. (2024) show that explicit self-translation into English can outperform direct non-English inference, while Saji et al. (2026) argues that English-pivoted reasoning introduces "Lost in Translation" errors that compound with task complexity. Closest to our setting, Kang et al. (2025) attributes most of the multilingual reasoning gap to input understanding rather than the reasoning process itself. We revisit this decomposition through an additional ablation that varies only the input language across our native specialists, which naturally continue reasoning in their training language without any constraint, isolating the contribution of reasoning language from input understanding.

3 Data and Benchmarks

3.1 Dataset Creation

Training open-reasoning models in non-English languages requires large native corpora with long-CoT reasoning traces, which remain scarce. We address this gap by constructing such a corpus through automatic translation from an English source, with samples up to 32,768 tokens across five target languages.

Source corpus We start from *allenai/Dolci-Think-SFT-32B* (Olmo et al., 2025), a decontaminated English post-training dataset covering mathematics, code, instruction following, science, safety, general chat, and structured data. We sample ∼500k examples uniformly, preserving the cate-

gory distribution (Table 3, Appendix).

Languages We target five languages spanning diverse typological and resource settings: French, German, and Spanish (high-resource European, Latin script), Chinese (high-resource, non-alphabetic, typologically distant), and Swahili (low-resource). Together with the original English subset, this yields six per-language corpora of ∼500k samples each (Table 4, Appendix).

Translation Motivated by prior evidence that translated long-CoT data outperforms direct native distillation (Barua et al., 2025), we translate the English corpus into the five target languages with `google/gemma-3-27b-it`, chosen for its strong multilingual coverage. Single-pass translation exhibited two failure modes: on the longest samples (input plus output up to 64K tokens), the model entered a degraded long-context regime with frequent looping; and, independently of length, on a small but consistent fraction of samples it silently dropped the reasoning trace while still translating the question and final answer, a failure that persisted under alternative `<think>` delimiters. We therefore translate each sample component-wise, splitting question, reasoning trace, and final answer into ∼2k-token chunks at sentence or paragraph boundaries, translating each chunk independently, and recomposing, at some cost in global coherence. We manually inspected a subset of translations to verify quality and reasoning-trace preservation.

Filtering We apply two filtering passes. Before translation, we remove English samples that explicitly reference translating into or answering in a specific language, since they become self-contradictory after translation. After translation, we discard (i) empty outputs, (ii) samples whose zlib (Gailly and Adler, 2012) compression ratio against the source deviates anomalously from the dataset mean, which catches degenerate or repeated outputs, (iii) samples whose translated-to-source length ratio similarly deviates, flagging truncations or over-generation, and (iv) samples whose total length exceeds the 32K-token training context, which disproportionately affects languages with less efficient tokenization (e.g. Swahili) or higher verbosity (e.g. French, Spanish). Chunk-level translation limits per-call context and keeps translation error rates low.

3.2 Benchmarks

We evaluate across four different domains:

- **Mathematics:** *MGSM-Rev2* (Peter et al., 2025), a revised multilingual version of the grade-school mathematics benchmark *GSM8K* (Shi et al., 2022) that corrects translation errors; and multilingual versions of *AIME24* and *AIME25* (Qi et al., 2025) for competition-level hard reasoning problems.
- **Science:** A multilingual version (Qi et al., 2025) of *GPQA-Diamond* (Rein et al., 2023), a PhD-level science QA benchmark.
- **Knowledge:** *Global-MMLU-Lite* (Singh et al., 2024), a curated multilingual MMLU benchmark that addresses cultural and translation biases in the original version.
- **Code:** We translate *HumanEvalPlus* (Liu et al., 2023) with `google/gemma-3-27b-it` into our five target languages.

Quality control The translated reasoning benchmarks (AIME24, AIME25, and GPQA-Diamond)[2] initially contained a small number of translation artifacts that biased evaluation against non-English languages. To mitigate this, we performed an LLM-as-a-judge review using Claude Opus 4.7, which identified and rewrote a few malformed samples in each target language. Evaluation was conducted using a forked version of `lm-eval-harness` (Gao et al., 2024); the per-language prompts are listed in Appendix A.3.

Evaluation protocol We evaluate models using a temperature of 1, as lower values caused frequent decoding loops in our models, top-p 0.95, top-k 20, and min-p 0 across all benchmarks. We report mean accuracy over multiple random seeds, using 10 sampled runs by default, except for AIME24 and AIME25, which we average over 30 runs to reduce sampling variance on these smaller test sets.

4 The Native Reasoning Gap Under Matched Supervision

4.1 Experimental Setup

We run distributed training on multiple H100 nodes with TRL (von Werra et al., 2020). At a 32K sequence length, full fine-tuning of an 8B model exceeds single-H100 memory. To mitigate the is-

[2] https://huggingface.co/collections/lightonai/multilingual-reasoning

sue, we combine DeepSpeed ZeRO-3 (Rajbhandari et al., 2020) parameter sharding with Ulysses sequence parallelism (Jacobs et al., 2023) that splits attention heads across GPUs, together with FlashAttention-3 (Shah et al., 2024) and sequence packing for efficiency. All specialists are fully fine-tuned from `Qwen/Qwen3-8B-Base` (Table 11).

4.2 Scaling Across Languages

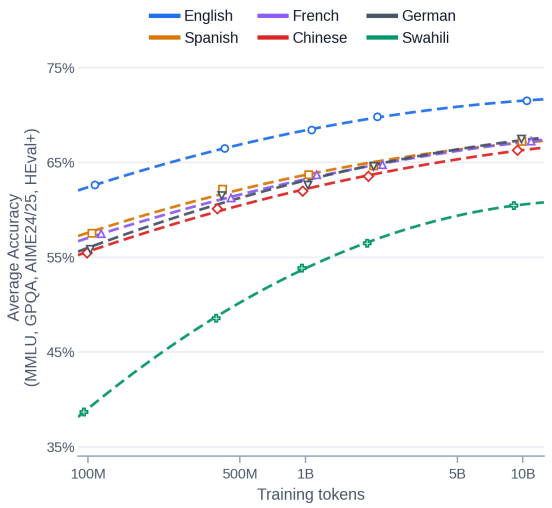

Figure 3: Scaling curves of native reasoning performance across six languages, fine-tuned from `Qwen/Qwen3-8B-Base`: average accuracy on MGSM-Rev2, Global-MMLU-Lite, GPQA-Diamond, AIME 24/25, and HumanEvalPlus in the training language, as a function of the SFT-token budget.

Before measuring the native reasoning gap under matched supervision, we verify that our translated corpus produces useful training signal in every language. We fine-tune each of the six per-language corpora at training-data budgets ranging from ∼100M to ∼10B tokens, holding all other hyperparameters fixed, and evaluate each resulting specialist on our five benchmarks in its training language. Figure 3 shows that average accuracy increases monotonically with the budget in every language. This confirms that the translated corpus carries usable training signal in every language. We adopt the ∼10B-token budget for the matched-supervision experiment that follows (Table 12, Appendix).

The curves separate into three resource tiers (Figure 3). English forms the upper envelope at every budget. The high-resource cluster of French, German, Spanish, and Chinese tracks it closely, with a gap that stays in a narrow 4.5–6% band across budgets. Swahili sits below this cluster but shows the largest absolute gains, closing its gap to the high-resource tier from roughly 18% at the ∼100M-token budget to 7% at ∼10B.

4.3 Matched-Supervision Experiments

To compare native and English-pivoted reasoning under matched supervision, we train, for each non-English language XX ∈ {FR, DE, ES, ZH, SW}, two specialists from `Qwen/Qwen3-8B-Base` on the same per-language question–answer pairs, differing only in the CoT language: a native specialist `Qwen3-8B-XX` whose reasoning trace is also in XX, and an English-pivoted specialist `Qwen3-8B-XX-Pivot-EN` whose questions and answers remain in XX but whose reasoning trace is generated in English. Both are trained with comparable per-language token budgets of 10–11B and identical hyperparameters (Table 12, Appendix). Note that sample counts differ slightly because samples exceeding the 32K-token training context are dropped, which disproportionately affects non-English versions due to less efficient tokenization (e.g. Swahili) or higher verbosity (e.g. French, Spanish). We also include the English specialist `Qwen3-8B-EN` for reference, but the controlled comparison remains `Qwen3-8B-XX` vs. `Qwen3-8B-XX-Pivot-EN`: comparing `Qwen3-8B-XX` directly to `Qwen3-8B-EN` would conflate the CoT language with the fact that `Qwen3-8B-EN` has seen no XX-language supervision.

Under this setting, the native reasoning gap is substantially smaller than prior reports suggest. On the five-benchmark average, `Qwen3-8B-XX` trails `Qwen3-8B-XX-Pivot-EN` by 2.7%, 3.1%, 1.9%, 3.1%, and 3.5% on French, German, Spanish, Chinese, and Swahili respectively (Table 1). The gap broadly tracks the resource tier: smallest on the high-resource languages (1.9–3.1%) and largest on low-resource Swahili (3.5%). Per benchmark, AIME 24/25 dominates the gap, with native specialists trailing their English-pivoted counterparts by 5–7% on French, German, Spanish, and Chinese and by 12% on Swahili; the other benchmarks contribute much smaller differences (around 1–4% on average), and HumanEvalPlus is essentially tied across languages. The lower-resource end is also where native SFT contributes the most in absolute terms: `Qwen3-8B-EN` scores only 37.96% on the Swahili five-benchmark average, and the matched native specialist nearly doubles this to 66.98%, whereas the corresponding gain over `Qwen3-8B-EN` is marginal for the high-resource Latin-script trio.

Eval Lang	Model	CoT	Avg	MGSM	G-MMLU	GPQA-D	AIME 24/25	HEval+
French	Qwen3-8B-FR	FR	72.36	92.80 ±1.71	76.45 ±1.73	53.59 ±2.85	55.67 ±3.93	83.31 ±1.56
	Qwen3-8B-FR-Swap	FR	74.74	**97.40** ±0.60	76.57 ±1.54	54.55 ±2.74	59.11 ±4.08	**86.06** ±2.06
	Qwen3-8B-FR-Pivot-EN	EN	**75.04**	94.52 ±1.28	**78.37** ±0.81	**54.65** ±2.48	**62.78** ±4.84	84.88 ±1.53
	Qwen3-8B-EN	EN	74.27	95.72 ±0.68	77.50 ±1.84	52.53 ±2.16	61.39 ±4.61	84.19 ±2.04
German	Qwen3-8B-DE	DE	72.59	93.12 ±0.94	75.15 ±1.19	55.20 ±2.30	54.56 ±3.33	84.94 ±1.12
	Qwen3-8B-DE-Swap	DE	75.15	**96.96** ±0.89	77.35 ±1.24	56.16 ±2.49	58.28 ±4.12	**87.00** ±1.24
	Qwen3-8B-DE-Pivot-EN	EN	**75.67**	93.76 ±0.93	**78.05** ±0.98	**57.68** ±2.64	**62.06** ±3.81	86.81 ±1.63
	Qwen3-8B-EN	EN	73.53	95.88 ±0.89	75.80 ±1.29	55.45 ±2.97	57.94 ±4.46	82.56 ±1.49
Spanish	Qwen3-8B-ES	ES	72.41	93.20 ±0.98	76.58 ±1.37	55.15 ±2.54	56.11 ±4.18	81.00 ±2.59
	Qwen3-8B-ES-Swap	ES	**74.80**	**97.08** ±1.05	77.10 ±0.94	55.15 ±2.87	58.50 ±4.11	**86.19** ±1.40
	Qwen3-8B-ES-Pivot-EN	EN	74.32	95.08 ±0.98	78.20 ±1.24	**56.57** ±2.33	**61.33** ±4.43	80.44 ±1.84
	Qwen3-8B-EN	EN	74.41	94.76 ±0.93	**78.55** ±1.52	54.44 ±2.14	61.06 ±4.32	83.25 ±2.02
Chinese	Qwen3-8B-ZH	ZH	70.80	88.92 ±2.13	74.85 ±1.46	50.71 ±2.29	53.89 ±4.30	85.62 ±1.59
	Qwen3-8B-ZH-Swap	ZH	71.62	88.24 ±2.22	**76.42** ±1.25	52.58 ±3.14	55.17 ±3.77	**85.69** ±1.92
	Qwen3-8B-ZH-Pivot-EN	EN	**73.89**	**94.84** ±1.80	76.15 ±1.49	**54.19** ±2.44	**59.06** ±3.91	85.19 ±0.93
	Qwen3-8B-EN	EN	66.49	76.04 ±2.55	75.00 ±1.01	47.53 ±2.97	50.00 ±4.89	83.88 ±1.93
Swahili	Qwen3-8B-SW	SW	66.98	93.16 ±1.55	61.98 ±2.23	49.39 ±1.83	47.67 ±3.35	82.69 ±1.02
	Qwen3-8B-SW-Swap	SW	69.09	**96.12** ±0.98	64.10 ±2.16	49.29 ±1.74	50.33 ±3.59	**85.62** ±2.10
	Qwen3-8B-SW-Pivot-EN	EN	**70.52**	89.68 ±0.62	**66.00** ±2.26	**52.73** ±3.31	**59.67** ±4.00	84.50 ±1.28
	Qwen3-8B-EN	EN	37.96	35.88 ±1.68	33.88 ±1.72	36.82 ±3.13	24.78 ±3.44	58.44 ±1.36

Table 1: Detailed per-language evaluation, for XX ∈ {FR, DE, ES, ZH, SW}: Qwen3-8B-XX (native reasoning in XX), Qwen3-8B-XX-Pivot-EN (English-pivoted reasoning, same XX Q&A pairs as Qwen3-8B-XX), Qwen3-8B-XX-Swap (Layer Swap from Qwen3-8B-EN into Qwen3-8B-XX; see §5), and Qwen3-8B-EN (English specialist, reference). Bold marks the best score per column within each language group. ± denotes the sample standard deviation across runs.

5 Layer Swap

5.1 Method and Layer Selection

Layer Swap (Bandarkar et al., 2024; Bandarkar and Peng, 2025) is a training-free model-composition technique (Ilharco et al., 2023). Starting from a shared base model, two *experts* are independently fine-tuned and recomposed into a single hybrid model by replacing a contiguous range of transformer layers in one expert with the corresponding layers from the other, while the remaining layers are kept unchanged. Bandarkar et al. (2024) apply this to short-form math instruction tuning, combining a math-knowledge expert trained in English with a target-language fluency expert trained on generic instructions. To our knowledge, the technique has not been applied to long-CoT reasoning models, nor to a pair of experts that share the same reasoning skills but differ in their training language, the setting we study here. The two experts are the native specialist (e.g. Qwen3-8B-FR) and the English specialist (Qwen3-8B-EN); we recompose them by transferring a contiguous range of the English specialist's transformer layers into the native specialist.

Weight-space analysis We measure how the N per-language specialists' weight updates $\Delta\theta_L^{(\mathrm{XX})} = \theta_L^{(\mathrm{XX})} - \theta_{\mathrm{base}}$ (for language XX $\in$ {EN, FR, DE, ES, ZH, SW} at layer L) agree across languages, layer by layer, using two complementary statistics; in our setup $N = 6$ (English, French, German, Spanish, Chinese, and Swahili). The first is the mean pairwise cosine of the N deltas,

$$\bar{c}_L = \binom{N}{2}^{-1} \sum_{\mathrm{XX}<\mathrm{YY}} \frac{\langle \Delta\theta_L^{(\mathrm{XX})}, \Delta\theta_L^{(\mathrm{YY})} \rangle}{\|\Delta\theta_L^{(\mathrm{XX})}\| \, \|\Delta\theta_L^{(\mathrm{YY})}\|}, \quad (1)$$

which captures local alignment between pairs of language-specific deltas. The second is the top-1 SVD variance share: stacking the N deltas as rows of a matrix $D_L \in \mathbb{R}^{N \times P_L}$ (where P_L is the number of parameters in layer L) whose singular values are $\sigma_1 \geq \cdots \geq \sigma_N$,

$$s_L = \frac{\sigma_1^2}{\sum_{k=1}^{N} \sigma_k^2}, \quad (2)$$

capturing the fraction of cross-language variance concentrated along a single direction in weight space. Both statistics are low at the two ends of the stack, rise sharply between L9 and L13, and reach a high range across L13–L22, where $\bar{c}_L \approx 0.6$ and s_L captures 65–74% of the cross-language variance (Figure 4; (a)), with per-layer delta norms remaining comparable across the stack (Appendix A, Figure 6). In this window, the six per-language SFT updates align along a shared cross-language

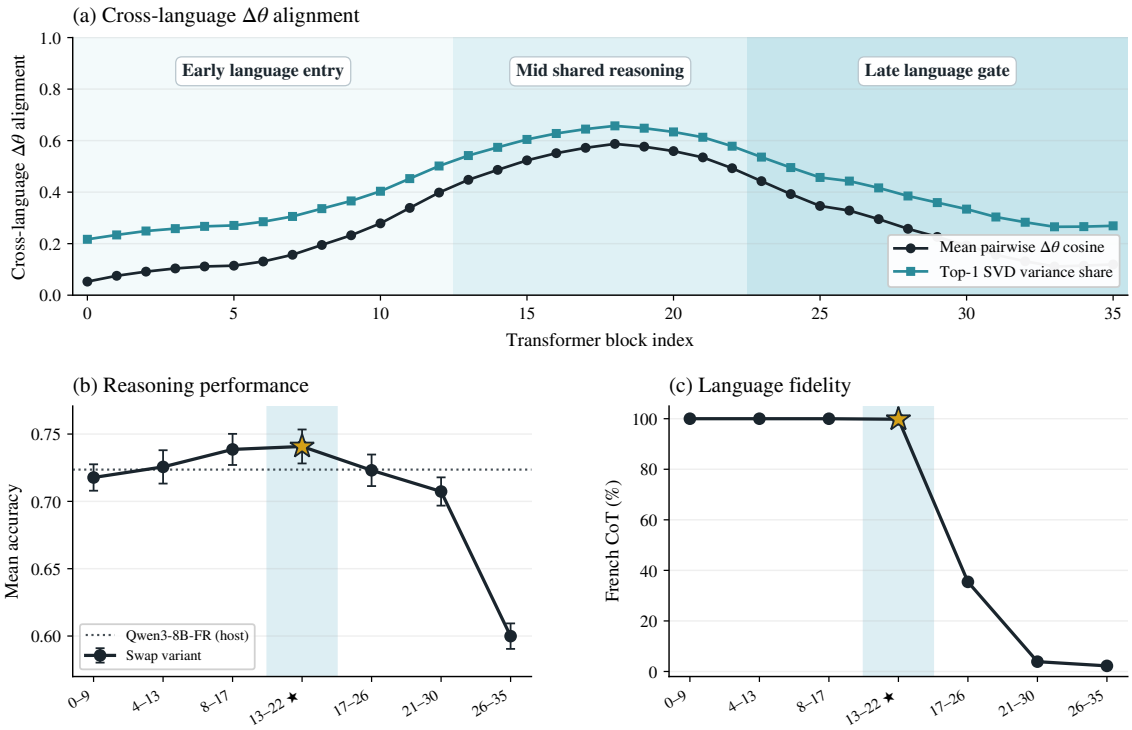

Figure 4: **(a)** Cross-language alignment of the per-language SFT updates $\Delta\theta_L^{(i)}$ as a function of transformer layer index L: mean pairwise cosine $\bar{c}_L$ and top-1 SVD variance share s_L across the six per-language specialists. **(b)** Mean accuracy on the five French benchmarks for the seven Swap variants, each transferring a contiguous 10-layer window from `Qwen3-8B-EN` into `Qwen3-8B-FR`. **(c)** Language fidelity, the fraction of generated reasoning traces classified as French, for each Swap variant.

direction (a largely language-agnostic component of the SFT update) while at the early and late ends of the stack they diverge along language-specific directions. This depth structure is consistent with prior weight-space analyses of multilingual layer specialization (Bandarkar et al., 2024; Tang et al., 2024); we verify it here for long-CoT native reasoning specialists across six languages.

Motivation This mid-stack cross-lingual alignment has a natural interpretation. Current multilingual LLMs appear to route reasoning through English-aligned intermediate representations (Wendler et al., 2024; Schut et al., 2025; Etxaniz et al., 2024; Zhao et al., 2024). In an English specialist, the middle layers may therefore develop reasoning circuits directly on top of representations already aligned with the pretraining distribution, which is not the case for a native specialist. This asymmetry may limit how effectively the native specialist's central layers specialize for reasoning during post-training, yielding a stronger reasoning core in the English specialist's middle layers. Transferring only the English middle layers into the native specialist should therefore install part of the English reasoning advantage while leaving the native edge layers, and with them the target-language CoT, intact. We test whether this swap improves reasoning while preserving target-language CoT generation.

Layer-range ablation We conduct the layer-range ablation on the French specialist as a case study. Transferring layers from the English specialist (`Qwen3-8B-EN`) into the French specialist (`Qwen3-8B-FR`), we sweep the transferred layer range from L0 to L35 in contiguous windows and report two quantities per configuration: average accuracy on the French benchmarks (Figure 4; (b)), and *language fidelity* (Figure 4; (c)), the fraction of generated reasoning traces classified as French by a FastText language identifier (Grave et al., 2018). Fidelity stays at ∼100% for any window confined to L0–L22 and collapses for windows extending further, placing the CoT-language gate around L22. Accuracy improves materially only when the English specialist's mid-stack is included; early-only swaps preserve fidelity but transfer little of its reasoning capability. The L13–L22 window satisfies both criteria, preserving native CoT generation while capturing the English specialist's reasoning advantage. Replacing the English source specialist with either the Chinese or the German specialist eliminates the improvement (Table 17, Appendix A.8), confirming that the effect originates specifically from the English specialist's middle-

layer representations.

5.2 Results

In Table 1, we compare `Qwen3-8B-XX-Swap`, obtained by transplanting a mid-stack window from the English specialist into the native specialist, against the three models introduced in §4.3. For French, German, Spanish, and Swahili, we use the L13–L22 window identified above. For Chinese, however, this window causes roughly 60% of traces to revert to English mid-generation, suggesting that the target-language gate is located slightly earlier in the stack. We therefore use L13–L20 instead, which fully restores Chinese CoT fidelity (Appendix A.9).

In particular, the swapped models retain the target reasoning language while closing most of the native reasoning gap relative to English pivoting. Measured as the fraction of the native-vs-pivoted gap closed on the five-benchmark average, Swap closes 89% of the gap on French, 83% on German, 27% on Chinese, 60% on Swahili, and matches the English-pivoted ceiling on Spanish. In absolute terms, the remaining gap to `Qwen3-8B-XX-Pivot-EN` shrinks to between 0.0% and 2.3% across languages. The gap is essentially closed for the Latin-script European trio, while the remaining differences on Chinese and Swahili reflect, respectively, typological distance from English and lower-resource status.

Across benchmarks, Layer Swap leads on HumanEvalPlus in all five languages and on MGSM-Rev2 in most of them, while the remaining gap concentrates on the more reasoning-intensive tasks (Global-MMLU-Lite, GPQA-Diamond, and AIME). On AIME 24/25 specifically, Layer Swap narrows the gap to just ∼3% on French, German, Spanish, and Chinese, with Swahili remaining the outlier at ∼9%. Language fidelity remains at ∼100% across all five Swap variants. We also test whether transplanting the English specialist's mid-layers introduces English-centric bias on culturally grounded content using *Global-PIQA* (Chang et al., 2025), a commonsense benchmark written natively by native-speaker researchers and containing many culturally specific items. Swap matches or slightly outperforms native specialists across all five languages while preserving target-language CoT (Appendix A.6). These results suggest that mid-layer transplantation does not introduce noticeable English-centric contamination.

6 The Understanding Gap

As an additional ablation enabled by our per-language specialists, we isolate the input-understanding component of the multilingual reasoning gap, a question Kang et al. (2025) address with inference-time interventions on `Qwen/Qwen3-4B` but that remains open under large-scale native SFT. We evaluate each specialist on the native and English versions of every benchmark; each model continues to reason in its training language in both conditions without any prompting or constraint, so only the input language varies.

Every non-English specialist scores higher on English inputs than on native inputs despite never seeing English in SFT (Table 2; per-benchmark numbers in Table 14, Appendix). The improvement grows sharply with the language's distance from English in the base model, a ∼4× effect on Swahili relative to the European trio. Two effects drive this. **(i)** Pretraining data imbalance: the Qwen3 base most likely saw considerably more English and Latin-script European tokens than Swahili during pretraining (the exact mixture is not disclosed), and our 10B-token SFT only retargets the reasoning policy on top of the input-side representations inherited from that pretraining. **(ii)** Intermediate representations in English-centric base models are more strongly aligned with the English representation space for high-resource Latin-script languages than for typologically distant or low-resource ones (Wendler et al., 2024; Schut et al., 2025), so the representational gap between input-side embeddings and the English-aligned reasoning mid-stack is largest precisely for Swahili.

Model	Evaluation language		Δ
	English	Native	
`Qwen3-8B-EN`	**77.00**	–	–
`Qwen3-8B-ES`	**73.72**	72.41	−1.31
`Qwen3-8B-DE`	**73.94**	72.59	−1.35
`Qwen3-8B-FR`	**73.78**	72.36	−1.42
`Qwen3-8B-ZH`	**74.54**	70.80	−3.74
`Qwen3-8B-SW`	**72.43**	66.98	−5.45

Table 2: Average accuracy of each per-language specialist, averaged over MGSM-Rev2, Global-MMLU-Lite, GPQA-Diamond, AIME 24/25, and HumanEvalPlus, on the English and native versions of each benchmark.

7 Conclusion

We revisited the native reasoning gap in multilingual LLMs under matched supervision. We trained native and English-pivoted specialists from `Qwen/Qwen3-8B-Base` on the same per-language Q&A pairs using $\sim$10B-token budgets and 32k context windows across French, German, Spanish, Chinese, and Swahili. Across five evaluation benchmarks, the performance gap shrank to 1.9–3.5% on average, substantially smaller than prior reports suggested, with the remaining differences concentrated in complex reasoning-intensive mathematics tasks. A weight-space analysis revealed a largely language-agnostic reasoning core in the mid-stack, motivating a training-free Layer Swap that transferred these English layers into each native specialist while preserving $\sim$100% language fidelity. On the five-benchmark average, the gap was fully closed on Spanish, reduced by 83–89% on French and German, and by 27–60% on Chinese and Swahili. Since English-pivoted reasoning is the default in essentially all open reasoning models, the required English specialist is available off the shelf, making the recipe directly applicable to other model families.

8 Limitations

While this work establishes a controlled study of the native reasoning gap, several limitations remain. First, our main analysis was limited to a single model family and parameter scale. As an initial cross-architecture probe, we replicated the French pipeline on `HuggingFaceTB/SmolLM3-3B-Base`, and observed a native-vs-pivoted gap and a Layer Swap behaviour consistent with what we report on `Qwen/Qwen3-8B-Base` (Appendix A.7). Extending the study to a wider range of architectures, scales, and languages would nonetheless strengthen the findings. The recipe also presupposed a multilingually pretrained base and covered only six languages, with Swahili as the sole lower-resource representative. In addition, tokenization efficiency varied across languages, meaning that an equivalent token budget may have corresponded to less effective supervision for languages such as Swahili. Our primary control was a matched token-budget comparison rather than a matched-example comparison. A stricter design would train both regimes on the intersection of examples retained in both native and English-pivoted form; we leave this paired control to future work.

Second, we did not perform an extensive hyperparameter sweep for our $\sim$10B-token SFT runs, nor did we exhaustively explore the Layer Swap window, and finer-grained sweeps over its position and width could identify a better configuration. While our training budget was substantial for native long-CoT supervision, it remained below production-scale post-training settings, where larger token budgets and multi-stage curricula may have led to different outcomes.

Finally, as is common in multilingual long-CoT post-training, our corpus is machine-translated, so the reported native-vs.-English-pivoted gap should be interpreted within this translated supervision setting. However, because translation noise primarily affects the native specialists, stronger translation models or higher-quality data would likely narrow the gap further; the reported 1.9–3.5% range should therefore be viewed as an upper bound. Our evaluation focused on mathematics, science, general knowledge, and code generation; extending it to tasks with stronger cultural or sociopragmatic components would provide a broader assessment of multilingual reasoning. Moreover, our specialist models were trained solely with SFT, and combining the Swap models with RL or preference tuning remains a natural direction for future work.

9 Ethical Considerations

In terms of broader impact, our specialists maintain CoT in the input language across six languages, including the lower-resource Swahili, making reasoning models more interpretable for non-English users and reducing reliance on English as a mandatory intermediate language. On the risk side, the released specialists inherit biases from both `Qwen/Qwen3-8B-Base` and the translation model used to construct our corpus. They have not undergone additional safety tuning, and may therefore exhibit unsafe or undesirable behaviors outside the scope of our evaluation setting.

10 Acknowledgments

We thank Adrien Cavaillès, Théo Lasnier, and Armel Randy Zebaze for helpful comments and discussions. We also gratefully acknowledge the EuroHPC Joint Undertaking for awarding us Fast Lane allocations on MareNostrum 5, hosted by the Barcelona Supercomputing Center, under applications EHPC-AIF-2025FL01-571 and EHPC-AIF-2026FL01-212, which provided the compu-

tational resources used in this work. This project is also supported by the OpenEuroLLM project, co-funded by the Digital Europe Programme under GA no. 101195233. For more information see https://openeurollm.eu.

References

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

Lucas Bandarkar and Nanyun Peng. 2025. The unreasonable effectiveness of model merging for cross-lingual transfer in llms. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 131–148.

Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. 2025. Long chain-of-thought reasoning across languages. *arXiv preprint arXiv:2508.14828*.

Tyler A Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarimayum Meerajita Sharma, and 1 others. 2025. Global piqa: Evaluating commonsense reasoning across 100+ languages and cultures. *arXiv preprint arXiv:2510.24081*.

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.

Jean-Loup Gailly and M Adler. 2012. zlib compression library.

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.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. A survey of multilingual reasoning in language models. *arXiv preprint arXiv:2502.09457*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Daniil Gurgurov, Tom Röhr, Sebastian von Rohrscheidt, Josef van Genabith, Alexander Löser, and Simon Ostermann. 2026. Reasonxl: Shifting llm reasoning language without sacrificing performance. *arXiv preprint arXiv:2604.12378*.

Shulin Huang, Yiran Ding, Junshu Pan, and Yue Zhang. 2025. Beyond english-centric training: How reinforcement learning improves cross-lingual reasoning in llms. *arXiv preprint arXiv:2509.23657*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. 2023. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim, and Gary Geunbae Lee. 2025. Why do multilingual reasoning gaps emerge in reasoning language models? *arXiv preprint arXiv:2510.27269*.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in neural information processing systems*, 36:21558–21572.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Cheonbok Park, Jeonghoon Kim, Joosung Lee, Sanghwan Bae, Jaegul Choo, and Kang Min Yoo. 2025. Cross-lingual collapse: How language-centric foundation models shape reasoning in large language models. *arXiv preprint arXiv:2506.05850*.

Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, and Markus Freitag. 2025. Mind the gap... or not? how translation errors and evaluation details skew multilingual results. *arXiv preprint arXiv:2511.05162*.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle S. Bitterman, and Arianna Bisazza. 2025. When models reason in your language: Controlling thinking trace language comes at the cost of accuracy. *Preprint*, arXiv:2505.22888.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pages 1–16. IEEE.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Alan Saji, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. 2026. The reasoning lingua franca: A double-edged sword for multilingual ai. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 329–344.

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

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Guijin Son, Donghun Yang, Hitesh Laxmichand Patel, Amit Agarwal, Hyunwoo Ko, Chanuk Lim, Srikant Panda, Minhyuk Kim, Nikunj Drolia, Dasol Choi, and 1 others. 2025. Pushing on multilingual reasoning models with language-mixed chain-of-thought. *arXiv preprint arXiv:2510.04230*.

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.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning. https://github.com/huggingface/trl.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Think natively: Unlocking multilingual reasoning with consistency-enhanced reinforcement learning. *arXiv preprint arXiv:2510.07300*.

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

A Appendix

A.1 Dataset Statistics

Table 3 reports the category mix of the English source corpus, dominated by mathematics and code with long mean trace lengths. Table 4 gives the per-language sample counts and total trained tokens after filtering using the Qwen3 tokenizer.

Category	Count	%	Mean Tok.
math	184,284	37.93	19,369
code	169,102	34.80	7,340
instr.-follow.	69,543	14.31	2,310
science	25,290	5.21	9,089
safety	19,351	3.98	739
general-chat	17,262	3.55	3,076
struct.-data	1,041	0.21	2,228

Table 3: Category distribution and mean token length of the English subset of *Dolci-Think-SFT-32B*.

Language	Samples	Tokens	
		Mean	Total
English	485,873	10,848	5.27B
Chinese	484,078	10,890	5.27B
Spanish	479,748	13,400	6.43B
French	483,588	13,702	6.63B
German	480,529	13,846	6.65B
Swahili	473,582	16,208	7.68B

Table 4: Per-language statistics of the native reasoning SFT dataset: number of samples, mean sample length, and total token count.

A.2 Translation

Table 5 gives the chunk-level translation prompt applied independently to question, reasoning trace, and final answer at ∼2k-token boundaries. We perform translation with `google/gemma-3-27b-it` at a temperature of 0.15. For reference, translating 1B tokens with `google/gemma-3-27b-it` required approximately 750 H100 GPU-hours in our setup.

A.3 Evaluation Prompts

We list the language-specific instruction templates passed to `lm-eval-harness` for each benchmark family. Prompts were translated and adapted from the English originals while preserving the answer-extraction format used by the harness. Table 6 covers AIME 24/25, Table 7 MGSM-Rev2, Table 8 Global-MMLU-Lite, Table 9 GPQA-Diamond, and Table 10 HumanEvalPlus.

```
You are a translation engine. Your ONLY task
is to translate the text inside <source>
into <LANGUAGE>.

HARD RULES (must follow):
- Output ONLY the translation, and nothing
else.
- Do NOT answer the message or continue the
conversation.
- Do NOT add advice, role labels, extra
lines, or formatting.
- Translate EVERYTHING inside <source>
completely.
- Use informal address unless formality or
plurality is clearly required.
- Preserve meaning, intent, tone, and
punctuation as naturally as possible.

SPECIAL CASES:
- For code, do NOT translate identifiers.
Translate only comments and human-language
strings, preserving whitespace and
indentation.
- For units, convert to metric while keeping
calculations consistent.

INPUT:
<source>
{message_text}
</source>

OUTPUT: <LANGUAGE> translation only.
```

Table 5: Prompt template used for our translation pipeline.

Lang	Suffix appended after `{{problem}}\n`
en	Please reason step by step, and put your final answer within `\boxed{}`.
de	Bitte denke Schritt für Schritt nach und gib deine endgültige Antwort in `\boxed{}` an.
fr	Veuillez raisonner étape par étape, et mettre votre réponse finale dans `\boxed{}`.
es	Por favor, razona paso a paso y pon tu respuesta final en `\boxed{}`.
zh	请一步步推理，并将最终答案放在 `\boxed{}` 内。
sw	Tafadhali fikiri hatua kwa hatua, na uweke jibu lako la mwisho ndani ya `\boxed{}`.

Table 6: Prompt template used to evaluate the multilingual AIME24 and AIME25 benchmarks in `lm-eval-harness`.

A.4 Training Settings

Table 11 lists the SFT hyperparameters used to train every per-language specialist, held identical across the native and English-pivoted regimes.

A.5 Detailed Training Results

Table 12 reports the unique sample count and total training-token count for each specialist. Table 13 expands the scaling curves of Figure 3 with per-budget mean accuracies in every language. Table 14 extends the cross-lingual evaluation of Table 2 with per-benchmark scores on both the English and native versions of every benchmark.

Lang	Prefix preceding `\n{{question}}`
en	Solve this math problem. Give the reasoning steps before giving the final answer on the last line by itself in the format of "Answer:". Do not add anything other than the integer answer after "Answer:".
de	Löse dieses Mathematikproblem. Gib die Schritte zur Begründung an, bevor du die endgültige Antwort in der letzten Zeile alleine im Format "Antwort:" gibst. Füge nichts anderes als die ganzzahlige Antwort nach "Antwort:" hinzu.
fr	Résolvez ce problème de mathématiques. Donnez les étapes de raisonnement avant de fournir la réponse finale sur la dernière ligne elle-même dans le format de "Réponse:". N'ajoutez rien d'autre que le nombre de la réponse après "Réponse:".
es	Resuelve este problema matemático. Proporciona los pasos de razonamiento antes de dar la respuesta final en la última línea por sí misma en el formato de "Respuesta:". No añadas nada más que la respuesta entera después de "Respuesta:".
zh	解决这个数学问题。在最后一行给出答案前，请提供推理步骤。最后一行应该以"答案："的形式独立给出答案。在"答案："后不要添加除整数答案之外的任何内容。
sw	Suluhisha tatizo hili la hesabu. Toa hatua za mantiki kabla ya kutoa jibu la mwisho kwenye mstari wa mwisho peke yake katika muundo wa "Jibu:". Usiongeze chochote kingine isipokuwa jibu la integer baada ya "Jibu:".

Table 7: Prompt template used to evaluate the MGSM-Rev2 benchmark in `lm-eval-harness`.

Lang	Prefix preceding the question and four options
en	Answer the following multiple choice question. On the last line of your response, place your final answer letter within `\boxed{}` (e.g. `\boxed{A}`), ensuring only the letter is inside. Think step by step before answering.
de	Beantworten Sie die folgende Multiple-Choice-Frage. Setzen Sie in der letzten Zeile Ihrer Antwort den Buchstaben Ihrer endgültigen Antwort in `\boxed{}` (z.B. `\boxed{A}`). Denken Sie Schritt für Schritt, bevor Sie antworten.
fr	Répondez à la question suivante à choix multiples. Sur la dernière ligne de votre réponse, placez la lettre de votre réponse finale dans `\boxed{}` (par exemple, `\boxed{A}`). Réfléchissez étape par étape avant de répondre.
es	Responda la siguiente pregunta de opción múltiple. En la última línea de su respuesta, coloque la letra de su respuesta final dentro de `\boxed{}` (por ejemplo, `\boxed{A}`). Piense paso a paso antes de responder.
zh	回答以下多项选择题。在回答的最后一行，将你的最终答案字母放在 `\boxed{}` 中（例如 `\boxed{A}`），确保框内只有字母。请在回答前逐步思考。
sw	Jibu swali lifuatalo la chaguo nyingi. Kwenye mstari wa mwisho wa jibu lako, weka herufi ya jibu lako la mwisho ndani ya `\boxed{}` (kwa mfano, `\boxed{A}`), hakikisha herufi pekee iko ndani. Fikiria hatua kwa hatua kabla ya kujibu.

Table 8: Prompt prefix used to evaluate the Global-MMLU benchmark in `lm-eval-harness`; the question and four options A/B/C/D follow on subsequent lines.

Lang	Suffix appended after `{{problem}}\n\n`
en	Think step by step before answering.
de	Denken Sie Schritt für Schritt nach, bevor Sie antworten.
fr	Réfléchissez étape par étape avant de répondre.
es	Piensa paso a paso antes de responder.
zh	请一步步思考后再回答。
sw	Fikiria hatua kwa hatua kabla ya kujibu.

Table 9: Prompt template used to evaluate the GPQA-Diamond benchmark in `lm-eval-harness`. The dataset's `problem` field already contains the question and four options A/B/C/D.

A.6 Global-PIQA Cultural Evaluation

As a check that Layer Swap does not introduce English-centric bias on culturally grounded content, we evaluate on *Global-PIQA* (Chang et al., 2025), a commonsense benchmark natively authored by over 350 native-speaker researchers, with more than half of its items referencing local customs and culture. Table 15 shows that Swap matches or exceeds the native specialist across all five languages, indicating no English-centric contamination from the transplanted mid-layers.

A.7 SmolLM3-3B French Replication

The same pattern replicates on `SmolLM3-3B` (Table 16): Pivot-EN yields the strongest French average, and Layer Swap closes most of the gap while leading on MGSM by preserving the base model's French arithmetic reasoning.

A.8 Layer-Swap Source Language

Table 17 verifies that the Layer Swap gain originates specifically from the English specialist's mid-layers: replacing `Qwen3-8B-EN` with the Chinese specialist `Qwen3-8B-ZH` as the swap source erases the improvement on French.

A.9 Chinese Layer-Swap Window

Figure 5 repeats the layer-range ablation of Figure 4 on `Qwen3-8B-ZH` and motivates the L13–L20 window adopted for `Qwen3-8B-ZH-Swap` since the L13–L22 window leaks ∼60% of Chinese traces back into English.

Lang	doc_to_text ({{ prompt }} is the function signature + docstring)
en	Write a solution to the following problem and make sure that it passes the tests:\n```python\n{{ prompt }}\n```\n Please reason step by step, and put your final answer within a ```python code layer.\n
de	Schreiben Sie eine Lösung für das folgende Problem und stellen Sie sicher, dass sie die Tests besteht:\n```python\n{{ prompt }}\n```\n Bitte denken Sie Schritt für Schritt und geben Sie Ihre endgültige Antwort in einem ```python Codelayer an.\n
fr	Écrivez une solution au problème suivant et assurez-vous qu'elle passe les tests :\n```python\n{{ prompt }}\n```\n Veuillez raisonner étape par étape, et mettre votre réponse finale dans un bloc de code ```python.\n
es	Escribe una solución al siguiente problema y asegúrate de que pase las pruebas:\n```python\n{{ prompt }}\n```\n Por favor, razona paso a paso y coloca tu respuesta final dentro de un bloque de código ```python.\n
zh	编写以下问题的解决方案，并确保它通过测试：\n```python\n{{ prompt }}\n```\n请逐步推理，并将最终答案放在 ```python 代码块中。\n
sw	Andika suluhisho la tatizo lifuatalo na uhakikishe kuwa linapita majaribio:\n```python\n{{ prompt }}\n```\n Tafadhali fikiria hatua kwa hatua, na uweke jibu lako la mwisho ndani ya layer ya msimbo ```python.\n

Table 10: Prompt template used to evaluate the HumanEvalPlus benchmark in `lm-eval-harness`.

Hyperparameter	Setting
Base model	Qwen3-8B-Base
Training parallelism	ZeRO-3 & Ulysses SP
Attention backend	FlashAttention 3
Maximum sequence length	32,768 tokens
Global batch size	∼1M tokens
Optimizer	AdamW
Learning rate	5×10^{-5}
Warmup ratio	0.1
Weight decay	0.01
Training epochs	2
Numerical precision	bf16
Gradient checkpointing	Enabled
Sequence packing	Enabled
Loss masking	Assistant-only
Liger kernels	Enabled

Table 11: SFT hyperparameters, identical across all per-language specialists and across the native and English-pivoted regimes.

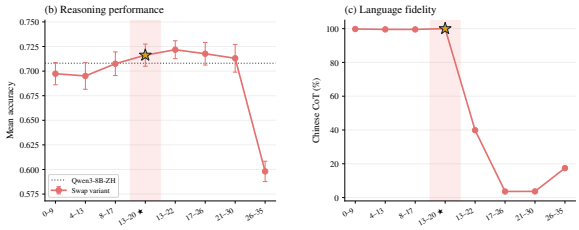

Figure 5: Same ablation as Figure 4 (b,c) applied to `Qwen3-8B-ZH`: mean accuracy across the five Chinese benchmarks and Chinese language fidelity, as a function of the transferred layer window.

A.10 Per-Layer Update Magnitudes

Figure 6 reports the per-layer L2 norm of the per-language SFT updates. Norms remain comparable across the stack, confirming that the mid-stack agreement of Figure 4 (a) reflects directional alignment rather than vanishing updates.

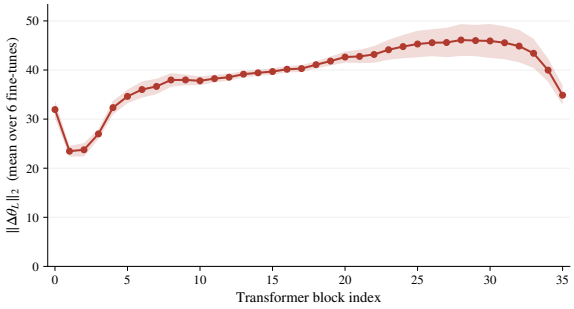

Figure 6: Per-layer L2 norm of the language-specific SFT update $\Delta\theta_L^{(\mathrm{XX})} = \theta_L^{(\mathrm{XX})} - \theta_{\mathrm{base}}$ across the six per-language specialists $\mathrm{XX} \in \{\mathrm{EN}, \mathrm{FR}, \mathrm{DE}, \mathrm{ES}, \mathrm{ZH}, \mathrm{SW}\}$; mean with $\pm$ std.

Lang	Model	CoT	Unique Samples	Trained Tokens
English	Qwen3-8B-EN	EN	480,625	10.40B
French	Qwen3-8B-FR	FR	447,650	10.92B
	Qwen3-8B-FR-Pivot-EN	EN	478,255	10.53B
German	Qwen3-8B-DE	DE	443,100	10.85B
	Qwen3-8B-DE-Pivot-EN	EN	475,205	10.54B
Spanish	Qwen3-8B-ES	ES	448,984	10.90B
	Qwen3-8B-ES-Pivot-EN	EN	474,463	10.48B
Chinese	Qwen3-8B-ZH	ZH	478,622	10.39B
	Qwen3-8B-ZH-Pivot-EN	EN	478,811	10.39B
Swahili	Qwen3-8B-SW	SW	400,844	10.03B
	Qwen3-8B-SW-Pivot-EN	EN	468,213	10.49B

Table 12: Training-data composition per specialist: number of unique samples and total trained tokens, after filtering and the 32K-token context cap.

Language	∼100M tokens	∼400M tokens	∼1B tokens	∼2B tokens	∼10B tokens
English	62.62 ±1.31	66.47 ±1.52	68.42 ±1.54	69.83 ±1.23	**71.51** ±1.31
French	57.48 ±1.55	61.24 ±1.40	63.70 ±1.47	64.74 ±1.40	67.26 ±1.35
German	55.83 ±1.14	61.51 ±1.51	62.59 ±1.55	64.59 ±1.38	67.46 ±1.09
Spanish	57.52 ±1.29	62.16 ±1.30	63.70 ±1.53	64.61 ±1.28	67.21 ±1.43
Chinese	55.46 ±1.43	60.09 ±1.36	61.95 ±1.50	63.54 ±1.36	66.27 ±1.33
Swahili	38.67 ±1.14	48.56 ±1.52	53.82 ±1.27	56.47 ±1.42	60.43 ±1.13

Table 13: Mean accuracy of the six per-language specialists at each SFT-token budget, averaged across MGSM-Rev2, Global-MMLU-Lite, GPQA-Diamond, AIME 24/25, and HumanEvalPlus in the training language. Token counts are approximate. ± denotes the sample standard deviation across runs.

Model	Eval Lang	CoT	Avg	MGSM	G-MMLU	GPQA-D	AIME 24/25	HEval+
Qwen3-8B-EN	EN	EN	77.00	98.96 ±0.47	81.72 ±1.30	55.66 ±2.05	62.89 ±4.22	85.75 ±1.88
Qwen3-8B-FR	FR	FR	72.36	92.80 ±1.71	76.45 ±1.73	53.59 ±2.85	55.67 ±3.93	83.31 ±1.56
	EN	FR	**73.78**	**94.20** ±1.53	**79.30** ±1.15	**55.25** ±2.58	**56.61** ±4.02	**83.56** ±2.06
Qwen3-8B-DE	DE	DE	72.59	93.12 ±0.94	75.15 ±1.19	**55.20** ±2.30	54.56 ±3.33	84.94 ±1.12
	EN	DE	**73.94**	**94.76** ±1.12	**79.83** ±1.22	54.04 ±2.14	**55.94** ±3.91	**85.13** ±1.79
Qwen3-8B-ES	ES	ES	72.41	93.20 ±0.98	76.58 ±1.37	**55.15** ±2.54	56.11 ±4.18	81.00 ±2.59
	EN	ES	**73.72**	**94.48** ±1.45	**79.90** ±1.39	**55.15** ±2.28	**56.67** ±3.47	**82.38** ±1.21
Qwen3-8B-ZH	ZH	ZH	70.80	88.92 ±2.13	74.85 ±1.46	50.71 ±2.29	53.89 ±4.30	**85.62** ±1.59
	EN	ZH	**74.54**	**95.96** ±1.34	**79.45** ±1.39	**54.14** ±2.14	**58.61** ±3.69	84.56 ±0.84
Qwen3-8B-SW	SW	SW	66.98	93.16 ±1.55	61.98 ±2.23	49.39 ±1.83	47.67 ±3.35	82.69 ±1.02
	EN	SW	**72.43**	**95.04** ±1.00	**78.98** ±1.30	**51.87** ±2.08	**51.56** ±3.81	**84.69** ±1.45

Table 14: Detailed evaluation of each per-language specialist on the English and native versions of MGSM-Rev2, Global-MMLU-Lite, GPQA-Diamond, AIME 24/25, and HumanEvalPlus. Bold marks the best score per column within each model. ± denotes the sample standard deviation across runs.

Variant	French	German	Spanish	Chinese	Swahili
Qwen3-8B-XX	**77.6**	74.2	83.0	**75.0**	80.8
Qwen3-8B-XX-Pivot-EN	77.5	76.0	**85.1**	74.2	81.5
Qwen3-8B-XX-Swap	76.9	**76.3**	84.7	74.8	**83.1**

Table 15: Global-PIQA accuracy for each per-language specialist and its Pivot-EN and Layer-Swap variants, averaged over 10 runs. Each column reports accuracy in the language named at the top. Bold marks the best variant per language.

Model	Avg	MGSM	G-MMLU	GPQA-D	AIME 24/25	HEval+
SmolLM3-3B-FR	53.6	84.4	56.0	30.9	27.2	69.4
SmolLM3-3B-FR-Pivot-EN	**55.5**	82.8	**59.7**	**32.5**	**31.7**	**70.9**
SmolLM3-3B-FR-Swap	55.0	**89.0**	58.9	29.2	28.0	70.0

Table 16: Comparison of SmolLM3-3B-FR variants across reasoning, knowledge, and code benchmarks. Bold marks the best score per column.

Model	Source	Avg	MGSM	G-MMLU	GPQA-D	AIME 24/25	HEval+
Qwen3-8B-FR-Swap	EN	**74.74**	**97.40** ± 0.60	**76.57** ± 1.54	**54.55** ± 2.74	**59.11** ± 4.08	**86.06** ± 2.06
–	ZH	72.34	94.08 ± 1.25	76.08 ± 1.77	53.33 ± 2.37	54.44 ± 3.46	83.75 ± 1.98
–	DE	72.12	92.24 ± 1.25	76.08 ± 1.36	52.78 ± 2.47	56.11 ± 4.09	83.38 ± 2.13

Table 17: Layer-Swap source-language ablation on French: layers L13–L22 transferred into Qwen3-8B-FR from either the English specialist Qwen3-8B-EN (source EN), the Chinese specialist Qwen3-8B-ZH (source ZH), or the German specialist Qwen3-8B-DE (source DE). Bold marks the best score per column. $\pm$ denotes the sample standard deviation across runs.