

Marco-MoE: Open Multilingual Mixture-of-Expert Language Models with Efficient Upcycling

Fan Jiang^{1,2*}, Yu Zhao¹, Chenyang Lyu¹, Tianqi Shi¹, Yichao Du^{3,1}, Feihu Jiang¹, Longyue Wang¹, Weihua Luo¹

¹Alibaba Group, ²The University of Melbourne, ³Wuhan University
fan.jiang1@student.unimelb.edu.au, wanglongyue.wly@alibaba-inc.com

Abstract

We present Marco-MoE, a suite of fully open multilingual sparse Mixture-of-Experts (MoE) models. Marco-MoE features a highly sparse design in which only 5% of the total parameters are activated per input token. This extreme sparsity, combined with upcycling from dense models, enables efficient pre-training on 5T tokens. Our models surpass similarly-sized competitors on English and multilingual benchmarks, achieving a best-in-class performance-to-compute ratio. Our analysis reveals that Marco-MoE learns structured expert activation patterns shared across related languages, while maintaining highly specialized utilization for linguistically isolated ones. We further show that Marco-MoE allows for scalable language expansion without the interference typical of dense models. Finally, we establish a robust post-training recipe to adapt Marco-MoE into Marco-MoE-Instruct, which consistently surpasses state-of-the-art open-weight instruct models of comparable scale. To support the community, we disclose our full training datasets, recipes, and model weights.

 [ATH-MaaS/Marco-MoE](#)  [ATH-MaaS/Marco-LLM](#)

1 Introduction

The rapid advancement of Large Language Models (LLMs) has transformed natural language understanding, yet achieving a balance between extensive linguistic coverage and high per-language proficiency remains a critical hurdle, often referred to as the "curse of multilinguality" (Conneau et al., 2020). This challenge stems from the fact that expanding a model's language coverage within a fixed parameter budget often degrades performance on individual languages due to capacity bottlenecks and cross-lingual interference. While leading models such as Qwen3 (Yang et al., 2025) and Gemma3 (Team et al., 2025) attempt to bypass these constraints through massive-scale pre-training on up to 36T tokens or sophisticated strong-to-weak distillation (Comanici et al., 2025), such strategies remain computationally prohibitive for most pre-training scenarios. Meanwhile, existing small-scale multilingual models, such as Tiny-Aya (Salamanca et al., 2026), typically use dense architectures that struggle to reconcile broad linguistic breadth with deep task proficiency, as they lack the inherent architectural flexibility to scale effectively across diverse language families without compromising performance.

Mixture-of-Experts (MoE) architectures (Shazeer et al., 2017) offer a promising solution to this capacity bottleneck by leveraging conditional computation, which enhances model capacity without increasing the number of activated parameters. However, the computational overhead of training large-scale MoE models from scratch has spurred the emergence of MoE Upcycling (Komatsuzaki et al., 2023), a paradigm that initializes sparse models using pre-trained dense checkpoints. Yet current approaches predominantly rely on coarse-grained expert replication, in which entire feed-forward networks (FFNs) of a dense transformer are replicated. This monolithic duplication deviates from existing wisdom that uses fine-grained

*Work done while at Alibaba Group; Now at Amazon.

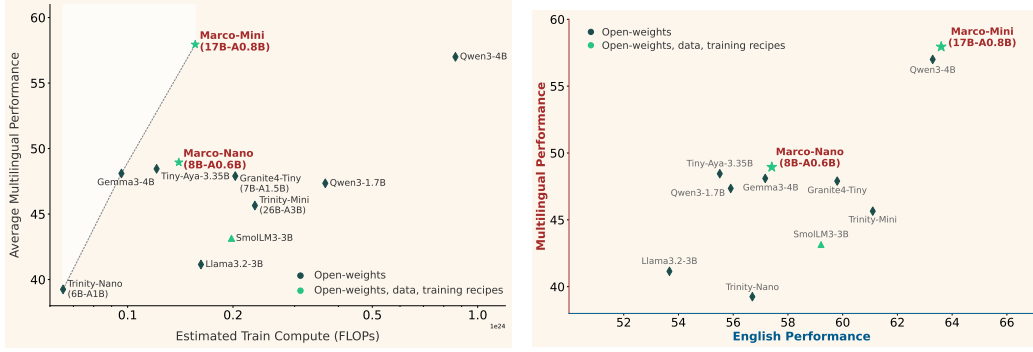


Figure 1: Comparison of performance and efficiency. Our Marco-MoE models demonstrate a superior performance-to-compute ratio in (a) and set the state-of-the-art for simultaneous proficiency in both English and multilingual capabilities in (b).

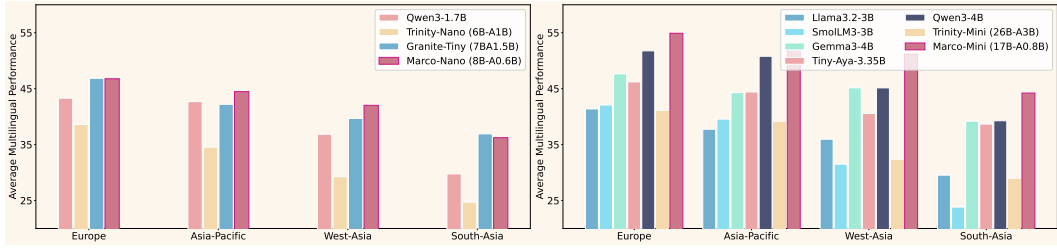


Figure 2: Multilingual benchmark performance across geographic regions.

experts (Dai et al., 2024), thereby impeding the emergence of specialized representations necessary for nuanced multilingual reasoning.

To bridge this gap, we introduce Marco-MoE, a family of compact, highly sparse multilingual MoE models designed to break these traditional trade-offs in multilinguality. Our work distinguishes itself from previous efforts in three fundamental ways:

1. **First Sparse Multilingual Upcycling:** To the best of our knowledge, this is the first work to leverage the MoE upcycling paradigm specifically to optimize multilingual performance in compact model sizes. By repurposing pre-trained dense representations into a fine-grained MoE framework, we significantly reduce computational overhead while enhancing model capacity.
2. **Fine-Grained Expert Specialization:** Unlike conventional coarse-grained upcycling methods that replicate entire FFN blocks, we employ a sub-matrix splitting technique to initialize a large number of fine-grained experts. This approach, combined with Drop-Upcycling (Nakamura et al., 2025), catalyzes specialized expert convergence and minimizes the redundancy bottlenecks typical of monolithic replication.
3. **Full Transparency and Openness:** In contrast to many models that keep their pre-training data and recipes proprietary, Marco-MoE is fully transparent. We disclose the entirety of our pre-training datasets, our data synthesis methodologies, and our rigorous four-stage pre-training curriculum spanning 5.1T tokens. By open-sourcing our models, data, and training recipes, Marco-MoE sets a new standard for accessible, high-performance multilingual LLM development.

Our empirical results demonstrate that Marco-Mini-Base (0.86B activated parameters) and Marco-Nano-Base (0.6B activated parameters) achieve a superior performance-to-compute ratio (Figure 1). Notably, our models set a new state-of-the-art for simultaneous proficiency in both English and multilingual capabilities, particularly excelling in long-tail and low-resource languages where the performance gap in specific geographic regions (e.g., West and South Asia) over dense counterparts widens significantly (Figure 2).

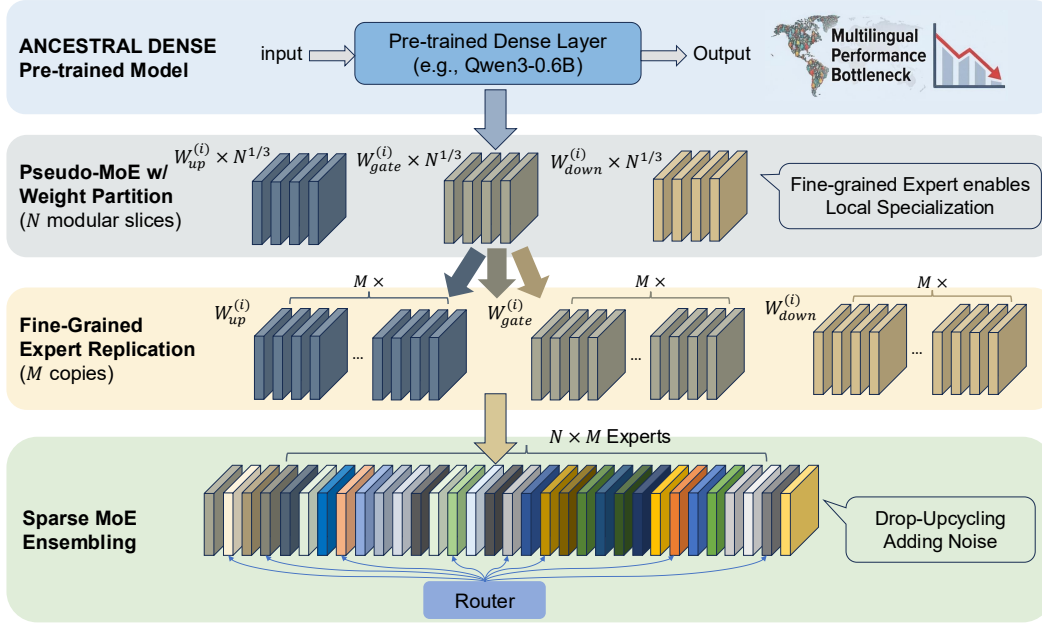


Figure 3: The overview of upcycling dense models to MoE models with fine-grained experts.

2 MoE Upcycling with Fine-Grained Experts

Marco-MoE LLM leverages a decoder-only Transformer architecture (Vaswani et al., 2023) with sparse Mixture-of-Experts (MoE) layer following Yang et al. (2025). Detailed configurations are summarized in Appendix Table 6. Specifically, the Marco-Nano-Base variant comprises 8B total parameters with 0.6B activated (7.5% active ratio), while the Marco-Mini-Base Base contains 17.3B total parameters with 0.86B activated (5% active ratio). Both of them are upcycled from the Qwen3-0.6B-Base model using the method detailed below.

To initialize our MoE architecture efficiently, we adopt the upcycling paradigm (Komat-suzaki et al., 2023), which repurposes pre-trained dense representations to reduce computational overhead. While naive replication of Feed-Forward Network (FFN) layers often results in expert homogeneity and sub-optimal specialization, we mitigate this by incorporating Drop-Upcycling (Nakamura et al., 2025). This strategy involves randomly dropping and re-initializing expert parameters based on the Gaussian statistics of the original weights, thereby inducing the gradient variance necessary for expert diversification.

Our method specifically adapts this approach to a fine-grained MoE setting (Figure 3). We first reconfigure the ancestral dense model into a pseudo-MoE framework by partitioning the dense FFN weights, $W_{up} \in \mathbb{R}^{H \times H_{dense}}$, $W_{gate} \in \mathbb{R}^{H \times H_{dense}}$, and $W_{down} \in \mathbb{R}^{H_{dense} \times H}$, into $N = H_{dense} / H_{expert}$ modular slices:

$$W_{up} \rightarrow [W_{up}^{(1)}, \dots, W_{up}^{(N)}], \quad W_{gate} \rightarrow [W_{gate}^{(1)}, \dots, W_{gate}^{(N)}], \quad W_{down} \rightarrow [W_{down}^{(1)}, \dots, W_{down}^{(N)}]$$

Each i -th expert is then initialized with the corresponding slice pair $(W_{up}^{(i)}, W_{gate}^{(i)}, W_{down}^{(i)})$.

A critical challenge in this transition is the functional mismatch between the dense model’s aggregate sum ($Y = \sum E_i(x)$) and the MoE’s weighted combination ($Y_{MoE} = \sum p_i E_i(x)$) (He et al., 2025a). Under a uniform routing assumption where $p_i \approx 1/N$, the MoE output magnitude is diminished by a factor of N . To preserve numerical integrity and ensure the pseudo-MoE is mathematically equivalent to the source dense model, we apply a scaling factor $\lambda = N^{1/3}$ directly to the sliced weight matrices.¹ This preserves the original signal’s

¹We empirically find that this weight scaling makes training more stable in the early pre-training stage and reduces loss spikes. Please refer to Appendix Figure 9 for evidence.

Model	Multilingual Diverse QA	Multilingual STEM			Cultural & Regional Data		
	BELEBELE	Global-MMLU	MMMLU	MGSM	INCLUDE	IndoMMLU	TurkishMMLU
Qwen3-4B-Multilingual	82.2	64.2	60.7	73.7	62.3	60.1	60.1
+ Multilingual Synthetic	83.5	65.6	62.2	81.6	63.5	61.2	64.0

Table 1: The effects of including our various curated multilingual synthetic data into the pre-training mixture.

dynamic range without altering the forward-pass logic. Finally, we expand this pseudo-MoE into a sparse architecture by replicating the N modular slices M times (e.g., $N \times M$ experts) with zero-initialized routers, using Drop-Upcycling to catalyze the convergence of specialized experts.

3 Pre-training Data

3.1 High-Quality English, Reasoning, and Instruction Data

To enhance the general language understanding of Marco-MoE, we leverage Nemotron-CC-v2 (NVIDIA et al., 2025a) as the primary source of high-quality English data, specifically utilizing the **High** and **High-Synthetic** partitions. Furthermore, to bolster targeted capabilities, we incorporate the **Diverse question-answer (QA)** pairs from Nemotron-CC-v2. These pairs encompass various formats, including yes/no, open-ended, and multiple-choice questions, centered on factual information within the text. Besides, we incorporate a diverse array of open-source datasets spanning the STEM, Coding, and instruction-following domains, which provide a robust foundation of reasoning-intensive synthetic problems, high-quality domain-specific documents, and diverse instruction data. Details are in Appendix B.1.

3.2 High-Quality Multilingual Data

3.2.1 Web-Crawled Data

We aggregate multilingual web-crawled data for 28 target languages, sourcing from open-source repositories. Our primary resource is Fineweb-2 (Penedo et al., 2025), which provides corpora for over a thousand languages. To ensure data quality, we prioritize the FineWeb2-HQ variant (Messmer et al., 2025), which utilizes a model-based classifier to filter out low-quality documents for 20 languages. We use this dataset when a high-quality version is available for a language. For the remaining languages that lack high-quality subsets, we employ a rephrasing strategy to mitigate the inherent noise in web-crawled content. Specifically, we employ Qwen3-30B-A3B-Instruct (Yang et al., 2025) to generate synthetic data using four prompting templates derived from Su et al. (2025) and Maini et al. (2024). We specifically curate the training data for Chinese by sourcing from the high-quality partition of Fineweb-Edu-Chinese-V2.1 dataset (Yu et al., 2025b).

3.2.2 Synthetic Data

Multilingual QA Data Building on evidence that high-quality multilingual question-answering (QA) data boosts downstream multilingual performance (NVIDIA et al., 2025b), we curate an expanded dataset by extending Nemotron-CC-v2 with Qwen3-30B-A3B-Instruct translations of the English Diverse QA set to the remaining 15 languages. Our translation pipeline employs line-level processing and curated prompts (Appendix Prompt D.3.3) to preserve the integrity of technical syntax, such as $\text{\LaTeX}$ and code snippets, while utilizing language identification and heuristic n -gram filtering to ensure quality in low-resource settings. To validate this dataset, we conduct an ablation study following the annealing methodology of OLMo et al. (2025) with a 100B-token budget. By continuously pre-training a Qwen3-4B base model on a mixture of 50% synthesized QA data and 50% web-crawled text, we observe a consistent 1.4% average improvement (Table 1) across 29 languages on the BELEBELE benchmark (Bandarkar et al., 2024) relative to a web-only baseline.

Multilingual STEM Data STEM corpora are characterized by high information density and have been shown to bolster both the general and task-specific capabilities of LLMs. Nev-

ertheless, such resources remain scarce across most non-English languages. To address this disparity, we employ a translation-based data synthesis pipeline to project English STEM knowledge into various target languages. Specifically, for general-domain STEM content, we leverage the [Nemotron-Pretraining-SFT-v1](#) dataset as our primary source and apply our synthesis framework to generate high-fidelity multilingual data across 28 languages. Furthermore, we curate a specialized multilingual mathematical dataset by translating the OpenMathInstruct-2 corpus ([Toshniwal et al., 2024](#)). To evaluate the curated multilingual STEM dataset, we conduct the same ablation with a 100B-token budget with performance measured on Global-MMLU ([Singh et al., 2025](#)), [MMMLU](#), and [MGSM](#) ([Shi et al., 2023](#)). Table 1 indicates that incorporating multilingual STEM data significantly enhances performance, yielding an average absolute gain of up to 8%.

Cultural and Regional Data To capture nuanced socio-cultural contexts within a multilingual framework, we address the scarcity of culturally salient data through two distinct curation strategies. First, we develop **Fineweb2-Culture** by filtering web-crawled data from Fineweb2 for high cultural density. Using an LLM-based evaluator (see Appendix Prompt D.3.3), we retain high-quality documents (score > 7) where the predicted geographic origin aligns with the primary language. These documents are further enhanced through rephrasing and QA-generation strategies to maximize token utility ([Team et al., 2026](#)).

Second, we introduce **Synthetic-Regional-MCQs**, a pipeline for generating diverse, culture-specific synthetic data. We employ Gemini-2.5-Pro ([Comanici et al., 2025](#)) to perform hierarchical expansion, decomposing coarse cultural themes into fine-grained sub-topics. These sub-topics are then used to generate a vast array of multiple-choice questions (MCQs) following prompting techniques adapted from [Bercovich et al. \(2025\)](#) and [Nathawani et al. \(2025\)](#) (see Appendix Prompt D.3.3 for the exact prompt and Appendix Example D.3.3 for specific examples).²

The efficacy of these datasets is validated through an ablation study during a 10B-token annealing phase. Evaluation across regional benchmarks, including INCLUDE ([Romanou et al., 2025](#)), IndoMMLU ([Koto et al., 2023](#)), and TurkishMMLU ([Yüksel et al., 2024](#)), demonstrates that the inclusion of specialized cultural data consistently improves performance on tasks requiring localized cultural knowledge (Table 1).

4 Experiments

4.1 Four-Stage Pre-training Curriculum

We adopt a four-stage data-mixture strategy to pre-train Marco-MoE models across 29 languages using 5.1T tokens. The data mixtures used in each stage are shown in Appendix Figure 7 and Table 11,³ and the detailed hyperparameters are in Appendix B.3.

1. **Stage 1 (0 – 2.4T tokens):** The foundational training stage comprises a high-quality blend of English, reasoning, and instruction corpora, alongside multilingual web and QA data covering 19 languages (Appendix B.2).
2. **Stage 2 (2.4T – 4.1T tokens):** To optimize the curriculum, we adjust the data mixture by upsampling reasoning corpora and downsampling English web data. We simultaneously upsample Chinese data to improve targeted linguistic performance.
3. **Stage 3 (4.1T – 4.6T tokens):** The sampling weights for English, reasoning, and instruction domains are reduced to allocate token bandwidth for another nine lan-

²For data decontamination, we use the target cultural and regional benchmarks to detect and remove synthetic samples exhibiting excessive n -gram overlap. To safeguard against sophisticated rephrasing that surface-level metrics might miss, we further follow [Yang et al. \(2023\)](#) to identify and exclude semantically contaminated data.

³Due to computational resource constraints, the data mixture ratios were determined via empirical heuristics and ablations on smaller-scale models, aligning with methods established in recent literature ([Akter et al., 2026](#); [Xiaomi et al., 2025](#); [Abdin et al., 2024](#)).

guages. We concurrently upsample medium-resource languages to ensure balanced cross-lingual efficacy.

4. **Stage 4 (4.6T – 5.1T tokens):** We further decay the sampling weights of the English, reasoning, and instruction corpora, integrating a diverse array of curated multilingual synthetic data to explicitly target and boost specific reasoning and linguistic capabilities across all languages.

4.2 Evaluation Settings

Evaluation Benchmarks Unless otherwise specified, all model evaluations are conducted using the Light-Eval framework (Habib et al., 2023) to ensure a standardized and fair comparison. The evaluation suite is divided into two primary dimensions: English proficiency and multilingual capability. To measure multilingual performance, we evaluate the model on both general multilingual benchmarks and datasets that assess cultural and regional knowledge. Detailed benchmarks are in Appendix B.4.

Baselines We compare Marco-MoE models against a diverse set of open-source baselines, including both open-weight and open-weight & open-data models, selected for their explicit multilingual focus and comparable model scale and training compute. The comparison set includes Qwen3 1.7B and 4B (Yang et al., 2025), Granite4-Tiny, Llama3.2-3B (Grattafiori et al., 2024), SmolLM3-3B (Bakouch et al., 2025), Gemma3-4B (Team et al., 2025), Tiny-Aya-3.35B (Salamanca et al., 2026), and Trinity Nano and Mini (Singh et al., 2026).

4.3 Main Results

Table 2 contains the results of pre-trained Marco-MoE models.

English results. On English benchmarks, Marco-Mini-Base achieves a leading average score of 63.7, surpassing dense baselines such as Qwen3-4B while utilizing significantly less training compute (1.56×10^{23} vs. 8.64×10^{23} FLOPs) and MoE models with way less activated and total parameters (i.e., Trinity-Mini). Its performance is particularly robust in reasoning and commonsense tasks, including ARC-Challenge and CommonsenseQA. Meanwhile, it remains near the top on knowledge-intensive and math-oriented benchmarks. Similarly, Marco-Nano-Base demonstrates high parameter efficiency: with only 0.6B active parameters, it outperforms several 3B–4B-class models, including Llama3.2-3B and Gemma3-4B. These results establish the proposed models as top-tier general-purpose English LLMs that maintain strong reasoning and knowledge-intensive capabilities without the typical performance trade-offs associated with compact architectures or multilingual optimizations.

Multilingual general results. The multilingual general benchmarks show the clearest advantage of the Marco-MoE family. Marco-Mini-Base achieves the best average score (50.9), outperforming all baselines, including Qwen3-4B-Base (48.3). It ranks first on a broad set of tasks covering multilingual knowledge, reasoning, and translation, including GlobalMMLU (64.2), MMMLU (62.0), mHellaSwag (58.6), mARC-Challenge (45.4), FLORES-200, and WMT24++ in both directions. On MGSM, it achieves 75.6, effectively matching the strongest baseline. This consistent performance indicates that Marco-Mini-Base is a well-balanced multilingual model rather than one optimized for a narrow subset of tasks. Marco-Nano-Base is also highly competitive, achieving an average of 43.7. Notably, it achieves the best scores on BELEBELLE and MGSM among all models compared, while remaining strong on the others. Given its small activated footprint, these results highlight particularly efficient cross-lingual reasoning.

Multilingual cultural and regional results. Evaluation on cultural and regional benchmarks demonstrates that the Marco family excels at localized reasoning and domain-specific knowledge. Marco-Mini-Base nearly rivals Qwen3-4B-Base while significantly outperforming other baselines across diverse contexts, including Arabic, Greek, and Turkish MMLU. Similarly, Marco-Nano-Base exhibits impressive efficiency, surpassing larger models like Llama3.2-3B and Gemma3-4B. Although Qwen3-4B maintains an edge in specialized Chinese tasks (CMMLU and C-Eval), the findings highlight Marco’s robust cultural grounding and its ability to capture regional nuances that go beyond generic multilingual transfer.

Benchmark (Metric)	# Shots	Qwen3 1.7B	Trinity Nano	Granite4 Tiny	Marco Nano	Llama3.2 3B	SmolLM3 3B	Gemma3 4B	Tiny-Aya 3.35B	Qwen3 4B	Trinity Mini	Marco Mini
# Activated Params	-	1.7B	1.09B	1.47B	0.6B	3B	3B	4B	3.35B	4B	3.85B	0.86B
# Total Params	-	1.7B	6B	7B	8B	3B	3B	4B	3.35B	4B	26.5B	17.3B
# Train FLOPs ($\times 10^{23}$)	-	3.67	0.654	2.03	1.40	1.62	1.98	0.96	1.21	8.64	2.31	1.56
English												
MMLU (Acc)	5-shot	<u>65.1</u>	64.7	69.1	64.7	57.6	62.6	61.1	58.6	75.2	71.4	<u>72.8</u>
MMLU-Redux (Acc)	0-shot	61.2	60.1	65.8	<u>62.9</u>	56.9	58.4	57.7	51.7	71.3	68.2	<u>68.8</u>
MMLU-Pro (Acc)	5-shot	<u>33.2</u>	32.0	32.1	35.9	26.0	35.1	28.8	26.9	45.9	41.3	<u>45.3</u>
AGIEval (Acc)	0-shot	35.9	31.4	<u>36.1</u>	38.4	31.2	34.5	32.6	29.0	44.0	39.7	<u>41.9</u>
BBH (EM)	3-shot	<u>54.5</u>	49.3	59.9	53.5	47.1	60.0	52.2	46.8	72.3	57.6	<u>65.1</u>
ARC-Easy (Acc)	0-shot	69.3	<u>77.9</u>	78.5	75.3	71.8	78.5	82.6	76.5	75.0	80.6	<u>82.4</u>
ARC-Challenge (Acc)	0-shot	42.8	53.5	<u>52.3</u>	49.4	46.0	52.6	54.1	47.4	49.9	57.8	<u>56.3</u>
HellaSwag (Acc)	0-shot	66.6	<u>77.4</u>	77.9	69.2	75.6	76.1	76.7	71.0	74.4	82.8	<u>77.4</u>
WinoGrande (Acc)	0-shot	<u>57.1</u>	57.1	58.6	53.4	58.6	58.9	61.4	56.6	59.6	<u>60.8</u>	57.7
BoolQ (Acc)	0-shot	74.6	<u>71.5</u>	63.5	71.2	75.2	79.3	<u>76.6</u>	74.6	74.2	72.5	74.2
CommonsenseQA (Acc)	0-shot	49.5	54.1	55.9	<u>55.7</u>	60.4	55.4	<u>61.1</u>	60.4	52.9	57.7	61.5
OpenBookQA (Acc)	0-shot	36.4	<u>42.0</u>	43.6	39.4	42.2	40.4	<u>42.6</u>	40.4	42.6	44.8	<u>44.6</u>
PIQA (Acc)	0-shot	75.5	69.6	80.6	<u>76.5</u>	78.2	79.1	<u>80.3</u>	76.9	77.4	71.7	81.1
SIQA (Acc)	0-shot	47.8	<u>52.7</u>	53.0	46.0	51.0	49.8	50.4	49.9	53.0	<u>52.5</u>	49.4
GSM8K (EM)	5-shot	69.1	57.8	70.7	<u>69.7</u>	27.3	67.4	39.3	58.0	81.7	57.5	<u>76.4</u>
Average	-	55.9	56.7	59.8	<u>57.5</u>	53.7	59.2	57.2	55.5	<u>63.3</u>	61.1	63.7
Multilingual – General												
GlobalMMLU (Acc)	5-shot	49.6	43.6	54.8	<u>52.2</u>	43.2	46.7	50.8	50.0	<u>61.6</u>	52.6	64.2
MMMLU (Acc)	0-shot	48.6	41.2	<u>52.3</u>	52.6	44.0	47.3	47.4	44.5	<u>59.3</u>	50.9	62.0
MMLU-ProX-Lite (Acc)	5-shot	27.2	20.3	30.1	<u>28.9</u>	22.4	28.3	24.3	24.3	<u>38.5</u>	32.2	39.2
BELEBELE (Acc)	0-shot	<u>67.5</u>	54.5	61.2	73.8	60.1	54.3	65.7	65.4	81.5	67.6	<u>79.8</u>
mHellaSwag (Acc_norm)	0-shot	43.9	42.5	53.2	<u>48.8</u>	49.0	49.6	<u>55.2</u>	53.5	53.2	51.5	58.6
mARC-Challenge (Acc_norm)	0-shot	34.7	30.9	39.9	<u>36.9</u>	34.2	36.1	41.5	37.2	<u>42.5</u>	37.5	45.4
FLORES-200 (En-Xx) (BLEU)	5-shot	18.6	15.1	25.4	<u>24.7</u>	23.5	19.7	<u>32.1</u>	30.2	25.4	13.7	32.3
FLORES-200 (Xx-En) (BLEU)	5-shot	31.5	31.1	36.7	<u>33.6</u>	34.6	30.3	<u>39.7</u>	37.3	36.8	24.1	40.1
WMT24++ (En-Xx) (BLEU)	5-shot	18.3	15.0	21.9	<u>20.7</u>	16.4	17.8	<u>27.7</u>	26.1	23.9	7.5	28.1
WMT24++ (Xx-En) (BLEU)	5-shot	<u>28.3</u>	28.0	30.7	28.1	28.9	27.4	<u>34.0</u>	32.7	32.9	10.6	34.4
MGSM (EM)	8-shot	<u>58.8</u>	40.6	56.7	65.3	22.4	50.8	36.6	38.4	76.0	57.2	<u>75.6</u>
Average	-	38.8	33.0	<u>42.1</u>	42.3	34.4	37.1	41.4	39.9	<u>48.3</u>	36.9	50.9
Multilingual – Cultural & Regional												
INCLUDE (Acc)	5-shot	51.2	43.9	<u>52.1</u>	53.2	45.5	46.2	52.6	53.9	<u>61.4</u>	51.9	61.7
Global-PIQA (Acc_norm)	0-shot	60.3	52.3	<u>64.0</u>	64.3	62.2	60.9	<u>69.4</u>	67.9	65.4	57.2	72.3
CMMLU (Acc)	5-shot	66.1	49.6	53.5	<u>55.5</u>	44.1	50.1	50.2	58.8	76.2	58.6	<u>68.0</u>
C-Eval (Acc)	5-shot	65.1	47.6	50.9	<u>56.0</u>	43.1	47.9	48.5	57.6	76.6	57.1	<u>66.0</u>
ArabicMMLU (Acc)	3-shot	<u>57.6</u>	44.0	60.5	55.8	48.9	60.6	61.6	63.2	<u>67.0</u>	57.1	67.1
TurkishMMLU (Acc)	5-shot	<u>47.9</u>	29.6	41.8	48.9	36.7	28.4	43.7	45.2	<u>60.6</u>	43.0	62.7
GreekMMLU (Acc)	5-shot	58.1	52.2	<u>62.3</u>	64.1	56.4	64.0	63.4	66.3	<u>69.4</u>	59.7	70.3
KazakhMMLU (Acc)	5-shot	52.1	43.1	<u>52.6</u>	53.1	44.7	47.4	52.1	47.1	<u>62.3</u>	49.6	62.6
IndoMMLU (Acc)	0-shot	51.0	41.5	<u>49.0</u>	51.0	47.0	43.7	48.5	52.0	60.1	51.0	<u>59.9</u>
IndoCareer (Acc)	3-shot	53.9	46.7	<u>53.0</u>	52.1	48.6	47.7	53.4	56.6	61.5	55.2	61.5
IndoCulture (Acc)	0-shot	<u>51.6</u>	49.8	51.3	57.4	50.1	44.5	59.1	58.5	<u>61.1</u>	57.6	62.3
Average	-	55.9	45.5	53.7	<u>55.6</u>	47.9	49.2	54.8	57.0	65.6	54.4	<u>65.0</u>

Table 2: Comparison among Marco-Nano-Base & Marco-Mini-Base and other open-source base models of comparable sizes and training costs. The best and second best results are marked in **bold** and underline.

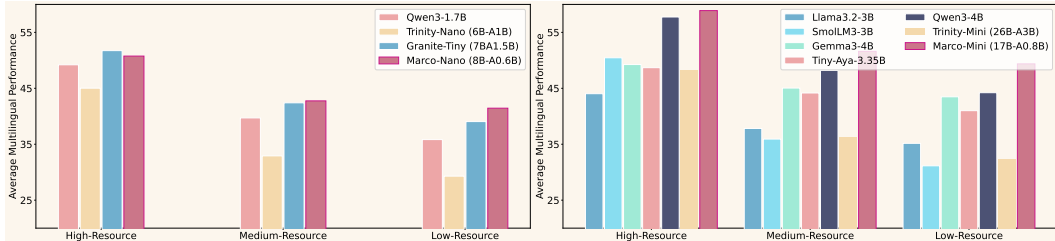


Figure 4: Comparison between Marco-MoE and baselines across language resource levels.

4.4 Analysis

Marco-MoE models excel in long-tail languages. Figure 4 demonstrates the superior multilingual capabilities of the Marco-MoE models, particularly in long-tail languages. In the approximately 1B-parameter category, Marco-Nano-Base consistently outperforms counterparts such as Granite4-Tiny and Qwen3-1.7B, which have more activated parameters, and the performance gap widens significantly as language resource availability decreases. A similar trend is observed at the 3B-4B scale, where Marco-Mini-Base achieves the highest overall performance across all resource tiers, with a substantial lead over established models in low-resource languages.

	Stage-1	Stage-2	Stage-3	Stage-4
English	59.4	63.5	63.7	63.7
General Multilingual	40.6	46.3	50.3	52.1
Cultural & Regional	54.0	59.8	63.2	65.0

Table 3: Average results of Marco-Mini-Base on different benchmark categories after each pre-training stage. Per-benchmark results are shown in Appendix Table 12.

	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Marco Mini Global Base
English	57.2	55.5	63.3	63.6
General Multilingual	39.5	37.3	44.5	48.1
Cultural & Regional	54.6	56.9	65.1	64.7

Table 4: Comparing Marco-Mini-Global-Base on different benchmark categories with baselines on 64 languages. Per-benchmark results are shown in Appendix Table 13.

Multi-stage pre-training gradually improves downstream performance. As shown in Table 3, the four-stage pre-training curriculum yields consistent improvements across all evaluation dimensions. From Stage-1 to Stage-2, performance increases substantially on every benchmark category, highlighting the importance of the learning-rate decay phase for enhancing overall model capability. From Stage-3 onward, both general multilingual and cultural & regional multilingual performance continue to improve, while English performance largely saturates, suggesting that the data-mixture switch introduced in Stage-3 effectively strengthens multilingual capability without degrading English performance. Overall, these results confirm the effectiveness of the multi-stage curriculum: each stage provides complementary gains, and the later stages are particularly important for improving multilingual coverage and culturally grounded knowledge.

Scaling up to 64 Languages To evaluate the scalability of the Marco-MoE framework, we expand the model’s linguistic coverage from 29 core languages to 64 languages. Following the multi-stage curriculum, we branch from the Stage-2 checkpoint and recalibrate the data mixtures in Stages 3 and 4 to integrate pre-training corpora for a set of 35 newly introduced languages (See Appendix B.2). This variant, designated as Marco-Mini-Global-Base, was trained on an additional 1.4T tokens and consistently demonstrates superior performance over comparable baselines as shown in Table 4. Notably, Marco-Mini-Global-Base preserves robust English proficiency, achieving 63.6 compared to the 63.7 average of the original Marco-Mini-Base, while simultaneously increasing the performance advantage in multilingual tasks from 2.6% to 3.6% relative to Qwen3-4B-Base.

MoE expert activation patterns resemble language similarity. To investigate whether MoE routing dynamics align with established linguistic taxonomies, we formalize the Language-Expert Signature. For each language L , we extract routing statistics by feeding documents sampled from the FineWeb-2 corpus to the model to define a specialization score, $\text{Specialization}(E_{i,l}, L)$, representing the proportion of tokens from L routed to expert E_i at layer l . This procedure yields a language-specific activation matrix $\mathbf{M}_L$ across all layers and experts, which we flatten and compare using the Pearson Correlation Coefficient to quantify the functional similarity in expert activation patterns between any two languages.

Our analysis reveals that these signatures largely mirror established linguistic families. Pairwise correlations (Figure 5) show distinct, high-correlation clusters for Romance, East Asian, and Austronesian, while low-correlation pairs highlight specialized expert utilization.

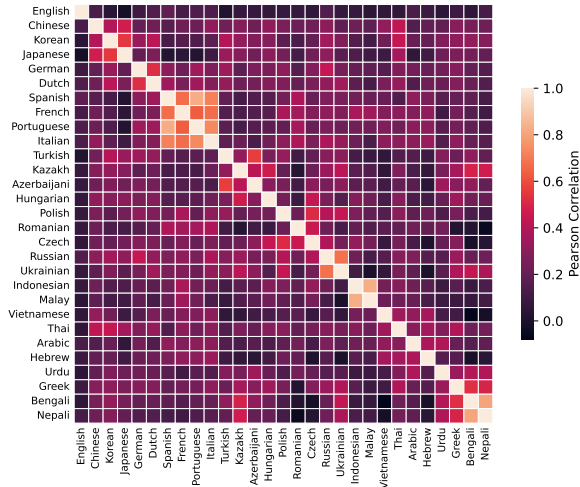


Figure 5: Pearson correlation heatmap between language-specific expert activation patterns. Structured experts are shared among related languages (e.g., Romance, East Asian, and Austronesian), while low-correlation pairs highlight specialized expert utilization.

Benchmark	Qwen3-1.7B	Qwen3-VL-2B	Ministral3-3B	LFM2-8B-A1B	Granite4-Tiny	Marco-Nano
English	53.7	52.9	59.2	62.5	44.4	62.8
Multilingual - General	31.7	32.5	22.7	31.7	30.6	38.9
Multilingual - Cultural	50.3	51.0	39.5	45.4	41.3	59.1

Benchmark	Qwen3-4B	Ministral3-8B	Gemma3-12B	Granite4-Small	LFM2-24B-A2B	Marco-Mini
English	73.3	70.9	65.8	64.8	66.7	75.5
Multilingual - General	47.9	36.8	48.6	44.5	36.9	50.8
Multilingual - Cultural	69.1	54.5	67.7	63.1	49.9	71.0

Table 5: Comparing Marco-MoE-Instruct’s performance with other open-weight Instruct models of similar sizes.

Slavic, Austronesian, and Indic languages, suggesting the router exploits cross-lingual transfer by leveraging shared morphological and syntactic structures. Conversely, languages with unique scripts or isolated grammars, such as Thai and Arabic, exhibit low correlations, indicating that the model effectively partitions specialized expert subsets to minimize interference. Notably, English remains relatively isolated, and hierarchical clustering of these signatures (Appendix C Figure 6) closely recovers known linguistic family structures.

4.5 Post-Training Recipe and Evaluation

Full descriptions of post-training are in Appendix D.

Post-training settings. To evaluate the downstream capabilities of Marco-MoE models, we implement a robust post-training recipe on open-source datasets to obtain Marco-MoE-Instruct models. Specifically, our pipeline consists of two consecutive stages: Supervised Fine-tuning (SFT) and on-policy knowledge distillation (OPD) (Lu & Lab, 2025; Yang et al., 2025):

1. **SFT Stage:** We fine-tune the base models on a curated mixture of open-source datasets, including Dolci-Instruct (Olmo et al., 2025), Nemotron-Cascade-2 (Yang et al., 2026), and translation data from NLLB (Team et al., 2022).
2. **OPD Stage:** Following SFT, we perform on-policy distillation utilizing stronger teacher models from the Qwen series, specifically, Qwen3-30B-A3B and Qwen3-Next-80B-A3B. The training data spans open-source datasets, NLLB translation pairs, and synthetically generated queries tailored to cultural knowledge.

Evaluation results of Marco-MoE-Instruct. We evaluate the post-trained Marco-MoE-Instruct on English proficiency, general multilingual capabilities, and culturally specific tasks. Table 5 contains the downstream results. When evaluated against various state-of-the-art baselines, we find Marco-MoE-Instruct models consistently achieve superior performance and the highest overall accuracy across all benchmarks despite being significantly more parameter-efficient than their respective competitors.

5 Related Work

MoE Model Upcycling. The computational overhead of training large-scale Mixture-of-Experts (MoE) models from scratch has spurred the emergence of MoE Upcycling (Komatuzaki et al., 2023), a paradigm that initializes sparse models using pre-trained dense checkpoints. Prominent methods, such as Sparse Upcycling (e.g., Mixtral 8x7B (Jiang et al., 2024)), typically utilize coarse-grained expert construction. In these frameworks, the feed-forward networks (FFNs) of a dense transformer are replicated in their entirety to form the expert pool, followed by continued pre-training. Because all experts are identical at initialization, this approach may impede the emergence of specialized representations. While recent strategies seek to break this symmetry by injecting random noise during replication (Yang et al., 2024; Nakamura et al., 2025), the model’s fundamental units remain monolithic.

While computationally convenient, coarse-grained upcycling stands in tension with recent architectural trends suggesting that a large number of fine-grained experts is more effective than a few coarse-grained experts under an equivalent parameter budget (Dai et al., 2024).

Our method shifts toward fine-grained expert upcycling. Rather than duplicating entire FFN blocks, we decompose dense weights into modular sub-components via sub-matrix splitting before expanding them into an expert ensemble. Our method mitigates the redundancy bottlenecks inherent in monolithic block replication and facilitates expert specialization.

Multilingual LLMs. The development of multilingual Large Language Models (LLMs) is fundamentally constrained by the "curse of multilinguality" (Conneau et al., 2020), in which increasing language coverage within a fixed-parameter budget degrades performance due to capacity bottlenecks and cross-lingual interference. While cross-lingual transfer can benefit low-resource languages (Joshi et al., 2020; Nigatu et al., 2024), it is most effective between typologically related languages (He et al., 2025c), making it difficult for small-scale models to maintain high performance across diverse families (Üstün et al., 2024).

Recent models have adopted varying strategies to mitigate these constraints: Qwen3 (Yang et al., 2025) and Gemma3 (Team et al., 2025) utilize massive pre-training (up to 36T tokens) and sophisticated strong-to-weak distillation, claiming to support 119 and 140+ languages, respectively. In contrast, models such as Marco-LLM (Ming et al., 2024) and SmolLM3 (Bakouch et al., 2025) prioritize per-language depth over breadth, focusing exclusively on a narrow subset of higher-resource languages. Other efforts, including SEA-LION (Ong & Limkonchotiwat, 2023), Sailor2 (Dou et al., 2025), EuroLLM (Martins et al., 2024; Ramos et al., 2026), and Franken-Adapter (Jiang et al., 2025), address capacity limitations through regional specialization. While extreme cases like Apertus (Apertus et al., 2025) expand coverage to over 1,000 languages, they often encounter substantial performance trade-offs on standard benchmarks.

Despite these advances, most existing small-scale multilingual models rely on dense architectures, which inherently struggle to balance breadth and proficiency. Our work introduces the first compact, highly-sparse Multilingual MoE models with a fully open pre-training recipe. By leveraging the conditional computation of MoE, our method effectively breaks the curse of multilinguality, scaling support to 64 languages without the performance degradation typical of dense counterparts. This architectural shift facilitates specialized expert activation patterns that preserve the integrity of individual languages while maximizing cross-lingual transfer across related languages.

6 Conclusion and Discussion

In this work, we introduced Marco-MoE, a family of fully open-source, multilingual, and highly sparse MoE models that effectively mitigate the capacity bottlenecks inherent in dense architectures. By employing a fine-grained upcycling strategy, we successfully converted pre-trained dense models into highly sparse MoE layers that activate only 5% of parameters per token. Our models, Marco-Nano-Base and Marco-Mini-Base, demonstrate state-of-the-art performance across English and multilingual benchmarks, significantly surpassing similarly sized dense and MoE baselines while maintaining a superior performance-to-compute ratio. Furthermore, our analysis confirms that Marco-MoE naturally learns expert activation patterns that mirror linguistic family structures, enabling scalable expansion to 64 languages without language interference. By releasing our entire four-stage training curriculum, datasets, and weights, we aim to provide a transparent and high-performance foundation for future multilingual research.

While we have demonstrated the efficient scaling of Marco-MoE to 64 languages, this coverage encompasses only a small fraction of global linguistic diversity, particularly compared to massively multilingual models that target over 1,000 languages. Our reliance on translation-based and synthetic data pipelines to boost performance in low-resource settings, while empirically validated, underscores the persistent scarcity of authentic, high-quality multilingual reasoning data in naturalistic environments. Future research will focus on broadening Marco-MoE’s linguistic scope while investigating more sophisticated routing mechanisms to enhance expert specialization for extremely long-tail languages. Furthermore, because our current framework necessitates retraining the full model to integrate new languages, developing methods for modular, incremental expansion remains a compelling direction for efficient language scaling.

Ethics Statement

Data Sourcing and Privacy. The development of Marco-MoE relies on open-source web-crawled and synthesized datasets. We prioritize high-quality partitions and employ model-based classifiers to filter out low-quality content. While we extensively use large-scale web data, we acknowledge the inherent risks of such datasets, including potential biases and sensitive information. To mitigate potential issues, we disclose the full training datasets and recipes for generating synthetic data.

Linguistic Fairness and Cultural Representation. Marco-MoE is particularly designed to boost the performance in long-tail and low-resource languages. By incorporating specialized cultural and regional data, we managed to capture nuanced socio-cultural contexts that are often overlooked in standard multilingual models. However, we recognize that synthetic data generated by LLMs may reflect the biases of the teacher models used in the synthesis pipeline.

Transparency. We are committed to full transparency and the democratization of high-performance LLMs. We disclose our model weights and training recipes to support the broader research community. However, we acknowledge that open-weight models can be repurposed for malicious use and we encourage users of Marco-MoE to implement responsible fine-tuning and deployment practices.

LLM Usage Statement. We used LLMs for the grammatical refinement and structural polishing of this paper. In addition, coding agents were employed to support experiments. The final manuscript and all technical implementations were thoroughly checked by the authors.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL <https://arxiv.org/abs/2412.08905>.
- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.298. URL <https://aclanthology.org/2023.emnlp-main.298/>.
- Syeda Nahida Akter, Shrimai Prabhumoye, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, Yejin Choi, and Bryan Catanzaro. Front-loading reasoning: The synergy between pretraining and post-training data. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=VmEkHV2yCX>.
- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgrén, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL <https://arxiv.org/abs/2502.02737>.
- Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhiran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, Ramin Hasani, Saniya Karwa, Yuri Khrustalev, Maxime Labonne, Mathias Lechner, Valentine Lechner, Simon Lee, Zetian Li, Noel Loo, Jacob Marks, Edoardo Mosca, Samuel J. Paech,

Paul Pak, Rom N. Parnichkun, Alex Quach, Ryan Rogers, Daniela Rus, Nayan Saxena, Bettina Schlager, Tim Seyde, Jimmy T. H. Smith, Aditya Tadimeti, and Neehal Tumma. Lfm2 technical report, 2025. URL <https://arxiv.org/abs/2511.23404>.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Inés Altemir Mariñas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponskshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao, Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoeffler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. Apertus: Democratizing open and compliant llms for global language environments, 2025. URL <https://arxiv.org/abs/2509.14233>.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL <https://arxiv.org/abs/2511.21631>.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. <https://huggingface.co/blog/smollm3>, 2025.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL <https://aclanthology.org/2024.acl-long.44/>.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. Careful selection of knowledge to solve open book question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6120–6129, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1615. URL <https://aclanthology.org/P19-1615>.

- Akhiad Bercovich et al. Llama-nemotron: Efficient reasoning models, 2025. URL <https://arxiv.org/abs/2505.00949>.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL <https://arxiv.org/abs/1911.11641>.
- Tyler A. Chang et al. Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures, 2025. URL <https://arxiv.org/abs/2510.24081>.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL <https://aclanthology.org/N19-1300>.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL <https://arxiv.org/abs/1803.05457>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.
- Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL <https://arxiv.org/abs/2507.06261>.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747>.
- Damai Dai, Chengqi Deng, Chenggang Zhao, R.x. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.k. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.70. URL <https://aclanthology.org/2024.acl-long.70/>.
- Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks, 2017. URL <https://arxiv.org/abs/1612.08083>.
- Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, Haonan Wang, Jiaheng Liu, Yongchi Zhao, Xiachong Feng, Xin Mao, Man Tsung Yeung, Kunat Pipatanakul, Fajri Koto, Min Si Thu, Hynek Kydlíček, Zeyi Liu, Qunshu Lin, Sittipong Sripaisarnmongkol, Kridtaphad Sae-Khow, Nirattisai Thongchim, Taechawat Konkaew, Narong Borijindargoon, Anh Dao, Matichon Maneegard, Phakphum Artkaew, Zheng-Xin Yong, Quan Nguyen, Wannaphong Phatthiyaphaibun, Hoang H. Tran, Mike Zhang, Shiqi Chen, Tianyu Pang, Chao Du, Xinyi Wan, Wei Lu, and Min Lin. Sailor2: Sailing in south-east asia with inclusive multilingual llms, 2025. URL <https://arxiv.org/abs/2502.12982>.
- Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. Are LLMs breaking MT metrics?

- results of the WMT24 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 47–81, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.2. URL <https://aclanthology.org/2024.wmt-1.2/>.
- Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 5069–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.262. URL <https://aclanthology.org/2025.naacl-long.262/>.
- Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL <https://arxiv.org/abs/2506.04178>.
- Nathan Habib, Cl  mentine Fourrier, Hynek Kydl  cek, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023. URL <https://github.com/huggingface/lighteval>.
- Ethan He, Abhinav Khattar, Ryan Prenger, Vijay Korthikanti, Zijie Yan, Tong Liu, Shiqing Fan, Ashwath Aithal, Mohammad Shoeybi, and Bryan Catanzaro. Upcycling large language models into mixture of experts, 2025a. URL <https://arxiv.org/abs/2410.07524>.
- Jujie He, Jiakai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork open reasoner 1 technical report, 2025b. URL <https://arxiv.org/abs/2505.22312>.
- Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. Scaling laws for multilingual language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4257–4273, Vienna, Austria, July 2025c. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.221. URL <https://aclanthology.org/2025.findings-acl.221/>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021b. URL <https://arxiv.org/abs/2103.03874>.

- Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, dahai li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=3X2L2TFr0f>.
- Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*, 2025.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=f0rm2rGX2r>.
- IBM Research. Granite 4.0 language models. <https://github.com/ibm-granite/granite-4.0-language-models>, 2025. Accessed: 2025-10-01.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mixtral of experts, 2024. URL <https://arxiv.org/abs/2401.04088>.
- Fan Jiang, Honglin Yu, Grace Chung, and Trevor Cohn. Franken-adapter: Cross-lingual adaptation of llms by embedding surgery, 2025. URL <https://arxiv.org/abs/2502.08037>.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL <https://aclanthology.org/2020.acl-main.560>.
- Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=T5nUQDrM4u>.
- Fajri Koto. Cracking the code: Multi-domain LLM evaluation on real-world professional exams in Indonesia. In Weizhu Chen, Yi Yang, Mohammad Kachuee, and Xue-Yong Fu (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 938–948, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-194-0. doi: 10.18653/v1/2025.naacl-industry.69. URL <https://aclanthology.org/2025.naacl-industry.69/>.
- Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12359–12374, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.760. URL <https://aclanthology.org/2023.emnlp-main.760/>.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Al-raeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.),

- Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5622–5640, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.334. URL <https://aclanthology.org/2024.findings-acl.334/>.
- Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719, 2024b. doi: 10.1162/tacl_a_00726. URL <https://aclanthology.org/2024.tacl-1.92/>.
- Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 318–327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.28. URL <https://aclanthology.org/2023.emnlp-demo.28/>.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in Chinese. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11260–11285, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.671. URL <https://aclanthology.org/2024.findings-acl.671/>.
- Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. Self-alignment with instruction backtranslation. In *The Twelfth International Conference on Learning Representations*, 2024b. URL <https://openreview.net/forum?id=1oiJHJBRsT>.
- Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyt, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yasmine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL <https://arxiv.org/abs/2601.08584>.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. <https://thinkingmachines.ai/blog/on-policy-distillation>.

- Rabeeh Karimi Mahabadi, Sanjeev Satheesh, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc-math: A 133 billion-token-scale high quality math pretraining dataset, 2025. URL <https://arxiv.org/abs/2508.15096>.
- Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14044–14072, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.757. URL <https://aclanthology.org/2024.acl-long.757/>.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. Eurollm: Multilingual language models for europe, 2024. URL <https://arxiv.org/abs/2409.16235>.
- Bettina Messmer, Vinko Sabolčec, and Martin Jaggi. Enhancing multilingual llm pretraining with model-based data selection. *arXiv*, 2025. URL <https://arxiv.org/abs/2502.10361>.
- Lingfeng Ming, Bo Zeng, Chenyang Lyu, Tianqi Shi, Yu Zhao, Xue Yang, Yefeng Liu, Yiyu Wang, Linlong Xu, Yangyang Liu, Xiaohu Zhao, Hao Wang, Heng Liu, Hao Zhou, Huifeng Yin, Zifu Shang, Haijun Li, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-llm: Bridging languages via massive multilingual training for cross-lingual enhancement, 2024. URL <https://arxiv.org/abs/2412.04003>.
- Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoe: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=xXTkbTBmqq>.
- Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. Drop-upcycling: Training sparse mixture of experts with partial re-initialization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=gx1wHnf5Vp>.
- Dhruv Nathawani, Igor Gitman, Somshubra Majumdar, Evelina Bakhturina, Ameya Sunil Mahabaleshwarkar, Jian Zhang, and Jane Polak Scowcroft. Nemotron-Post-Training-Dataset-v1, July 2025.
- Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Tamar Solorio, and Monojit Choudhury. The Zeno’s paradox of ‘low-resource’ languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17753–17774, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.983. URL <https://aclanthology.org/2024.emnlp-main.983/>.
- NVIDIA. Nemotron 3 Nano: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for Agentic reasoning, 2025. URL <https://research.nvidia.com/labs/nemotron/files/NVIDIA-Nemotron-3-Nano-Technical-Report.pdf>. Technical report.
- NVIDIA et al. Nemotron-h: A family of accurate and efficient hybrid mamba-transformer models, 2025a. URL <https://arxiv.org/abs/2504.03624>.
- NVIDIA et al. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model, 2025b. URL <https://arxiv.org/abs/2508.14444>.
- 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, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL <https://arxiv.org/abs/2512.13961>.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL <https://arxiv.org/abs/2501.00656>.
- David Ong and Peerat Limkonchotiawat. SEA-LION (Southeast Asian languages in one network): A family of Southeast Asian language models. In Liling Tan, Dmitrijs Milajevs, Geeticka Chauhan, Jeremy Gwinnup, and Elijah Rippeth (eds.), *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pp. 245–245, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.nlposs-1.26. URL <https://aclanthology.org/2023.nlposs-1.26/>.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL <https://arxiv.org/abs/2406.17557>.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language. In *Second Conference on Language Modeling*, 2025. URL <https://openreview.net/forum?id=jnRBe6zatP>.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher, 2022. URL <https://arxiv.org/abs/2112.11446>.
- Miguel Moura Ramos, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M. Guerreiro, Ricardo Rei, Nicolas Boizard, Amin Farajian, Mateusz Klimaszewski, José G. C. de Souza, Barry

- Haddow, François Yvon, Pierre Colombo, Alexandra Birch, and André F. T. Martins. Eurollm-22b: Technical report, 2026. URL <https://arxiv.org/abs/2602.05879>.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL <https://arxiv.org/abs/2311.12022>.
- Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia soltani moakhar, Ayush Kumar Tarun, Azmine Toughik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=k3gCieTXeY>.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.
- Alejandro R. Salamanca, Diana Abagyan, Daniel D’souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaei, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, Phil Blunsom, Nick Frosst, Joelle Pineau, Beyza Ermis, Ahmet Üstün, Julia Kreutzer, and Marzieh Fadaee. Tiny aya: Bridging scale and multilingual depth, 2026. URL <https://arxiv.org/abs/2603.11510>.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL <https://aclanthology.org/D19-1454>.
- John Schulman and Thinking Machines Lab. Lora without regret. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250929. <https://thinkingmachines.ai/blog/lora/>.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarsz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL <https://arxiv.org/abs/1701.06538>.
- 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. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=fr3wGCK-IXp>.
- Shivalika Singh, Angelika Romanou, Clémentine Fourier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume*

- 1: *Long Papers*), pp. 18761–18799, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.919. URL <https://aclanthology.org/2025.acl-long.919/>.
- Varun Singh, Lucas Krauss, Sami Jaghouar, Matej Sirovatka, Charles Goddard, Fares Obied, Jack Min Ong, Jannik Straube, Fern, Aria Harley, Conner Stewart, Colin Kealty, Maziyar Panahi, Simon Kirsten, Anushka Deshpande, Anneketh Vij, Arthur Bresnu, Pranav Veldurthi, Raghav Ravishankar, Hardik Bishnoi, DatologyAI Team, Arcee AI Team, Prime Intellect Team, Mark McQuade, Johannes Hagemann, and Lucas Atkins. Arcee trinity large technical report, 2026. URL <https://arxiv.org/abs/2602.17004>.
- Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeibi, and Bryan Catanzaro. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2459–2475, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.123. URL <https://aclanthology.org/2025.acl-long.123/>.
- Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL <https://arxiv.org/abs/2104.09864>.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL <https://aclanthology.org/2023.findings-acl.824>.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL <https://aclanthology.org/N19-1421>.
- Gemma Team et al. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.
- Kimi Team et al. Kimi k2: Open agentic intelligence, 2026. URL <https://arxiv.org/abs/2507.20534>.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. URL <https://arxiv.org/abs/2207.04672>.
- Changxin Tian, Kunlong Chen, Jia Liu, Ziqi Liu, Zhiqiang Zhang, and Jun Zhou. Towards greater leverage: Scaling laws for efficient mixture-of-experts language models, 2025. URL <https://arxiv.org/abs/2507.17702>.
- Mukhammed Togmanov, Nurdaulet Mukhituly, Diana Turmakhan, Jonibek Mansurov, Maiya Goloburda, Akhmed Sakip, Zhuohan Xie, Yuxia Wang, Bekassyl Syzdykov, Nurkhan Laiyk, Alham Fikri Aji, Ekaterina Kochmar, Preslav Nakov, and Fajri Koto. KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge

- of Kazakhstan. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14403–14416, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.701. URL <https://aclanthology.org/2025.acl-long.701/>.
- Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024.
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.845. URL <https://aclanthology.org/2024.acl-long.845/>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL <https://arxiv.org/abs/1706.03762>.
- Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL <https://doi.org/10.1145/2629489>.
- Yiming Wang, Pei Zhang, Jialong Tang, Hao-Ran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. Polymath: Evaluating mathematical reasoning in multilingual contexts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL <https://openreview.net/forum?id=B1vCImy6yI>.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=y10DM6R2r3>.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=gEZrGCozdqR>.
- LLM-Core Xiaomi, :, Bingquan Xia, Bowen Shen, Cici, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, Liang Zhao, Peidian Li, Peng Wang, Shihua Yu, Shimao Chen, Weikun Wang, Wenhan Ma, Xiangwei Deng, Yi Huang, Yifan Song, Zihan Jiang, Bowen Ye, Can Cai, Chenhong He, Dong Zhang, Duo Zhang, Guoan Wang, Hao Tian, Haochen Zhao, Heng Qu, Hongshen Xu, Jun Shi, Kainan Bao, Kai Fang, Kang Zhou, Kangyang Zhou, Lei Li, Menghang Zhu, Nuo Chen, Qiantong Wang, Shaohui Liu, Shicheng Li, Shuhao Gu, Shuhuai Ren, Shuo Liu, Sirui Deng, Weiwei Zhuang, Weiwei Lv, Wenyu Yang, Xin Zhang, Xing Yong, Xing Zhang, Xingchen Song, Xinzhe Xu, Xu Wang, Yihan Yan, Yu Tu, Yuanyuan Tian, Yudong Wang, Yue Yu, Zhenru Lin, Zhichao Song, and Zihao Yue. Mimo: Unlocking the reasoning potential of language model – from pretraining to posttraining, 2025. URL <https://arxiv.org/abs/2505.07608>.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse

- Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1513–1532, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.79. URL <https://aclanthology.org/2025.emnlp-main.79/>.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL <https://arxiv.org/abs/2407.10671>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples, 2023. URL <https://arxiv.org/abs/2311.04850>.
- Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, Renjie Pi, Grace Lam, Nayeon Lee, Alexander Bukharin, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nemotron-cascade 2: Post-training llms with cascade rl and multi-domain on-policy distillation, 2026. URL <https://arxiv.org/abs/2603.19220>.
- Qiyong Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025a. URL <https://arxiv.org/abs/2503.14476>.
- Yijiong Yu, Ziyun Dai, Zekun Wang, Wei Wang, Ran Chen, and Ji Pei. Opencsg chinese corpus: A series of high-quality chinese datasets for llm training, 2025b. URL <https://arxiv.org/abs/2501.08197>.
- Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7035–7055, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.413. URL <https://aclanthology.org/2024.findings-emnlp.413/>.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting*

- of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL <https://aclanthology.org/P19-1472>.
- Biao Zhang and Rico Sennrich. Root mean square layer normalization, 2019. URL <https://arxiv.org/abs/1910.07467>.
- Yang Zhang, Mersin Konomi, Christos Xypolopoulos, Konstantinos Divriotis, Konstantinos Skianis, Giannis Nikolentzos, Giorgos Stamou, Guokan Shang, and Michalis Vazirgiannis. Greekmmlu: A native-sourced multitask benchmark for evaluating language models in greek, 2026. URL <https://arxiv.org/abs/2602.05150>.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL <https://arxiv.org/abs/2506.05176>.
- Wanjuan Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2299–2314, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.149. URL <https://aclanthology.org/2024.findings-naacl.149/>.
- Fan Zhou, Zengzhi Wang, Nikhil Ranjan, Zhoujun Cheng, Liping Tang, Guowei He, Zhengzhong Liu, and Eric P. Xing. Megamath: Pushing the limits of open math corpora. *arXiv preprint arXiv:2504.02807*, 2025. Preprint.
- Zilin Zhu, Chengxing Xie, Xin Lv, and slime Contributors. slime: An llm post-training framework for rl scaling. <https://github.com/THUDD/slme>, 2025. GitHub repository. Corresponding author: Xin Lv.
- Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models, 2022. URL <https://arxiv.org/abs/2202.08906>.

Model	Marco-Nano-Base	Marco-Mini-Base
Num Layers	28	28
Model Dimension	1024	1024
FFN Intermediate Dimension	3072	3072
Q-heads	16	16
KV-heads	8	8
Head Dimension	128	128
Expert Dimension	384	768
Total Experts	232	256
Number of Activated Experts	8	8
Tie Embedding	True	True

Table 6: Summary of the Marco-MoE LLM architectures.

A Model Architecture

Marco-MoE LLM leverages a decoder-only Transformer architecture (Vaswani et al., 2023), substituting conventional Feed-Forward Network (FFN) layers with sparse Mixture-of-Experts (MoE) layers (Shazeer et al., 2017). This substitution facilitates enhanced model capacity and accuracy while significantly reducing the number of activated parameters. To optimize performance, we implement a granular MoE architecture following Dai et al. (2024). Other architectural refinements, including Grouped-Query Attention (GQA) (Ainslie et al., 2023), RMSNorm (Zhang & Sennrich, 2019), SwiGLU activation (Dauphin et al., 2017), and Rotary Positional Embeddings (RoPE) (Su et al., 2023), are aligned with the Qwen framework (Yang et al., 2025). Detailed configurations are summarized in Table 6. Specifically, the Marco-Nano-Base variant comprises 8B total parameters with 0.6B activated (7.5% active ratio), while the Marco-Mini-Base Base contains 17.3B total parameters with 0.86B activated (5% active ratio). Both of them are upcycled from the Qwen3-0.6B-Base model (Yang et al., 2025) using the proposed method.

B Detailed Experimental Setup

B.1 Public Pre-training Datasets

For synthetic data spanning reasoning, math, coding and instruct-following domains, we leverage Nemotron-Pretraining-SFT-v1 (NVIDIA et al., 2025b), Nemotron-Pretraining-Specialized-v1 (NVIDIA, 2025), Nemotron-CC-Math-v1 (Mahabadi et al., 2025), FineMath (Allal et al., 2025), MegaMath (Zhou et al., 2025), OpenThoughts3-1.2M (Guha et al., 2025), and FLAN (Wei et al., 2022; OLMo et al., 2025). Their detailed usage is shown in Table 11.

B.2 Pre-training Languages

Marco-Nano-Base and Marco-Mini-Base support 29 **core-languages**: English, Chinese, Arabic, German, Spanish, French, Korean, Japanese, Portuguese, Turkish, Indonesian, Italian, Dutch, Polish, Russian, Vietnamese, Thai, Bengali, Czech, Hebrew, Ukrainian, Malay, Urdu, Kazakh, Greek, Romanian, Hungarian, Nepali, Azerbaijani.

We further expand the number of supporting languages to 64 with another 35 languages to train Marco-Mini-Global-Base: Danish, Swedish, Norwegian, Catalan, Galician, Welsh, Irish, Basque, Croatian, Latvian, Lithuanian, Slovak, Slovenian, Estonian, Finnish, Serbian, Bulgarian, Persian, Maltese, Hindi, Marathi, Gujarati, Punjabi, Tamil, Telugu, Tagalog, Javanese, Khmer, Lao, Burmese, Amharic, Swahili, Yoruba, Igbo, Zulu.

B.3 Hyperparameters

We pre-train the Marco-MoE models on a total of 5.1T tokens, employing the Warmup-Stable-Decay learning rate (LR) schedule (Hu et al., 2024). Specifically, the LR is linearly warmed up over 8.4B tokens to its peak value and is then maintained at that value throughout

Marco-Nano-Base	Stage-1	Stage-2	Stage-3	Stage-4
Learning Rate Schedule	Constant	Linear Decay	Linear Decay	Linear Decay
LR Warmup steps	890	0	0	0
Peak LR	4.9505×10^{-4}	4.9505×10^{-4}	1×10^{-5}	6×10^{-6}
Final LR	4.9505×10^{-4}	1×10^{-5}	6×10^{-6}	0
Batch size (# instances)	1,152	1,584	2,304	2,304
Sequence length	8,192	8,192	8,192	8,192
Total training tokens	2.4T	1.7T	500B	500B

Marco-Mini-Base	Stage-1	Stage-2	Stage-3	Stage-4
Learning Rate Schedule	Constant	Linear Decay	Linear Decay	Linear Decay
LR Warmup steps	890	0	0	0
Peak LR	4.6854×10^{-4}	4.6854×10^{-4}	1×10^{-5}	6×10^{-6}
Final LR	4.6854×10^{-4}	1×10^{-5}	6×10^{-6}	0
Batch size (# instances)	1,152	1,584	2,304	2,304
Sequence length	8,192	8,192	8,192	8,192
Total training tokens	2.4T	1.7T	500B	500B

Table 7: Pre-training hyperparameters for each stage of Marco-Nano-Base and Marco-Mini-Base.

Stage 1. Peak learning rates are determined based on our training FLOPs (Tian et al., 2025), yielding 4.9505×10^{-4} for Marco-Nano-Base and 4.6854×10^{-4} for Marco-Mini-Base. During Stage 2, the LR undergoes a linear decay to 1.0×10^{-5} over 1.7T tokens. In Stages 3 and 4, the LR is further linearly decayed to zero over an additional 1T tokens. Initial pre-training utilizes a sequence length of 8,192 and a batch size of 1,152, resulting in approximately 9.4M tokens per batch. This batch size is later expanded to 1,584 (roughly 13M tokens per batch) to accommodate a hardware scaling from 128 to 176 GPUs in Stage 2. In Stages 3 and 4, we further increase the batch size to 2,304 (roughly 19M tokens per batch). Following Muennighoff et al. (2025), we stabilize MoE training by incorporating a Load Balancing Loss (Shazeer et al., 2017) with a coefficient of 0.01 and a Router Z-loss (Zoph et al., 2022) with a coefficient of 0.001. Optimization is performed using AdamW (Loshchilov & Hutter, 2019) with a weight decay of 0.1, $\beta_1 = 0.9$, and $\beta_2 = 0.95$. Detailed hyperparameters are summarized in Table 7. The detailed pre-training logs of Marco-Mini-Base are shown in Figure 8.

B.4 Evaluation Benchmarks

We evaluate Marco-MoE models on a diverse range of English and multilingual benchmarks:

1. English Proficiency: We assess general language understanding and reasoning via a broad array of benchmarks, including MMLU (Hendrycks et al., 2021a), MMLU-Redux (Gema et al., 2025), MMLU-Pro (Wang et al., 2024), BBH (Suzgun et al., 2023), AGIEval (Zhong et al., 2024), ARC-Easy/Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2019), and several commonsense reasoning tasks such as BoolQ (Clark et al., 2019), CommonsenseQA (Talmor et al., 2019), OpenBookQA (Banerjee et al., 2019), PIQA (Bisk et al., 2019), SIQA (Sap et al., 2019), and GSM8K (Cobbe et al., 2021).
2. Multilingual Capability: To measure performance across diverse languages, we evaluate the model on both general multilingual benchmarks and datasets featuring cultural and regional knowledge assessment. General datasets include GlobalMMLU (Singh et al., 2025), MMLU-ProX-Lite (Xuan et al., 2025), BELEBELE (Bandarkar et al., 2024), multilingual HellaSwag & ARC (Lai et al., 2023), MMMLU, MGSM (Shi et al., 2023), and translation tasks with FLORES-200 (Team et al., 2022) and WMT24++ (Freitag et al., 2024). For cultural and regional benchmarks, we evaluate the models on INCLUDE (Romanou et al., 2025), Global-PIQA (Chang et al., 2025), CMMLU (Li et al., 2024a), C-Eval (Huang et al., 2023), ArabicMMLU (Koto et al., 2024a), TurkishMMLU (Yüksel et al., 2024), GreekMMLU (Zhang et al., 2026),

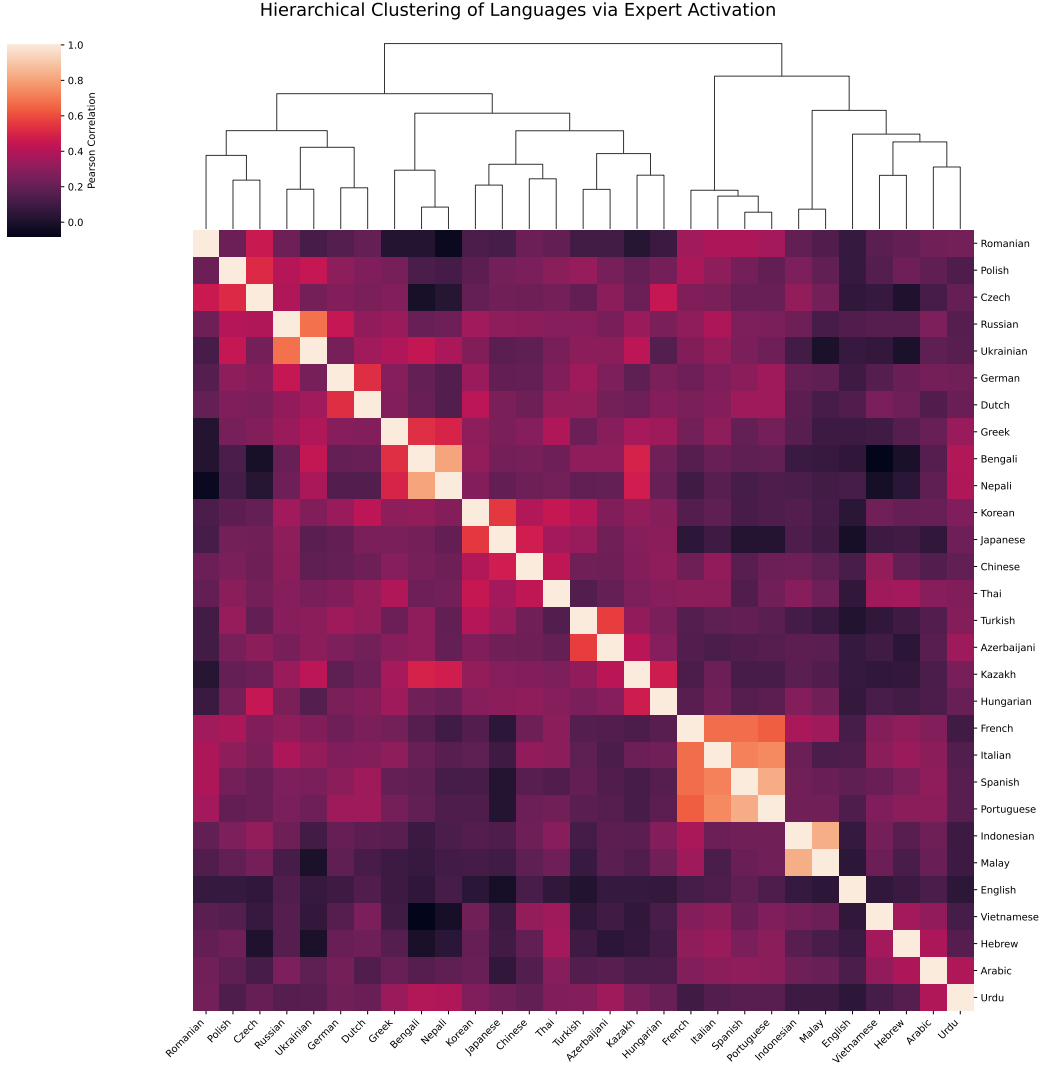


Figure 6: Hierarchical clustering of expert activation patterns, demonstrating that the model’s neural routing trajectories tightly mirror established linguistic family structures. The emergence of distinct blocks for Romance, Slavic, Germanic, and East Asian groups illustrates how the MoE architecture naturally discovers and exploits cross-lingual commonalities.

KazakhMMLU (Togmanov et al., 2025), IndoMMLU (Koto et al., 2023), IndoCareer (Koto, 2025), and IndoCulture (Koto et al., 2024b).

C Language Clustering through Expert Activation Patterns

To further corroborate that expert routing in Marco-MoE encodes linguistically meaningful structure, we perform agglomerative hierarchical clustering based on the Language-Expert Signatures shown in Figure 5. Specifically, we treat each language’s flattened activation vector $\mathbf{M}_L$ as a point in expert-activation space, adopt $1 - \rho(\mathbf{M}_{L_i}, \mathbf{M}_{L_j})$ (with ρ the Pearson correlation coefficient) as the pairwise distance between language L_i and L_j , and apply average-linkage clustering to construct the dendrogram shown in Figure 6. The resulting hierarchy tightly mirrors established linguistic taxonomies: Romance, Slavic, Germanic, and East Asian languages each coalesce into compact subtrees at small merge distances,

while at deeper levels phylogenetically related branches further converge into broader super-clusters that echo the Indo-European macro-family. Languages with isolated typology or unique scripts, in contrast, attach to the tree only at large distances, reflecting the router’s allocation of dedicated expert subsets to accommodate their distinctive morphology. This unsupervised rediscovery of language families offers mechanistic evidence that Marco-MoE has learned to exploit cross-lingual regularities in a manner that mirrors human-curated linguistic phylogenies, promoting positive transfer among related languages while isolating typologically distant ones from mutual interference.

D Post-Training Details

We employ a two-stage post-training pipeline, comprising supervised fine-tuning (SFT) and on-policy distillation (Lu & Lab, 2025) from high-capacity teacher models, to develop *general* INSTRUCT models. This pipeline produces models for low-latency responses to standard user queries by bypassing explicit, computationally intensive reasoning processes.

D.1 Supervised Fine-Tuning

D.1.1 Data

General Instruction Data. Our SFT process leverages the Dolci-Instruct dataset (Olmo et al., 2025), originally employed to train the OLMO3-Instruct models. To further enhance the instruction-following capabilities of Marco-MoE, we augment this corpus with additional data sourced from Yang et al. (2026).

Knowledge-Intensive Data. We observe that scientific data is underrepresented within the Dolci-Instruct dataset. To address this gap, we augment the corpus by sourcing scientific prompts from Yang et al. (2026) and distilling the corresponding responses from Gemini3-Flash.

Translation Data. For translation data, our data curation pipeline begins by sourcing parallel corpora from the web-mined NLLB translation dataset (Team et al., 2022). To ensure high data fidelity, we implement a multi-stage filtering strategy. First, we apply heuristic text filters adapted from Gopher (Rae et al., 2022) and FineWeb (Penedo et al., 2024) to the English segments to exclude low-quality pairs. To further enhance semantic alignment, we leverage the Qwen3-Embedding-8B model (Zhang et al., 2025) to compute relevance scores for each translation pair. Ultimately, we retain only the top 10,000 highest-scoring pairs per language.

Multilingual & Cultural Data. To generate culturally-grounded data, we leverage Wikidata (Vrandečić & Krötzsch, 2014) as our primary knowledge source within a multi-stage generation pipeline. Initially, we identify a taxonomy of topics pertinent to regional and cultural contexts. For each target language, we then filter entities within these topics, selecting only those whose "country" property aligns with the corresponding linguistic region. Utilizing these retrieved entities and their associated metadata, we employ Gemini3-Flash to synthesize coherent textual descriptions. Building on these descriptions, we adapt a back-translation approach (Li et al., 2024b) to prompt the model to derive corresponding instructions. Recognizing that descriptions derived solely from structured properties can be overly formulaic or information-sparse, we further task Gemini3-Flash with augmenting these instructions by recalling relevant background knowledge, thereby producing more natural and comprehensive responses. This process yields a high-quality, diverse dataset of instruction-response pairs tailored to the cultural nuances of each language. In addition, we augment these pairs by transforming them into multiple-choice question-answering (QA) formats with concise step-by-step chain-of-thoughts to further enhance training data diversity.

D.1.2 Implementation Details

We fine-tune Marco-MoE utilizing the SLIME framework (Zhu et al., 2025). The learning rate is initialized at 1×10^{-5} and follows a cosine decay schedule, reaching a minimum of

Category	Datasets	Mixture Ratio
Instruction Following	Nemotron-RL-instruction-following Nemotron-RL-instruction-following-structured-outputs	25%
Knowledge & Reasoning	Nemotron-RL-ReasoningGym-v1 Nemotron-RL-knowledge-mcqa	25%
Alignment	Nemotron-Cascade-RL-RLHF	10%
Math	DAPO-Math-17k Skywork-OR1-RL-Data	10%
Multilingual	Translation Multilingual & Cultural Data Nemotron-SFT-Multilingual-v1	30%

Table 8: Datasets and mixture ratios for the on-policy distillation stage.

1×10^{-6} . The model is trained for a single epoch with a global batch size of 512, totaling approximately 4,000 training steps. We set the maximum context length to 8,192 tokens. Optimization is performed using AdamW with a weight decay of 0.1, $\beta_1 = 0.9$, and $\beta_2 = 0.95$. The whole SFT process takes around 24 hours using 64 GPUs.

D.2 On-Policy Distillation from Stronger Models

On-policy learning involves training a model on samples generated by its own current policy, fundamentally enabling it to correct its own errors. While this approach is a cornerstone of reinforcement learning (RL), it is often constrained by sparse feedback; RL typically yields a limited amount of information per training episode regardless of sequence length (Schulman & Lab, 2025), resulting in significant sample inefficiency and high computational overhead. In contrast, off-policy learning seeks to *distill* the behavior of a superior teacher model. By leveraging supervision for every token produced by the teacher, off-policy methods provide substantially denser learning signals. However, this paradigm is susceptible to distribution shift: the student model is trained on trajectories native to the teacher rather than those it will encounter during its own inference. This can lead to compounding errors, where early deviations from the teacher’s distribution cause the model to diverge into unfamiliar states. To bridge these two paradigms, On-Policy Distillation (OPD) (Lu & Lab, 2025; Yang et al., 2025) utilizes trajectories sampled from the student model while incorporating the dense per-token supervision characteristic of off-policy distillation. Consequently, OPD allows the model to learn from its own generations while maintaining the high signal density required for efficient optimization.

D.2.1 Data

We employ a diverse set of data sources to post-train the SFT checkpoint using OPD, with the exact data mixtures shown in Table 8.

Instruction Following. We sample instruction-following prompts from the [Nemotron-RL-instruction-following](#) dataset. To further enhance the ability of Marco-MoE in following specific output formats and generating structured data, we further augment this set with prompts from [Nemotron-RL-instruction-following-structured-outputs](#).

Knowledge & Reasoning. We gather a diverse collection of datasets focused on reasoning and knowledge grounding. Specifically, we sample reasoning tasks from [Nemotron-RL-ReasoningGym-v1](#), which encompasses a wide array of domains, including algebra, geometry, and graph theory. Furthermore, we incorporate [Nemotron-RL-knowledge-mcqa](#), a dataset comprising knowledge-intensive multiple-choice question-answering (MCQA) samples across various scientific disciplines.

Alignment. To facilitate human preference alignment and enhance conversational proficiency, we utilize alignment prompts from the [Nemotron-Cascade-RL-RLHF](#) dataset to enhance Marco-MoE.

Math. While Marco-MoE is not explicitly optimized for mathematical reasoning, we incorporate the DAPO-Math-17k (Yu et al., 2025a) and Skywork-OR1-RL-Data (He et al., 2025b) datasets to enhance the model’s general problem-solving proficiency.

Multilingual. The multilingual corpus incorporates a subsample of our curated translation and cultural data. We further augment this with the [Nemotron-SFT-Multilingual-v1](#) dataset, which provides high-quality translations of mathematical and scientific problems.

D.2.2 Implementation Details

We implement OPD within the SLIME framework and adopt a cascaded distillation strategy that progressively transfers knowledge from increasingly stronger teacher models. Specifically, for both Marco-Nano-Instruct and Marco-Mini-Instruct models, we first use [Qwen3-30B-A3B-Instruct-2507](#) as the initial teacher. After the student model converges, we switch to the stronger [Qwen3-Next-80B-A3B-Instruct](#) and continue distillation. Training uses a constant learning rate of 1×10^{-6} and a global batch size of 512. To improve training efficiency, we sample only two responses per prompt, resulting in 1,024 sequences per training step. The maximum sequence lengths are set to 2,048 tokens for prompts and 8,192 tokens for responses. The complete OPD procedure for Marco-Nano-Instruct involves approximately 1,900 steps with [Qwen3-30B-A3B-Instruct-2507](#) followed by 1,000 steps with [Qwen3-Next-80B-A3B-Instruct](#); Marco-Mini-Instruct undergoes 1,900 steps for each phase. The total training process requires roughly 110 hours on a cluster of 64 GPUs.

D.3 Instruct Model Evaluation

D.3.1 Experimental Settings

Evaluation Benchmarks. Following base mode evaluation settings, all models are conducted using the Light-Eval framework ([Habib et al., 2023](#)) to ensure fair comparisons. The evaluation suite is divided into two primary dimensions:

1. **English Proficiency:** We assess general language understanding and reasoning via diverse benchmarks, including MMLU ([Hendrycks et al., 2021a](#)), MMLU-Redux ([Gema et al., 2025](#)), MMLU-Pro ([Wang et al., 2024](#)), AGIEval ([Zhong et al., 2024](#)), GPQA-Diamond ([Rein et al., 2023](#)), GSM8K ([Cobbe et al., 2021](#)), and MATH ([Hendrycks et al., 2021b](#)).
2. **Multilingual Capability:** To measure performance across diverse languages, we evaluate the model on both general multilingual benchmarks and datasets assessing cultural and regional knowledge. General datasets include GlobalMMLU ([Singh et al., 2025](#)), MMLU-ProX-Lite ([Xuan et al., 2025](#)), MMMLU, MGQA ([Huang et al., 2025](#)), MGSM ([Shi et al., 2023](#)), PolyMath ([Wang et al., 2025](#)) and translation tasks with FLORES-200 ([Team et al., 2022](#)) and WMT24++ ([Freitag et al., 2024](#)). For cultural and regional benchmarks, we evaluate the models on INCLUDE ([Romanou et al., 2025](#)), Global-PIQA ([Chang et al., 2025](#)), CMMLU ([Li et al., 2024a](#)), C-Eval ([Huang et al., 2023](#)), ArabicMMLU ([Koto et al., 2024a](#)), TurkishMMLU ([Yüksel et al., 2024](#)), GreekMMLU ([Zhang et al., 2026](#)), KazakhMMLU ([Togmanov et al., 2025](#)), IndoMMLU ([Koto et al., 2023](#)), IndoCareer ([Koto, 2025](#)), and IndoCulture ([Koto et al., 2024b](#)).

Baselines. We compare Marco-Nano-Instruct and Marco-Mini-Instruct against various open-source INSTRUCT models. The comparison set includes dense models: Qwen3-1.7B-Instruct (i.e., non-thinking), Qwen3-4B-Instruct-2507 ([Yang et al., 2025](#)), Qwen3-VL-2B-Instruct ([Bai et al., 2025](#)), Ministral-3-3B-Instruct-2512 & Ministral-3-8B-Instruct-2512 ([Liu et al., 2026](#)), and Gemma3-12B-Instruct ([Team et al., 2025](#)); MoE models: Granite4-Tiny & Granite4-Small ([IBM Research, 2025](#)), and LFM2-8B-A1B & LFM2-24B-A2B ([Amini et al., 2025](#)). All models are evaluated using a sampling temperature of 1.0 and a top-p value of 1.0. For most benchmarks, we generate eight responses per test instance with a maximum response length of 8,192 and report the mean accuracy. However, for GlobalMMLU and MMMLU, we only sample a single response per query to optimize computational efficiency, given the substantially larger scale of these datasets.

D.3.2 Main Results

We demonstrate the results of INSTRUCT models in Table 9.

Benchmark (Metric)	Qwen3 1.7B Instruct	Qwen3-VL 2B Instruct	Minstral3 3B Instruct	LFM2 8B-A1B Instruct	Granite4 Tiny Instruct	Marco Nano Instruct	Qwen3 4B Instruct	Minstral3 8B Instruct	Gemma3 12B Instruct	Granite4 Small Instruct	LFM2 24B-A2B Instruct	Marco Mini Instruct
# Activated Params	1.7B	2B	3.84B	1.5B	1.47B	0.6B	4B	8.8B	12B	9B	2B	0.86B
# Total Params	1.7B	2B	3.84B	8.3B	7B	8B	4B	8.8B	12B	32B	24B	17.3B
English												
MMLU (Acc)	62.4	62.1	69.8	<u>72.1</u>	50.8	73.2	<u>80.8</u>	79.8	76.2	76.7	74.9	83.4
MMLU-Redux (Acc)	62.4	62.2	69.6	<u>71.9</u>	51.2	73.3	<u>80.9</u>	79.9	76.2	76.7	74.9	83.5
MMLU-Pro (Acc)	35.2	38.3	<u>49.5</u>	<u>49.5</u>	25.3	54.5	<u>66.9</u>	63.9	55.8	57.1	57.6	70.7
AGIEval (Acc)	39.6	33.0	<u>44.7</u>	<u>45.2</u>	30.7	49.8	51.7	<u>52.4</u>	43.6	44.7	49.0	55.4
GPQA-Diamond (Acc)	27.5	21.0	<u>31.6</u>	31.9	28.3	22.2	50.8	44.8	35.2	38.6	39.7	<u>50.3</u>
GSM8K (EM)	77.9	79.7	79.0	<u>84.6</u>	71.1	86.7	88.6	89.5	<u>89.7</u>	83.9	87.2	93.1
MATH (EM)	70.6	73.7	70.2	82.6	53.4	<u>79.6</u>	93.4	86.2	83.8	75.7	83.9	<u>91.8</u>
Average	53.7	52.9	59.2	<u>62.5</u>	44.4	62.8	<u>73.3</u>	70.9	65.8	64.8	66.7	75.5
Multilingual – General												
GlobalMMLU (Acc)	46.3	45.9	38.4	<u>49.0</u>	43.0	58.7	<u>70.2</u>	55.4	69.2	67.4	57.0	73.3
MMMLU (Acc)	49.0	49.0	39.4	<u>56.5</u>	44.1	59.9	<u>71.3</u>	56.4	69.4	68.1	62.3	73.7
MMLU-ProX-Lite (Acc)	28.6	30.3	26.7	<u>33.8</u>	22.1	43.2	<u>58.3</u>	43.3	51.3	51.6	43.3	61.2
MGPQA (Acc)	25.3	22.3	18.8	<u>27.2</u>	21.6	21.6	<u>41.0</u>	30.5	32.8	35.0	32.7	41.8
FLORES-200 (En-Xx) (BLEU)	12.7	15.3	8.3	14.9	22.5	<u>22.3</u>	22.1	17.5	35.6	31.9	19.2	<u>30.6</u>
FLORES-200 (Xx-En) (BLEU)	28.2	28.6	18.9	20.1	<u>30.4</u>	31.1	33.5	31.0	40.3	32.2	22.7	<u>36.8</u>
WMT24++ (En-Xx) (BLEU)	13.2	14.6	4.4	14.6	18.9	<u>18.7</u>	20.9	14.4	32.1	26.6	16.0	<u>26.8</u>
WMT24++ (Xx-En) (BLEU)	<u>26.4</u>	26.2	8.3	17.9	25.1	27.3	29.9	24.2	35.5	27.5	18.8	<u>31.3</u>
MGSM (EM)	63.6	<u>67.6</u>	47.0	56.5	55.3	76.5	<u>84.4</u>	68.7	84.0	75.7	67.8	87.4
PolyMath (EM)	23.4	25.5	16.3	<u>26.5</u>	18.7	29.6	47.2	26.4	35.5	28.9	29.3	<u>44.7</u>
Average	31.7	<u>32.5</u>	22.7	31.7	30.6	38.9	47.9	36.8	<u>48.6</u>	44.5	36.9	50.8
Multilingual – Cultural & Regional												
INCLUDE (Acc)	<u>44.9</u>	44.4	35.4	43.5	38.6	54.3	63.8	50.7	<u>65.0</u>	60.3	49.1	65.6
Global-PIQA (Acc)	62.0	<u>65.8</u>	50.6	60.8	63.3	70.7	79.6	61.3	<u>82.2</u>	80.2	69.0	84.2
CMMLU (Acc)	<u>60.4</u>	63.3	48.9	52.7	39.2	60.0	78.6	67.4	60.8	59.6	56.7	<u>75.3</u>
C-Eval (Acc)	58.7	63.2	50.6	50.8	39.4	<u>60.8</u>	80.4	68.0	59.7	59.4	56.7	<u>75.4</u>
ArabicMMLU (Acc)	<u>48.8</u>	46.9	22.7	56.5	43.4	56.5	66.0	41.4	70.1	66.3	61.3	<u>67.8</u>
TurkishMMLU (Acc)	<u>42.7</u>	39.6	38.6	26.3	31.6	59.9	<u>71.6</u>	48.2	64.4	57.9	33.4	74.7
GreekMMLU (Acc)	<u>48.7</u>	48.0	38.4	40.0	44.8	61.6	68.6	49.5	77.7	71.7	44.7	<u>72.5</u>
KazakhMMLU (Acc)	46.0	<u>47.1</u>	41.4	39.6	39.6	56.3	66.6	59.1	<u>66.8</u>	63.5	47.6	68.8
IndoMMLU (Acc)	48.8	<u>49.3</u>	35.2	41.1	37.2	56.3	64.4	52.4	<u>65.3</u>	59.6	42.7	65.7
IndoCareer (Acc)	<u>46.1</u>	45.7	36.0	41.7	34.7	54.9	62.2	53.4	<u>63.2</u>	56.3	43.7	64.4
IndoCulture (Acc)	45.8	<u>47.7</u>	37.2	45.9	42.8	59.1	58.7	47.8	69.6	59.3	44.2	<u>67.1</u>
Average	50.3	<u>51.0</u>	39.5	45.4	41.3	59.1	<u>69.1</u>	54.5	67.7	63.1	49.9	71.0

Table 9: Comparison among Marco-Nano-Instruct & Marco-Mini-Instruct and other open-source Instruct models of comparable sizes. The best and second best results are marked in **bold** and underline. Avg@8 accuracies are reported, except for GlobalMMLU and MMMLU where Acc@1 is reported.

English Results. On English benchmarks, Marco-Mini-Instruct achieves an average score of 75.5, effectively surpassing Qwen3-4B-Instruct (73.3) and significantly outperforming Minstral3-8B-Instruct (70.9) and Gemma3-12B-Instruct (65.8), despite utilizing a fraction of the activated parameters. This efficiency is even more pronounced when compared to other sparse models; Marco-Mini-Instruct significantly outperforms LFM2-24B-A2B (66.7) and Granite-4-Small (64.8), even though these baselines activate 2B and 9B parameters, respectively. Similarly, Marco-Nano-Instruct defines a highly efficient frontier, achieving an average score of 62.8. It notably surpasses LFM2-8B-A1B (62.5) while requiring only 40% of its activated parameter budget (0.6B *v.s.* 1.5B).

Multilingual General Results. The advantages of the Marco-MoE models are most pronounced in general multilingual capabilities, where Marco-Mini-Instruct ranks first among all evaluated models with an average score of 50.8. It not only surpasses Qwen3-4B-Instruct but also demonstrates nearly double the efficiency of comparable MoE competitors. For instance, Marco-Mini-Instruct outperforms LFM2-24B-A2B by an absolute point of 12.4, despite having 38% fewer total parameters and 57% fewer activated parameters. Marco-Nano-Instruct also shows remarkable cross-lingual transfer, achieving 38.9 and outperforming not only baselines with comparable size but also the significantly larger LFM2-24B-A2B variant.

Multilingual Cultural and Regional Results. Regarding localized knowledge and culturally grounded reasoning, Marco-Mini-Instruct emerges as the top-performing model with an average of 71.0, outperforming Qwen3-4B-Instruct (69.1) and the substantially larger Gemma3-12B-Instruct (67.7). The performance gap between Marco-MoE and other MoE baselines widens substantially in this category, with Marco-Mini-Instruct outperforming LFM2-24B-A2B-Instruct (49.9) by 21.1 points. Even the compact Marco-Nano-Instruct (59.1) demonstrates superior cultural proficiency relative to both LFM2-8B-A1B and LFM2-24B-

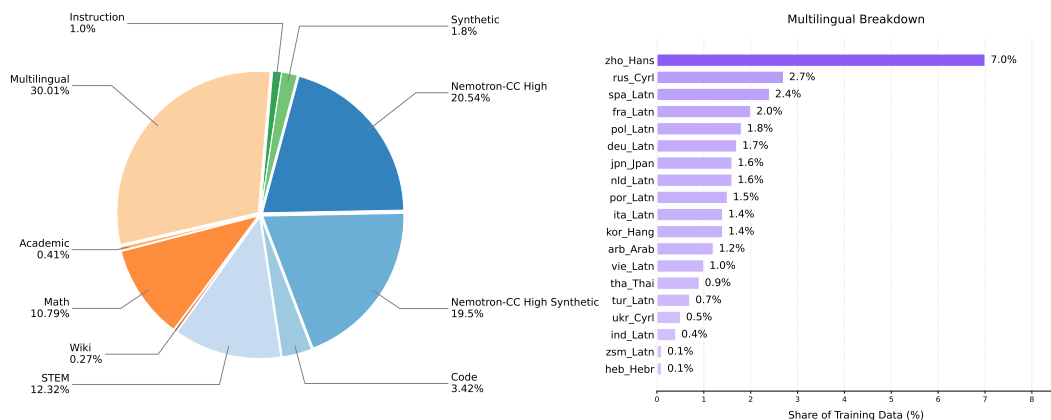
	Marco-Nano-Instruct			Marco-Mini-Instruct		
	Stage-1	Stage-2	Δ	Stage-1	Stage-2	Δ
English	60.5	62.8	+2.3	73.7	75.5	+1.8
General Multilingual	38.0	38.9	+0.9	49.3	50.8	+1.5
Cultural & Regional	58.6	59.1	+0.5	69.9	71.0	+1.1

Table 10: Performance progression across the two-stage cascaded OPD. The Δ indicates the absolute improvement from Stage-1 (utilizing the 30B-A3B teacher) to Stage-2 (transitioning to the 80B-A3B teacher) across three benchmark categories.

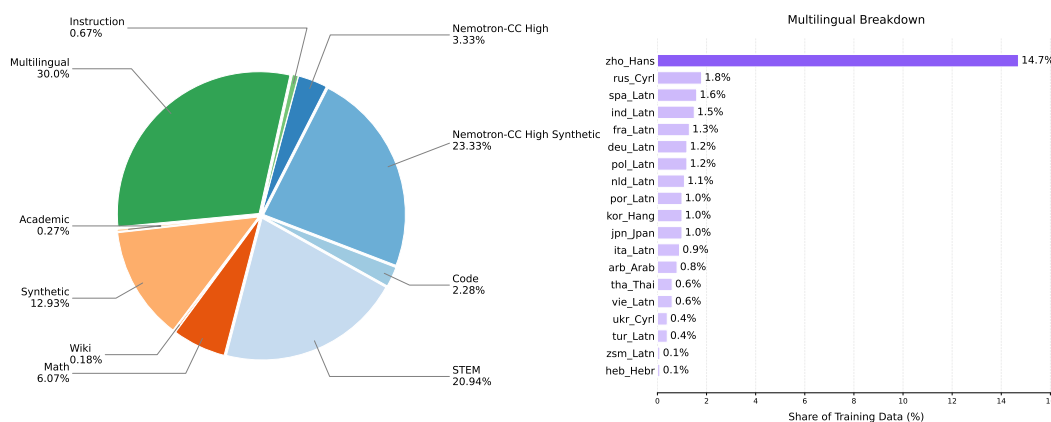
A2B models. Nevertheless, Marco-MoE models still lag behind Qwen3 models in highly localized Chinese knowledge reasoning, as evidenced by lower relative performance on the CMMLU and C-Eval benchmarks.

D.3.3 The Effects of Cascaded Distillation

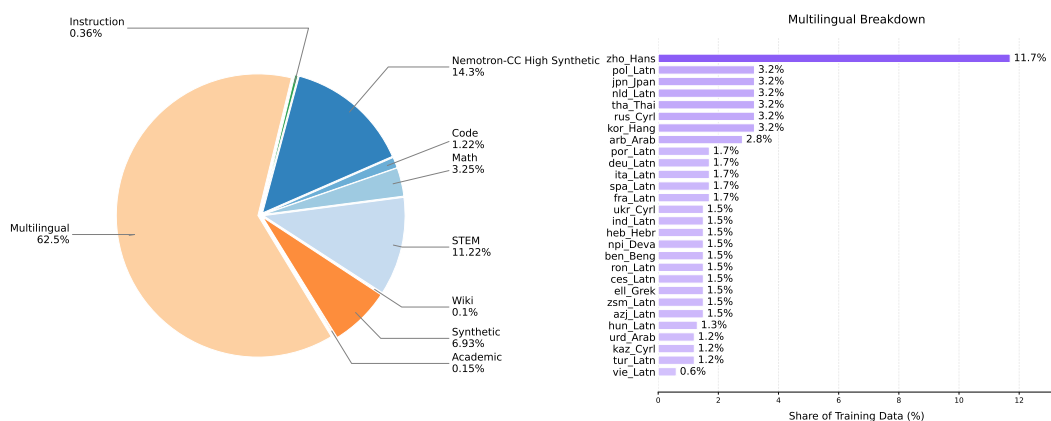
Table 10 demonstrate the efficacy of our cascaded distillation strategy, which progressively transfers knowledge from increasingly capable teacher models. Both Marco-Nano-Instruct and Marco-Mini-Instruct models achieve consistent gains across all evaluation axes upon transitioning from the 30B-A3B to the 80B-A3B teacher. This progressive improvement validates that the student models have not reached their capacity limits after the first stage. Instead, OPD from a stronger teacher provides higher-quality reward signals that allow the model to correct more sophisticated errors.



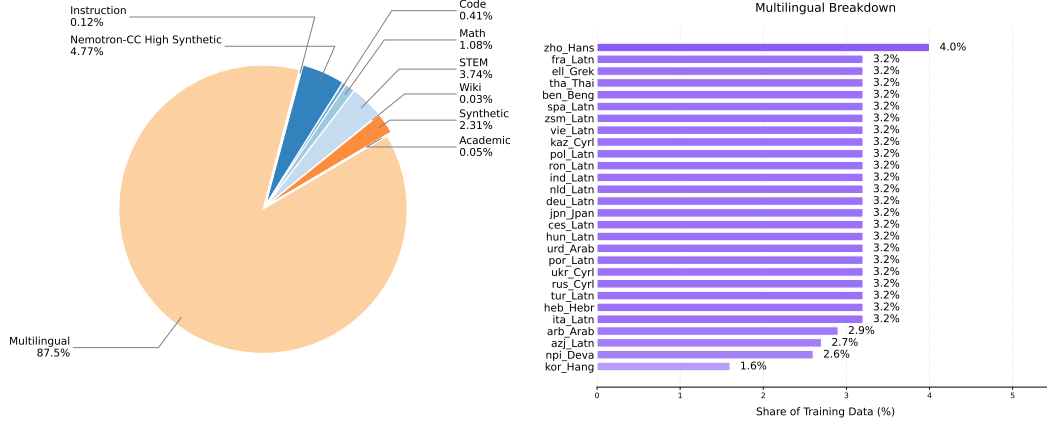
(a) Data mixture of Phase 1.



(b) Data mixture of Phase 2.



(c) Data mixture of Phase 3.



(d) Data mixture of Phase 4.

Figure 7: Data mixtures for each pre-training phase.

Dataset	Phase1-Mix	Phase2-Mix	Phase3-Mix	Phase4-Mix
Nemotron-CC-v2/High-Quality	20.54%	6.5%	-	-
Nemotron-CC-v2/Diverse-QA	18.5%	18.33%	14.13%	6.5%
Nemotron-CC-v2/High-Quality-Synthetic	0.94%	5.17%	-	-
Nemotron-Pretraining-SFT-v1/Nemotron-SFT-General	6.0%	2.02%	1.18%	0.30%
Nemotron-Pretraining-Specialized-v1/Nemotron-Pretraining-STEM-SFT	2.61%	7.70%	4.50%	1.15%
Nemotron-Pretraining-Specialized-v1/Nemotron-Pretraining-RQA	2.44%	10.37%	6.06%	1.56%
Nemotron-Pretraining-Specialized-v1/Nemotron-Pretraining-Wiki-Rewrite	0.27%	0.17%	0.1%	0.026%
Nemotron-Pretraining-Specialized-v1/Nemotron-Pretraining-InfiniByte-Reasoning	1.0%	0.64%	0.37%	0.096%
Nemotron-Pretraining-Specialized-v1/Nemotron-Pretraining-Math-Textbooks	0.87%	0.70%	0.41%	0.10%
Nemotron-CC-v2.1/High-Quality-DQA	0.27%	0.17%	0.1%	0.026%
peS2o	0.41%	0.26%	0.15%	0.039%
Nemotron-Pretraining-SFT-v1/Nemotron-SFT-MATH	6.93%	4.45%	2.60%	0.67%
Nemotron-CC-Math-v1/4plus	1.30%	0.69%	0.40%	0.10%
OpenMathInstruct-2	0.50%	2.25%	1.31%	0.37%
Finemath/finemath-4plus	0.33%	-	-	-
Finemath/finwebmath-4plus	0.30%	-	-	-
Megamath/megamath-web-pro	0.45%	-	-	-
Megamath/megamath-qa/qwen2.5	0.11%	-	-	-
Nemotron-Pretraining-SFT-v1/Nemotron-SFT-Code	1.90%	1.22%	0.71%	0.18%
Megamath/megamath-text-code-block	1.43%	0.92%	0.54%	0.14%
Nemotron-Pretraining-Specialized-v1/Nemotron-Pretraining-Scientific-Coding	0.04%	0.026%	0.015%	0.004%
Stackexchange	0.04%	0.028%	0.016%	0.004%
OpenThoughts3-1.2M	1.80%	2.69%	1.57%	0.40%
FLAN	1.0%	0.64%	0.37%	0.096%
Fineweb-Edu-Chinese-V2.1	6.69%	14.74%	10.0%	0.89%
Fineweb-2 & Fineweb-2-HQ	13.99%	9.15%	30.0%	-
Nemotron-CC-v2/Translated-Diverse-QA	9.33%	6.11%	22.5%	14.3%
Dolma3-dolmino-mix-100B-1125	-	5.06%	2.96%	0.79%
Fineweb2-Culture	-	-	-	25.0%
Translated-Nemotron-SFT-General	-	-	-	25.0%
Translated-OpenMathInstruct-2	-	-	-	12.5%
Synthetic-Regional-MCQs	-	-	-	9.82%

Table 11: Pre-training datasets and mixture ratios for the four pre-training stages.

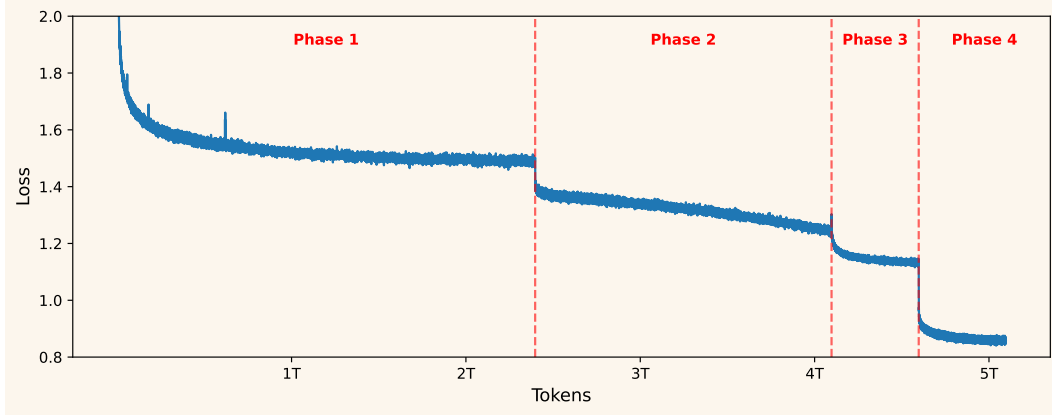


Figure 8: The training loss graph for Marco-Mini-Base. We indicate the boundaries between training phases where data mixtures are switched.

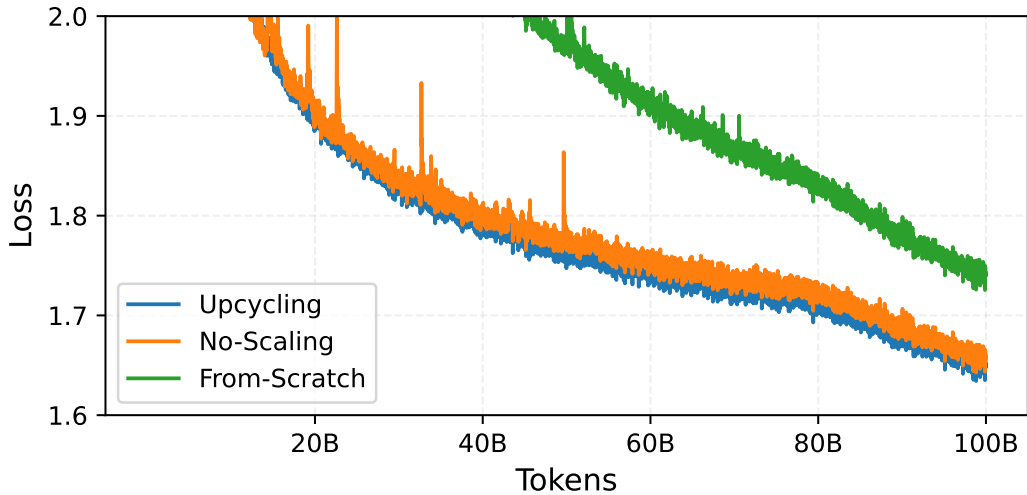


Figure 9: Ablation on weight scaling and upcycling. We train MoE models with 0.8B active, 17B total parameters on 100B tokens. Upcycling significantly speeds up pre-training, leading to faster convergence. Weight scaling on expert parameters helps reduce loss spikes while achieving similar final training loss.

Benchmark (Metric)	Stage-1	Stage-2	Stage-3	Stage-4
English				
MMLU _(Acc)	66.1	72.8	73.5	72.8
MMLU-Redux _(Acc)	63.0	69.4	68.9	68.8
MMLU-Pro _(Acc)	35.6	42.7	44.8	45.3
AGIEval _(Acc)	37.5	40.6	41.9	41.9
BBH _(EM)	55.6	63.7	62.0	65.1
ARC-Easy _(Acc)	80.1	82.5	82.7	82.4
ARC-Challenge _(Acc)	53.1	57.4	57.4	56.3
HellaSwag _(Acc)	73.0	77.6	77.5	77.4
WinoGrande _(Acc)	55.5	59.0	58.0	57.7
BoolQ _(Acc)	74.5	74.9	76.1	74.2
CommonsenseQA _(Acc)	59.1	61.3	61.5	61.5
OpenBookQA _(Acc)	44.2	44.6	44.8	44.6
PIQA _(Acc)	79.2	79.9	80.7	81.1
SIQA _(Acc)	47.9	50.6	49.0	49.4
GSM8K _(EM)	66.3	75.0	75.9	76.4
Average	59.4	63.5	63.7	63.7
Multilingual – General				
GlobalMMLU _(Acc)	52.4	59.5	62.6	64.2
MMMLU _(Acc)	51.5	59.3	60.5	62.0
MMLU-ProX-Lite _(Acc)	29.2	34.2	38.2	39.2
BELEBELE _(Acc)	63.4	70.5	80.1	79.8
mHellaSwag _(Acc_norm)	51.4	55.2	58.2	58.6
mARC-Challenge _(Acc_norm)	39.1	42.3	44.8	45.4
FLORES-200 (En-Xx) _(BLEU)	22.3	25.6	31.7	32.3
FLORES-200 (Xx-En) _(BLEU)	30.2	34.5	40.0	40.1
WMT24++ (En-Xx) _(BLEU)	19.8	23.8	27.7	28.1
WMT24++ (Xx-En) _(BLEU)	26.5	30.9	34.4	34.4
MGSM _(EM)	61.0	73.1	75.1	75.6
Average	40.6	46.3	50.3	52.1
Multilingual – Cultural & Regional				
INCLUDE _(Acc)	51.5	56.7	61.0	61.7
Global-PIQA _(Acc_norm)	64.9	66.6	70.6	72.3
CMMLU _(Acc)	58.6	63.8	64.2	68.0
C-Eval _(Acc)	57.1	62.9	63.5	66.0
ArabicMMLU _(Acc)	60.3	65.8	66.6	67.1
TurkishMMLU _(Acc)	50.1	60.1	61.2	62.7
GreekMMLU _(Acc)	41.8	51.6	68.1	70.3
KazakhMMLU _(Acc)	47.9	54.3	60.5	62.6
IndoMMLU _(Acc)	51.1	56.3	59.1	59.9
IndoCareer _(Acc)	54.8	60.1	60.3	61.5
IndoCulture _(Acc)	55.6	60.0	60.5	62.3
Average	54.0	59.8	63.2	65.0

Table 12: Per-benchmark model performance across the four pre-training stages.

Benchmark (Metric)	# Shots	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Marco Mini Global Base
# Activated Params	-	4B	3.35B	4B	0.86B
# Total Params	-	4B	3.35B	4B	17.3B
# Train FLOPs ($\times 10^{23}$)	-	0.96	1.21	8.64	1.584
English					
MMLU (Acc)	5-shot	61.1	58.6	75.2	72.9
MMLU-Redux (Acc)	0-shot	57.7	51.7	71.3	68.9
MMLU-Pro (Acc)	5-shot	28.8	26.9	45.9	44.5
AGIEval (Acc)	0-shot	32.6	29.0	44.0	41.0
BBH (EM)	3-shot	52.2	46.8	72.3	65.0
ARC-Easy (Acc)	0-shot	82.6	76.5	75.0	82.4
ARC-Challenge (Acc)	0-shot	54.1	47.4	49.9	57.0
HellaSwag (Acc)	0-shot	76.7	71.0	74.4	77.2
WinoGrande (Acc)	0-shot	61.4	56.6	59.6	58.3
BoolQ (Acc)	0-shot	76.6	74.6	74.2	75.6
CommonsenseQA (Acc)	0-shot	61.1	60.4	52.9	61.2
OpenBookQA (Acc)	0-shot	42.6	40.4	42.6	45.0
PIQA (Acc)	0-shot	80.3	76.9	77.4	80.7
SIQA (Acc)	0-shot	50.4	49.9	53.0	48.4
GSM8K (EM)	5-shot	39.3	58.0	81.7	76.4
Average	-	57.2	55.5	63.3	63.6
Multilingual – General					
GlobalMMLU (Acc)	5-shot	49.1	48.4	57.8	60.9
MMMLU (Acc)	0-shot	45.0	42.8	54.8	58.2
MMLU-ProX-Lite (Acc)	5-shot	23.3	23.5	35.6	36.2
BELEBELE (Acc)	0-shot	62.3	62.5	74.0	76.0
mHellaSwag (Acc_norm)	0-shot	51.9	50.3	48.5	54.4
mARC-Challenge (Acc_norm)	0-shot	39.3	35.7	39.3	41.2
FLORES-200 (En-Xx) (BLEU)	5-shot	27.9	25.6	25.8	29.5
FLORES-200 (Xx-En) (BLEU)	5-shot	39.2	37.2	33.4	40.2
WMT24++ (En-Xx) (BLEU)	5-shot	26.0	24.4	19.6	26.0
WMT24++ (Xx-En) (BLEU)	5-shot	34.4	32.9	31.2	34.5
MGSM (EM)	8-shot	35.7	36.6	69.1	71.7
Average	-	39.5	37.3	44.5	48.1
Multilingual – Cultural & Regional					
INCLUDE (Acc)	5-shot	52.3	53.5	60.0	61.1
Global-PIQA (Acc_norm)	0-shot	67.8	66.7	61.8	70.2
CMMLU (Acc)	5-shot	50.2	58.8	76.2	67.9
C-Eval (Acc)	5-shot	48.5	57.6	76.6	66.2
ArabicMMLU (Acc)	3-shot	61.6	63.2	67.0	66.6
TurkishMMLU (Acc)	5-shot	43.7	45.2	60.6	63.1
GreekMMLU (Acc)	5-shot	63.4	66.3	69.4	70.4
KazakhMMLU (Acc)	5-shot	52.1	47.1	62.3	61.8
IndoMMLU (Acc)	0-shot	48.5	52.0	60.1	59.5
IndoCareer (Acc)	3-shot	53.4	56.6	61.5	61.8
IndoCulture (Acc)	0-shot	59.1	58.5	61.1	62.5
Average	-	54.6	56.9	65.1	64.7

Table 13: Per-benchmark model performance comparison when scaling up the supporting languages to 64.

Translation Prompt

Translate the following text into [Language], ensuring that all LaTeX symbols, code snippets, and any related formatting structures remain unchanged and intact. Only translate the non-code and non-symbol portions of the text, preserving all technical formatting exactly as presented. Pay careful attention to maintaining the original layout and ensuring the translated text integrates seamlessly with non-translated elements.

[Document]

Regional Web Document Annotation Prompt**System Prompt:**

You are given a [Document] from the web.

Your goal is to:

[Document]: {}

Task 1: Training Value Score (1–10)

Analyze the content and assign a score based on quality and trustworthiness for LLM training:

1. **10 (Pristine):** Factually impeccable, encyclopedic (e.g., peer-reviewed papers).
2. **9 (Excellent):** Expert-level, well-researched (e.g., reputable news).
3. **8 (Very Good):** Reliable, solid blog posts or documentation.
4. **7 (Good):** Standard "how-to" guides, clear information.
5. **6 (Acceptable):** Average, coherent but superficial (e.g., forum posts).
6. **5 (Borderline):** Mixed quality, requires heavy cleaning.
7. **4 (Low Quality):** Problematic, high fluff, questionable facts.
8. **3 (Very Low Quality):** SEO spam, keyword stuffing, machine-generated.
9. **2 (Harmful/Junk):** Hate speech, dangerous theories, malicious code.
10. **1 (Toxic):** Unintelligible, entirely devoid of value.

Task 2: Subject Classification

Subject Classification: Assign a subject/category to the document based on its content. Choose from the following categories:

- Accounting - Agriculture - Anthropology - Architecture and Design - Arts & Humanities - Biology - Business administration - Business ethics - Business - Chemistry - Computer Science - Culturology - Earth science - Economics - Education - Engineering - Environmental studies and forestry - Family and consumer science - Finance - Geography - Health - History - Human physical performance and recreation - Industrial and labor relations - International trade - Journalism - media studies - and communication - Language - Law - Library and museum studies - Literature - Logic - Management - Marketing - Math - Medicine - Military Sciences - Multiple exams - Performing arts - Philosophy - Physics - Political sciences - Psychology - Public Administration - Public Policy - Qualimetry - Religious studies - Risk management and insurance - Social work - Sociology - STEM - Transportation - Visual Arts - Driving License - Marine License - Medical License - Professional Certifications - Others

Task 3: Country/Region Classification

Decide if the content is culturally specific (e.g., Japan) or "No specific country."

Output Format:

1. Text Quality Score (1-10):
2. Subject Category:
3. Country/Region:

Only generate the final result without additional descriptions.

Regional MCQs Generation Prompt

Write a question on [country], more specifically on [sub_topic], with up to four choices.

Desired difficulty level: [difficulty]. The required language: [language]

- For questions marked "intermediate", the question should require applying basic concepts independently or in simple combinations. The relationships between concepts should be straightforward and directly tied to the problem. Solutions should be accessible with a foundational understanding and methodical effort.

- For questions marked "complex", the question should require identifying and combining multiple clear but non-obvious connections between concepts. The relationships may span different areas within the topic, requiring structured reasoning and careful application

of techniques. Focus on recognizing patterns and synthesizing information rather than straightforward computation.

- For questions marked "hard", the question should involve intricate or abstract relationships between concepts. Solving them should demand advanced reasoning skills, mastery of topic-specific techniques, and the ability to deal with ambiguities or edge cases. Solutions often require navigating multiple layers of logic or constraints.

- For questions marked "brutal", the question should push problem-solving to the limit, requiring mastery of nuanced relationships across diverse concepts. It should demand precise reasoning, significant abstraction, and fluency in advanced techniques. Problems may include nested or multi-step dependencies, where overlooking a detail can derail the solution.

- For questions marked "borderline unsolvable", the question should challenge even experts, involving exceptionally intricate or unconventional problem structures. It should combine deep abstraction, subtle constraints, and highly interconnected concepts. While solutions exist, they often require extensive expertise and innovative approaches. Avoid speculative or unsolvable topics (e.g., "Does God exist?" or "How can time travel be achieved?").

Ensure that one of the four choices is correct. Your question must include a valid answer within the given choices; you should not specify which one is correct.

Your response should not contain any headers, explanations, or your answer to the question.

Question: [Question]

(A): [Option A]

(B): [Option B]

(C): [Option C]

(D): [Option D]

Regional MCQ Example

Question: 在《素问·痹论》“痛者，寒气多也，有寒故痛也”的理论基础上，若患者表现为关节剧烈疼痛、如锥刺刀割，且痛处固定不移，遇寒则痛甚，得热则稍减，局部皮色不红、触之不温，甚至出现明显的“筋脉拘急”导致关节屈伸不利。从寒邪“收引凝滞”的病理特性及其对“营卫气血运行受阻”的深度逻辑推导，下列关于“痛痹（寒痹）”发病机制的描述最准确且全面的是：

- A: 寒邪凝滞导致血流减慢，其深层逻辑在于寒邪易于诱发“血瘀化热”的病理转型。因此，痛痹的治疗重点应在于清热化瘀，以缓解因热扰筋脉所致的局部红肿与活动受限。
- B: 寒为阴邪，极易损伤人体阳气，导致卫外之气涣散，使得风邪乘虚深袭经络。其病机关键在于风寒相搏引起的“气行游走”，导致关节局部出现一过性的拘挛与酸楚感。
- C: 寒邪性主收引，入经则使经脉绌急，入络则使血行凝滞；其剧烈疼痛的根源不仅在于“凝则不通”，更在于寒邪遏阻阳气，使营卫之气在受阻部位产生激烈的“郁遏交争”与脉络极度挛缩。
- D: 寒邪侵袭导致阳气受损，其病理重点在于肾阳虚衰而不能温煦经脉，导致水饮内停并流注于骨节。临床辨证核心应聚焦于“水饮溢于经络”所致的关节肿胀与沉重感。

Answer: 根据《素问·痹论》的理论及中医病理逻辑，对“痛痹（寒痹）”的发病机制进行如下推理：

1. 寒邪的特性分析：

- 收引 (Constriction): 寒邪侵袭，导致经络、筋脉缩急。正如文中所述“筋脉拘急”，这是因为寒性收缩，使脉道、筋膜挛缩，从而出现关节屈伸不利。
- 凝滞 (Stagnation): 寒则气血凝涩不通，“不通则痛”。寒邪入于脉络，使血流减慢甚至瘀阻，形成固定不移的剧烈疼痛。

2. 对营卫气血的影响：

- 营卫受阻：营行脉中，卫行脉外。寒邪侵入后，卫阳被遏，营血凝滞。
- 郁遏交争：剧烈疼痛（如锥刺刀割）的深层逻辑在于，寒邪作为阴邪遏制了局部阳气的运行，阳气试图冲破寒邪的阻碍，在受阻部位产生剧烈的搏击、交争（即“郁遏交争”），加之脉络极度挛缩，导致疼痛感远重于其他类型的痹症。

3. 选项辨析：

- A项：误导在于“血瘀化热”和“清热化瘀”。虽然长期瘀滞可能化热，但题目明确描述为“皮色不红、触之不温、遇寒痛甚”，属于典型的寒证，治疗应温经散寒，而非清热。
- B项：误导在于“风邪深袭”和“气行游走”。“游走”是风痹（行痹）的特点，而本题强调“痛处固定不移”，是寒邪凝滞的表现。

- **C项**：精准地抓住了寒邪“收引”导致“经脉绌急”、“凝滞”导致“血行不畅”的要点。同时，它深入解释了剧烈疼痛的根源在于寒邪遏阻阳气，使营卫在局部产生激烈的郁遏交争，符合“痛者，寒气多也”的深度病理逻辑。
- **D项**：误导在于“水饮内停”和“关节肿胀沉重”。这描述的是湿痹（着痹）的特征，而非痛痹。

C

Language	GlobalMMLU	MMMLU	MMLU-ProX-Lite	BELEBELE	mHellaSwag	mARC-Challenge	FLORES-200	WMT24++	MGSM	INCLUDE	Global-PIQA
eng_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
zho_Hans	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
por_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
fra_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
rus_Cyrl	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
deu_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ita_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
spa_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
jpn_Jpan	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
kor_Hang	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
arb_Arab	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ind_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
vie_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ron_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
tha_Thai	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
zsm_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ces_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
nld_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ukr_Cyrl	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
pol_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
tur_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ell_Grek	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
hun_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
heb_Hebr	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ben_Beng	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
npi_Deva	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
urd_Arab	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
kaz_Cyrl	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
azj_Latn	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 14: Languages included in each of the evaluated benchmark.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	45.5	36.3	51.6	49.0	39.3	49.5	48.1	48.5	<u>58.2</u>	44.0	62.9
ben_Beng	39.4	33.0	47.2	43.1	34.0	31.6	46.1	44.9	<u>50.6</u>	38.1	55.6
ces_Latn	48.2	42.2	56.5	51.6	42.9	41.8	51.3	50.2	<u>61.8</u>	51.7	64.4
deu_Latn	52.6	47.5	58.1	54.7	45.8	53.3	52.8	51.6	65.3	57.6	<u>64.9</u>
eng_Latn	61.6	61.7	65.2	62.9	54.6	60.1	58.7	56.6	71.8	68.5	<u>71.4</u>
fra_Latn	54.0	51.0	59.6	55.6	47.1	54.4	53.3	51.8	66.3	60.6	<u>66.1</u>
heb_Hebr	41.2	35.5	48.4	47.2	35.7	33.5	46.7	47.6	<u>51.9</u>	43.1	60.8
ind_Latn	51.8	44.3	55.8	53.3	44.9	45.8	52.0	51.1	<u>63.9</u>	52.8	65.6
ita_Latn	53.9	48.2	59.0	55.6	47.0	54.8	54.1	52.5	<u>66.0</u>	57.7	66.8
jpn_Jpan	49.7	43.3	53.7	51.4	41.1	48.0	48.6	49.0	<u>60.7</u>	53.2	63.4
kor_Hang	47.5	40.0	52.2	49.9	40.2	46.9	48.1	49.6	<u>59.3</u>	49.9	61.8
nld_Latn	50.7	46.3	58.0	54.3	45.6	46.3	52.6	50.4	<u>63.9</u>	54.6	66.4
npi_Deva	37.5	32.4	47.8	43.6	34.5	35.9	45.2	45.5	<u>49.7</u>	39.9	54.7
pol_Latn	48.5	42.5	53.7	51.8	42.9	42.3	50.7	48.8	<u>60.9</u>	51.3	65.0
por_Latn	55.0	50.1	59.7	56.3	48.0	54.9	53.6	52.8	<u>66.2</u>	60.3	66.8
ron_Latn	50.1	45.4	56.4	53.4	44.7	45.1	52.4	50.4	<u>63.5</u>	53.8	65.3
rus_Cyrl	50.8	44.5	55.9	53.3	43.5	51.4	51.4	49.7	<u>64.1</u>	55.3	66.4
spa_Latn	55.0	50.5	60.1	56.3	48.2	55.3	54.3	53.0	<u>66.9</u>	60.7	67.5
tur_Latn	46.0	38.1	49.9	48.3	41.6	38.8	49.1	48.2	<u>57.7</u>	46.9	62.5
ukr_Cyrl	47.1	40.5	53.7	50.5	41.8	43.1	49.9	48.0	<u>60.7</u>	50.0	64.7
vie_Latn	49.9	39.8	49.7	51.7	43.7	46.8	49.8	48.9	<u>61.1</u>	49.5	63.9
zho_Hans	56.0	49.4	54.8	54.1	44.1	50.2	50.8	50.8	66.0	58.6	<u>65.1</u>
zsm_Latn	48.1	41.4	53.1	52.7	42.6	43.3	49.8	49.2	<u>60.6</u>	50.6	64.6
Avg	49.6	43.6	54.8	52.2	43.2	46.7	50.8	50.0	<u>61.6</u>	52.6	64.2

Table 15: Per-language model performance comparison on GlobalMMLU.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	42.9	32.7	47.9	48.6	39.7	47.1	43.7	42.8	<u>53.7</u>	41.1	60.3
ben_Beng	37.3	30.0	44.0	44.6	34.6	32.2	39.9	37.9	<u>46.3</u>	36.8	55.1
deu_Latn	50.5	43.2	54.0	54.0	46.8	50.4	49.7	46.1	<u>62.0</u>	53.1	63.9
fra_Latn	51.1	46.2	56.0	54.6	47.7	51.8	50.4	45.8	<u>62.8</u>	57.6	63.3
ind_Latn	49.2	41.4	52.0	53.9	45.1	43.9	49.1	45.2	<u>60.3</u>	50.8	62.4
ita_Latn	51.5	44.3	55.5	55.9	46.4	51.2	50.5	46.0	<u>63.2</u>	54.7	64.1
jpn_Jpan	47.4	39.2	50.4	51.2	41.6	46.3	42.7	44.0	<u>58.3</u>	49.0	61.2
kor_Hang	45.8	38.1	49.9	49.9	41.2	44.9	45.7	43.6	<u>55.3</u>	46.3	60.0
por_Latn	52.5	46.2	56.2	56.0	47.7	51.9	51.1	46.5	<u>61.9</u>	56.9	63.8
spa_Latn	52.1	47.6	56.2	55.9	48.3	52.8	52.0	46.1	<u>63.9</u>	57.3	65.1
zho_Hans	54.3	44.0	53.2	54.2	44.7	47.9	47.2	45.7	64.0	56.1	<u>62.7</u>
Avg	48.6	41.2	52.3	52.6	44.0	47.3	47.5	44.5	<u>59.2</u>	50.9	62.0

Table 16: Per-language model performance comparison on MMMLU.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	69.1	52.2	63.4	73.2	60.6	62.9	66.3	63.8	<u>81.8</u>	69.2	<u>82.9</u>
azj_Latn	55.7	37.3	47.9	65.3	45.1	35.6	55.4	44.2	<u>69.6</u>	48.3	<u>73.7</u>
ben_Beng	52.7	38.0	54.4	63.8	46.8	33.2	58.0	60.7	<u>70.1</u>	53.0	<u>71.7</u>
ces_Latn	68.9	55.2	62.6	78.7	61.0	52.8	66.6	67.7	<u>83.2</u>	69.4	<u>82.4</u>
deu_Latn	73.1	64.8	66.1	80.3	64.6	64.3	69.1	71.0	<u>87.0</u>	76.0	<u>83.9</u>
ell_Grek	66.6	43.2	61.3	75.0	62.0	63.2	68.6	69.7	<u>80.2</u>	68.1	<u>82.2</u>
eng_Latn	81.4	76.9	75.1	81.9	74.8	71.0	74.8	82.2	<u>89.2</u>	84.3	<u>87.1</u>
fra_Latn	76.3	66.9	68.8	79.3	68.2	63.8	70.6	72.1	<u>87.2</u>	77.3	<u>84.2</u>
heb_Hebr	62.0	44.7	57.7	71.2	49.2	44.4	64.3	65.3	<u>74.3</u>	62.6	<u>78.2</u>
hun_Latn	63.3	48.1	56.1	71.6	59.2	42.4	64.6	61.7	<u>79.9</u>	61.2	<u>78.2</u>
ind_Latn	71.9	59.7	60.6	75.1	62.2	57.2	67.0	68.2	<u>83.1</u>	70.1	<u>79.2</u>
ita_Latn	73.7	63.6	66.8	78.0	67.2	62.7	68.0	65.7	<u>85.3</u>	73.7	<u>80.6</u>
jpn_Jpan	69.6	55.7	59.7	69.8	58.1	55.4	63.2	65.6	<u>79.0</u>	73.3	<u>76.1</u>
kaz_Cyrl	50.9	34.2	45.0	61.6	42.6	32.1	56.1	33.7	<u>67.0</u>	39.9	<u>71.7</u>
kor_Hang	71.4	56.9	65.4	74.4	63.9	61.4	68.4	69.1	<u>83.6</u>	74.9	<u>80.1</u>
nld_Latn	69.3	60.3	64.1	77.0	64.1	54.6	68.6	69.7	<u>84.0</u>	71.6	<u>81.7</u>
npi_Deva	48.1	37.4	51.6	60.2	43.6	36.4	56.7	58.9	<u>67.1</u>	50.4	<u>70.6</u>
pol_Latn	68.4	53.1	56.1	74.4	59.1	51.9	64.9	63.0	<u>83.3</u>	67.1	<u>80.6</u>
por_Latn	75.1	68.0	69.4	77.9	68.4	66.6	71.2	70.4	<u>84.0</u>	75.2	<u>84.1</u>
ron_Latn	69.1	60.0	63.3	79.2	60.6	55.1	68.7	67.0	<u>84.0</u>	74.1	<u>83.4</u>
rus_Cyrl	73.9	65.4	68.9	79.7	67.2	65.3	70.4	71.0	<u>85.7</u>	75.7	<u>82.8</u>
spa_Latn	74.6	67.8	68.4	77.2	68.8	66.3	68.6	68.6	<u>85.2</u>	75.8	<u>82.9</u>
tha_Thai	66.0	44.3	50.6	70.1	57.9	58.4	62.0	65.3	<u>78.2</u>	60.8	<u>76.1</u>
tur_Latn	64.7	50.1	56.4	72.0	60.6	44.9	65.7	64.7	<u>77.3</u>	62.1	<u>76.8</u>
ukr_Cyrl	67.9	52.0	62.7	75.2	61.3	55.4	69.3	67.4	<u>83.9</u>	68.3	<u>82.4</u>
urd_Arab	53.7	39.4	55.1	65.6	46.3	39.3	56.1	56.1	<u>72.3</u>	56.1	<u>74.0</u>
vie_Latn	71.6	55.0	64.7	75.9	65.2	57.8	66.0	68.8	<u>83.8</u>	71.6	<u>80.9</u>
zho_Hans	79.9	71.1	72.7	79.2	70.9	63.7	68.4	76.6	<u>86.6</u>	80.3	<u>82.6</u>
zsm_Latn	70.1	58.7	61.0	76.8	63.7	55.9	68.0	67.4	<u>81.8</u>	70.1	<u>82.2</u>
Avg	67.6	54.5	61.2	73.8	60.1	54.3	65.7	65.4	<u>80.6</u>	67.6	<u>79.8</u>

Table 17: Per-language model performance comparison on BELEBELE.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
ben_Beng	40.0	18.0	37.2	51.2	10.4	12.0	31.2	28.8	<u>66.0</u>	36.8	<u>64.8</u>
deu_Latn	58.0	45.6	60.0	66.4	24.0	60.0	35.6	38.8	<u>76.0</u>	63.6	<u>79.2</u>
eng_Latn	74.0	58.4	70.4	78.4	34.0	74.0	46.0	52.4	<u>85.2</u>	60.4	<u>83.2</u>
fra_Latn	62.0	52.0	63.2	69.6	27.2	61.2	42.0	48.8	<u>75.6</u>	63.2	<u>77.6</u>
jpn_Jpan	46.4	23.6	46.0	49.6	14.0	34.8	25.6	23.6	<u>64.0</u>	46.4	<u>68.8</u>
rus_Cyrl	66.0	46.8	64.0	74.0	22.8	59.2	36.8	44.0	<u>84.0</u>	63.6	<u>83.2</u>
spa_Latn	69.6	57.6	66.0	71.2	27.6	64.4	42.4	44.4	<u>83.2</u>	67.6	<u>76.4</u>
tha_Thai	52.8	20.0	42.8	62.0	20.4	49.2	36.0	30.0	<u>74.0</u>	45.2	<u>70.4</u>
zho_Hans	60.4	43.2	60.8	65.6	21.6	42.8	34.0	35.2	<u>75.6</u>	67.6	<u>76.4</u>
Avg	58.8	40.6	56.7	65.3	22.4	50.8	36.6	38.4	<u>76.0</u>	57.2	<u>75.6</u>

Table 18: Per-language model performance comparison on MGSM.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	39.7	34.8	<u>51.1</u>	44.1	41.6	50.7	50.3	49.7	46.7	43.1	<u>55.2</u>
ben_Beng	30.4	29.7	<u>36.8</u>	33.4	32.3	29.0	36.7	36.4	34.7	32.6	<u>38.2</u>
deu_Latn	46.1	45.3	<u>59.7</u>	51.5	53.3	<u>59.7</u>	59.0	57.4	56.4	56.5	<u>63.5</u>
fra_Latn	51.3	53.8	<u>64.5</u>	56.2	57.5	63.9	63.8	62.3	61.1	64.3	<u>67.3</u>
hun_Latn	35.8	33.1	42.1	41.5	42.9	32.2	<u>49.4</u>	47.4	44.0	39.6	<u>49.9</u>
ind_Latn	45.9	43.1	52.1	51.9	51.7	45.2	<u>58.3</u>	57.8	53.9	53.3	<u>61.9</u>
ita_Latn	48.9	47.6	60.9	54.6	55.0	<u>62.3</u>	61.4	59.8	58.6	58.5	<u>65.3</u>
nld_Latn	44.1	43.0	59.0	51.7	53.8	45.0	<u>60.8</u>	56.5	54.0	54.1	<u>64.3</u>
npi_Deva	30.2	29.4	<u>34.8</u>	31.9	32.1	31.2	34.7	34.5	32.6	31.9	<u>35.5</u>
por_Latn	50.8	51.4	62.9	55.7	56.8	63.3	<u>63.5</u>	61.7	60.0	61.5	<u>67.1</u>
ron_Latn	41.2	39.9	51.3	50.3	48.2	40.4	<u>58.2</u>	55.7	51.3	49.7	<u>60.3</u>
rus_Cyrl	46.8	45.7	54.6	50.5	51.0	56.4	<u>57.0</u>	54.4	54.4	56.0	<u>62.0</u>
spa_Latn	52.4	52.9	65.4	56.7	59.3	<u>65.8</u>	64.9	62.4	62.1	64.3	<u>68.3</u>
ukr_Cyrl	40.7	38.4	51.6	46.8	46.3	42.5	<u>53.8</u>	51.6	49.4	48.8	<u>58.1</u>
vie_Latn	45.5	39.9	48.0	48.0	49.4	49.0	<u>54.7</u>	53.6	52.2	49.3	<u>57.5</u>
zho_Hans	53.0	52.6	56.7	52.6	51.9	56.3	57.1	54.4	59.8	60.7	<u>62.4</u>
Avg	43.9	42.5	53.2	48.6	48.9	49.6	<u>55.2</u>	53.5	52.0	51.5	<u>58.5</u>

Table 19: Per-language model performance comparison on mHellaSwag.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	32.5	25.4	37.6	34.6	29.6	36.4	39.0	35.6	38.9	30.8	43.4
ben_Beng	27.2	22.7	31.1	28.7	27.1	24.3	29.7	29.1	31.1	25.1	33.8
deu_Latn	36.2	32.3	44.4	38.1	35.5	42.3	45.5	38.7	44.9	41.5	48.8
fra_Latn	37.4	37.4	44.5	41.2	38.3	44.1	45.7	41.4	45.9	44.1	51.2
hun_Latn	29.7	25.7	33.8	32.9	32.4	25.5	39.0	35.5	38.7	31.5	41.2
ind_Latn	37.8	30.9	41.4	38.3	35.5	33.8	46.9	41.1	45.1	39.0	47.8
ita_Latn	38.8	35.9	45.3	41.4	36.8	45.3	44.1	40.5	45.3	42.7	52.6
nld_Latn	33.4	31.2	41.9	39.3	33.8	31.7	43.5	35.8	41.2	37.8	48.4
npi_Deva	23.9	23.6	26.4	25.7	23.8	24.6	26.9	26.3	28.1	25.5	30.0
por_Latn	40.8	37.0	47.4	41.0	38.8	45.8	47.6	42.6	45.9	45.8	50.8
ron_Latn	35.0	29.6	37.2	39.3	35.1	30.8	41.4	38.8	41.0	38.4	45.2
rus_Cyrl	36.9	31.8	41.6	39.6	35.6	40.4	40.4	36.9	43.6	40.7	48.7
spa_Latn	37.3	36.2	46.8	39.4	39.0	46.8	49.3	42.6	46.0	43.4	51.3
ukr_Cyrl	33.8	30.4	38.9	36.9	35.3	30.3	40.8	34.6	39.5	36.0	44.5
vie_Latn	34.9	27.7	36.7	34.4	33.0	35.2	40.6	36.2	42.2	33.5	42.4
zho_Hans	39.6	36.6	43.7	40.1	37.4	40.7	44.4	39.0	48.2	44.8	46.6
Avg	34.7	30.9	39.9	36.9	34.2	36.1	41.5	37.2	41.6	37.5	45.4

Table 20: Per-language model performance comparison on mARC-Challenge.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	24.1	16.3	26.2	24.3	23.3	29.4	21.3	22.8	34.2	27.7	39.5
ben_Beng	21.6	13.3	25.3	23.6	19.9	19.2	20.7	23.8	31.8	15.1	34.2
ces_Latn	28.2	21.4	31.8	29.6	20.9	25.7	23.3	23.1	39.8	32.0	40.8
deu_Latn	26.9	22.1	30.8	30.6	22.3	30.3	24.7	24.3	38.8	35.7	40.5
eng_Latn	33.7	32.3	35.9	34.7	27.2	36.7	28.4	27.0	44.6	42.3	45.4
fra_Latn	29.4	23.6	31.8	31.6	23.3	32.5	25.3	24.5	40.3	38.8	41.5
hun_Latn	26.4	16.2	26.7	27.0	22.3	24.3	24.0	22.1	36.6	27.2	35.9
ind_Latn	28.6	21.9	29.4	27.9	22.4	26.4	26.0	25.5	40.8	31.8	41.2
ita_Latn	29.9	22.3	32.8	32.0	23.8	31.0	27.4	24.0	40.8	36.1	41.2
jpn_Jpan	26.5	19.4	30.1	27.7	20.2	28.9	23.8	24.3	37.6	33.3	37.1
kor_Hang	26.7	18.5	29.8	27.6	21.6	27.0	22.3	23.3	37.6	30.8	35.7
por_Latn	27.7	22.4	30.3	30.6	22.6	31.3	25.7	27.0	40.6	39.5	40.3
rus_Cyrl	26.7	21.3	31.1	32.1	24.1	28.9	25.5	25.9	40.8	35.9	40.0
spa_Latn	29.3	24.7	32.8	31.1	23.5	31.6	25.0	24.7	41.0	38.9	42.0
tha_Thai	25.9	14.1	26.9	24.7	21.3	27.2	21.9	22.1	33.3	21.8	37.1
ukr_Cyrl	26.5	16.5	31.8	28.9	20.9	25.5	24.5	24.1	38.3	30.3	38.9
urd_Arab	21.9	17.5	28.6	25.0	20.9	24.7	22.8	22.6	34.4	25.9	35.9
vie_Latn	29.3	20.1	28.2	29.6	23.1	28.2	24.7	26.5	40.3	32.8	40.3
zho_Hans	27.4	21.4	31.3	30.4	21.8	27.9	24.7	24.1	39.3	35.7	38.1
Avg	27.2	20.3	30.1	28.9	22.4	28.2	24.3	24.3	38.5	32.2	39.2

Table 21: Per-language model performance comparison on MMLU-ProX-Lite.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	49.1	38.6	51.1	50.4	43.5	52.0	57.2	56.9	61.8	49.3	62.1
azj_Latn	47.8	38.3	45.1	47.1	38.5	33.0	46.2	41.6	51.8	43.6	56.6
ben_Beng	40.5	34.9	49.5	44.0	34.3	32.1	48.4	45.6	53.3	42.0	50.4
deu_Latn	42.4	43.2	47.5	46.0	36.0	39.6	41.0	41.7	56.1	48.9	51.8
ell_Grek	44.9	38.0	45.5	48.2	38.2	44.4	45.1	49.3	55.6	40.8	60.0
fra_Latn	54.4	48.7	63.0	55.1	54.9	61.6	55.8	58.7	65.9	63.0	68.0
heb_Hebr	56.9	53.3	57.8	63.3	48.4	47.8	59.1	66.0	65.3	61.8	69.1
hun_Latn	41.6	32.2	43.5	46.4	38.5	34.5	42.5	49.8	52.7	38.5	50.2
ind_Latn	59.5	53.8	52.9	57.5	53.5	51.5	58.2	61.3	65.5	56.5	68.5
ita_Latn	65.1	56.2	68.1	63.7	58.9	63.3	64.2	64.4	73.0	65.7	72.1
jpn_Jpan	63.1	52.5	65.3	61.5	47.1	55.3	58.7	61.7	78.2	67.9	77.4
kaz_Cyrl	40.0	33.8	41.0	48.6	39.2	35.0	41.4	33.2	45.0	35.6	54.6
kor_Hang	58.0	43.0	56.4	56.4	46.6	52.0	55.2	60.8	63.8	54.6	64.2
nld_Latn	52.8	48.6	61.0	58.1	53.7	47.0	56.4	58.6	66.4	59.2	71.0
npi_Deva	37.0	32.6	47.4	49.4	37.6	36.8	48.6	51.6	49.0	42.4	54.2
pol_Latn	46.7	40.1	48.7	47.6	41.8	41.2	50.7	51.1	60.4	51.3	58.9
por_Latn	54.1	49.9	53.9	52.3	48.1	53.2	47.4	51.5	66.2	56.1	61.9
rus_Cyrl	57.6	47.1	51.1	53.8	48.4	51.4	52.4	53.1	63.0	53.8	60.7
spa_Latn	59.8	53.3	64.5	60.5	55.8	58.9	59.8	61.1	69.3	62.4	70.9
tur_Latn	46.4	40.9	51.3	53.6	47.4	40.0	57.5	53.3	58.6	45.6	58.8
ukr_Cyrl	57.1	48.7	54.9	58.9	51.3	51.5	62.5	63.6	66.5	54.2	63.6
urd_Arab	36.6	38.9	37.5	37.2	36.6	32.7	44.6	39.8	44.6	43.8	42.3
vie_Latn	52.9	38.5	42.7	57.8	44.7	45.3	54.4	55.5	64.4	52.9	65.6
zho_Hans	65.5	49.9	51.6	56.0	49.4	54.7	52.8	59.6	76.1	56.7	64.0
zsm_Latn	49.5	42.5	50.5	55.5	45.3	40.3	54.7	56.5	62.3	50.7	65.3
Avg	51.2	43.9	52.1	53.2	45.5	46.2	52.6	53.9	61.4	51.9	61.7

Table 22: Per-language model performance comparison on INCLUDE.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	53.0	47.0	58.0	53.0	55.0	50.0	<u>62.0</u>	65.0	60.0	48.0	59.0
azj_Latn	52.0	45.0	<u>59.0</u>	48.0	52.0	53.0	54.0	53.0	57.0	50.0	63.0
ben_Beng	45.0	36.0	59.0	55.0	49.0	38.0	60.0	63.0	44.0	41.0	63.0
ces_Latn	54.0	40.0	63.0	62.0	60.0	54.0	71.0	<u>70.0</u>	61.0	45.0	69.0
deu_Latn	66.0	63.0	64.0	61.0	<u>71.0</u>	68.0	73.0	70.0	65.0	70.0	66.0
ell_Grek	64.0	58.0	64.0	58.0	62.0	57.0	61.0	60.0	63.0	58.0	58.0
eng_Latn	74.0	71.0	<u>79.0</u>	73.0	72.0	<u>79.0</u>	78.0	75.0	76.0	72.0	82.0
fra_Latn	65.5	59.5	<u>73.5</u>	69.0	69.5	<u>73.5</u>	72.5	<u>73.5</u>	70.0	66.0	77.0
heb_Hebr	48.0	44.0	60.0	53.0	51.0	50.0	56.0	58.0	51.0	46.0	60.0
hun_Latn	50.0	41.0	52.0	58.0	48.0	48.0	<u>61.0</u>	54.0	53.0	42.0	64.0
ind_Latn	70.0	53.0	61.0	72.0	68.0	52.0	<u>75.0</u>	<u>79.0</u>	70.0	59.0	80.0
ita_Latn	62.0	55.0	69.0	71.0	70.0	<u>80.0</u>	79.0	<u>80.0</u>	78.0	56.0	83.0
jpn_Jpan	72.0	63.0	76.0	73.0	68.0	71.0	81.0	<u>79.0</u>	75.0	68.0	77.0
kaz_Cyrl	65.0	45.0	63.0	75.0	68.0	57.0	<u>78.0</u>	55.0	64.0	49.0	80.0
kor_Hang	53.0	47.0	56.0	56.0	52.0	50.0	51.0	<u>58.0</u>	57.0	53.0	64.0
nld_Latn	57.0	47.0	69.0	65.0	64.0	54.0	<u>75.0</u>	70.0	61.0	51.0	83.0
npi_Deva	54.0	44.0	59.0	63.0	54.0	47.0	66.0	<u>67.0</u>	58.0	50.0	69.0
pol_Latn	53.0	45.0	59.0	65.0	56.0	51.0	<u>71.0</u>	70.0	64.0	50.0	78.0
por_Latn	56.5	51.0	60.5	58.0	58.5	64.0	<u>68.0</u>	65.0	69.5	61.0	<u>68.0</u>
ron_Latn	63.0	63.0	69.0	69.0	63.0	65.0	<u>71.0</u>	70.0	68.0	68.0	76.0
rus_Cyrl	71.0	59.0	71.0	73.0	77.0	<u>82.0</u>	81.0	77.0	79.0	64.0	85.0
spa_Latn	73.3	61.7	73.7	76.3	76.3	<u>77.7</u>	<u>80.7</u>	80.3	78.3	70.7	82.3
tha_Thai	60.0	47.0	53.0	62.0	58.0	65.0	61.0	<u>66.0</u>	64.0	48.0	68.0
tur_Latn	46.0	42.0	44.0	58.0	56.0	40.0	<u>68.0</u>	56.0	53.0	47.0	70.0
ukr_Cyrl	70.0	54.0	76.0	74.0	70.0	67.0	<u>83.0</u>	76.0	78.0	61.0	87.0
urd_Arab	57.0	52.0	70.0	71.0	63.0	55.0	<u>76.0</u>	<u>77.0</u>	72.0	58.0	81.0
vie_Latn	65.0	53.0	66.0	67.0	65.0	63.0	79.0	<u>75.0</u>	72.0	52.0	<u>76.0</u>
zho_Hans	56.0	54.0	56.5	<u>57.5</u>	53.0	55.5	54.0	52.0	59.5	56.0	56.0
zsm_Latn	49.0	49.0	55.0	53.0	51.0	57.0	57.0	<u>63.0</u>	47.0	60.0	67.0
Avg	59.5	51.4	63.4	63.8	61.4	59.4	<u>69.1</u>	67.5	64.4	55.9	72.1

Table 23: Per-language model performance comparison on Global-PIQA.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	16.6	7.2	31.5	22.6	18.8	30.5	<u>34.0</u>	34.8	24.4	6.6	33.2
azj_Latn	1.1	0.3	1.9	10.7	6.0	0.3	<u>11.0</u>	1.0	3.5	1.3	16.3
ben_Beng	4.6	2.3	21.3	14.7	12.6	0.6	26.0	21.3	13.3	2.9	<u>23.7</u>
ces_Latn	16.4	9.2	<u>34.4</u>	27.8	27.6	8.7	34.3	35.8	25.7	13.3	33.3
deu_Latn	30.5	32.5	39.7	32.5	34.6	39.0	<u>41.1</u>	39.3	37.0	25.9	41.3
ell_Grek	10.0	5.0	17.9	21.7	22.2	28.3	<u>31.9</u>	33.0	19.9	8.3	28.1
fra_Latn	39.8	44.0	48.3	43.0	43.5	46.9	<u>49.3</u>	48.4	46.1	33.4	50.3
heb_Hebr	7.9	6.5	25.9	20.5	19.6	2.1	<u>32.9</u>	34.6	16.3	8.7	30.4
hun_Latn	9.2	3.4	13.2	18.9	20.6	1.8	28.6	25.3	19.0	5.2	<u>25.5</u>
ind_Latn	32.1	24.1	33.7	37.8	35.3	23.4	45.0	46.2	39.2	23.5	<u>45.6</u>
ita_Latn	25.9	23.8	31.7	27.9	29.4	32.1	32.9	<u>33.9</u>	30.8	19.3	34.1
jpn_Jpan	17.1	14.0	23.7	19.4	18.4	20.7	<u>26.0</u>	22.8	23.8	8.1	27.0
kaz_Cyrl	1.6	0.3	2.1	8.2	4.6	0.3	<u>15.1</u>	0.3	6.2	1.4	18.2
kor_Hang	9.4	6.5	16.1	13.2	12.6	14.8	21.7	<u>22.8</u>	16.7	6.3	23.4
nld_Latn	20.8	15.8	29.0	27.1	27.4	17.1	<u>31.4</u>	29.6	26.5	16.8	32.1
npi_Deva	2.0	1.7	13.5	9.4	6.2	2.0	<u>19.2</u>	23.7	9.1	3.8	17.9
pol_Latn	15.9	12.8	21.2	22.2	22.8	10.0	<u>28.7</u>	26.6	20.9	12.3	29.3
por_Latn	40.5	40.2	47.5	42.6	44.3	46.5	<u>49.9</u>	49.6	47.5	32.3	50.7
ron_Latn	26.9	29.9	37.5	36.3	31.9	12.3	42.7	40.6	34.5	26.7	<u>41.0</u>
rus_Cyrl	27.1	28.1	33.4	27.5	30.3	32.8	<u>36.0</u>	33.6	31.9	21.7	36.6
spa_Latn	26.4	26.1	29.4	27.2	29.3	30.2	31.3	<u>31.2</u>	29.9	20.3	31.1
tha_Thai	22.6	9.7	19.1	29.3	20.2	27.0	<u>35.4</u>	27.3	30.3	12.8	36.5
tur_Latn	12.3	4.3	12.2	18.7	16.8	4.8	30.1	26.1	21.4	7.4	<u>29.1</u>
ukr_Cyrl	15.0	12.1	27.9	24.8	25.8	13.7	<u>33.4</u>	31.0	25.0	12.3	33.6
urd_Arab	1.7	1.9	12.6	11.4	8.1	0.6	<u>16.4</u>	15.5	7.3	3.6	19.7
vie_Latn	32.5	16.9	28.4	33.8	32.5	30.0	<u>40.9</u>	41.1	37.8	16.9	40.8
zho_Hans	32.0	28.3	30.8	27.6	27.1	29.7	33.1	30.9	36.7	18.2	<u>36.1</u>
zsm_Latn	22.9	16.7	27.1	34.4	29.0	17.5	<u>39.9</u>	40.6	30.5	14.6	39.8
Avg	18.6	15.1	25.4	24.7	23.5	18.7	<u>32.1</u>	30.2	25.4	13.7	32.3

Table 24: Per-language model performance comparison on FLORES En-Xx translation direction.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	33.3	30.3	40.3	35.1	35.7	40.6	<u>42.4</u>	41.6	40.1	22.7	42.9
azj_Latn	15.4	10.6	20.7	20.5	21.1	8.9	<u>25.8</u>	17.0	22.2	8.3	27.1
ben_Beng	22.7	19.6	32.4	25.6	25.5	5.8	34.3	31.7	31.0	13.7	<u>33.4</u>
ces_Latn	36.8	37.0	42.0	38.9	39.8	32.8	<u>43.3</u>	42.3	41.2	26.6	43.6
deu_Latn	42.6	45.4	46.9	43.2	45.1	46.3	<u>48.4</u>	46.6	45.9	36.7	49.2
ell_Grek	28.9	28.6	35.4	33.4	36.0	38.2	40.0	39.8	35.5	20.9	<u>40.0</u>
fra_Latn	44.4	48.5	<u>49.5</u>	45.6	46.3	48.3	49.2	48.0	47.8	38.8	50.5
heb_Hebr	30.8	32.4	40.5	37.4	38.3	21.4	<u>46.0</u>	44.5	38.8	22.7	46.2
hun_Latn	29.8	28.4	31.5	29.1	34.4	21.4	38.5	37.3	36.4	21.2	<u>38.1</u>
ind_Latn	40.6	40.9	44.0	41.7	42.2	39.2	<u>47.1</u>	46.3	45.4	32.4	47.1
ita_Latn	34.7	37.3	38.9	35.4	38.3	38.6	<u>40.0</u>	39.1	38.1	28.8	40.8
jpn_Jpan	25.2	26.9	29.4	25.3	25.4	27.5	<u>31.6</u>	29.7	30.4	20.3	31.8
kaz_Cyrl	19.6	7.3	22.2	25.5	22.9	5.6	<u>32.3</u>	8.0	27.8	8.0	33.8
kor_Hang	25.1	25.1	30.3	25.1	26.3	27.2	32.9	31.0	31.4	20.4	<u>32.8</u>
nld_Latn	32.1	33.7	36.5	33.9	34.4	31.2	<u>36.8</u>	35.6	34.8	26.4	36.9
npi_Deva	22.4	19.6	33.9	26.8	23.2	15.9	38.1	<u>36.7</u>	31.4	18.8	<u>36.7</u>
pol_Latn	30.0	30.1	33.7	30.6	31.7	27.9	<u>34.6</u>	34.1	31.7	23.6	35.8
por_Latn	46.1	50.4	52.5	48.6	50.1	50.6	<u>53.0</u>	51.8	51.9	38.4	53.7
ron_Latn	40.4	42.7	46.6	42.4	44.2	39.0	<u>47.5</u>	45.8	43.5	32.1	48.2
rus_Cyrl	33.9	36.2	38.4	34.7	36.3	38.0	<u>39.5</u>	37.3	36.8	28.0	41.5
spa_Latn	32.2	33.5	34.6	33.5	35.0	35.7	37.2	36.5	35.1	25.1	<u>36.7</u>
tha_Thai	26.9	20.9	27.6	28.8	29.7	31.7	<u>34.6</u>	33.9	32.3	18.1	36.0
tur_Latn	29.2	27.9	34.6	29.8	33.7	23.3	40.2	37.8	37.1	19.6	<u>39.6</u>
ukr_Cyrl	35.1	36.0	40.5	38.0	39.3	35.2	<u>43.1</u>	41.5	40.0	28.0	44.5
urd_Arab	21.4	17.9	31.7	25.7	26.1	10.8	34.7	33.0	29.5	17.6	<u>34.3</u>
vie_Latn	34.8	32.0	36.7	34.5	36.2	36.3	<u>40.0</u>	38.9	38.3	24.0	40.1
zho_Hans	30.0	30.8	32.5	28.9	30.0	30.6	32.9	31.9	<u>34.6</u>	22.9	34.9
zsm_Latn	39.1	39.8	44.2	41.7	42.4	39.1	<u>47.4</u>	46.8	44.4	31.1	48.1
Avg	31.6	31.1	36.7	33.6	34.6	30.3	<u>39.7</u>	37.3	36.9	24.1	40.2

Table 25: Per-language model performance comparison on FLORES Xx-En translation direction.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	9.1	4.8	13.9	11.4	7.1	13.1	14.9	15.7	11.5	3.2	<u>15.0</u>
ben_Beng	6.4	2.8	16.0	12.7	9.5	1.0	18.4	16.6	12.6	3.0	<u>17.8</u>
ces_Latn	14.4	9.4	23.7	18.8	13.9	7.0	<u>27.2</u>	28.6	21.8	6.0	26.0
deu_Latn	23.1	24.5	28.8	24.5	20.2	28.9	<u>30.6</u>	29.0	28.9	12.9	32.0
ell_Grek	11.1	7.5	17.2	18.7	16.3	30.0	<u>35.4</u>	36.2	22.0	4.4	30.7
fra_Latn	32.4	28.5	34.3	31.8	29.0	38.0	<u>40.6</u>	38.7	38.3	12.4	41.6
heb_Hebr	8.0	7.7	18.5	15.9	10.9	3.4	24.4	<u>25.5</u>	14.3	4.7	25.8
hun_Latn	10.1	3.9	9.6	15.1	12.0	2.3	22.3	19.7	16.7	4.0	<u>20.5</u>
ind_Latn	23.3	17.7	23.0	28.0	18.5	15.8	33.1	<u>33.6</u>	29.9	7.5	33.7
ita_Latn	29.5	25.6	34.7	30.3	23.8	31.4	<u>37.0</u>	<u>34.2</u>	34.3	12.1	38.5
jpn_Jpan	12.0	8.2	14.6	10.7	6.6	12.8	15.4	12.6	<u>15.5</u>	4.6	18.1
kor_Hang	9.4	6.5	13.6	10.7	8.1	11.7	16.9	<u>17.8</u>	15.8	5.1	19.1
nld_Latn	22.2	18.9	29.4	25.6	20.5	16.6	<u>31.6</u>	29.8	27.3	10.1	34.2
pol_Latn	14.6	11.7	18.6	17.4	13.1	8.0	<u>22.6</u>	21.1	18.9	5.9	24.4
por_Latn	29.1	29.2	33.7	29.9	25.3	32.5	<u>35.2</u>	33.1	33.8	12.0	36.2
ron_Latn	22.0	22.1	28.6	30.0	20.8	9.3	36.7	33.7	31.2	9.4	<u>36.6</u>
rus_Cyrl	20.5	20.1	24.1	19.6	17.9	24.3	26.5	24.5	24.2	8.3	<u>25.5</u>
spa_Latn	33.2	34.7	34.5	33.5	33.3	38.5	<u>41.2</u>	39.8	39.7	14.6	42.0
tha_Thai	17.3	7.1	12.3	20.2	11.6	19.7	27.0	20.6	23.6	5.1	<u>26.1</u>
tur_Latn	10.8	4.1	8.0	12.8	8.6	3.5	25.5	19.4	18.5	4.9	<u>24.2</u>
ukr_Cyrl	13.4	10.1	21.0	19.4	15.3	9.2	29.0	26.2	21.0	6.7	<u>28.5</u>
urd_Arab	3.0	2.5	12.4	12.0	8.3	1.5	22.3	18.0	10.2	2.5	<u>20.2</u>
vie_Latn	26.6	16.2	22.4	26.1	19.7	21.4	31.3	32.9	30.4	8.1	33.4
zho_Hans	20.5	17.3	18.6	16.5	13.4	17.5	20.2	18.5	22.9	7.6	<u>21.9</u>
Avg	17.6	14.2	21.3	20.5	16.0	16.6	<u>27.7</u>	26.1	23.5	7.3	28.0

Table 26: Per-language model performance comparison on WMT24++ En-Xx translation direction.

Language	Qwen3 1.7B Base	Trinity Nano Base	Granite4 Tiny Base	Marco Nano Base	Llama3.2 3B Base	SmolLM3 3B Base	Gemma3 4B Base	Tiny-Aya 3.35B Base	Qwen3 4B Base	Trinity Mini Base	Marco Mini Base
arb_Arab	17.1	15.2	18.4	15.1	18.4	19.7	23.2	22.8	21.3	6.3	21.6
ben_Beng	20.2	14.0	24.1	21.6	16.2	4.6	30.7	27.7	27.1	7.3	<u>28.1</u>
ces_Latn	30.4	30.3	34.2	30.0	31.7	25.3	35.4	34.9	<u>35.6</u>	11.4	36.9
deu_Latn	32.2	32.9	35.4	31.3	31.6	34.6	36.0	34.2	<u>36.1</u>	12.0	36.3
ell_Grek	31.9	29.4	35.1	33.0	36.0	39.8	42.4	40.5	38.9	10.1	<u>41.1</u>
fra_Latn	37.4	40.1	39.3	36.9	38.0	40.0	<u>41.0</u>	40.1	40.5	15.1	42.1
heb_Hebr	26.4	27.5	31.9	30.2	27.8	21.4	<u>35.2</u>	34.6	32.0	9.5	36.3
hun_Latn	24.8	22.3	27.3	24.3	26.1	16.5	<u>29.1</u>	28.2	<u>29.1</u>	9.4	30.5
ind_Latn	30.3	30.7	32.9	30.8	32.4	29.7	34.6	<u>34.7</u>	34.0	12.8	35.8
ita_Latn	37.9	40.1	41.0	37.2	38.1	42.0	43.5	42.1	42.3	14.6	<u>43.3</u>
jpn_Jpan	18.9	19.5	18.9	17.4	18.4	19.5	23.7	21.5	<u>23.2</u>	6.2	23.1
kor_Hang	22.0	19.9	23.1	18.7	20.6	22.3	26.9	25.5	25.8	6.6	<u>26.4</u>
nld_Latn	33.3	33.6	35.3	33.2	33.0	31.6	35.6	37.1	<u>37.2</u>	12.5	38.8
pol_Latn	25.6	26.2	29.8	27.0	27.4	23.1	<u>32.1</u>	30.4	30.3	10.5	33.9
por_Latn	35.6	38.2	39.7	35.5	36.5	36.3	39.3	39.3	39.6	14.2	41.2
ron_Latn	34.0	37.4	38.4	36.3	36.8	32.4	43.3	40.5	40.6	12.4	<u>42.7</u>
rus_Cyrl	24.8	27.1	27.2	24.8	25.3	26.9	<u>29.7</u>	27.7	28.5	10.7	29.9
spa_Latn	39.4	41.1	41.2	38.1	41.7	42.5	<u>45.3</u>	43.7	44.1	16.5	45.6
tha_Thai	22.7	16.5	21.0	22.2	23.1	26.2	29.2	27.6	27.8	7.6	<u>29.1</u>
tur_Latn	24.1	21.1	25.8	23.5	24.6	16.5	<u>32.7</u>	30.4	30.2	8.5	32.8
ukr_Cyrl	29.6	27.8	33.1	29.4	28.3	28.0	<u>35.9</u>	32.8	33.6	10.2	36.9
urd_Arab	25.0	21.3	29.3	27.5	27.7	15.2	36.7	34.1	32.0	8.1	<u>34.6</u>
vie_Latn	27.2	25.8	27.7	27.7	27.3	26.2	<u>31.7</u>	31.3	31.5	9.7	32.8
zho_Hans	25.2	26.5	26.8	23.1	25.1	26.0	28.0	26.2	<u>29.9</u>	8.9	30.1
Avg	28.2	27.7	30.7	28.1	28.8	26.9	<u>34.2</u>	32.8	33.0	10.5	34.6

Table 27: Per-language model performance comparison on WMT24++ Xx-En translation direction.