ShapleyLaw: A Game-Theoretic Approach to Multilingual Scaling Laws

Anonymous ACL submission

Abstract

In multilingual pretraining, the test loss of a pretrained model is heavily influenced by the proportion of each language in the pretraining data, namely the *language mixture ratios*. Multilingual scaling laws can predict the test loss under different language mixture ratios and can therefore be used to estimate the optimal ratios. However, the current approaches to multilingual scaling laws do not measure the *cross-lingual transfer* effect, resulting in suboptimal mixture ratios. In this paper, we consider multilingual pretraining as a cooperative game in which each language acts as a player that jointly contributes to pretraining, gaining the resulting reduction in test loss as the payoff. Consequently, from the perspective of cooperative game theory, we quantify the cross-lingual transfer from each language by its contribution in the game, and propose a game-theoretic multilingual scaling law called *ShapleyLaw*. Our experiments show that ShapleyLaw outperforms baseline methods in model performance prediction and language mixture optimization.[1]

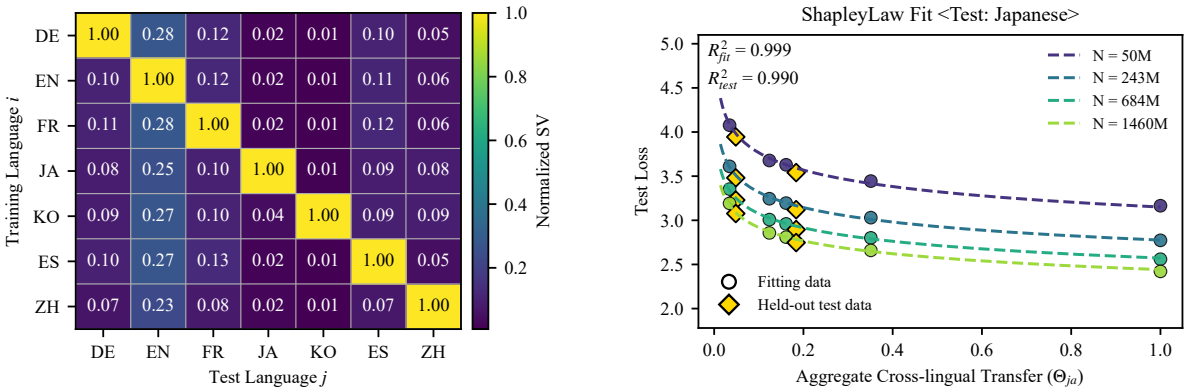
1 Introduction

As large language models (LLMs) serve increasingly diverse users, multilingual pretraining on mixed-language corpora has become standard practice for modern foundation models (Liu et al., 2020; Lin et al., 2020; Huang et al., 2023; Zhao et al., 2024; Xu et al., 2025; Wang et al., 2025a). The performance of a pretrained model on a given language j depends not only on the model size N and the pretraining data size D, but also critically on the proportion of each language i in the pretraining corpus, i.e., the mixture ratios $\mathbf{p}$. This is because knowledge learned from language i can transfer to language j, a phenomenon known as *cross-lingual transfer* (Weiss et al., 2016; Conneau and Lample, 2019; Schuster et al., 2019; Pfeiffer et al., 2020; Fernandes et al., 2023; Isonuma and Titov, 2024; He et al., 2025; Wang et al., 2025b). Therefore, identifying an optimal set of language mixture ratios that minimizes the test loss is a crucial problem for multilingual pretraining.

Recently, He et al. (2025) proposed using *multilingual scaling laws* to predict the test loss of pretrained models under different language mixture ratios, thereby optimizing language mixture. A central challenge in constructing such laws is how to measure cross-lingual transfer and relate mixture ratios to test loss. To simplify the complicated transfer effect among languages, the authors grouped languages by language family and assumed independence across families: *cross-lingual transfer occurs only among languages within the same family.* Then, they constructed multilingual scaling laws for each language family, referred to as *FamilyLaw*. However, FamilyLaw can merely be used to optimize mixture ratios across language families rather than across individual languages; the authors simply used the smoothed sampling strategy to allocate language mixture ratios within each family.

Moreover, the family-independence assumption underlying FamilyLaw does not hold. In our experiments, we find that cross-lingual transfer also exists between languages from different families, and that its strength varies substantially. For example, as Japanese and Chinese share substantial lexical overlap, although Chinese, Japanese, and Spanish belong to different language families, Japanese data improves performance on the Chinese task much more than Spanish data does (see Section 2.2). This result is consistent with the findings in Lin et al. (2019), which showed that the cross-lingual transfer effect from language i to language j largely depends on linguistic distances such as geographic distance and syntactic distance. By ignoring cross-family transfer, FamilyLaw may inform language mixtures that deviate substantially from the optimum.

[1] Our code and data are available at https://anonymous.4open.science/r/ShapleyLaw-DE3F/.



(a) Cross-lingual transfer measured by the SV.

(b) Test loss prediction by ShapleyLaw.

Figure 1: **(a)** Cross-lingual transfer effects quantified by normalized SVs $\phi_{i,j}^{NSV}$ estimated from small-scale models ($N = 50$M, $D = 50$B), shown as an asymmetric matrix. In contrast, the FamilyLaw (He et al., 2025) baseline yields zero off-diagonal entries. **(b)** Across diverse pre-training mixtures and model scales $N \in \{50\text{M}, 243\text{M}, 684\text{M}, 1460\text{M}\}$ with total data size $D = 100$B, the proposed ShapleyLaw $\mathcal{L}_j(N, D, \Theta_{ja})$ accurately fits and predicts Japanese performance.

In this work, we propose quantifying cross-lingual transfer from the perspective of cooperative game theory (CGT). We formulate multilingual pretraining as a cooperative game, in which each language is treated as a player and the payoff is defined by the reduction in test loss on a target evaluation language. We then use the Shapley value (SV) to measure the contribution of each training language to the target language, and interpret this contribution as cross-lingual transfer, as illustrated in Figure 1a.

Based on this transfer measure, we propose *ShapleyLaw*, a game-theoretic multilingual scaling law that predicts test loss from model size, data size, and language mixture ratios. To capture the effect of multilingual mixture ratios on a target language, we introduce an *aggregate cross-lingual transfer* quantity that summarizes the transfer effect of a mixture through SV-based transfer strengths. In this way, ShapleyLaw extends standard scaling law formulations with a language-level characterization of cross-lingual transfer. As shown in Figure 1b, it accurately predicts test loss on various mixtures and model scales, enabling mixture optimization in individual languages.

We summarize our contributions as follows.

- **SV-based measure of cross-lingual transfer.** We formulate multilingual pretraining as a cooperative game, which allows us to apply the SV to quantify the cross-lingual transfer effect across languages.
- **ShapleyLaw for multilingual scaling laws.** ShapleyLaw is the first framework to characterize the functional relationship between language-level mixture ratios $\mathbf{p}$ and test loss, thereby informing optimal mixture ratios across languages.
- **Empirical validation.** We evaluate ShapleyLaw on multilingual pretraining tasks covering 7 languages. The results show that ShapleyLaw consistently outperforms baseline methods in test loss prediction, mixture ratios optimization, and downstream task performance prediction.

2 Preliminaries

2.1 Problem Statement

We consider pretraining a multilingual model $\theta_{\mathcal{K}}$ that supports a given set of languages $\mathcal{K} = \{1, \ldots, K\}$. For each language $i \in \mathcal{K}$, we have a corpus of size $\bar{D}_i$. To construct a pretraining dataset $\mathcal{D}_{\mathcal{K}}$ of size D for pretraining $\theta_{\mathcal{K}}$, for each language i, we sample a training subset $\mathcal{D}_i$ from the corpus according to a mixture ratio p_i, where the subset size is $D_i = p_i \cdot D$. The union of all the languages' training subsets forms the pretraining dataset, i.e., $\mathcal{D}_{\mathcal{K}} = \bigcup_{i=1}^{K} \mathcal{D}_i$.

Performance prediction. Because pretraining a large language model is computationally expensive, we aim to predict the test loss of the pretrained model $\theta_{\mathcal{K}}$ on a target evaluation task $\mathcal{V}_{\mathcal{K}}$ by constructing a multilingual scaling law. Specifically, the target task $\mathcal{V}_{\mathcal{K}} = \{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$ consists of multiple language-specific tasks, each denoted by $\mathcal{V}_j$, where $\mathcal{V}_j$ evaluates the model's performance on language $j \in \mathcal{K}$. For each task $\mathcal{V}_j$, we estimate a language-specific scaling law $\mathcal{L}_j(N, D, \mathbf{p})$ to predict the loss of model $\theta_{\mathcal{K}}$ on that task, where N is

the model size, and $\mathbf{p} = [p_1, \ldots, p_K]$ is the vector of all the language mixture ratios. Then, we can predict the model's performance on the target task $\mathcal{V}_{\mathcal{K}}$ by aggregating all the language-specific scaling laws into a multilingual scaling law:

$$\mathcal{L}(N, D, \mathbf{p}) = \sum_{j=1}^{K} w_j \cdot \mathcal{L}_j(N, D, \mathbf{p})$$

where w_j denotes a weight representing the relative importance of task $\mathcal{V}_j$. For example, we can determine $w_1, \ldots, w_K$ by estimating the real-world distribution of language tokens across languages.

Mixture optimization. In addition to predicting test loss, which is the traditional function of scaling laws, multilingual scaling laws can further be used to inform optimal mixture ratios. Specifically, given a model size N, a total data size D, and importance weights $\mathbf{w} = [w_1, \ldots, w_K]$, we can estimate the optimal mixture ratios $\mathbf{p}^\star$ by solving the following optimization problem:

$$\mathbf{p}^\star = \arg\min_{\mathbf{p}} \sum_{j=1}^{K} w_j \cdot \mathcal{L}_j(N, D, \mathbf{p}). \quad (1)$$

In summary, a multilingual scaling law should (1) accurately predict test losses for different parameter configurations $(N, D, \mathbf{p})$, and (2) enable the model pretrained with the optimized mixture ratios $\mathbf{p}^\star$ to achieve a test loss as lower as possible.

2.2 Limitations of the FamilyLaw

He et al. (2025) proposed the current SOTA multilingual scaling law, which we refer to as FamilyLaw. To simplify the modeling of cross-lingual transfer, FamilyLaw groups the languages into M language families and predicts the test loss $\mathcal{L}_m^{\mathbf{fam}}$ for each family $m \in [M]$ as follows:

$$\mathcal{L}_m^{\mathbf{fam}}(N, D, \mathbf{p}) = (E_m + \frac{A_m}{N^{\alpha_m}} + \frac{B_m}{D^{\beta_m}})(\mathbf{p}_m)^{-\gamma_m}$$

where $\mathbf{p}_m$ is the sum of the mixture ratios for languages in family m, and $A_m, B_m, \alpha_m, \beta_m$ are fitting parameters.

However, we find that FamilyLaw has the following three limitations. **(1)**, for performance prediction, FamilyLaw does not construct language-specific scaling laws, making it difficult to predict model performance on single-language tasks. **(2)**, for mixture ratio optimization, FamilyLaw only optimizes mixture ratios across language families, but does not inform the optimal mixture ratios within each family (He et al. (2025) simply applied smoothed sampling to determine the intra-family ratios). **(3)**, FamilyLaw relies on a core assumption that cross-lingual transfer occurs only within the same language family. However, as shown in Figure 2, this assumption does not hold in practice: Although both Japanese and Spanish belong to language families different from Chinese, Japanese data exhibits substantially stronger cross-lingual transfer for the Chinese task because Japanese and Chinese share substantial lexical overlap.

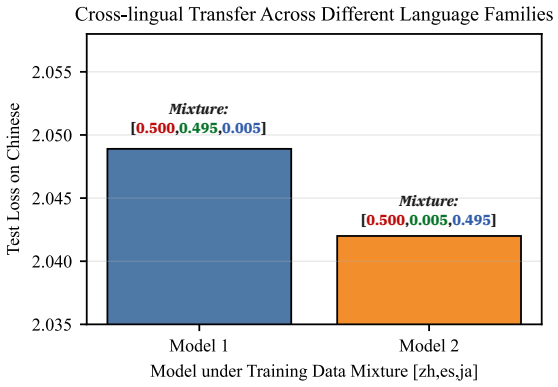


Figure 2: Consider pretraining two models for a Chinese task with training data mixtures in Chinese (zh), Japanese (es), and Spanish (es), which belong to different language families. We first pretrain Model 1 by sampling Chinese, Spanish, and Japanese data with ratios $[0.500, 0.495, 0.005]$, respectively. We then swap the mixture ratios of Japanese and Spanish to pretrain Model 2. Although Models 1 and 2 share the same model size N=684M and data size D=100B, they exhibit significantly different losses on the Chinese task.

Our objectives. Considering the above limitations, we aim to develop a multilingual scaling law that (1) predicts language-specific test loss, (2) informs optimal mixture ratios across individual languages rather than language families, and (3) models cross-lingual transfer across languages.

3 Shapley Value-Based Scaling Law

In this section, we propose a multilingual scaling law designed from the perspective of CGT.

3.1 Multilingual Pretraining Game

The core challenge in developing multilingual scaling laws lies in quantifying cross-lingual transfers across languages. We address this challenge using CGT, which studies how to fairly assess the contribution of each player when multiple players

jointly complete a task. In the context of multilingual pretraining, each language $i \in \mathcal{K}$ can likewise be viewed as a player. These players jointly pretrain a model, which is used to complete an evaluation task $\mathcal{V}_j$ in a target language j. This formulation allows us to use contribution metrics from CGT to quantify the contribution of each language i to the target language j, and thereby assess cross-lingual transfer. We refer to this cooperative game as *multilingual pretraining game*, defined as follows.

Definition 3.1 (Multilingual Pretraining Game (MPG))**.** *An MPG $\mathcal{G} = (\mathcal{K}, N, D, \mathcal{V})$ is specified by a language set $\mathcal{K}$, a model size N, a data size D, and a target evaluation task $\mathcal{V}$. For any language subset $S \subseteq \mathcal{K}$, the payoff $v_j^{\mathcal{G}}(S)$ of S on any language-specific task $\mathcal{V}_j \in \mathcal{V}$ is the decreased test loss by pretraining $\theta_{\mathcal{K}}$ using a pretraining dataset $\mathcal{D}_S^{uni}$, where $\mathcal{D}_S^{uni}$ consists of training data from languages S with uniform mixture ratios. A contribution metric ϕ to the game $\mathcal{G}$ returns a $K \times K$ contribution matrix $\phi(\mathcal{G}) = \{\phi_{i,j(\mathcal{G})}\}_{i,j\in\mathcal{K}}$ that distributes the payoff $v_j^{\mathcal{G}}(\mathcal{K})$ among all the languages $\mathcal{K}$ for all task $\mathcal{V}_j \in \mathcal{V}$, i.e., $\forall j, \sum_{i=1}^{K} \phi_{i,j}(\mathcal{G}) = v_j^{\mathcal{G}}(\mathcal{K})$.*

Intuitively, in the MPG $\mathcal{G}$, all the languages $\mathcal{K}$ as players participate in pretraining with the same amount of data. For each task $\mathcal{V}_j$, they gain a payoff of $v_j^{\mathcal{G}}(\mathcal{K})$, which is the reduction in the test loss of the pretrained model. The contribution of language i to task $\mathcal{V}_j$, denoted as $\phi_{i,j}(\mathcal{G})$, measures the portion of this test loss reduction attributable to language i. Consequently, we can measure the cross-lingual transfer from language i to language j as the contribution $\phi_{i,j}$ in MPG.

3.2 Shapley Value for Measuring Cross-Lingual Transfer

The Shapley value (SV) is a widely-adopted contribution metric because it is the only metric that satisfies four important properties: *linearity*, *efficiency*, *symmetry*, and *null player* (see Appendix A.1 for details). In MPG, we can also employ the SV ϕ^{SV} to compute the contribution matrix as follows:

$$\phi_{i,j}^{SV}(\mathcal{G}) = \textstyle\sum_{S\subseteq\mathcal{K}\setminus\{i\}} W(S) \cdot \Delta_{i,j}^{\mathcal{G}}(S)$$

where $W(S) = \frac{|S|!(K-|S|-1)!}{K!}$, and $\Delta_{i,j}^{\mathcal{G}}(S) = v_j^{\mathcal{G}}(S \cup \{i\}) - v_j^{\mathcal{G}}(S)$.

Intuitively, the SV measures the marginal contribution of language i to the reduction in test loss on task $\mathcal{V}_j$, averaged over all possible subsets of languages. Specifically, the SV (1) enumerates all language subsets $S \subseteq \mathcal{K} \setminus \{i\}$, (2) calculates the marginal improvement $\Delta_{i,j}^{\mathcal{G}}(S)$ in the test loss by adding language i into S for pretraining, and (3) finally aggregates the improvements by language i across all subsets S as language i's contribution to task $\mathcal{V}_j$, where $W(S)$ is the weight for aggregation. To ensure comparability across different tasks, we further apply exponential normalization to compute the normalized SV as follows:

$$\phi_{i,j}^{NSV}(\mathcal{G}) = \exp\left(\phi_{i,j}^{SV}(\mathcal{G}) - \max_{i'} \phi_{i',j}^{SV}(\mathcal{G})\right)$$

which ensures that the language i with the largest contribution to task $\mathcal{V}_j$ has a normalized SV of 1, while the values for other languages lie in the range $(0, 1]$. Consequently, the normalized SV can serve as a measure of the cross-lingual transfer from any language i to any language j. For simplicity, we denote $\phi_{i,j}^{NSV}(\mathcal{G})$ as $\phi_{i,j}^{NSV}$ in the following, when no ambiguity arises.

3.3 Construction of ShapleyLaw

Finally, we construct a multilingual scaling law, named *ShapleyLaw*, based on the SV. Formally, given a set of parameter configurations $(N, D, \mathbf{p})$ and a set of normalized SVs $\{\phi_{i,j}^{NSV}\}$ that represent cross-lingual transfers among languages in $\mathcal{K}$, ShapleyLaw predicts test loss on each task $\mathcal{V}_j$ as follows:

$$\mathcal{L}_j(N, D, \mathbf{p} \mid \{\phi_{i,j}^{NSV}\}) = \mathcal{L}_j^{\mathbf{chin}}(N, D) \cdot (\Theta_j)^{-\gamma_j}$$

where $E_j, A_j, B_j, \alpha_j, \beta_j, \gamma_j$ are fitting parameters; $\mathcal{L}_j^{\mathbf{chin}}(N, D) = E_j + \frac{A_j}{N^{\alpha_j}} + \frac{B_j}{D^{\beta_j}}$ corresponds to the Chinchilla scaling law (Hoffmann et al., 2022) that does not consider multilingual mixture ratios; and $\Theta_j = \sum_{i\in\mathcal{K}} p_i \cdot \phi_{i,j}^{NSV}$ measures the *aggregate cross-lingual transfer* from languages $\mathcal{K}$ to language j weighted by their mixture ratios. Intuitively, ShapleyLaw rescales the Chinchilla scaling law using Θ_j, thereby incorporating the effect of cross-lingual transfer on pretraining.

Note that SV-based cross-lingual transfer exhibits transferability. This implies that we can construct an MPG $\mathcal{G} = (\mathcal{K}, N', D', \mathcal{V})$ using a small model size N' and data size D' to compute the normalized SVs. These normalized SVs can then be treated as fixed parameters when constructing ShapleyLaw for larger model sizes $N \gg N'$ and data sizes $D \gg D'$. As illustrated in Figure 1a, we compute the normalized SVs under the setting $N' = 50$M and $D' = 50$B; ShapleyLaw accurately fits the test losses across in the settings with $N \in \{243\text{M}, 684\text{M}, 1460\text{M}\}$ and $D = 100$B.

Besides predicting language-specific test loss, ShapleyLaw can also be used to estimate *optimal mixture ratios* by solving the optimization problem in Formula 1 (see Appendix A.2). Therefore, ShapleyLaw achieves our three objectives.

SV approximation. Since the SV $\phi_{i,j}^{SV}$ enumerates all non-empty language subsets $S \subseteq \mathcal{K}$ to evaluate their payoff $v_j^{\mathcal{G}}(S)$, we need to pretrain $2^K - 1$ models for SV computation, which is NP-hard. To accelerate SV computation, we can employ existing SV approximation algorithms (Ghorbani and Zou, 2019; Jia et al., 2019) (see Appendix A.3) to reduce the number of pretrained models, thereby achieving polynomial computational complexity.

4 Experiments

Setup. We describe the basic experimental setup here. More details are provided in Appendix B. We study $K = 7$ languages from five language families, as summarized in Table 4. We sample 0.1% of the corpora as a fixed test set for evaluating all pretrained models. We consider four model scales with non-embedding parameter counts $N \in \{50\text{M}, 243\text{M}, 684\text{M}, 1460\text{M}\}$, following the open-source llm-jp configurations. All models share the same architecture family and tokenizers. We consider the cross-entropy (CE) loss for evaluation. The SVs are computed in an MPG with model size $N = 50\text{M}$ and $D = 50\text{B}$ across all experiments.

Research questions. Our experiments address the following research questions. **Q1** (Performance Prediction): Can ShapleyLaw more accurately predict the model's test loss? **Q2** (Mixture Optimization): Can ShapleyLaw achieve more optimal language mixtures? **Q3** (Downstream Tasks): Can ShapleyLaw's ability to predict model performance generalize to downstream tasks? **Q4** (SV Approximation): How is the fitting performance of ShapleyLaw based on approximate SVs?

4.1 Experiments on Performance Prediction

Experiment design. We pretrain models with $N \in \{50\text{M}, 243\text{M}, 684\text{M}, 1460\text{M}\}$ and $D = 100\text{B}$. At each model scale, we consider both *controlled mixtures* and *empirical mixtures*. Controlled mixtures isolate the effect of language mixtures under fixed ratios, including monolingual training, uniform mixtures over selected languages (e.g., zh–ja–ko and en–de–fr–es, abbreviated as *ZJK* and *EDFS*), and the uniform mixture over all languages (*Uni*). Empirical mixtures mirror multilingual data distributions adopted in prior LLM studies, including the language distributions of Common Crawl[2] and FineWeb-2 (Penedo et al., 2025), the English-centric multilingual model LLaMA-2 (Touvron et al., 2023), and the Japanese-focused pretraining corpus LLM-JP-V4 (LLM-jp et al., 2024) (abbreviated as *CC*, *FW2*, *L2*, and *JPv4*).

In addition, we examine 50 randomly sampled language mixtures under $N = 243\text{M}$ to further evaluate generalization to unseen mixture ratios. To ensure reliable fitting, we exclude those random mixtures where a training language i has subset size $D_i \leq 1.5\text{B}$. This regime corresponds to extremely low data availability for language i, causing optimization to become unstable even for $N = 243\text{M}$ models. This threshold remains practically relevant: in FineWeb-scale corpora, a language budget of roughly 1.5B tokens already extends to many low-resource languages considered in current multilingual LLM studies, including Basque and Armenian. We then split the remaining runs into fitting and held-out evaluation subsets.

Results. Figure 3 shows the fitted results for Chinese, Spanish, and Japanese. On curated multilingual mixtures across four model scales, ShapleyLaw achieves strong held-out performance, with $R^2_{\text{eval}} = 0.974$ for Chinese, 0.912 for Spanish, and 0.990 for Japanese. On randomly sampled mixtures at $N = 243\text{M}$, the fitted curves remain smooth with respect to Θ, with training-set fits of $R^2 = 0.975, 0.942$, and 0.987 for Chinese, Spanish, and Japanese, respectively. These results indicate that the proposed method makes stable predictions across sampling regimes and typologically diverse languages. Additional results for French, German and Korean are provided in Appendix C.1.

We next compare ShapleyLaw with the family-level baseline FamilyLaw in a controlled setting that isolates cross-family transfer. As shown in Figure 4, we run experiments on the *ZJS* mixture. We fix the Chinese mixture ratio at $p_{\text{zh}} = 0.05$ and vary only the Japanese–Spanish proportions, keeping the total token budget fixed. Under FamilyLaw, both Japanese and Spanish lie outside the Chinese language family, so changing their ratios should not affect Chinese performance; accordingly, the fitted relation is nearly constant and fails to explain the observed variation ($R^2 \leq 0$). In contrast, ShapleyLaw captures the cross-language transfer and fits the loss

[2]https://commoncrawl.org

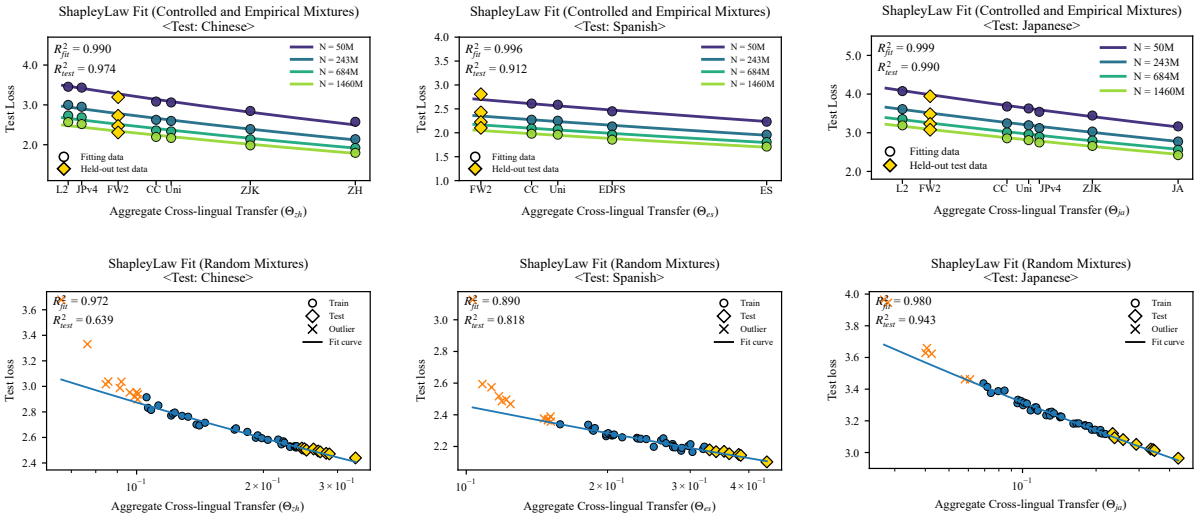


Figure 3: **Fitting ShapleyLaw across multilingual mixtures and target languages. Top row**: fits with respect to aggregate cross-lingual transfer Θ on curated multilingual mixtures for Chinese (ZH), Spanish (ES), and Japanese (JA) across four model scales $\{50\text{M}, 243\text{M}, 684\text{M}, 1460\text{M}\}$; circles denote fitting points and yellow diamonds denote held-out evaluation points. **Bottom row**: fits under randomly sampled mixtures at model scale $N = 243\text{M}$; circles indicate fitting points, yellow diamonds indicate held-out points, and orange crosses denote removed outliers.

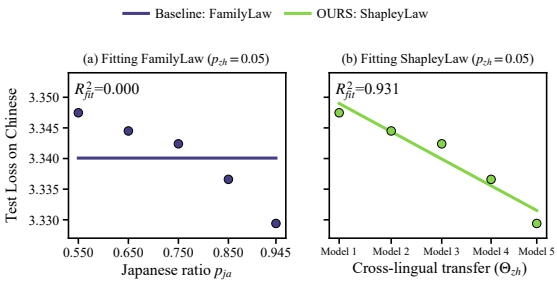


Figure 4: **Controlled comparison between FamilyLaw and ShapleyLaw on a Chinese evaluation task.** We pretrain multilingual models with $N = 50\text{M}$ and $D = 100\text{B}$ using Chinese, Japanese, and Spanish corpora. Starting from the mixture $[0.05, 0.55, 0.40]$, we vary the Spanish–Japanese proportions up to $[0.05, 0.945, 0.05]$ while keeping the Chinese ratio fixed. Performance is evaluated on the Chinese test set. **Left:** FamilyLaw, where x-axis stands for language ratio, yields a poor fit. **Right:** ShapleyLaw, where x-axis stands for different Chinese aggregate transfer Θ, fits accurately.

on Chinese much more accurately ($R^2 = 0.931$). This result shows that family-level aggregation is too coarse to model important cross-lingual interactions in multilingual pretraining, such as shared script and lexical overlap.

4.2 Experiments on Mixture Optimization

Experiment design. We compare the ShapleyLaw-based mixture optimization against four representative baselines chosen to reflect the main strategies used in multilingual data allocation, including *Uniform Sampling*, *Smoothed Sampling*, *FamilyLaw Optimization*, and *Empirical Mixtures* (see Appendix B for details). We pretrain a model ($N = 243\text{M}$ and $D = 200\text{B}$) using the mixture ratios obtained by each method.

We consider two settings of w_j for weighting the importance of each language-specific task $\mathcal{V}_j$. **Unweighted Sum:** Each task $\mathcal{V}_j$ receives equal weight, so the objective is to minimize the average multilingual loss. **Normalized Sum:** Each task $\mathcal{V}_j$ is weighted by the inverse of its monolingual loss. This ensures that intrinsically harder languages do not dominate the aggregate objective in Formula 1 and provides a more balanced evaluation of multilingual trade-offs across languages.

Results. Table 1 shows the optimized mixtures and their resulting CE loss. Under both preference settings, the mixtures obtained from ShapleyLaw outperform all baselines in the final weighted loss. The gains are especially meaningful because they are obtained under actual pretraining rather than only from the fitted objective, indicating that the learned law transfers effectively from prediction to optimization. By contrast, the mixture obtained from FamilyLaw is consistently less effective in the per-language setting, suggesting that family-level aggregation is too coarse to capture the trade-offs needed for fine-grained multilingual allocation.

4.3 Experiments on Downstream Tasks

Experiment design. We focus on three representative multilingual benchmarks that cover comple-

Table 1: Comparison of mixture optimization strategies and their CE loss under the Unweighted Sum (**Left**) and Normalized Sum (**Right**) settings.

Strategy	Language Mixture Ratio (%)							CE Loss ↓
	en	de	fr	es	zh	ja	ko	$\mathcal{L}$
Uniform	14.3	14.3	14.3	14.3	14.3	14.3	14.2	18.354
$\alpha = 0.5$	14.8	14.8	16.2	14.1	15.2	21.4	12.8	18.349
Common Crawl	44.4	12.1	9.8	9.5	12.0	10.7	1.5	18.679
FineWeb-2	81.3	3.3	2.3	2.7	6.8	3.1	0.5	19.357
FamilyLaw	8.1	8.1	8.1	8.1	24.2	23.1	20.3	18.435
ShapleyLaw	1.8	16.9	12.6	9.6	20.1	19.8	19.2	**18.289**

Strategy	Language Mixture Ratio (%)							CE Loss ↓
	en	de	fr	es	zh	ja	ko	$\mathcal{L}$
Uniform	14.3	14.3	14.3	14.3	14.3	14.3	14.2	8.269
$\alpha = 0.5$	14.8	14.8	16.2	14.1	15.2	21.4	12.8	8.258
Common Crawl	44.4	12.1	9.8	9.5	12.0	10.7	1.5	8.285
FineWeb-2	81.3	3.3	2.3	2.7	6.8	3.1	0.5	8.616
FamilyLaw	8.3	8.3	8.3	8.2	26.3	19.5	21.1	8.327
ShapleyLaw	0.1	18.7	15.0	11.5	20.8	15.3	18.6	**8.199**

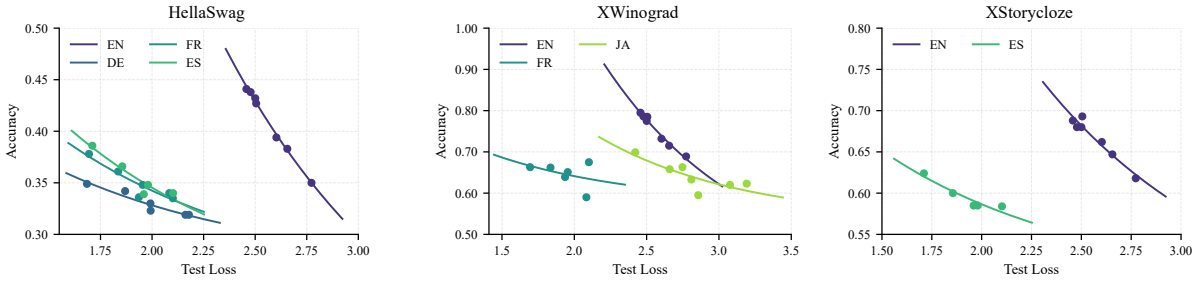


Figure 5: **Fitting Relationship between Downstream Performance and Test CE Loss.** Experiments are conducted on five downstream tasks for English, German, French, and Spanish. Across all tasks, we observe strong *negative correlations* ($|r| > 0.90$, with an average $R^2 = 0.931$), indicating that lower test loss consistently corresponds to higher downstream performance.

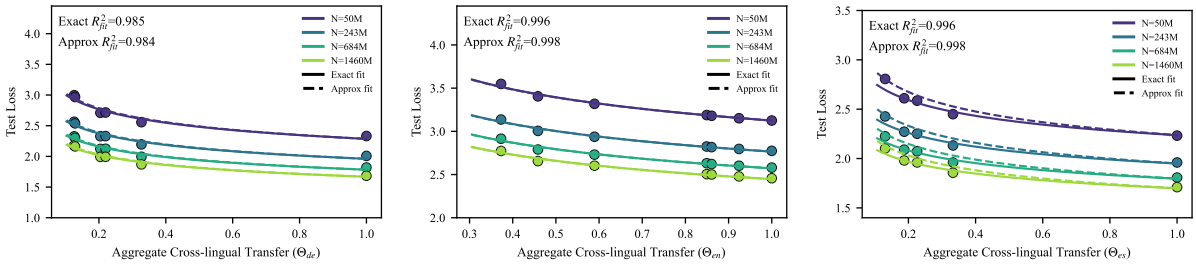


Figure 6: **Fitting results with exact and MC–approximated SVs.** To reduce the workload of SV computation, we use the MC approximation with $T = 50$ sampled permutations and compare with standard SV fitting curves.

mentary NLU abilities: **HellaSwag** (Zellers et al., 2019; Lai et al., 2023) for commonsense reasoning through contextually appropriate ending prediction, **XStoryCloze** (Lin et al., 2022; Mostafazadeh et al., 2017) for cross-lingual story ending prediction, and **XWinograd** for cross-lingual coreference resolution (Tikhonov and Ryabinin, 2021).

We evaluate pretrained models under the same setting as Section 4.1, using $N = 1.46$B and $D = 100$B, and the same multilingual mixture families (controlled and empirical mixtures). We report 5-shot accuracy on the standard benchmark splits. Specifically, we evaluate HellaSwag in {en, fr, es, de}, XWinograd in {en, fr, ja}, and XStoryCloze in {en, es}. For each task-language pair, we compare downstream accuracy against the corresponding pretraining test cross-entropy on the same language, measured on the held-out test corpus. This yields one calibration line per task-language pair.

Results. We fit a simple calibration model

$$s = f(\ell) = a \cdot g(\ell) + b, \tag{2}$$

where ℓ is the test loss, s is downstream accuracy, and a, b are fitting parameters (Chen et al., 2024). We choose $g(\ell) = \exp(-\ell)$ as the transformation function $g(\cdot)$ (see more details in Appendix C.2). Figure 5 shows the resulting fits for the three multilingual benchmarks, with one line per task-language pair and points colored by language. Across all evaluated languages, lower pretraining loss is consistently associated with higher downstream accuracy, and the relationship is generally well approximated by a linear fit after transformation. The fitted results further show that this trend holds not only within a single benchmark, but across multilingual commonsense reasoning, story completion, and coreference resolution tasks. Table 2 reports the corresponding R^2 values for each task-language pair.

Table 2: Corresponding R^2 values for the fitted relationship between downstream performance and test CE loss for each task–language pair. English achieves the highest R^2, while Japanese is the lowest. The results consistently show a negative correlation between downstream performance and test CE loss.

Task	de	en	fr	es	ja
HellaSwag	0.918	0.996	0.864	0.897	–
XWinograd	–	0.980	0.975	–	0.675
XStorycloze	–	0.947	–	0.905	–

These results show that for the multilingual NLU benchmarks considered here, language-specific pretraining loss remains strongly informative for downstream accuracy across diverse mixtures. As ShapleyLaw predicts this loss accurately across multilingual mixtures, the proposed scaling law provides a useful intermediate signal for downstream-oriented mixture design.

4.4 Approximate Shapley Value Estimation

Experiment Design. To improve the scalability of SV computation, we employ the Monte Carlo (MC) method that estimates SVs by averaging marginal contributions over T random language permutations (see Appendix A.3 for details). We use the resulting approximate SVs to fit ShapleyLaw.

We evaluate MC approximation in the same MPG as in the main experiments. In this setting, we use $T = 50$ random permutations to estimate the SV matrix (see Figure 11), and compare the fitted curves against those obtained with exact SVs.

Results. Figure 6 shows that the fitted curves obtained with MC-approximated SVs closely match those obtained with exact SVs, yielding comparable fitting quality in terms of R^2. In the 7-language setting, exact SV estimation requires evaluating $2^7 - 1 = 127$ pretrained models, whereas MC approximation with $T = 50$ requires 102 pretrained models. This result shows that substantial savings can already be achieved without noticeably degrading fit quality. As increasing the sampling budget improves the approximation accuracy, we further investigate the relationship between m and the mean absolute error of the SVs estimated by MC in Appendix A.3. These findings suggest that ShapleyLaw does not rely on exact SV computation to remain effective in practice. Instead, approximate SV estimation offers a practical path toward scaling the framework to larger multilingual settings.

5 Related Work

Neural scaling laws describe predictable relationships between language model performance and scale variables such as model parameters, training tokens, and compute (Kaplan et al., 2020; Hoffmann et al., 2022), but standard formulations do not account for *data composition*. Recent work extends scaling-law analysis to data mixtures in domain-specific pretraining (Que et al., 2024; Shukor et al., 2025). These works highlight the importance of data composition, but they do not provide a multilingual scaling law that directly optimizes *language-level* mixture ratios. In multilingual settings, prior work either models mixture ratios at the language-family level (He et al., 2025) or studies bilingual transfer in solely multilingual machine translation (Fernandes et al., 2023; Isik et al., 2025). To the best of our knowledge, ShapleyLaw is the first multilingual scaling law that enables optimization of language-level mixture ratios through explicit modeling of cross-lingual transfer between individual languages.

6 Conclusion

In this paper, we proposed *ShapleyLaw*, a game-theoretic multilingual scaling law that explicitly incorporates aggregate cross-lingual transfer into multilingual pre-training. Across experiments on seven languages from multiple language families, we show that the proposed law accurately predicts per-language pre-training loss across diverse model sizes, data budgets, and language mixtures, while capturing transfer effects that are missed by prior family-level approaches. We further demonstrate that ShapleyLaw supports practical optimization of multilingual sampling ratios, producing better mixtures than heuristic and prior-law baseline under real pre-training. In addition, the predicted language-specific loss was shown to remain informative for downstream multilingual NLU performance, highlighting the practical utility of the framework beyond pre-training loss alone. We also examine the efficiency of approximate Shapley value estimation and show that it preserves strong empirical performance, highlighting the potential scalability of our method to larger multilingual settings. Taken together, these findings suggest that multilingual scaling behavior is better understood through fine-grained, transfer-aware language interactions, and that our method provide a useful foundation for efficient multilingual mixture design.

Limitations

A limitation of ShapleyLaw is that computing the (exact) SVs requires pretraining $2^K - 1$ models. Therefore, compared to FamilyLaw, fitting ShapleyLaw involves additional pretraining workload. However, as discussed in Section 3.3, these models used for SV computation can be pretrained at a very small model scale and data size. The resulting SVs can then be treated as constants when fitting ShapleyLaw across arbitrary model scales and data sizes. In addition, the number of pretrained models required for computing SVs can be further reduced through SV approximation. In Section A.3, we apply the widely adopted Monte Carlo method (Jia et al., 2019) for SV approximation, which reduces the required number of pretrained models from $O(2^K)$ to $O(K^2 \log(K))$ while maintaining nearly identical fitting performance. Therefore, fitting ShapleyLaw remains computationally tractable.

Ethical Statements

We do not anticipate any ethical concerns arising from this work.

Use of Generative AI

AI assistants, such as ChatGPT, were used in this work for grammar checking and language editing.

References

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. 2024. Scaling laws for predicting downstream performance in llms. *arXiv preprint arXiv:2410.08527*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation. In *International Conference on Machine Learning*, pages 10053–10071. PMLR.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR.

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. 2025. Scaling laws for multilingual language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4257–4273, Vienna, Austria. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. 2025. Scaling laws for downstream task performance in machine translation. In *The Thirteenth International Conference on Learning Representations*.

Masaru Isonuma and Ivan Titov. 2024. Unlearning traces the influential training data of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6312–6325.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. 2019. Towards efficient data valuation based on the shapley value. In *The 22nd international conference on artificial intelligence and statistics*, pages 1167–1176. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and 1 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9019–9052.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, and 1 others. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, and 64 others. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *Preprint*, arXiv:2407.03963.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd workshop on linking models of lexical, sentential and discourse-level semantics*, pages 46–51.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7654–7673.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, and 1 others. 2024. D-cpt law: Domain-specific continual pre-training scaling law for large language models. *Advances in Neural Information Processing Systems*, 37:90318–90354.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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)*, pages 3795–3805.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. 2025. Scaling laws for optimal data mixtures. *arXiv preprint arXiv:2507.09404*.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 3534–3546.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, and 1 others. 2025a. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Ifeoluwa Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2025b. Multilingual language model pretraining using machine-translated data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28075–28095.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3(1):9.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319.

A More Technical Details

A.1 Theoretical Properties of the SV

The SV uniquely satisfies the following properties.

- *Efficiency:* The payoff (i.e., loss reduction) $v_j^{\mathcal{G}}(\mathcal{K})$ for each task $\mathcal{V}_j$ is fully distributed among all the languages $\mathcal{K}$, i.e., $\forall j, \sum_{i=1}^{K} \phi_{i,j}^{SV}(\mathcal{G}) = v_j^{\mathcal{G}}(\mathcal{K})$.
- *Linearity:* The SVs of each language to different evaluation tasks are linearly additive. Let $\phi_{i,j+j'}^{SV}$ denotes the Shapley value of language i for task $\mathcal{V}_{j+j'} = \mathcal{V}_j \cup \mathcal{V}_{j'}$. Then, for any language i, and for any two tasks $\mathcal{V}_j$ and $\mathcal{V}_{j'}$, we have $\phi_{i,j+j'}^{SV}(\mathcal{G}) = \phi_{i,j}^{SV}(\mathcal{G}) + \phi_{i,j'}^{SV}(\mathcal{G})$. That means that the contribution of language i to a combined task $\mathcal{V}_{j+j'}$ should be the sum of its contributions to the individual tasks $\mathcal{V}_j$ and $\mathcal{V}_{j'}$.
- *Symmetry:* If two languages i, i' are equivalent for improving model performance, then their SVs are identical. Formally, if $v_j^{\mathcal{G}}(S \cup \{i\}) = v_j^{\mathcal{G}}(S \cup \{i'\})$ for any language subset $S \subseteq \mathcal{K}$, then $\phi_{i,j}^{SV} = \phi_{i',j}^{SV}$.
- *Null player:* If a language i never improves the model performance when its data is included for pretraining, its contribution should be zero. That is, if $v_j^{\mathcal{G}}(S \cup \{i\}) = v_j^{\mathcal{G}}(S)$ for any language subset $S \subseteq \mathcal{K}$, then $\phi_{i,j}^{SV} = 0$.

A.2 ShapleyLaw Mixture Optimization

Optimal mixture ratios for language preferences. In practice, multilingual pre-training is often guided by an application-specific evaluation suite spanning multiple languages, where different languages may carry different importance. To model this, suppose we have nonnegative preference weights $\{w_j\}_{j=1}^{K}$ (e.g., proportional to user demand or benchmark priority) and consider the weighted objective

$$\mathcal{J}(\mathbf{p}) \;=\; \sum_{j=1}^{K} w_j \, \mathcal{L}_j(N, D, \mathbf{p})$$

where $\mathcal{L}_j$ is the test loss for target language j. Introducing ShapleyLaw into Eq. A.2 yields

$$\mathcal{J}(\mathbf{p}) \;=\; \sum_{j=1}^{K} w_j \cdot C_j \cdot \big(\Theta_j\big)^{-\gamma_j} \qquad (3)$$

where $C_j \triangleq E_j + \frac{A_j}{N^{\alpha_j}} + \frac{B_j}{D^{\beta_j}}$, which makes explicit how mixture influences the weighted loss through Θ_j.

Convexity and optimality conditions. For $\gamma_j > 0$, the function $x \mapsto x^{-\gamma_j}$ is convex over $x > 0$; since Θ_j is affine in $\mathbf{p}$, $\mathcal{J}(\mathbf{p})$ is a convex objective over the simplex whenever $\Theta_j > 0$ for all j. Therefore, the mixture selection problem

$$\min_{\mathbf{p} \in \Delta^{K-1}} \; \mathcal{J}(\mathbf{p})$$

admits a global optimum and can be solved efficiently. Using a Lagrangian with simplex constraints and Karush-Kuhn-Tucker (KKT) conditions, the gradient of Eq. 3 is

$$\frac{\partial \mathcal{J}}{\partial p_i} \;=\; -\sum_{j=1}^{K} w_j C_j \gamma_j \, \phi_{i,j}^{NSV} \, \big(\Theta_j\big)^{-(\gamma_j+1)}.$$

At an interior optimum (all $p_i > 0$), these gradients are equal up to a shared Lagrange multiplier, implying an intuitive *marginal utility equalization* principle: optimal mixtures allocate probability mass so that the expected reduction in weighted loss per unit increase of p_i is balanced across languages.

Closed-form solution in a diagonal (self-dominant) approximation. A useful and interpretable special case arises when each target language is primarily influenced by its own data, i.e., $\phi_{i,j}^{NSV}$ is approximately diagonal so that $\Theta_j \approx p_j \, r_j$ with $r_j \triangleq \phi_{j,j}^{NSV}$. Then Eq. 3 becomes separable:

$$\mathcal{J}(\mathbf{p}) \approx \sum_{j=1}^{K} w_j C_j (p_j r_j)^{-\gamma_j}.$$

The KKT conditions yield a closed-form optimal ratio:

$$p_i^{\star} \;=\; \frac{\Big(w_i C_i \gamma_i \, r_i^{-\gamma_i}\Big)^{\frac{1}{\gamma_i+1}}}{\sum_{m=1}^{K} \Big(w_m C_m \gamma_m \, r_m^{-\gamma_m}\Big)^{\frac{1}{\gamma_m+1}}}. \qquad (4)$$

First-order correction under cross-lingual transfer. For realistic multilingual settings, cross-lingual transfer introduces non-negligible off-diagonal terms in ϕ^{NSV}. Let the Shapley matrix be decomposed as

$$\phi^{NSV} = I + E$$

where E captures cross-lingual interactions with relatively small off-diagonal entries. Then

$$\Theta = \phi^{NSV}\mathbf{p} = \mathbf{p} + E\mathbf{p}.$$

Expanding $\Theta_i^{-(\gamma_i+1)}$ to first order around the diagonal solution $p^{(0)}$ from Eq. 4 yields

$$p_i^\star \approx p_i^{(0)} \left(1 - \frac{1}{\gamma_i + 1} \sum_{j, j \neq i} \phi_{i,j}^{NSV}\right) \quad (5)$$

followed by renormalization to satisfy the simplex constraint. The correction term depends on the column sum $\sum_{j \neq i} \phi_{i,j}^{NSV}$, which measures how strongly language i contributes to other languages through cross-lingual transfer. Languages with larger column sums effectively provide training signal to many other languages, and therefore require less direct sampling mass in the optimal mixture. Conversely, languages with weaker transfer connections receive higher allocation to maintain balanced marginal utility across tasks.

In practice, the available corpus size for each language may be limited. Let $\bar{D}_i$ denote the maximum usable token budget for language i, and let the total training budget be D. Then the mixture ratio must satisfy the additional box constraint

$$p_i \;\leq\; \bar{p}_i \quad \text{where} \quad \bar{p}_i \;=\; \frac{\bar{D}_i}{D}. \quad (6)$$

The mixture optimization problem thus becomes

$$\min_{\mathbf{p} \in \Delta^{K-1}} \mathcal{J}(\mathbf{p}) \quad \text{s.t.} \quad 0 \leq p_i \leq \bar{p}_i.$$

Because $\mathcal{J}(\mathbf{p})$ is convex, the constrained problem remains convex and can be characterized via KKT conditions with complementary slackness. However, a simple and effective approximation can be derived by projecting the unconstrained optimum onto the feasible region.

Clipped closed-form approximation. Let $p_i^{(0)}$ denote the unconstrained solution from Eq. 4 (or its cross-lingual correction in Eq. 5). Define the *active set* of data-limited languages

$$\mathcal{A} \;=\; \{\, i \mid p_i^{(0)} > \bar{p}_i \,\}.$$

For languages in $\mathcal{A}$, the optimal solution saturates the constraint:

$$p_i^\star \;=\; \bar{p}_i, \qquad i \in \mathcal{A}.$$

The remaining probability mass is

$$M \;=\; 1 - \sum_{i \in \mathcal{A}} \bar{p}_i$$

which is redistributed among the unconstrained languages $\mathcal{F} = \{1, \ldots, K\} \setminus \mathcal{A}$ proportionally to their original optimal ratios:

$$p_i^\star \;\approx\; M \cdot \frac{p_i^{(0)}}{\sum_{m \in \mathcal{F}} p_m^{(0)}}, \qquad i \in \mathcal{F}. \quad (7)$$

If redistribution causes additional violations ($p_i^\star > \bar{p}_i$ for some $i \in \mathcal{F}$), the clipping-and-redistribution step can be repeated. Since each iteration adds at least one index to $\mathcal{A}$, the procedure converges in at most K steps.

This approximation corresponds to projecting the unconstrained optimum onto the truncated simplex defined by Eq. 6. Intuitively, languages with insufficient data are trained to exhaustion, while the remaining training budget is allocated among data-rich languages according to their relative marginal utility. Notably, the structure of Eq. 7 preserves the optimal trade-off encoded by Shapley-aware scaling laws while ensuring practical feasibility under corpus-size limitations.

A.3 Monte Carlo Approximation of Shapley Values

Why approximation is needed. For a MPG, the definition of SV of language i on task j is

$$\phi_{i,j}^{SV}(\mathcal{G}) = \textstyle\sum_{S \subseteq \mathcal{K} \setminus i} W(S) \cdot \Delta_{i,j}^{\mathcal{G}}(S)$$

where $W(S) = \frac{|S|!(K-|S|-1)!}{K!}$ and $\Delta_{i,j}^{\mathcal{G}}(S) = v_j^{\mathcal{G}}(S \cup i) - v_j^{\mathcal{G}}(S)$. Computing such a SV exactly requires utility evaluation on all non-empty coalitions, which entails pre-training $2^K - 1$ models for K language candidates. This cost is already substantial for moderate K, and becomes prohibitive in multilingual pre-training where each utility evaluation itself requires training and validating a model. Therefore, approximation is necessary for scalable SV estimation.

Monte Carlo (MC) Approximation. A standard approximation strategy is permutation-based MC sampling. Instead of enumerating all subsets, we sample m random permutations of the K languages and estimate each language's SV by averaging its marginal contribution over the sampled permutations.

Formally, let π be a uniformly random permutation of the language set $\mathcal{K}$, and let P_i^π denote the set of languages appearing before language i in π. Then the SV of i can be written as the expectation

$$\phi_{i,j}^{SV} = \mathbb{E}_\pi \left[v_j^{\mathcal{G}}(P_i^\pi \cup \{i\}) - v_j^{\mathcal{G}}(P_i^\pi) \right]. \quad (8)$$

Algorithm 1 Monte Carlo Approximation

Require: Language set $\mathcal{K}$, payoff function $v_j^{\mathcal{G}}(\cdot)$, number of permutations T
Ensure: Approximate SVs $\hat{\phi}_{1,j}, \ldots, \hat{\phi}_{K,j}$

```
for i = 1 to K do
    φ̂_{i,j} ← 0
end for
for t = 1 to T do
    Sample a random permutation π_t of K
    S ← ∅
    for k = 1 to K do
        i ← π_t(k)
        Δ ← v_j^G(S ∪ {i}) − v_j^G(S)
        φ̂_{i,j} ← φ̂_{i,j} + Δ
        S ← S ∪ {i}
    end for
end for
for i = 1 to K do
    φ̂_{i,j} ← φ̂_{i,j} / T
end for
return φ̂_{1,j}, ..., φ̂_{K,j}
```

MC approximation replaces the expectation in Eq. 8 with an empirical average over T permutations:

$$\hat{\phi}_{i,j}^{SV} = \frac{1}{T} \sum_{t=1}^{T} \left[v_j^{\mathcal{G}}(P_i^{\pi_t} \cup \{i\}) - v_j^{\mathcal{G}}(P_i^{\pi_t}) \right].$$

This estimator is unbiased and converges to the exact SV as T increases.

Choice of T. Jia et al. (2019) found that when the number of permutations is

$$T(\epsilon, \delta) = [2(\bar{v}_j^{\mathcal{G}})^2 K/\epsilon^2]/\log(2K/\delta) \qquad (9)$$

where $\bar{v}_j^{\mathcal{G}}$ denotes the test loss on language j when the model has not been pretrained (i.e., the maximum possible payoff for task $\mathcal{V}_j$), the MC approximation achieves an (ϵ, δ)-approximation, i.e.,

$$\Pr\left[\left| \hat{\phi}_{\cdot,j}^{SV} - \phi_{\cdot,j}^{SV} \right|_2 \geq 1 - \delta \right] \leq \epsilon \qquad (10)$$

where $\phi_{\cdot,j}^{SV} = [\phi_{1,j}^{SV}, \ldots, \phi_{K,j}^{SV}]$. Therefore, given an approximation configuration (ϵ, δ), we can determine the number of permutations T using Formula 9.

Computational Complexity. For each permutation, the MC approximation requires evaluating K pretrained models. Therefore, when performing $T(\epsilon, \delta)$ permutations, the MC approximation requires evaluating a total of $K \cdot T(\epsilon, \delta) = O(K^2 \log(K))$ models. Thus, for MC approximation, we need to pretrain $O(K^2 \log(K))$ models for evaluation, which is significantly less complex than the $O(2^K)$ models required for exact SV computation. Note that there is repeated evaluation of the same pretrained models across different permutations. Therefore, the number of pretrained models required for MC approximation is usually significantly smaller than the number of model evaluations $K \cdot T(\epsilon, \delta)$.

B Experimental Setup Details

Corpora. We use the English $sample$-$350BT$ subset [3] from *FineWeb* (Penedo et al., 2024), and the German, French, Spanish, Chinese, Japanese, and Korean subsets from *FineWeb-2 v2.0.0* (Penedo et al., 2025). The total tokens are summarized in following Table 4.

Training Configuration. We pre-train multilingual LLMs with Megatron-LM (Shoeybi et al., 2019) on 8-node 8×H200 clusters. We adopt the open-source model architectures and training hyperparameters provided by the *llm-jp* project[4].

Table 3 summarizes the configurations of three transformer language models with increasing non-embedding parameter scales (50M, 243M, 684M, and 1.46B). As the model size grows, the number of layers and hidden dimensions increase, while the number of attention heads and context length remain constant. Larger models allocate substantially more parameters to both embedding and non-embedding components, with non-embedding parameters growing more rapidly, indicating that most additional capacity comes from deeper and wider transformer blocks rather than the embedding layer.

We adopt a decoder-only Transformer with RMSNorm, RoPE, and SwiGLU. Tokenization uses the LLaMA-2–style `llm-jp-tokenizer-100k.ver3.0b1`[5] tokenizer with a 100k multilingual vocabulary. Models are trained with learning rate 3×10^{-4}, sequence length 4,096, and global batch size 512 using Adam (β_1=0.9, β_2=0.95), cosine decay with warmup, and standard stabilization techniques (weight decay, gradient clipping, Z-loss). Training efficiency is improved through bf16 precision, FlashAttention, fused kernels, and Megatron-LM parallelism, and enabling stable large-scale multilingual pre-training.

Mixture Optimization Baselines. We consider the following baselines for mixture optimization.

[3] A subset randomly sampled from the whole dataset of around 350B gpt2 tokens (388GB).
[4] https://huggingface.co/llm-jp
[5] https://github.com/llm-jp/llm-jp-tokenizer

Table 3: Model configurations and parameter breakdown.

Params	Layers	Hidden size	Heads	Context length	Embedding parameters	Non-embedding parameters
150M	12	512	8	4,096	101,874,688	50,344,448
440M	16	1,024	8	4,096	203,749,376	243,303,424
980M	20	1,536	8	4,096	305,624,064	684,258,816
1.8B	24	2,048	16	4,096	407,498,752	1,459,718,144

Table 4: **Corpora used in our experiments.** The study covers seven languages from five language families, the available token counts is listed for each language.

Family	Languages	Tokens (B)
Germanic	English, German	373 & 450
Romance	French, Spanish	340 & 397
Sino-Tibetan	Chinese	788
Japonic	Japanese	281
Korean	Korean	52

- *Uniform Sampling:* All languages are assigned equal proportions. This balanced heuristic serves as the simplest non-parametric baseline.
- *Smoothed Sampling (Conneau and Lample, 2019):* This strategy downweights dominant languages while still reflecting corpus size through a smoothing parameter α, i.e., $p_i = \frac{(\bar{D}_i/D)^\alpha}{\sum_{i'=1}^{K}(\bar{D}_{i'}/D)^\alpha}$. We adopt the common setting $\alpha = 0.5$ for all the related experiments.
- *FamilyLaw Optimization (He et al., 2025):* This strategy first optimizes family-level mixture ratios using FamilyLaw and then allocates language-level mixture ratios using the Smoothed Sampling (see Appendix F in (He et al., 2025) for details).
- *Empirical Mixtures:* Mixture ratios are drawn from existing open-source corpora, such as *Common Crawl* and *FineWeb*. We include this setting to evaluate whether the proposed method improves over corpus-formed sampling strategies.

Downstream Tasks. We evaluate all seven pre-training languages on publicly available benchmarks covering linguistic competence, natural language understanding, and commonsense reasoning. The tasks are: *HellaSwag* (commonsense completion) (Zellers et al., 2019; Lai et al., 2023), *XStoryCloze* (cross-lingual story ending prediction) (Lin et al., 2022; Mostafazadeh et al., 2017), and *XWinograd* (multilingual coreference resolution) (Tikhonov and Ryabinin, 2021). Following standard multilingual evaluation practice, we use a *five-shot* setting and report accuracy.

C Additional Experimental Results

C.1 Additional Experiments on Performance Prediction.

To further validate the robustness of the proposed Shapley-based scaling law, we report additional fitting results on four target languages: EN, FR, DE, and KO. For each target language, we fit the proposed relation using both curated multilingual mixtures and randomly sampled mixtures, and evaluate whether the aggregate cross-lingual transfer Θ consistently explains the variation in held-out test CE loss. Representative examples for EN, and FR are shown in Figure 7: the top row presents fits across predefined mixture settings and model scales, while the bottom row shows results under random mixture sampling. In both cases, empirical losses align closely with the predicted curves, and held-out points remain near the fitted trend, indicating that the scaling law generalizes beyond the mixtures used for fitting.

The same pattern is observed for DE, and KO (see Figure 8), where losses across different model sizes and sampling strategies can also be well described by the ShapleyLaw. Across these additional languages, the fitted curves remain smooth and monotonic with respect to Θ, and the resulting high R^2 values ranging from 0.985 to 0.998. These results provide further evidence that the proposed parameterization is not specific to a single language or mixture family, but instead captures a common regularity in multilingual pre-training. Overall, the additional results strengthen the conclusion that aggregate cross-lingual transfer serves as an effective mixture variable for modeling per-language scaling behavior across diverse target languages.

C.2 Additional Experiments on Downstream Tasks

Results of Model Selection. To determine an appropriate form for $g(\ell)$, we evaluate five candidate transformations: $\{\ell, -\ell, \log \ell, \exp(-\ell), 1/\ell\}$. For each candidate, we fit a linear model of the form $s = a \cdot g(\ell) + b$ that maps loss values to predicted scores. Table 5 reports the resulting model com-

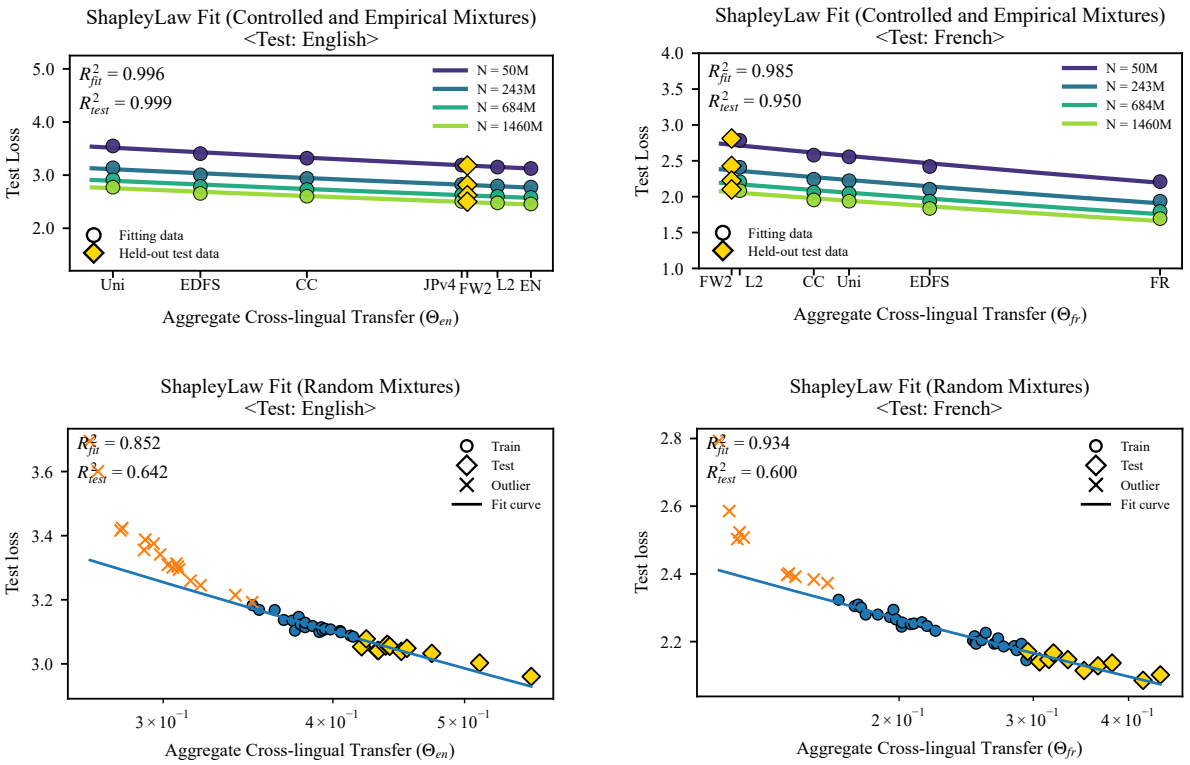


Figure 7: **Additional fitting results for Shapley-based scaling laws on <EN,FR>. Top row**: Fits on curated multilingual mixtures for EN, and FR across four model scales $\{50\text{M}, 243\text{M}, 684\text{M}, 1460\text{M}\}$; gray circles denote training points and yellow diamonds denote held-out points. **Bottom row**: Fits under randomly sampled mixtures under model scale N=243M. For both languages, the fitted curves remain smooth and monotonic with respect to Θ, and held-out points closely follow predicted trends, demonstrating strong generalization to unseen mixtures.

Table 5: **Model Selection.** Model comparison for $s = a \cdot g(\ell) + b$ using different transformations of loss. Numbers denote R^2. Across tasks except XStorycloze, $g(\ell) = \exp(-\ell)$ consistently achieves the highest explanatory power, suggesting that accuracy scales approximately linearly with token-level probability $p = \exp(-\ell)$.

Task	ℓ	$-\ell$	$\log \ell$	$\exp(-\ell)$	$1/\ell$
HellaSwag	0.994	0.994	0.995	**0.996**	**0.996**
XWinograd	0.971	0.971	0.975	**0.980**	0.978
XStorycloze	**0.947**	**0.947**	0.943	0.936	0.939

parisons. Among these transformations, $\exp(-\ell)$ always exhibits the strongest explanatory power, achieving the highest R^2 and consistently providing the best fit to the observed relationship between loss and score. These results indicate that an exponential transformation of loss most effectively captures the empirical mapping between the two quantities, and we therefore adopt $g(\ell) = \exp(-\ell)$ in the subsequent experiments.

Fitting Results on Downstream Tasks. We next model the relationship between pre-training loss ℓ and downstream performance using the functional form $s = a \cdot \exp(-\ell) + b$. Table 6 reports the fitting results across six English downstream tasks, comparing pre-training loss with few-shot (5-shot) accuracy. Across all tasks, we observe consistently strong *negative correlations*, with Pearson coefficients ranging from -0.909 to -0.997 (p-value < 0.01 in all cases), indicating that lower pre-training loss is strongly associated with higher downstream performance. The fitted models also achieve high explanatory power, with R^2 values between 0.843 and 0.996, suggesting that the exponential transformation $\exp(-\ell)$ captures the empirical relationship between loss and accuracy well. These results support the use of $s = a \cdot \exp(-\ell) + b$ as a simple yet effective mapping for predicting downstream task performance from pre-training loss.

Additional Fitting Results on Other Language Tasks. We conduct additional experiments in downstream tasks for Chinese and Japanese to explore feasibility of ShapleyLaw. In addition, we fit results of downstream performance vs test loss of 50 models ($N = 243$M, $D = 100$B) with randomly sampled language mixtures. As shown in Figure. 9, we report fitting results for Chinese, and Japanese.

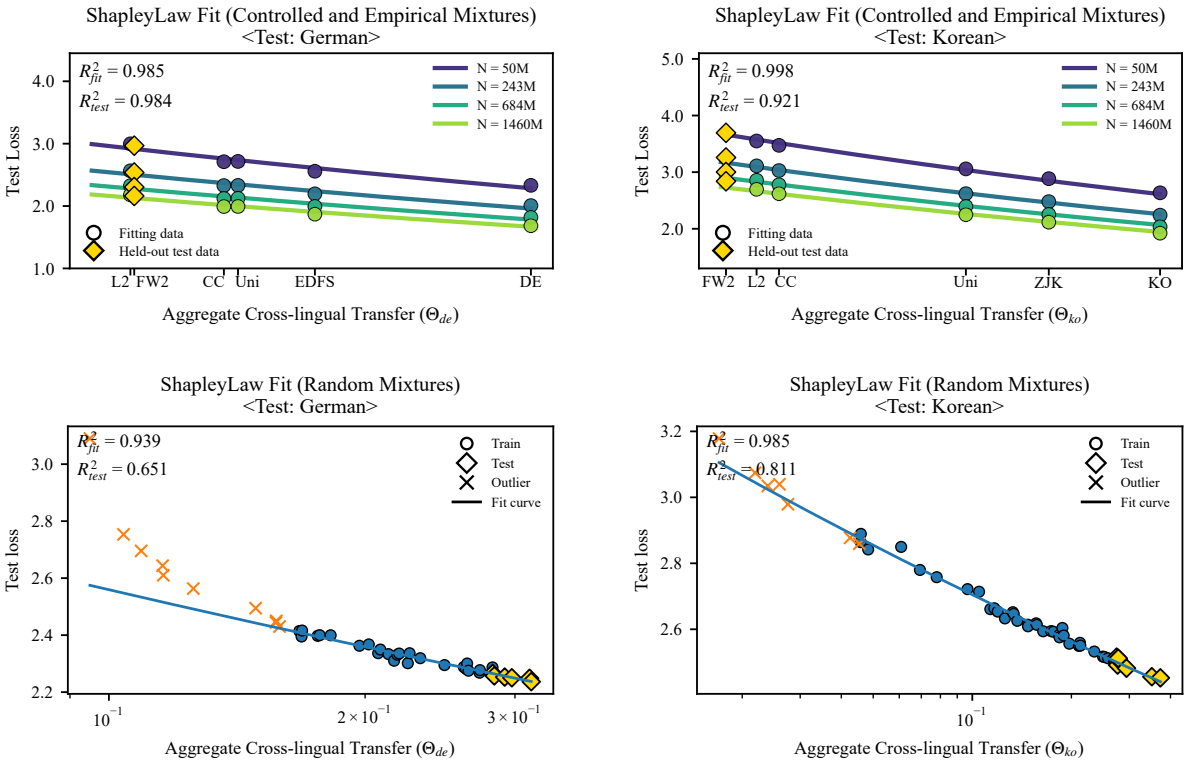


Figure 8: **Additional fitting results for Shapley-based scaling laws on <DE,KO>. Top row**: Fits on curated multilingual mixtures for DE and KO across the same four model scales. **Bottom row**: Fits under randomly sampled mixtures with model scale N=243M. For both languages, the fitted curves remain smooth and monotonic with respect to aggregate cross-lingual transfer Θ, and held-out points closely follow predicted trends, demonstrating strong generalization to unseen mixtures.

Table 6: **Fitted Results (English Downstream Tasks).** Pearson correlation between pretraining loss ℓ and few-shot accuracy. All tasks exhibit strong negative correlation ($|r| > 0.90$), indicating that lower loss consistently corresponds to higher downstream performance. Correlation between pretraining loss ℓ and few-shot accuracy, and best-fit results for $\hat{s} = ag(\ell) + b$. Across all tasks, $\exp(-\ell)$ consistently achieves the highest R^2.

Task	Pearson r	p-value	Best $g(\ell)$	R^2	a	b
HellaSwag	-0.997	8.15×10^{-7}	$\exp(-\ell)$	0.996	4.030	0.098
XWinograd	-0.985	4.92×10^{-5}	$\exp(-\ell)$	0.980	4.824	0.382
XStorycloze	-0.973	2.25×10^{-4}	$\exp(-\ell)$	0.980	3.034	0.433

C.3 Additional Experiments on SV Approximation

We apply MC approximation to the multilingual pre-training game with

$$\mathcal{K} = \{\texttt{en}, \texttt{de}, \texttt{fr}, \texttt{es}, \texttt{zh}, \texttt{ja}, \texttt{ko}\}$$

where $K = 7$. In the exact setting, computing the full SV matrix requires evaluating all $2^7 - 1 = 127$ non-empty coalitions, i.e., pre-training 127 models. Although this cost is still manageable for $K = 7$, it already illustrates the exponential growth of exact SV estimation and serves as a controlled benchmark for approximation quality.

In our comparison experiments, we first use the SV group $\phi^{SV}_{\texttt{ja}}$ as a representative case study and compare exact SVs against MC approximated SVs. First, we fix the number of samples to $m = 100$ and plot the approximated SVs against the exact SVs for $\phi^{SV}_{\texttt{ja}}$. Figure 10 directly shows that MC sampling can recover the overall SV pattern with substantially fewer utility evaluations than full coalition enumeration. Second, we vary the sampling number from $m = 20$ to $m = 120$ and plot the mean absolute error (MAE) $|\hat{\phi}^{SV(m)}_{i,\texttt{ja}} - \phi^{SV}_{i,\texttt{ja}}|$ to examine convergence behavior. The resulting trend shows that the approximation error decreases as the sampling budget increases, consistent with the concentration behavior in Eq. 10.

These results suggest that MC approximation is an effective surrogate for exact SV computation

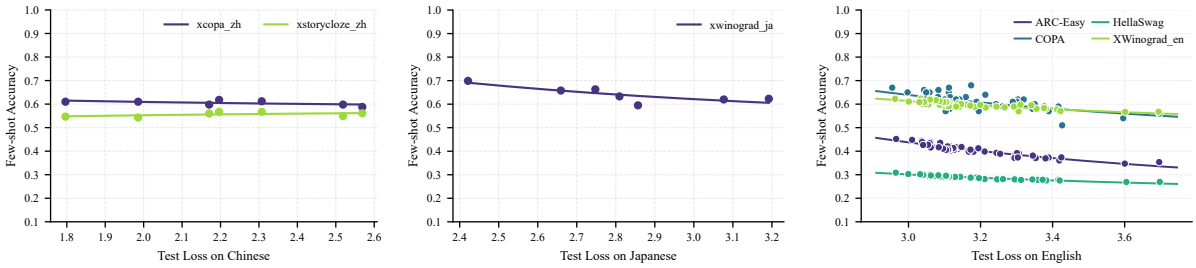


Figure 9: **Left.** Downstream performance versus test cross-entropy (CE) loss on Chinese tasks (*XCOPA_zh* and *XStoryCloze_zh*). **Middle.** Downstream performance versus test CE loss on the Japanese task (*XWinograd*). **Right.** Downstream performance for English versus test CE loss across 50 models ($N = 243$M, $D = 100$B) trained with randomly sampled data mixtures.

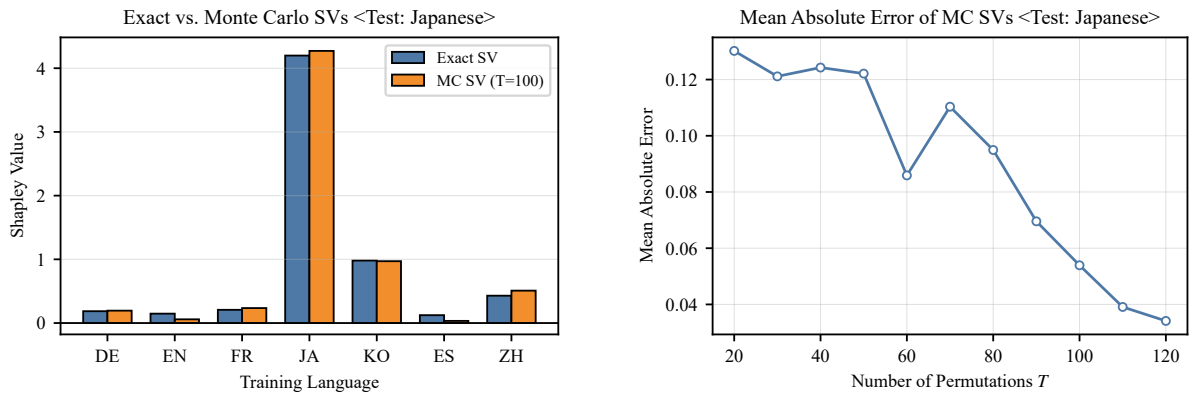


Figure 10: **Experimental Results of MC Approximation for SVs (Group: Japanese).** We use the SV group $\phi^{SV}_{\texttt{ja}}$ as a representative case study. **Left**: We compare exact SVs $\phi^{SV}_{\texttt{ja}}$ (blue bars) with MC approximations (orange bars) using $T = 100$ samples. The approximated values closely match the exact SVs (e.g., $\phi^{SV}_{\texttt{ja,ja}} = 4.199$ vs. $\hat{\phi}^{SV}_{\texttt{ja,ja}} = 4.272$) while preserving the same ranking. **Right**: We report the mean absolute error (MAE) as the sampling budget m varies from 20 to 120. The approximation error decreases steadily as the number of samples increases.

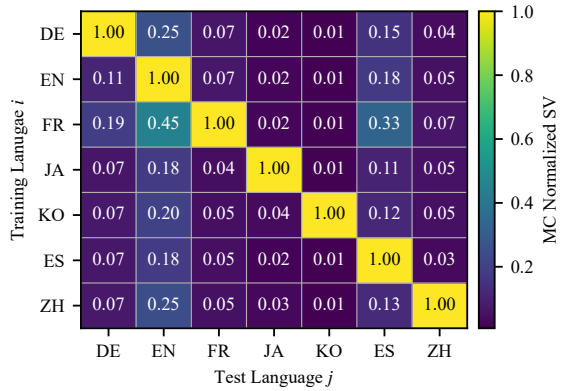


Figure 11: **Monte Carlo Approximated Normalized SVs Matrix.** The results show that MC–approximated SVs still capture cross-lingual transfer effects, with only a slight reduction in accuracy.

in multilingual pre-training. For small language sets such as $K = 7$, it provides a useful accuracy–efficiency trade-off and can be directly validated against exact SVs. For larger multilingual candidate pools, where exact enumeration becomes infeasible due to the $2^K - 1$ training cost, the benefit of MC approximation becomes even more pronounced. Overall, MC approximation offers a practical and scalable approach to SV estimation, especially when the total number of candidate languages is large.