

Appendix

Anonymous Author(s)
Affiliation
Address
email

1 A Details of Fitting Transfer Strength $\alpha_{j \rightarrow i}(D)$

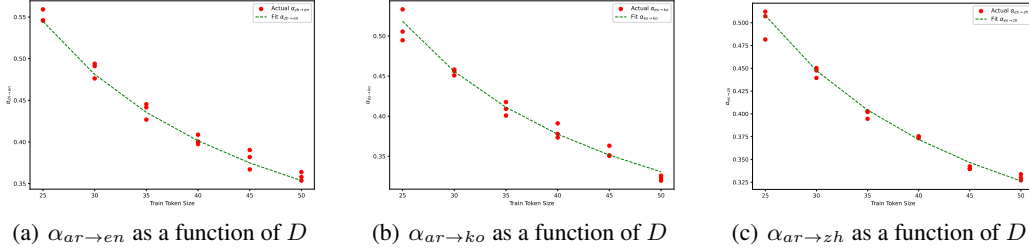


Figure 9: Illustration of cross-lingual interaction-aware language ratio (\tilde{r}_{ar}) and its dependency on original training proportions (r_{ar}).

2 As introduced in Figure 2 (c) of the main text, we observe that the curve relating \tilde{r}_i and r_i shifts
3 vertically depending on the total token budget D . Specifically, as D increases, the \tilde{r}_i versus r_i curve
4 tends to move downward, while smaller D values correspond to upward shifts. According to Equation
5 (4), the parameter $\alpha_{j \rightarrow i}(D)$ effectively acts as an intercept controlling this vertical shift.

6 To accurately characterize the relationship between $\alpha_{j \rightarrow i}$ and the data budget D , we adopt a two-step
7 procedure. First, we individually fit the relationship between \tilde{r}_i and r_i at different values of D using
8 Equation (4). This yields empirical estimates of $\alpha_{j \rightarrow i}$ at various token budgets. Figure 9 illustrates
9 the computed values of $\alpha_{j \rightarrow i}$ for three representative language pairs across different scales of D .

10 Moreover, our empirical findings suggest two critical properties for the $\alpha_{j \rightarrow i}(D)$ relationship:

- 11 • **Non-monotonicity:** $\alpha_{j \rightarrow i}$ does not continuously decrease with increasing D ; rather, it
12 converges towards a stable limiting value as D becomes sufficiently large.
- 13 • **Sign variability:** $\alpha_{j \rightarrow i}$ can be either positive or negative. Positive values indicate beneficial
14 cross-lingual transfer, whereas negative values reflect interference effects, where additional
15 data from language L_j eventually hinder the learning of language L_i .

16 Considering these empirical insights, we propose modeling $\alpha_{j \rightarrow i}(D)$ with the following parametric
17 form:

$$\alpha_{j \rightarrow i}(D) = b_{ji} + \frac{k_{ji}}{D}, \quad (1)$$

18 where b_{ji} represents the asymptotic transfer strength as $D \rightarrow \infty$, and k_{ji} controls the decay rate of
19 this transfer effect as the data budget increases.

20 The fitting results using this parametric form, depicted by the green curves in Figure 9, demonstrate
21 excellent agreement with the empirical $\alpha_{j \rightarrow i}$ - D relationships across various language pairs, validating
22 our choice of functional form.

23 B Derivation of Optimal Direction for Cross-Lingual Interaction-Aware 24 Ratios p_i

25 To compute the optimal direction of the Cross-Lingual Interaction-Aware Ratios $\{\tilde{r}_i\}$, we formulate
26 and solve the following uncoupled optimization subproblem:

$$\min_{\tilde{r}_i > 0} \sum_{i=1}^n \frac{B_i}{(D \tilde{r}_i)^{\beta_i}} \quad \text{s.t.} \quad \sum_{i=1}^n \tilde{r}_i = M, \quad (2)$$

27 where $M > 0$ is a fixed normalization constant, and B_i, β_i, D are known positive parameters.

28 Introducing a Lagrange multiplier λ , we construct the Lagrangian:

$$\mathcal{J}(\tilde{\mathbf{r}}, \lambda) = \sum_{i=1}^n \frac{B_i}{(D \tilde{r}_i)^{\beta_i}} + \lambda \left(\sum_{i=1}^n \tilde{r}_i - M \right). \quad (3)$$

29 Taking derivatives with respect to each \tilde{r}_i and setting them to zero, we obtain the first-order optimality
30 conditions:

$$-B_i \beta_i D^{-\beta_i} \tilde{r}_i^{-(\beta_i+1)} + \lambda = 0 \quad (4)$$

$$\implies \tilde{r}_i^{\beta_i+1} = \frac{B_i \beta_i}{\lambda D^{\beta_i}}. \quad (5)$$

31 Comparing the conditions for any two languages i, j , we have:

$$\frac{\tilde{r}_i}{\tilde{r}_j} = \left(\frac{B_i \beta_i}{B_j \beta_j} D^{\beta_j - \beta_i} \right)^{\frac{1}{\beta_i+1} / \frac{1}{\beta_j+1}}. \quad (6)$$

32 Thus, the optimal direction must satisfy:

$$\tilde{r}_i \propto (B_i \beta_i / D^{\beta_i})^{1/(\beta_i+1)}. \quad (7)$$

33 Applying the normalization constraint $\sum_i \tilde{r}_i = M$, we obtain the normalized optimal direction:

$$p_i = \frac{(B_i \beta_i)^{1/(\beta_i+1)} D^{-\beta_i/(\beta_i+1)}}{\sum_{k=1}^n (B_k \beta_k)^{1/(\beta_k+1)} D^{-\beta_k/(\beta_k+1)}}. \quad (8)$$

34 Since each term $B_i / (D \tilde{r}_i)^{\beta_i}$ is strictly convex in \tilde{r}_i and the constraint is linear, the stationary solution
35 derived above constitutes the unique global minimizer. This rigorous derivation justifies the Marginal-
36 Benefit Balancing approach presented in the main text, providing the closed-form solution for the
37 optimal direction $\{\tilde{r}_i\}$.

38 C Equivalence of Two-Stage Optimization with Direct Optimization

39 Here we provide a rigorous justification demonstrating that our proposed two-stage optimization
40 approach—first determining the optimal direction p_i and subsequently maximizing the magnitude of
41 effective data allocation—is equivalent to directly solving the original optimization problem.

42 **(i) Necessity of Optimizing the Direction:** Assume the direction of the cross-lingual interaction-
43 aware ratios $\{\tilde{r}_i\}$ deviates from the optimal direction p_i . Under any fixed effective data contribution
44 $\sum_i B_i / (D \tilde{r}_i)^{\beta_i}$, the total validation loss will always be greater than or equal to that obtained using
45 the optimal direction. Formally, the optimal direction condition is:

$$\frac{B_i \beta_i}{D^{\beta_i} \tilde{r}_i^{\beta_i+1}} = \frac{B_j \beta_j}{D^{\beta_j} \tilde{r}_j^{\beta_j+1}}, \quad \forall i, j. \quad (9)$$

46 Any deviation from this balanced proportionality condition disrupts marginal equilibrium, causing
47 certain languages to have unnecessarily higher marginal loss reductions, thus reducing overall
48 efficiency. Hence, identifying the direction $\{\tilde{r}_i\}$ by balancing marginal benefits ensures minimal total
49 loss given a fixed effective data contribution.

(ii) **Optimal Magnitude via Maximizing Effective Allocation:** Once the optimal direction p_i is fixed, we set $\tilde{r}_i = c \cdot p_i$, where c denotes the scaling magnitude of effective data allocation (with normalization $\sum_i \tilde{r}_i = c$). We then isolate the variable component of total loss as a function of c :

$$L_{\text{var}}(c) = \sum_i \frac{B_i}{(Dcp_i)^{\beta_i}} = \sum_i \frac{B_i}{D^{\beta_i} p_i^{\beta_i}} c^{-\beta_i}. \quad (10)$$

Differentiating with respect to c , we have:

$$\frac{dL_{\text{var}}}{dc} = - \sum_i \frac{\beta_i B_i}{D^{\beta_i} p_i^{\beta_i}} c^{-(\beta_i+1)} < 0, \quad (11)$$

provided that all $\beta_i > 0$. This negative derivative demonstrates a strictly monotonic decrease in loss as the magnitude c increases. Intuitively, larger c means greater effective data volumes $D\tilde{r}_i$ for each language, which consistently reduces loss due to the monotonicity of scaling laws. Therefore, to minimize the loss, we naturally aim to increase c as much as feasible—maximizing the total effective data contribution while maintaining the optimal relative proportions.

However, practical constraints limit the maximum achievable c . Given the normalization constraint $\sum r_i = 1$ and the implicit mapping from $\{r_i\}$ to $\{\tilde{r}_i\}$, the magnitude c has an upper bound c^* corresponding to feasible allocations.

In summary, stage 1 guarantees that adjusting the direction of ratios does not increase the loss, and stage 2 optimally maximizes effective data volume along this direction, ensuring minimal achievable loss. Thus, the two-stage solution is proven equivalent to directly solving the original optimization problem. This result aligns with previous studies on multilingual scaling laws, demonstrating the consistency and optimality of the two-stage optimization procedure.

D Training Details

Dataset Description

All experiments utilize data sampled from the Fineweb-2 corpus [7]. We further preprocess the dataset by training a custom Byte-Pair Encoding (BPE) tokenizer using the BBPE method, resulting in a vocabulary of 250k tokens for subsequent experiments.

Experimental Setup

We conduct multilingual experiments with various language combinations:

- **Bilingual Experiments:** {es-ko, en-zh, de-ar, ko-ja}
- **Trilingual Experiments:** {es-de-ar, es-ko-zh, en-zh-ja}
- **Five-language Experiment:** {es-de-ar-ko-ja}
- **Sixteen-language Experiment:** {de, en, nl, es, pt, fr, it, id, ja, ko, zh, ru, ar, th, vi, tr}

As detailed in Algorithm 1, for each multilingual setting, we first fix the proportion of one language and evenly distribute the remaining proportion among the other languages. For each selected language L_i , we systematically vary its proportion across the set {0.02, 0.025, 0.05, 0.1, 0.2, 0.25, 0.4, 0.5, 0.6, 0.75, 0.8, 0.9, 0.95, 0.975, 0.98} to establish comprehensive fitting functions. In the sixteen-language experiment, we follow Algorithm 1 for extrapolation and validation.

Model Configuration

We adopt a transformer-based architecture inspired by the LLaMA-2 [12] model, specifically configured with approximately 1.2 billion parameters. The detailed architecture settings are:

- Hidden size: 2048
- Vocabulary embedding dimension: 2048
- Intermediate layer dimension: 5504
- Attention heads: 16

- Layers: 24
- Maximum positional embeddings: 4096
- Layer normalization epsilon: 1.0×10^{-5}

All models are randomly initialized.

Training Hyperparameters

- Batch size: 3072
- Sequence length: 4096
- Optimizer: AdamW
- Learning rate schedule: Cosine decay to 10% of initial value
- Training steps: Varied according to total token budget D
- Precision: bf16 (mixed-precision training)

Computational Resources and Runtime

Each experiment is conducted using 64 H100 GPUs, with an average runtime of approximately 10 hours per experiment.

Evaluation Methodology

The validation datasets for each language are separately sampled from Fineweb-2, ensuring no overlap with training samples. Validation loss is computed by averaging the loss across the final three training steps of each run.

E Detailed Evaluation Protocols for Benchmarks

To rigorously assess the capabilities of our proposed model, we select benchmarks that span diverse evaluation dimensions, including natural language inference, commonsense reasoning, question answering, multilingual multitask understanding, and translation tasks. Recognizing that several benchmarks were originally developed only in English, we manually translated these datasets into multilingual versions (marked as \ddagger : XHS \ddagger , XARC-E \ddagger , XARC-C \ddagger , XGPQA \ddagger , XTQA \ddagger). Our translation approach involves encapsulating each benchmark component—prompts, questions, and answer choices—with explicit tags to maintain structural consistency. We then use GPT-based translation, ensuring strict validation of tag integrity post-translation, with any problematic samples retranslated. This systematic methodology guarantees accurate and faithful translations, supporting flexible future adaptations and mixed-language test scenarios. Below, we detail each evaluation benchmark grouped by task type.

Language Modeling and Natural Language Inference

XNLI (Cross-lingual Natural Language Inference) [2]: Extended from MultiNLI, XNLI evaluates cross-lingual sentence representations across 15 languages, measuring models’ inference capabilities.

XCOPA (Cross-lingual Choice of Plausible Alternatives) [8]: XCOPA tests models on causal commonsense reasoning across 11 languages, providing insights into multilingual causal reasoning capabilities.

XStoryCloze [5]: XStoryCloze assesses zero-shot and few-shot learning across 10 non-English languages, examining models’ narrative understanding and inference skills.

Commonsense Reasoning

HellaSwag (XHS \ddagger) [13]: Originally English-only, HellaSwag involves selecting the most plausible sentence ending from multiple choices, thereby testing commonsense reasoning.

XWinograd [6]: As a multilingual variant of the Winograd Schema Challenge, XWinograd evaluates pronoun resolution abilities in diverse linguistic contexts.

Question Answering

134 *ARC-Easy (XARC-E[‡]) / ARC-Challenge (XARC-C[‡])* [1]: ARC contains scientific multiple-choice
 135 questions designed for different complexity levels, evaluating reasoning from basic to advanced.

136 *GPQA (Graduate-Level Google-Proof Q&A, XGPQA[‡])* [9]: GPQA tests graduate-level understanding
 137 across domains like biology, physics, and chemistry, requiring deep comprehension beyond search-
 138 engine-based answers.

139 *TruthfulQA (XTQA[‡])* [4]: This dataset assesses the factual accuracy and common misconception
 140 avoidance of language models across diverse topics.

141 **Multitask Language Understanding (MMLU Series)**

142 *CMMLU (Chinese Massive Multitask Language Understanding)* [3]: Evaluates Chinese language
 143 models’ knowledge across multiple disciplines including natural sciences, engineering, and humani-
 144 ties.

145 *JMMLU (Japanese Massive Multitask Language Understanding)*¹: JMMLU assesses Japanese
 146 models on multitask language understanding, covering extensive topics.

147 *VMLU (Vietnamese Massive Language Understanding)*²: Focused on Vietnamese, VMLU evaluates
 148 broad academic and practical knowledge via a large set of multiple-choice questions.

149 *GMMLU (Global Massive Multitask Language Understanding)* [10]: GMMLU tests multilingual
 150 generalization capabilities across various languages and diverse tasks.

151 **Translation Tasks**

152 *FLORES (Facebook Low Resource Languages Evaluation Suite)* [11]: Supporting many-to-many
 153 translations, FLORES provides a high-quality benchmark suitable for assessing model performance
 154 on low-resource languages.

155 **F Detailed Per-Language Benchmark Results**

156 This appendix presents detailed, per-language evaluation results corresponding to the benchmarks
 157 summarized in Table 2. The following tables comprehensively report the performance of our CLIMB-
 158 derived multilingual allocation strategy across each evaluated language, facilitating an in-depth
 159 analysis and comparison against baseline methods.

Table 3: Detailed per-language performance on the **XWinograd** benchmark (5-shot accuracy).

Model / Method	EN	FR	JP	PT	RU	ZH
Open Source Multilingual LLMs						
LLaMA-3.2	93.65	71.25	67.17	72.09	73.75	77.13
Qwen-3	92.54	76.61	78.49	77.12	69.51	80.69
Gemma-3	77.60	65.56	62.95	62.86	64.29	68.38
Different Data Allocation Methods						
Uniform	82.93	72.45	71.39	73.80	67.89	71.58
Isolated	79.79	77.71	71.44	68.59	65.58	70.74
Natural	82.90	76.28	71.19	74.02	68.71	76.77
MSL	82.14	73.84	69.90	71.56	67.04	74.06
CLIMB	90.57	78.14	74.27	74.98	73.66	73.25

160 **G Limitations and Future Work**

161 While our experiments demonstrate strong performance using the proposed multilingual allocation
 162 strategy based on scaling laws, several limitations should be acknowledged. First, our parametric
 163 fitting and allocation strategies are primarily validated on a 1.2 billion-parameter (1.2B) model, and

¹<https://github.com/nlp-waseda/JMMLU>

²<https://vmlu.ai/>

Table 4: Detailed per-language performance on the **XStoryCloze** benchmark (0-shot accuracy).

Model / Method	AR	EN	ES	EU	HI	ID	MY	RU	SW	TE	ZH
Open Source Multilingual LLMs											
LLaMA-3.2	52.99	73.18	63.20	51.77	57.81	60.26	50.74	61.94	52.12	56.29	59.57
Qwen-3	56.96	74.71	65.52	53.32	58.07	62.47	53.32	63.26	51.65	60.33	66.00
Gemma-3	51.94	62.49	57.01	52.74	54.69	54.63	50.73	55.35	51.87	56.61	55.18
Different Data Allocation Methods											
Uniform	60.45	70.35	66.44	53.01	50.31	65.22	49.98	65.25	51.19	54.91	61.76
Isolated	59.87	71.34	64.97	52.27	50.82	63.92	50.25	65.87	51.06	54.67	61.09
Natural	59.19	67.96	62.06	51.26	52.19	61.62	49.82	61.33	50.19	54.31	61.24
MSL	60.19	69.16	63.30	51.42	52.59	62.49	49.30	62.21	50.13	54.51	62.04
CLIMB	62.39	73.09	66.22	53.74	55.59	64.49	51.11	66.20	52.57	58.10	62.43

Table 5: Detailed per-language performance on the **XCOPA** benchmark (5-shot accuracy).

Model / Method	ET	HT	ID	IT	QU	SW	TA	TH	TR	VI	ZH
Open Source Multilingual LLMs											
LLaMA-3.2	52.09	52.31	62.69	62.49	51.51	51.31	55.08	55.90	55.90	64.68	64.47
Qwen-3	52.57	53.17	66.63	65.07	49.77	53.17	54.38	57.81	57.81	70.03	74.64
Gemma-3	51.99	52.77	60.18	56.59	52.20	55.21	55.62	54.19	55.62	59.79	57.99
Different Data Allocation Methods											
Uniform	49.59	50.99	67.99	66.99	51.63	51.63	56.46	61.05	61.26	69.59	67.20
Isolated	49.86	51.66	70.62	64.80	50.60	50.21	56.60	61.78	61.78	69.94	65.77
Natural	50.76	51.48	64.31	59.09	49.97	51.85	54.44	58.76	58.49	62.85	59.93
MSL	52.32	52.77	66.29	61.15	51.18	52.88	55.45	59.54	60.00	64.75	62.33
CLIMB	54.21	53.80	68.06	63.89	52.18	54.12	56.73	60.95	61.50	67.46	66.90

Table 6: Detailed per-language performance on the **XNLI** benchmark (5-shot accuracy).

Model / Method	AR	DE	EN	ES	FR	RU	TH	TR	VI	ZH
Open Source Multilingual LLMs										
LLaMA-3.2	34.05	42.16	46.15	40.41	42.20	40.48	38.41	39.90	39.90	39.86
Qwen-3	33.83	42.38	47.43	43.58	43.58	42.38	39.70	37.44	41.10	41.90
Gemma-3	38.94	41.35	44.81	41.53	41.92	41.92	39.74	40.18	42.28	41.03
Different Data Allocation Methods										
Uniform	32.68	43.75	44.37	41.39	43.95	40.41	37.67	41.81	36.45	38.31
Isolated	31.15	40.95	42.76	40.16	42.84	40.22	36.53	41.60	35.64	37.46
Natural	34.26	40.61	43.19	40.23	41.53	39.40	37.50	39.58	35.74	38.46
MSL	32.88	39.57	43.00	40.23	40.95	38.88	37.62	38.66	35.47	38.14
CLIMB	35.14	43.01	48.18	43.93	44.41	42.72	40.87	41.76	38.92	37.56

Table 7: Detailed per-language performance on the **Global MMLU (GMMLU)** benchmark (5-shot accuracy).

Model / Method	AR	DE	EN	ES	FIL	FR	ID	IT	JA	KO	MS	NL	PT	TR	VI	ZH
Open Source Multilingual LLMs																
LLaMA-3.2	25.88	29.12	35.30	29.31	28.05	28.84	28.59	28.54	27.58	27.90	28.33	28.11	29.16	27.21	28.39	29.21
Qwen-3	29.62	34.79	43.92	35.77	31.23	35.68	33.94	34.85	32.75	32.21	32.15	33.23	35.74	31.04	33.63	37.94
Gemma-3	25.43	26.94	31.13	27.75	27.00	27.20	27.15	27.05	26.49	26.95	26.57	25.96	27.49	26.68	27.29	27.42
Different Data Allocation Methods																
Uniform	27.56	29.78	31.30	29.81	25.64	29.85	29.76	29.37	28.45	28.77	28.51	28.88	30.18	28.75	28.99	29.20
Isolated	26.80	29.20	30.96	29.56	25.66	29.17	29.44	29.03	28.27	28.35	28.21	28.73	29.61	28.21	28.60	28.45
Natural	28.84	31.18	33.38	31.83	27.15	31.09	31.11	30.51	29.43	28.64	28.31	30.25	31.47	29.81	29.72	30.96
MSL	28.00	29.93	32.47	30.55	26.46	30.43	29.80	28.84	27.51	27.38	26.69	29.01	30.47	28.30	28.58	29.60
CLIMB	30.53	32.91	36.26	33.85	28.86	33.95	33.04	32.11	30.70	30.33	29.68	31.48	33.92	30.79	31.16	28.91

Table 8: Detailed per-language performance on the **FLORES Translation** benchmark (5-shot chrF++ scores).

Model / Method	Translation to English (xx-en)																
	AR	DE	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TH	TL	TR	VI	ZH
LLaMA-3.2	46.47	57.15	51.99	58.86	53.57	53.20	39.51	39.20	52.00	50.82	61.54	50.82	42.53	42.56	42.26	48.77	44.12
Qwen-3	55.40	61.66	54.54	62.48	58.90	56.20	48.68	48.82	58.09	53.48	64.18	55.65	49.75	51.95	51.39	54.26	52.12
Gemma-3	43.14	53.40	38.32	49.88	40.80	46.72	36.66	28.78	42.00	43.14	52.78	45.00	35.26	37.90	38.32	38.06	40.29
Uniform	55.37	61.04	54.87	61.75	59.21	56.23	45.28	46.07	57.67	54.38	65.10	54.46	48.91	20.60	51.28	53.56	46.88
Isolated	55.57	60.91	54.24	61.77	59.62	55.73	45.86	46.43	57.82	54.92	64.74	54.64	49.19	20.68	51.81	53.53	46.38
Natural	56.99	61.56	55.39	62.87	60.74	56.99	46.47	47.01	58.48	54.79	65.69	55.60	49.82	22.59	52.61	54.99	47.57
MSL	56.15	60.65	54.35	61.85	59.94	56.15	45.96	46.43	57.34	53.97	64.40	54.47	48.76	23.12	51.95	54.23	47.02
CLIMB	58.99	63.65	57.13	65.55	63.43	59.08	48.89	49.32	60.26	56.75	66.66	57.12	51.36	25.46	54.77	57.01	46.89

Model / Method	Translation from English (en-xx)																
	AR	DE	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TH	TL	TR	VI	ZH
LLaMA-3.2	27.93	49.67	47.56	55.84	52.44	45.92	19.61	17.24	47.10	46.11	57.57	42.05	25.85	30.45	33.82	45.69	20.53
Qwen-3	36.82	54.06	50.86	61.53	59.39	50.22	26.69	23.64	52.19	46.81	62.44	47.53	35.09	37.32	39.47	53.24	30.23
Gemma-3	24.68	37.67	34.94	49.05	43.39	33.38	16.18	14.72	38.58	32.06	49.51	32.01	24.07	25.28	28.93	36.46	19.52
Uniform	42.48	51.98	47.80	57.92	60.45	47.63	23.59	24.38	54.57	48.72	59.52	44.10	35.14	8.21	43.24	52.00	20.95
Isolated	41.99	52.49	47.88	58.06	60.70	47.94	24.12	24.08	54.64	49.18	59.31	44.55	34.49	9.38	43.19	51.06	20.38
Natural	43.44	53.47	48.64	59.13	61.07	48.83	24.48	24.50	55.45	50.14	60.41	45.28	35.57	10.95	44.14	52.85	21.81
MSL	42.71	52.35	47.41	57.74	59.67	47.83	24.56	24.61	53.91	49.03	58.77	44.11	35.23	11.78	43.36	51.67	22.01
CLIMB	45.63	55.50	50.28	61.43	63.18	50.64	26.59	26.66	56.94	51.92	62.14	46.58	37.75	13.45	46.16	54.82	22.65

Table 9: Detailed per-language performance on the **ARC-Challenge** benchmark (25-shot accuracy).

Model / Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TA	TH	TR	VI	ZH
Open Source Multilingual LLMs																		
LLaMA-3.2	26.64	30.93	42.26	35.16	33.42	30.34	32.88	28.67	31.21	30.76	30.30	32.88	29.53	25.12	28.69	29.05	30.45	32.80
Qwen-3	34.00	42.03	54.39	43.51	41.31	41.22	43.49	35.42	36.33	37.58	38.32	43.11	39.92	28.09	32.77	33.03	37.52	45.54
Gemma-3	25.58	29.77	38.41	30.69	31.03	30.27	30.27	27.92	28.42	28.18	27.09	30.94	28.42	25.92	25.92	27.59	27.26	29.94
Different Data Allocation Methods																		
Uniform	33.46	35.60	40.10	38.47	35.60	35.77	38.59	34.48	35.17	35.34	35.17	37.72	38.23	22.78	32.00	35.85	35.09	37.97
Isolated	31.19	35.53	38.72	37.02	36.25	37.45	37.87	35.44	34.62	37.02	36.00	37.26	34.75	23.15	31.71	34.62	32.72	34.71
Natural	31.75	33.98	38.37	35.91	34.60	34.76	36.61	33.24	32.80	33.63	33.50	35.84	33.81	21.99	30.05	33.69	33.48	35.72
MSL	32.31	34.59	38.90	36.64	35.30	35.46	37.34	33.74	33.20	34.12	33.98	36.38	34.32	22.60	30.70	34.47	34.28	36.74
CLIMB	34.48	37.10	41.78	39.05	38.03	38.27	40.19	36.33	35.64	37.20	37.05	39.67	37.40	24.08	32.54	36.58	36.40	36.30

Table 10: Detailed per-language performance on the **ARC-Easy** benchmark (25-shot accuracy).

Model / Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TA	TH	TR	VI	ZH
Open Source Multilingual LLMs																		
LLaMA-3.2	39.16	50.09	70.21	54.98	52.23	48.35	51.18	40.29	41.26	44.43	47.09	52.19	48.56	35.74	37.63	42.37	46.96	49.40
Qwen-3	49.23	62.56	80.14	67.38	64.09	60.14	62.73	53.92	52.27	51.49	54.70	65.11	60.84	41.24	46.35	49.35	57.51	69.64
Gemma-3	41.68	49.67	70.80	55.19	53.55	50.56	54.01	48.37	47.44	43.70	48.62	52.49	47.02	39.15	39.61	44.08	46.18	55.31
Different Data Allocation Methods																		
Uniform	56.70	62.68	70.93	67.27	63.81	64.28	64.07	58.97	58.01	57.71	62.59	66.34	60.24	29.64	49.59	60.20	58.76	63.86
Isolated	55.12	62.07	69.93	65.59	63.45	63.74	61.73	56.77	56.98	56.93	61.73	63.21	58.62	30.29	48.31	59.46	56.56	63.08
Natural	53.88	59.60	67.83	63.25	61.68	61.16	60.30	53.99	53.45	52.40	57.84	62.68	57.09	29.54	47.44	56.40	55.10	60.12
MSL	55.21	60.93	68.99	64.61	63.06	62.34	61.63	55.14	54.64	53.74	59.17	64.14	58.49	30.41	48.57	57.73	56.42	61.55
CLIMB	57.75	63.84	72.47	67.74	66.25	65.45	64.96	58.17	57.80	57.09	62.03	67.11	61.45	32.24	50.96	60.64	59.12	63.02

Table 11: Detailed per-language performance on the **GPQA** benchmark (0-shot accuracy).

Model / Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TH	TL	TR	VI	ZH
Open Source Multilingual LLMs																		
LLaMA-3.2	23.54	25.67	26.39	23.30	26.39	23.30	22.81	22.81	25.20	23.79	23.30	23.30	24.02	23.79	25.67	23.30	23.79	24.50
Qwen-3	27.47	31.60	32.53	28.06	32.79	31.89	31.60	30.38	29.79	27.72	34.87	31.60	32.53	28.95	31.89	31.30	28.36	31.60
Gemma-3	24.58	23.30	25.45	23.03	24.44	23.49	25.25	23.49	22.76	24.91	23.49	24.91	22.05	24.58	21.25	22.76	26.60	25.45
Different Data Allocation Methods																		
Uniform	25.41	26.48	27.28	25.72	27.28	26.72	25.91	24.65	26.24	24.90	27.52	26.72	26.72	26.97	25.71	25.71	25.41	25.91
Isolated	23.51	22.77	26.27	21.76	24.24	25.03	25.03	23.51	24.51	25.24	25.24	25.03	23.27	28.64	25.03	23.51	24.80	25.03
Natural	24.97	26.40	28.24	26.19	27.77	26.94	26.64	25.17	25.02	25.60	26.78	27.15	26.26	25.66	26.56	26.47	25.70	27.16
MSL	24.23	25.43	27.06	25.34	26.61	25.94	25.64	24.19	24.00	24.72	25.89	26.27	25.10	24.44	25.50	25.39	24.53	26.20
CLIMB	25.96	27.09	28.60	27.24	28.46	27.71	27.44	25.88	25.71	26.41	27.59	28.00	26.67	26.13	27.28	27.18	26.20	26.98

Table 12: Detailed per-language performance on the **HellaSwag** benchmark (10-shot accuracy).

Model / Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TA	TH	TR	VI	ZH
Open Source Multilingual LLMs																		
LLaMA-3.2	36.12	42.69	67.10	47.45	46.86	43.14	45.06	36.78	37.02	40.66	43.35	46.28	42.38	35.30	35.21	36.46	42.35	43.48
Qwen-3	39.37	45.66	65.36	51.17	50.93	46.08	48.86	42.25	39.52	42.49	43.30	51.26	45.53	35.81	36.79	36.01	44.79	52.39
Gemma-3	35.91	40.54	58.37	42.84	45.18	42.15	43.81	37.61	36.51	38.84	41.17	44.27	39.03	34.94	33.76	34.87	38.20	41.35
Different Data Allocation Methods																		
Uniform	45.03	49.72	58.01	54.03	54.83	52.41	52.90	45.08	43.01	47.15	51.51	53.67	48.50	30.04	39.10	45.27	47.84	48.06
Isolated	44.67	49.70	58.02	53.31	54.35	52.06	52.25	45.70	43.34	47.23	51.35	53.44	48.33	30.61	39.68	45.99	48.34	47.58
Natural	42.64	46.95	55.16	51.21	51.95	49.75	50.12	42.98	41.06	44.48	48.50	50.71	45.73	29.08	37.87	43.24	45.31	45.52
MSL	43.80	48.06	56.71	52.67	53.37	51.11	51.45	44.12	42.19	45.69	49.92	52.23	46.93	30.37	39.32	44.90	46.82	46.73
CLIMB	45.71	49.93	58.99	54.39	55.31	53.20	53.38	46.05	44.05	47.67	51.79	54.15	48.65	32.10	40.95	46.75	48.51	45.92

Table 13: Detailed per-language performance on the **TruthfulQA** benchmark (0-shot accuracy).

Model / Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TH	TL	TR	VI	ZH
Open Source Multilingual LLMs																		
LLaMA-3.2	38.66	37.41	34.59	37.20	36.56	38.92	34.16	36.91	39.59	36.09	37.20	36.76	40.04	36.76	36.32	37.90	42.73	41.36
Qwen-3	49.16	50.27	47.94	50.01	50.54	48.24	50.01	51.56	47.18	47.70	46.15	52.98	51.17	47.35	42.46	46.15	52.03	48.96
Gemma-3	39.20	41.60	39.60	39.83	38.73	42.47	40.92	37.39	41.14	37.83	36.28	42.03	44.24	40.05	34.51	39.38	43.80	42.25
Different Data Allocation Methods																		
Uniform	41.42	38.07	38.28	41.63	41.42	39.09	40.74	35.97	41.42	39.09	39.73	39.94	39.94	38.07	37.02	38.28	41.42	41.63
Isolated	36.55	42.71	37.18	39.93	40.55	39.49	40.13	37.59	40.78	36.98	39.73	41.43	38.48	38.26	38.26	40.13	45.00	41.61
Natural	41.62	40.17	40.95	42.12	41.83	40.91	40.76	38.60	40.09	39.40	40.42	41.94	40.99	39.06	37.51	39.74	42.35	42.91
MSL	40.53	39.05	39.91	41.02	40.76	39.76	39.57	37.50	38.93	38.26	39.27	40.70	39.75	37.84	36.35	38.74	41.18	41.70
CLIMB	42.05	40.57	41.53	42.57	42.32	41.36	41.15	38.99	40.49	39.72	40.80	42.17	41.09	39.31	37.78	40.38	42.62	42.03

although Section 4.2 indicates robust performance at a larger scale (7B), explicitly incorporating model size (N) into the allocation optimization could potentially yield even more optimal data distributions. Exploring how scaling laws evolve explicitly with both dataset size (D) and model scale (N) thus remains an open area for future research.

Secondly, our current methodology exclusively considers cross-lingual transfer between languages included within the training dataset. An important and intriguing direction for future work involves extending our approach to account for potential transfer effects to and from languages not directly represented in the training set. Such an extension would enable more comprehensive and strategically informed allocation decisions, optimizing not just for immediate languages but also for broader linguistic coverage and potential downstream adaptability.

H Social Impact

CLIMB contributes positively by systematically enhancing multilingual performance in large language models (LLMs), thereby significantly improving global accessibility to advanced AI capabilities across diverse linguistic communities. Such improvements have the potential to reduce linguistic biases, bridge language gaps, and enhance equitable information access globally. However, there remain potential risks, including inadvertent reinforcement of cultural or linguistic biases inherent in training data and the possibility of over-reliance on optimized multilingual models leading to reduced human oversight and critical evaluation. It is crucial to responsibly deploy CLIMB-optimized models with ongoing evaluation and monitoring, actively addressing ethical considerations and biases to ensure equitable and inclusive benefits.

References

- [1] 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.
- [2] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.
- [3] 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, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [4] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

- [5] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [6] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [7] Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. Fineweb2: A sparkling update with 1000s of languages, December 2024.
- [8] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online, November 2020. Association for Computational Linguistics.
- [9] 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. In *First Conference on Language Modeling*, 2024.
- [10] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermiş, and Sara Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2025.
- [11] 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.
- [12] Hugo Touvron and Louis Martin et al. Llama 2: Open foundation and fine-tuned chat models, 2023.
- [13] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.