How to Improve LLMs' Performance on Specific Languages: A Perspective on LLM-Derived Language Similarity

Xinhe Shi[1], Qingcheng Zeng[2]*, Weihao Xuan[3,4]*, Linchao Zhu[1]†,

[1]Zhejiang University, [2]Northwestern University, [3]The University of Tokyo [4]RIKEN AIP,
shixinhe2024@outlook.com, qcz@u.northwestern.edu,
weihaoxuan@g.ecc.u-tokyo.ac.jp, zhulinchao@zju.edu.cn

Abstract

Large language models (LLMs) exhibit uneven performance across languages. In language-specific applications, practitioners often rely on target-language corpora or cross-lingual transfer to achieve better performance. However, traditional linguistic typology, commonly used as a transfer language selection strategy in previous studies, may not align with LLM's perception of language similarity. This work proposes **LLM-based language similarity** as a novel perspective for selecting effective fine-tuning languages. We construct a framework to quantify the similarity within each language pair through both the lenses of **language-specific performance patterns** and **cross-lingual transferability**, ultimately deriving three similarity score matrices. Moreover, we observe a counterintuitive phenomenon: **super-additive transfer effect**, where fine-tuning on a certain language yields higher performance than fine-tuning directly on the target language. Additionally, due to the absence of an existing dataset meeting our experimental requirements, we construct and release the **M4CQ-Pro** dataset, which features domain-diverse distribution of **135** tasks and content consistency across **31** languages (including over 20 medium- and low-resource languages), with 61518 manually reviewed high-quality questions per language. We evaluate our approach on representative multilingual LLMs and results show that all three LLM-based similarity measures effectively guide fine-tuning language selection, outperforming traditional linguistic similarity, with the integrated measure achieving the best results. Our approach provides not only **a novel perspective on language similarity**, but also **practical baselines for selecting fine-tuning languages**.

Project: https://github.com/LearnerSXH/LanguageSimilarity

*Co-second authors
†Corresponding author

1 Introduction

To improve the performance of large language models (LLMs) on specific target languages, a common approach is fine-tuning models on target languages or leveraging cross-lingual transfer with corpora of similar languages (Zubillaga et al., 2024; Rice et al., 2025; Bankula and Bankula, 2025).

However, acquiring corpora of the target language is not always feasible, and these traditional typological frameworks, based on synchronic (focusing on structural features like syntax/morphology) and diachronic (focusing on historical relationships) classifications (McMahon and McMahon, 2005; Brown et al., 2008; Dunn et al., 2011; Bouckaert et al., 2012; Pagel et al., 2013; Wichmann et al., 2022; Lewis et al., 2023), may not be directly applicable to LLMs, whose representations emerge from large-scale data-driven training rather than explicit linguistic typology. Additionally, although linguists have long studied language similarity, consensus remains elusive yet. For instance, the classification of Japanese remains disputed, with some considering it part of the Altaic language family (Ramstedt and Aalto, 1952; Murayama, 1957, 1962; Miller, 1967, 1971, 1975, 1979, 1980, 1983, 1985a,b; Street and Miller, 1973, 1975–77; Starostin, 1991; Robbeets, 2005), others the Austronesian family (Kawamoto, 1977, 1978, 1980; Benedict, 1990; Vovin, 1994; Hudson, 1999), and still others viewing it as an isolated language (Shibatani, 1990; Vovin, 2005; Tranter, 2012).

In this work, we investigate LLM-based language similarity from two complementary perspectives: (1) **task performance patterns** and (2) **cross-lingual transferability**. Then we integrate them to get a third perspective. Each perspective yields a quantitative similarity score for every language pair. These **similarity scores serve as baselines for selecting fine-tuning languages**.

Additionally, due to the lack of datasets meet-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15139–15164
July 2-7, 2026

ing our experimental requirements (detailed in Section 2.4), we constructed the **M4CQ-Pro** dataset. Along with covering 135 tasks across various domains and 31 languages (including over 20 medium- and low-resource languages (Joshi et al., 2020)), the M4CQ-Pro dataset has also undergone manual verification to remove language/culture-biased cases and ambiguous or wrong cases, ensuring the quality of data.

We verify our approaches on representative multilingual LLMs, including Qwen (Yang et al., 2025), Gemma (Team et al., 2025), and LLaMA (Grattafiori et al., 2024), and demonstrate the efficacy of LLM-based language similarity for guiding fine-tuning language selection. Results show the effectiveness of all three LLM-based similarity measures, and **the combined measure outperforms**, providing comparable performance with direct fine-tuning (fine-tuning directly on the target language).

2 Methodology

In this section, we present two perspectives to investigate LLM-based language similarity: 1. LLM-derived language features, which we define as LLMs' task performance patterns on each language. 2. Cross-lingual transferability. Each perspective produces quantitative similarity scores for every language pair, which are then organized into a similarity matrix.

Then we combine these two perspectives, so eventually three baselines (i.e., three similarity score matrices) will be obtained for language selection.

2.1 LLM-derived Language Features

We derive language features from **LLMs' task performance patterns** (i.e., LLMs' task performance on the specific language across multiple tasks). This approach constructs a high-dimensional feature space, through which we can explore language similarity patterns under a key hypothesis: similar languages induce similar LLM capability distributions in multitask scenarios.

Given a multitask dataset comprising parallel versions in multiple languages, let $\mathcal{L}$ denote the set of all languages in the dataset, let N denote the number of task categories in the dataset. For each language $L_e \in \mathcal{L}$, define its **language feature vector** as:

$$\mathbf{v}_e = [v_{e,1}, \ldots, v_{e,N}]^\top \in \mathbb{R}^N,$$

where $v_{e,j} \in [0, 1]$ represents the model's accuracy on task j for L_e.

The feature vector $\mathbf{v}_e$ reflects LLMs' multidimensional capabilities on the language L_e, with dual interpretability: 1) Each element $v_{e,j}$ quantifies task-specific LLM capability for language L_e; 2) Inter-vector geometric relationships reveal LLM-perceived language similarity in task performance distributions.

Then we can represent each language in the feature space using feature vector. For each language pair (L_a, L_b) where $L_a \in \mathcal{L}$ and $L_b \in \mathcal{L}$, the similarity score is defined as

$$S_{a,b}^{\text{feature}} = \text{CosineSimilarity}(\mathbf{v}_a, \mathbf{v}_b).$$

We obtain an $|\mathcal{L}| \times |\mathcal{L}|$ similarity score matrix

$$\mathbf{M}^{\text{feature}} = \left[m_{i,j}^{\text{feature}}\right]_{1 \le i,j \le |\mathcal{L}|},$$

where $m_{i,j}^{\text{feature}} = S_{i,j}^{\text{feature}}$, and L_i and L_j denote the i-th and j-th languages, respectively.

It's also feasible to define LLM's language feature as hidden representations in models. However, hidden representations mainly capture the latent distributional patterns inside models, which do not necessarily correlate with the actual performance on downstream tasks. In contrast, this performance-based approach directly reflects functional differences in how the model processes different languages, therefore more interpretable for practical applications such as cross-lingual transfer. Moreover, while hidden-state analyses are typically limited to single-task or single-corpus settings (van Aken et al., 2019; Conneau et al., 2020; Li et al., 2025a), task performance patterns naturally integrate information from diverse tasks, capturing linguistic differences at semantic, syntactic, and knowledge-related levels simultaneously.

2.2 Cross-Lingual Transferability

In this perspective, the core hypothesis is: the greater the performance improvement on evaluation language L_e achieved by fine-tuning on language L_f, the higher their similarity in LLM representation space.

Given a multitask dataset comprising parallel versions in multiple languages, let $\mathcal{L}$ denote the set of all languages in the dataset, let N denote the number of task categories in the dataset. For each language pair (L_f, L_e) where $L_f \in \mathcal{L}$ and $L_e \in \mathcal{L}$, the similarity quantification proceeds as:

Dataset	Languages	Tasks	Instances per Language	Evaluation Modality
MMMLU (OpenAI, 2024)	14	57	≈ 14k	Multiple-choice (4 choices)
P-MMEVAL (Zhang et al., 2024)	10	64	3038	Multiple-choice (4 choices) & Text Generation
Global MMLU (Singh et al., 2024)	42	57	≈ 14k	Multiple-choice (4 choices)
MMLU-ProX (Xuan et al., 2025)	29	14	≈ 12k	Multiple-choice (10 choices)
M4CQ	19	119	≈ 56k	Multiple-choice (2∼108 choices)
M4CQ-Pro (this work)	**31**	**135**	**≈ 62k**	**Multiple-choice (2∼108 choices)**

Table 1: Comparison of M4CQ-Pro and existing multilingual multitask datasets.

1. Fine-tuning: Use the same LoRA (Hu et al., 2021) configuration and fine-tune base model M_{base} on L_f and L_e respectively to obtain models M_f and M_e.

2. Cross-lingual testing: Evaluate M_f on L_e's N tasks, obtaining task accuracies $\{\text{acc}_{f,e,j}\}_{j=1}^{N}$, where $\text{acc}_{f,e,j}$ is M_f's accuracy on task j in L_e. Calculate the average accuracy of $\{\text{acc}_{f,e,j}\}_{j=1}^{N}$, obtaining $\text{Acc}_{f,e}$.

3. Baseline acquisition: Evaluate M_e on L_e's N tasks, obtaining baseline accuracies $\{\text{acc}_{e,e,j}\}_{j=1}^{N}$. Calculate the average accuracy of $\{\text{acc}_{e,e,j}\}_{j=1}^{N}$, obtaining $\text{Acc}_{e,e}$.

4. Metric computation: Calculate the similarity score as

$$S_{f,e}^{\text{transferability}} = \frac{\text{Acc}_{f,e}}{\text{Acc}_{e,e}}.$$

Then we obtain an $|\mathcal{L}| \times |\mathcal{L}|$ similarity score matrix

$$\mathbf{M}^{\text{transferability}} = \left[m_{i,j}^{\text{transferability}} \right]_{1 \le i,j \le |\mathcal{L}|},$$

where $m_{i,j}^{\text{transferability}} = S_{i,j}^{\text{transferability}}$, and L_i and L_j denote the i-th and j-th languages, respectively.

We define the similarity score as the ratio between the performance of cross-lingual fine-tuning and direct fine-tuning (fine-tuning directly on the target language), rather than using the raw cross-lingual performance itself. This design is motivated by practical scenarios where target-language resources are scarce and fine-tuning has to be conducted on an alternative language: the score quantifies how closely the alternative can approximate the performance of direct fine-tuning on the target language. Moreover, this ratio serves as a normalization step. Comparing similarity scores across different target languages makes no sense if raw cross-lingual performance is adopted, since models naturally achieve different performance on different languages.

This perspective highly overlaps with our initial motivation—enhancing LLMs' performance on specific languages—yet differs and does not constitute an unfair advantage. An important reason is that during the investigation of LLM-based language similarity, we deliberately use monolingual pre-trained models to avoid confounding variables introduced by multilingual pre-training. (There are plenty of constraints for model selection and dataset selection, see Section 2.4 for details.) It means up-to-date models are excluded in our experiments, but in validation, we will only involve up-to-date models.

2.3 Integration of Feature and Transferability Perspectives

To provide a unified view, we define another similarity score matrix based on the aforementioned matrices. Specifically, we first normalize the two matrices using Min-Max (Agarwal, 2013) respectively, and then combine them via element-wise product to obtain the integrated matrix:

$$\mathbf{M}^{\text{integration}} = \text{Norm}_{Min-Max}(\mathbf{M}^{feature}) \odot \text{Norm}_{Min-Max}(\mathbf{M}^{transferability}).$$

After Min-Max normalization, elements in $\mathbf{M}^{feature}$ and $\mathbf{M}^{transferability}$ will be mapped to the range $[0, 1]$. We adopt element-wise product rather than weighted summation to combine the two matrices, as it preserves only the mutually high similarity scores from both perspectives.

2.4 Model and Dataset Requirements

To ensure the validity and reliability of our methodology, the pre-trained model and the multilingual multitask dataset selected should satisfy the following specifications.

Model Requirements:

(I) **Monolingual pre-training.** We predict that models' language-specific performance patterns

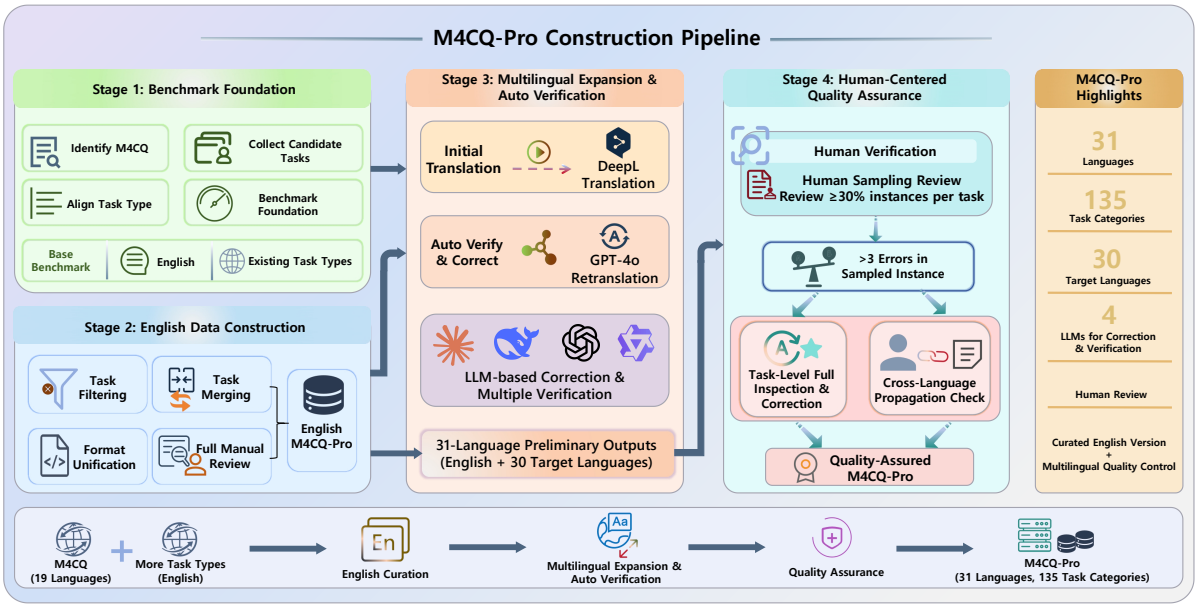

Figure 1: The construction workflow of M4CQ-Pro. Tasks inherited from M4CQ are frozen in their original language versions and do not undergo the construction pipeline shown here.

would be impacted by whether this language has been widely seen in pre-training. For instance, models' performance might be significantly higher on pre-training languages than others, then pre-training languages will be considered dissimilar to other languages. This gap is introduced by pre-training corpora, instead of the inherent language similarity. So we will exclude the pre-training languages from our research scope. The base model M_{base} should be pre-trained exclusively on one language, not only to retain more languages in our research scope, but also to minimize the potential cross-influence from multiple pre-training languages. ***Note***: For our experiments, the ideal model would be one that has been pre-trained on corpora which contain exactly all target languages (i.e., the languages present in the multilingual multitask dataset) with complete semantic equivalence across languages. If such a model were available, our experiments could be conducted without introducing any irrelevant variable in this dimension. However, given the absence of suitable open-source models and the considerable challenges in training such a model ourselves, we resort to using monolingual pre-trained models as a practical alternative. (II) **Adequate Context Length.** The model's context length must exceed the maximum token length of 99% of task instances (extremely long instances are excluded). (III) **Tokenization Compatibility.** The model's tokenizer should support all languages present in the dataset.

Dataset Requirements:

(I) **Task Parallelism.** Each language variant of the dataset must contain identical task categories. This guarantees consistent feature dimensions in performance vectors across languages for meaningful clustering comparisons. (II) **Task Domain Coverage.** Tasks in the dataset should cover comprehensive domains to better capture LLM-derived language features and cross-lingual transferability. (III) **Balanced Task Distribution.** Task categories should be evenly distributed throughout the dataset. For instance, if mathematical tasks constitute 90% of all tasks, they might disproportionately dominate the clustering result, thereby skewing the assessment of language similarity. (IV) **Content Equivalence Across Languages.** Each language variant of the multitask dataset should contain linguistically diverse but *semantically equivalent* task instances. This ensures direct comparability of task accuracy across languages. (V) **Language Neutrality.** Tasks within the dataset should not inherently favor any particular language. For instance, tasks involving code reasoning (biased towards English) or culturally-specific knowledge (such as the history of a specific country) are excluded, as they may introduce language-specific biases. (VI) **Moderate Difficulty.** Problems should not be too difficult, since our aim is to measure language capabilities rather than solve highly complex problems. (VII) **Task Sufficiency.** The dataset should contain a large number of tasks. A higher number of tasks

ensures more robust and reliable clustering results, enhancing the credibility of the language similarity analysis. (VIII) **Language Coverage.** The dataset should cover as many languages as possible. Including a wide variety of languages enhances the generalizability and applicability of our research.

Code	Language	Code	Language
AR	Arabic	BG	Bulgarian
CS	Czech	DA	Danish
DE	German	EL	Greek
EN	English	ES	Spanish
ET	Estonian	FI	Finnish
FR	French	HU	Hungarian
ID	Indonesian	IT	Italian
JA	Japanese	KO	Korean
LT	Lithuanian	LV	Latvian
NB	Norwegian Bokmål	NL	Dutch
PL	Polish	PT-BR	Portuguese (Brazilian)
PT-PT	Portuguese (Non-Brazilian Variants)	RO	Romanian
RU	Russian	SK	Slovak
SL	Slovenian	SV	Swedish
TR	Turkish	UK	Ukrainian
ZH	Chinese (Simplified)		

Table 2: Language Codes and Full Names in M4CQ-Pro Dataset

3 The M4CQ-Pro Dataset

Following a meticulous survey of existing open-source multilingual benchmarks, we identify M4CQ[1] as the dataset that most closely aligns with the requirements outlined in the previous section. M4CQ was explicitly designed to support comprehensive evaluation of model performance on individual languages and covers 119 task types that are evenly distributed across diverse domains, making it well suited for multi-task and cross-lingual analysis. However, while M4CQ is sufficient for evaluating per-language performance, it is insufficient for our downstream objective of analyzing LLM-derived language similarity, which requires substantially broader language coverage to enable robust and fine-grained comparative analysis. In particular, the original M4CQ benchmark includes only 19 languages, limiting its ability to characterize global language relationships.

To address this limitation, we construct **M4CQ-Pro**, extending M4CQ from 19 to 31 languages and augmenting it with additional task types. As a result, M4CQ-Pro comprises 135 distinct task categories, enabling more comprehensive coverage of linguistic phenomena across languages. Figure 1 illustrates the entire dataset construction process.

[1] https://huggingface.co/datasets/LearnerSXH/M4CQ

3.1 Dataset Overview

The M4CQ-Pro dataset consists of **31 languages** (covering high-/low-resource languages), each containing **135 task categories** in multiple-choice format. Table 2 provides a complete list of languages covered in M4CQ-Pro.

M4CQ-Pro's task distribution is shown in Figure 2, demonstrating a balanced spread across diverse domains. Each question has semantically equivalent versions in all languages. Refer to Appendix B for detailed information about these 135 tasks.

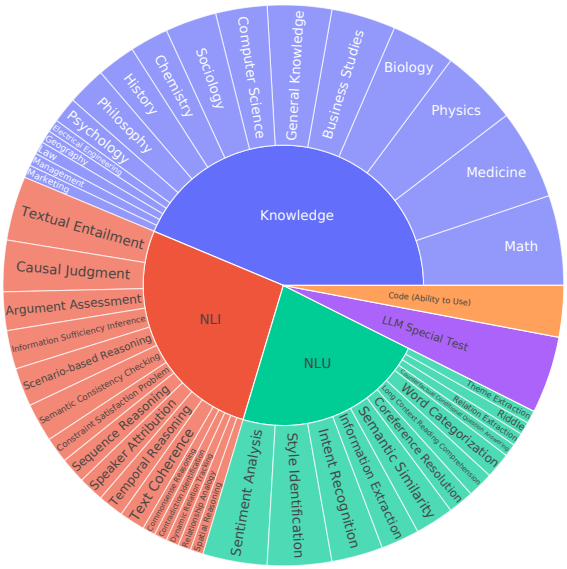

Figure 2: Domain distribution of 135 tasks in M4CQ-Pro, where each domain's angular span represents the proportion of its constituent tasks.

Table 1 details the comparison of M4CQ-Pro and other existing multilingual multitask datasets.

3.2 Data Curation

This process consists of multiple steps.

First, we selected candidate tasks from several publicly available English-language datasets (Johannes Welbl, 2017; Wang et al., 2018; Mihaylov et al., 2018; Wang et al., 2019; Bisk et al., 2020; Hendrycks et al., 2020; Sakaguchi et al., 2021; Srivastava et al., 2022), guided by the requirements specified in Section 2.4. Next, we merged tasks of the same type from different datasets to avoid redundant task formulations. Then we converted all curated tasks into a unified format consistent with M4CQ, facilitating subsequent procedures. To ensure data quality and ethical compliance, we performed instance-level inspection for each task, filtering out low-quality question–answer pairs in

which the correct answer is ambiguous, weakly supported, or unconvincing. In addition, we explicitly remove samples containing personally identifiable information or other sensitive content, following standard data governance practices. Finally, the newly curated tasks were combined with the task types inherited from the original M4CQ dataset.

Through this process, we obtained **the English version of M4CQ-Pro**, which comprises a total of **135 distinct task categories**.

3.3 Translation Pipeline

To ensure translation quality, we adopt the same translation workflow and engine selection used in the original M4CQ dataset construction.

Specifically, we first translate the English version of M4CQ-Pro into target languages using DeepL[2] as the primary translation engine. The translated outputs are then subjected to an automated verification and correction stage that jointly leverages Claude 3.5 Sonnet[3], DeepSeek-R1 (DeepSeek-AI et al., 2025), GPT-4o(OpenAI et al., 2024), and Qwen2.5-72B-Instruct (Qwen et al., 2025). These models are employed to detect and correct translation errors, including mistranslations, semantic drift, and inconsistencies in answer options, with the goal of reducing translation noise through automated validation.

Following automated correction, we conduct human verification to further ensure translation quality. For each task, at least 30% of instances are randomly sampled and manually reviewed by human annotators. If more than three translation errors are identified within the sampled instances of any given task, the entire task is subjected to full manual inspection and correction.

3.4 Other Potential Applications

Owing to the large-scale multilingual and multitask coverage, M4CQ-Pro supports a range of research applications beyond the scope of this work. Below, we outline several representative directions. **Granular Multilingual Model Benchmarking**: With its balanced task distribution and guaranteed semantic equivalence, M4CQ-Pro is an ideal benchmark for nuanced evaluation of Multilingual Language Models. It facilitates direct comparisons of performance across the 31 languages. More importantly, it allows researchers to decompose overall multilingual ability into fine-grained capabilities—assessing whether a model's superiority in one language generalizes across all 135 task types or is confined to specific domains like reasoning or generation. This enables more insightful model diagnostics and comparisons. **Research in Multitask & Cross-Task Learning**: The extensive taxonomy of 135 tasks across diverse domains (visualized in Figure 2) creates a rich testbed for research on multitask learning optimization. Researchers can utilize it to investigate multitask learning optimization and analyze inter-task correlations. **Advancing Under-Resourced Language Understanding**: M4CQ-Pro offers high-quality, translation-verified evaluation data for over twenty medium- and low-resource languages, including Estonian, Lithuanian, and Norwegian. This coverage enables research on language-specific data augmentation and pre-training for under-resourced languages.

These applications underscore M4CQ-Pro's value as a foundational resource.

4 Experiments and Results

See Appendix C for experiment settings.

4.1 LLM-Based Language Similarity

Thirty languages are covered in this investigation. Three language similarity score matrices are presented as heatmaps in Figure 3, with detailed values provided in Appendix D.

$\mathbf{M}^{\text{feature}}$ is a symmetric matrix because it is derived from the cosine similarity of language feature vectors. But $\mathbf{M}^{\text{transferability}}$ is asymmetric, as it is computed from cross-lingual transferability, where the transferability from L_a to L_b is not necessarily equal to that from L_b to L_a. The matrix $\mathbf{M}^{\text{integration}}$ is also asymmetric, inheriting this property from $\mathbf{M}^{\text{transferability}}$.

After obtaining these three matrices, we can find quantitative similarity scores for any language pair among 30×30 language pairs.

Additionally, in the experiment of cross-lingual transferability, there is a counter-intuitive finding: **Super-Additive Transfer Effect**, which means fine-tuning on a specific language provides a more effective "bridge" for knowledge transfer than direct fine-tuning on the target language. Specifically, many $S_{f,e}^{\text{transferability}}$ values exceed 1 when $L_f \neq L_e$ (e.g., FR→FI=1.0704, ZH→ET=1.0564), indicating that fine-tuning on a different but sim-

[2] https://www.deepl.com
[3] https://www.anthropic.com/

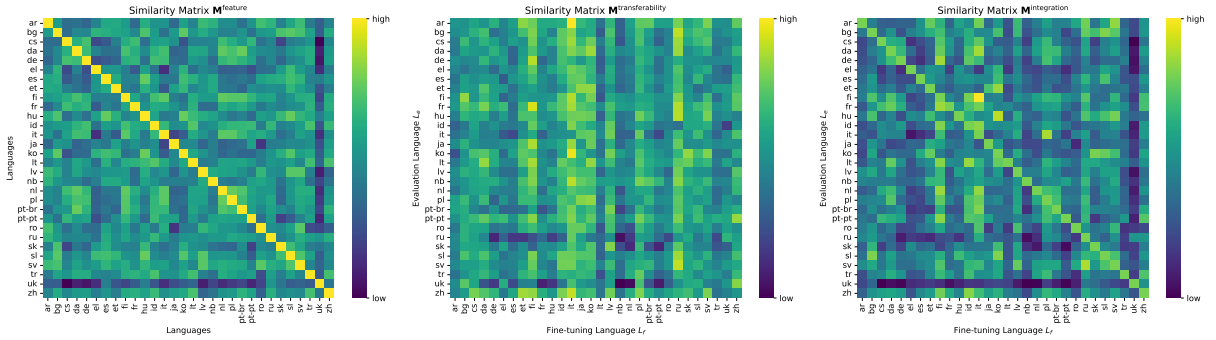

Figure 3: Similarity score matrices from three perspectives. **Left**: $\mathbf{M}^{\text{feature}}$, based on LLM-derived language features; **Middle**: $\mathbf{M}^{\text{transferability}}$, based on cross-lingual transferability; **Right**: $\mathbf{M}^{\text{integration}}$, based on the integration of the two perspectives.

L_f \ L_e	AR	CS	DE	EL	ES	FR	ID	IT	JA	KO	LT	NL	PL	PT	RO	RU	SV	TR	UK	ZH
UK	0.418	0.431	0.470	0.421	0.489	0.474	0.465	0.468	0.408	0.417	0.464	0.430	0.476	0.438	0.454	0.442	0.419	0.429	0.456	
LT	0.386	0.394	0.420	0.386	0.464	0.442	0.411	0.443	0.348	0.392	0.377	0.429	0.391	0.442	0.409	0.407	0.405	0.379	0.393	0.392
EN	0.453	0.468	0.484	0.433	0.509	0.486	0.466	0.486	0.426	0.447	0.416	0.491	0.476	0.500	0.468	0.461	0.476	0.447	0.453	0.460
ZH	0.448	0.473	0.495	0.464	0.514	0.496	0.482	0.498	0.450	0.466	0.423	0.505	0.478	0.506	0.481	0.476	0.485	0.458	0.456	0.478

Table 3: Multilingual performance of Gemma-4b fine-tuned by high- and low-resource languages. The rows denote fine-tuning languages and columns denote evaluation languages.

ilar language may yield better performance than direct fine-tuning on the target language. Additionally, this effect is not necessarily symmetrical (e.g., NB→RU=1.0501 vs. RU→NB=0.9195), highlighting directional transfer preferences.

What are the causes of the super-additive transfer effect? Pure language similarity alone cannot account for the super-additive transfer effect and its directionality. By observing circumstances where the super-additive transfer effect occurs, we hypothesize that cross-lingual transfer efficiency is jointly determined by both **linguistic similarity** and **language resource richness**.

To verify this hypothesis, we conduct additional experiments and analyses, finding that **fine-tuning on low-resource languages lead to larger parameter fluctuations and thus poorer performance**.

Table 3 compares the multilingual performance of models fine-tuned on low-resource languages and high-resource languages. As shown, models fine-tuned on low-resource languages (UK, LT) consistently underperform those fine-tuned on high-resource languages (EN, ZH) across all target languages. This finding is also consistent with prior literature. Li et al. found that fine-tuning had limited effectiveness for low-resource languages. Lee et al. also reported that direct SFT on low-resource data can induce destructive parameter fluctuations, which degrade models' multilingual performance. Our observations provide independent empirical support for this mechanism and suggest that the super-additive effect may partially arise from optimization stability differences across source languages.

Building on these experiments and analyses, we summarize a **mechanistic explanation**: When fine-tuning on a high-resource source language L_s that is linguistically similar to the target language L_t, the model inherits stable parameter configurations that have already learned robust feature representations. The linguistic similarity ensures that these stable configurations are relevant to L_t, while the resource richness guarantees that the optimization path remains smooth when adapting to L_t. This dual advantage (parameter stability plus feature similarity) explains why fine-tuning on L_s can outperform direct fine-tuning on L_t, particularly when L_t is a low-resource language (would induce destructive parameter fluctuations). Therefore, to allow the super-additive effect to occur, **the fine-tuning language should simultaneously satisfy two conditions: high-resource status and high linguistic similarity with the target language**. For instance, Portuguese (PT) → Turkish (TR), Russian (RU) → Ukrainian (UK), Arabic (AR) → Slovenian (SV) all exhibited the super-additive effect, where PT, RU and AR are not only high-resource but also selected with the highest similarity scores (see Table 12).

		AR	CS	DE	EL	ES	FR	ID	IT	JA	KO
Llama-3B											
Traditional Language Similarity	Acc.	-	0.383	0.406	-	**0.448**	0.446	-	0.439	-	-
	Selected L_f	-	PL	NL	-	IT, FR, RO, PT	IT, ES, RO, PT	-	FR, ES, RO, PT	-	-
LLM-Based Language Similarity (feature)	Acc.	0.352	0.403	**0.453**	0.363	0.443	**0.447**	**0.455**	**0.471**	0.388	0.385
	Selected L_f	RO	DE	CS	KO	KO	CS	LT	NL	PT	ES
LLM-Based Language Similarity (transferability)	Acc.	0.358	0.392	0.438	0.365	0.439	0.430	0.424	0.438	0.373	**0.394**
	Selected L_f	PT	ZH	PT	PT	ZH	ZH	ZH	AR	PL	LT
LLM-Based Language Similarity (integration)	Acc.	**0.359**	**0.410**	0.426	**0.372**	0.443	0.430	**0.455**	**0.471**	**0.393**	**0.394**
	Selected L_f	SV	FR	FR	SV	KO	ZH	LT	NL	FR	LT
Reference: Direct Fine-tuning	Acc.	0.358	0.403	0.439	0.374	0.444	0.435	0.438	0.453	0.368	0.369
Qwen-4B											
		AR	CS	DE	EL	ES	FR	ID	IT	JA	KO
Traditional Language Similarity	Acc.	-	0.585	0.617	-	**0.660**	0.647	-	0.657	-	-
	Selected L_f	-	PL	NL	-	IT, FR, RO, PT	IT, ES, RO, PT	-	FR, ES, RO, PT	-	-
LLM-Based Language Similarity (feature)	Acc.	0.550	0.600	0.630	0.549	0.653	**0.659**	0.632	**0.659**	0.594	**0.591**
	Selected L_f	RO	DE	CS	KO	KO	CS	LT	NL	PT	ES
LLM-Based Language Similarity (transferability)	Acc.	0.570	**0.610**	0.629	0.554	0.645	0.657	**0.633**	0.649	0.598	0.584
	Selected L_f	PT	ZH	PT	PT	ZH	ZH	ZH	AR	PL	LT
LLM-Based Language Similarity (integration)	Acc.	**0.573**	0.603	**0.639**	**0.557**	0.653	0.657	0.632	**0.659**	**0.613**	0.584
	Selected L_f	SV	FR	FR	SV	KO	ZH	LT	NL	FR	LT
Reference: Direct Fine-tuning	Acc.	0.568	0.597	0.641	0.539	0.661	0.664	0.627	0.656	0.602	0.593
Gemma-4B											
		AR	CS	DE	EL	ES	FR	ID	IT	JA	KO
Traditional Language Similarity	Acc.	-	0.446	0.482	-	0.506	0.496	-	0.493	-	-
	Selected L_f	-	PL	NL	-	IT, FR, RO, PT	IT, ES, RO, PT	-	FR, ES, RO, PT	-	-
LLM-Based Language Similarity (feature)	Acc.	0.443	0.464	0.491	0.412	0.479	0.489	0.411	**0.505**	0.438	**0.443**
	Selected L_f	RO	DE	CS	KO	KO	CS	LT	NL	PT	ES
LLM-Based Language Similarity (transferability)	Acc.	0.443	**0.473**	0.488	0.446	**0.514**	**0.496**	**0.482**	0.497	0.429	0.392
	Selected L_f	PT	ZH	PT	PT	ZH	ZH	ZH	AR	PL	LT
LLM-Based Language Similarity (integration)	Acc.	**0.450**	0.459	**0.501**	**0.460**	0.479	**0.496**	0.411	**0.505**	**0.453**	0.392
	Selected L_f	SV	FR	FR	SV	KO	ZH	LT	NL	FR	LT
Reference: Direct Fine-tuning	Acc.	0.450	0.484	0.493	0.455	0.502	0.485	0.486	0.486	0.460	0.422

Table 4: The validation results of four fine-tuning language selection strategies, conducted on the Global MMLU dataset using Llama-3B, Qwen-4B, and Gemma-4B. Each column represents a target language. The reported values are average task accuracies (Acc.) on the target language after fine-tuning the model with the selected language. "-" means there is no candidate language for fine-tuning within these 20 languages. **The integrated similarity measure performs** (in green), providing **comparable performance with direct fine-tuning** (in gray). Only results on first ten languages listed there, refer to Appendix E for the complete results.

4.2 Applying LLM-Based Language Similarity to Selecting Fine-Tuning Languages

We evaluate whether LLM-based language similarity can effectively guide the selection of fine-tuning languages to improve target-language performance. Specifically, we compare four strategies for choosing fine-tuning languages: three derived from LLM-based similarity measures (feature-based, transferability-based, and integrated) and one based on traditional linguistic similarity. We fine-tune LLMs according to each strategy and examine their performance on the target languages after fine-tuning.

We conduct experiments on the multilingual benchmark Global MMLU (Singh et al., 2024), which contains the largest language overlap (20 languages) with our study, allowing us to comprehensively validate the efficacy of our similarity-driven language selection strategies. And we randomly partition the test split of Global MMLU into a fine-tuning split and a new test split with an 8:2 ratio. To highlight the impact of fine-tuning, we deliberately avoid using models that already perform strongly prior to fine-tuning. Instead, we select smaller-scale models of comparable sizes from different organizations: Llama-3.2-3B (Grattafiori et al., 2024), denoted as Llama-3B; Qwen3-4B-Base (Yang et al., 2025), denoted as Qwen-4B; and Gemma-3-4B-PT (Team et al., 2025), denoted as Gemma-4B.

Table 4 presents the experimental results for the first ten languages, where the performance of direct fine-tuning has been added as a reference. Appendix E provides the complete results. Since traditional linguistic similarity lacks quantitative similarity scores and multiple languages often belong to the same language branch as the target language, contributing to multiple candidate languages under this strategy, we report the average performance using those candidate fine-tuning languages. Additionally, in cases where no candidate language exists among these 20 languages under the traditional similarity measure, we denote the entry with "–". For the mapping between language abbreviations and their full names, refer to Table 2.

Table 4 shows that LLM-based similarity measures outperform traditional similarity measures in almost all cases, with the best results among the four strategies highlighted in bold. Notably, these measures even surpass direct fine-tuning in many cases.

To provide an intuitive summary of these results, in Table 5, we further count the number

Strategy	Best Count / Total Evaluations
Traditional Language Similarity	6/33
LLM-Based Similarity (Feature)	23/60
LLM-Based Similarity (Transferability)	18/60
LLM-Based Similarity (Integration)	40/60

Table 5: Counts of best-performing cases for each fine-tuning language selection strategy across all evaluation settings. There are a total of 60 comparison opportunities (20 target languages × 3 models), but traditional similarity strategy is only applicable to 33 out of 60 evaluations due to missing candidate languages. The measure integrating two LLM perspectives outperforms.

Model	AR	CS	DE	EL	ES	FR	ID	IT	JA	KO	LT	NL	PL	PT	RO	RU	SV	TR	UK	ZH
Llama-3B	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Qwen-4B	✓	✓	-	✓	✓	–	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Gemma-4B	✓	–	✓	✓	✓	✓	-	✓	–	✓	✓	✓	✓	✓	–	✓	✓	✓	✓	✓

Table 6: Languages exhibiting the **super-additive transfer effect** for each model. "✓" indicates that fine-tuning on another language achieves higher accuracy than direct fine-tuning on the target language; "–" indicates the absence of this effect for the corresponding language–model pair.

of times each strategy achieves the highest. Table 5 shows that any of the strategies based on LLM-derived language similarity (their "best count" ranges from 18/60 to 40/60) outperform the strategy based on traditional similarity measure (only 6/33 "best count"). And among the three LLM-derived approaches, **the integrated similarity measure** performs best, achieving 40 "best counts" out of 60 total evaluations.

Moreover, this experiment further verifies that the **super-additive transfer effect** is a widespread phenomenon: among the 20 target languages, only few cases show that direct fine-tuning (fine-tuning directly on the target language) outperforms cross-lingual fine-tuning. Through Table 4, we can easily compare the performance of our methods and that of direct fine-tuning. Table 6 presents the detailed occurrences of the super-additive effect observed in our experiments, showing that this phenomenon appears in nearly all 20 languages across models.

5 Conclusion

In this work, we systematically investigate **LLM-based language similarity** through both the lenses of language-specific performance patterns and cross-lingual transferability. Our approach provides not only a novel perspective on language similarity, but also **practical baselines for selecting fine-tuning languages**. We verify the effectiveness LLM-based similarity measures on guiding fine-tuning language selection to improve LLMs' performance on target languages. Results show that **the integrated measure outperforms**, providing comparable performance with direct fine-tuning. We recommend researchers to adopt this strategy in fine-tuning language selection. Moreover, we observe an intriguing and widespread phenomenon: **super-additive transfer effects**, where fine-tuning on a certain language yields higher performance than fine-tuning directly on the target language.

Additionally, we introduce the multilingual multitask dataset **M4CQ-Pro**, which covers 135 tasks across various domains and 31 languages (including over 20 medium- and low-resource languages), with 61518 manually reviewed high-quality questions per language. We expect this resource to also facilitate future research in other directions of NLP.

Limitations

Language Coverage Limitations: While M4CQ-Pro includes 31 languages, it excludes many low-resource and morphologically complex languages critical for comprehensive analysis of language similarity.

Task Coverage Limitations: The 135-task taxonomy, though diverse, may not fully capture all dimensions of linguistic similarity.

Ethical Considerations

Our study adheres to the following ethical guidelines: (1) The M4CQ-Pro dataset excludes personally identifiable information (PII) and culturally sensitive content through automated filtering and manual review. (2) The dataset is released under CC BY-NC-SA 4.0 license with explicit prohibitions against military or surveillance applications. Potential biases in monolingual pre-training corpora remain an unresolved limitation.

Acknowledgments

This work was supported in part by General Program of National Natural Science Foundation of China (62372403) and "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2025C02032). This work was also supported by the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

Shivam Agarwal. 2013. Data mining: Data mining concepts and techniques. In *2013 International Conference on Machine Intelligence and Research Advancement*, pages 203–207.

Ajitesh Bankula and Praney Bankula. 2025. Cross-linguistic transfer in multilingual nlp: The role of language families and morphology. *Preprint*, arXiv:2505.13908.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Paul K. Benedict. 1990. *Japanese/Austro-Tai*. Karoma, Ann Arbor.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson. 2012. Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960.

C. H. Brown, E. W. Holman, S. Wichmann, and V. Velupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *STUF - Language Typology and Universals*, 61(4):285–308.

Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022. An exploration of hierarchical attention transformers for efficient long document classification. *arXiv preprint*.

Charles Condevaux and Sébastien Harispe. 2023. Lsg attention: Extrapolation of pretrained transformers to long sequences. In *Advances in Knowledge Discovery and Data Mining*, pages 443–454, Cham. Springer Nature Switzerland.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

M. Dunn, S. J. Greenhill, S. C. Levinson, and R. D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Mark Hudson. 1999. *Japanese and Austronesian: An Archeological Perspective on the Proposed Linguistic Links*, pages 267–279.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Takao Kawamoto. 1977. Toward a comparative japanese-austronesian i. *Bulletin of Nara University of Education*, 26(1).

Takao Kawamoto. 1978. *Minami kara kita Nihongo*. Sanseidō, Tōkyō.

Takao Kawamoto. 1980. *Nihongo no genryū*. Kōdansha, Tōkyō.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *ArXiv*, abs/2001.04451.

Jungseob Lee, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. 2025. Cross-lingual optimization for

language transfer in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15100–15119, Vienna, Austria. Association for Computational Linguistics.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, 26 edition. SIL International.

Daoyang Li, Haiyan Zhao, Qingcheng Zeng, and Mengnan Du. 2025a. Exploring multilingual probing in large language models: A cross-language analysis. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 61–70, Vienna, Austria. Association for Computational Linguistics.

Yue Li, Zhixue Zhao, and Carolina Scarton. 2025b. It's all about in-context learning! teaching extremely low-resource languages to LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29544–29559, Suzhou, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

April McMahon and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford University Press.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Roy A. Miller. 1967. *The Japanese Language*. University of Chicago Press, Chicago.

Roy A. Miller. 1971. *Japanese and the Other Altaic Languages*. University of Chicago, Chicago.

Roy A. Miller. 1975. Japanese-altaic lexical evidence and the proto-turkic "zetacism-sigmatism". In *Researches in Altaic Languages*, pages 157–172. Akadémiai Kiadé, Budapest.

Roy A. Miller. 1979. Old korean and altaic. *Ural-Altaische Jahrbucher*, 51:1–54.

Roy A. Miller. 1980. *Origins of the Japanese Language*. University of Washington, Seattle.

Roy A. Miller. 1983. Japanese evidence for some altaic denominal verb-stem derivational suffixes. *Acta Orientalia Hungarica*, 36:391–403.

Roy A. Miller. 1985a. Altaic connections of the old japanese negatives. *Central Asiatic Journal*, 29:35–84.

Roy A. Miller. 1985b. Externalizing internal rules: Lyman's law in japanese and altaic. *Diachronica*, 2(2):137–165.

Shichirō Murayama. 1957. Vergleichende betrachtung der kasus-suffixe im altjapanischen. In Julius von Farkas and Omeljan Pritsak, editors, *Studia Altaica: Festschrift für Nikolaus Poppe zum 60 Geburstag*, volume 5 of *Ural-Altaische Bibliothek*, pages 126–131. Otto Harrassowitz, Wiesbaden.

Shichirō Murayama. 1962. Nihongo no tingusugo teki yéso. *Minzokugaku kenkyū*, 26(3).

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024. Mmmlu dataset. https://huggingface.co/datasets/openai/MMMLU. Accessed: 2025-09-14.

M. Pagel, Q. D. Atkinson, A. S. Calude, and A. Meade. 2013. Ultraconserved words point to deep language ancestry across eurasia. *Proceedings of the National Academy of Sciences of the United States of America*, 110(21):8471–8476.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

G. J. Ramstedt and Pentti Aalto. 1952. Einführung in die altaische sprachwissenschaft.

Enora Rice, Ali Marashian, Hannah Haynie, Katharina von der Wense, and Alexis Palmer. 2025. Untangling the influence of typology, data and model architecture on ranking transfer languages for cross-lingual pos tagging. *Preprint*, arXiv:2503.19979.

Martine Robbeets. 2005. Is japanese related to korean, tungusic, mongolic and turkic?

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Masayoshi Shibatani. 1990. *The Languages of Japan*. Cambridge University Press, Cambridge.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek B Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 425 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615.

Sergei A. Starostin. 1991. *Altaiskaia problema i proiskhozhdenie iaponskogo iazyka*. Nauka, Moscow.

John Street and Roy Andrew Miller. 1975–77. *Altaic Elements in Old Japanese*, volume 1. Madison, Wisconsin.

John Charles Street and Roy Andrew Miller. 1973. Japanese and the other altaic languages. *Language*, 49:950.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Nicolas Tranter, editor. 2012. *The Languages of Japan and Korea*, 1st edition. Routledge.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2019. How does bert answer questions? *Proceedings of the 28th ACM International Conference on Information and Knowledge Management - CIKM '19*.

Alexander Vovin. 1994. Is japanese related to austronesian? *Oceanic Linguistics*, 33(2):369–390. Accessed 16 Mar. 2025.

Alexander Vovin. 2005. The end of the altaic controversy. *Central Asiatic Journal*, 49:71–132.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461.

Søren Wichmann, Eric W. Holman, and Cecil H. (eds.) Brown. 2022. The asjp database (version 20). Available at `http://asjp.clld.org/`.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *Preprint*, arXiv:2503.10497.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yidan Zhang, Yu Wan, Boyi Deng, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, Fei Huang, and Jingren Zhou. 2024. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. *Preprint*, arXiv:2411.09116.

Mikel Zubillaga, Oscar Sainz, Ainara Estarrona, Oier Lopez de Lacalle, and Eneko Agirre. 2024. Event extraction in Basque: Typologically motivated cross-lingual transfer-learning analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6607–6621, Torino, Italia. ELRA and ICCL.

A Dataset Format of M4CQ-Pro

The format of all tasks in M4CQ-Pro are consistent with that of M4CQ:

- **`idx`** (Int32): Unique instance identifier within each task.
- **`question`** (String): Task prompt.
- **`choices`** (List[String]): Options, supporting variable counts.
- **`answer`** (List[Int32]): Binary vector indicating correct options (1 for correct, 0 for incorrect), with the same length as options.

This schema supports flexible option counts and future extension to multi-answer questions, overcoming the rigidity of fixed-option formats (e.g., "A/B/C/D" constraints).

B Task Information of M4CQ-Pro

Table 7 shows domains and descriptions of 135 tasks in M4CQ-Pro.

Table 7: Task information of M4CQ-Pro

Domain	Task Name	Description
Knowledge - Math	elementary math	Math problems of elementary school difficulty.
	high school math	Math problems of high school difficulty.
	high school statistics	Statistics problems of high school difficulty.
	geometry intersect	The number of points of intersection of two geometric figures (expressed in coordinates).
	abstract algebra	Questions about abstract algebra.
	college math	Math problems of college difficulty.
	partial function	Math problems based on partial functions.
Knowledge - Physics	high school physics	Physics problems of high school difficulty.
	physical intuition	Physics questions that do not involve calculations.
	physics formula	Identify the most useful physics formula for a Physics problem.
	college general physics	Examination questions in general physics.
	astrophysics	Questions about astrophysics.
	conceptual physics	Questions about physics concepts.
Knowledge - Biology	high school biology	Biology problems of high school difficulty.
	college biology	Biology problems of college difficulty.
	genetics	Questions about genetics.
	virology	Questions about virology.
	cryobiology	Questions about cryobiology.
Knowledge - Chemistry	periodic elements	Questions about the Periodic Table.
	high school chemistry	Chemistry problems of high school difficulty.
	college chemistry	Chemistry problems of college difficulty.
Knowledge - Philosophy	college philosophy	Examination questions in college philosophy.
	logical fallacies	Questions about logical fallacies.
	moral disputes	Questions about moral disputes.

Table 7 continued from previous page

Domain	Task Name	Description
Knowledge - Sociology	college sociology	Examination questions in college sociology.
	world religions	Questions about world religions.
	security studies	Questions about security studies, related to environmental security, terrorism, weapons of mass destruction, etc.
	public relations	Questions about public relations.
Knowledge - Psychology	high school psychology	Psychology problems of high school difficulty.
	professional psychology	Psychology problems of college difficulty.
Knowledge - Geography	high school geography	Geography problems of high school difficulty.
Knowledge - History	high school world history	World history problems of high school difficulty.
	anachronisms	Given a description, answer if there are any items/phrases that appear out of place in the period context.
	prehistory	Questions about prehistory studies.
Knowledge - Law	international law	Questions about international law.
Knowledge - Medicine	anatomy	Questions about anatomy.
	nutrition	Questions about nutrition.
	diagnostics	Given a clinical diagnostic case, answer the relevant medical questions.
	organology	Questions about the functions of human organs.
	college medicine	Medical problems of college difficulty.
	human aging	Questions about human aging.
	human sexuality	Questions about human sexuality, related to gender differences, sexual orientation, pregnancy, etc.
Knowledge - Computer Science	high school computer science	Computer science questions of high school difficulty.
	college computer science	Computer science questions of college difficulty.
	computer security	Questions about computer security.
	machine learning	Questions about machine learning.

Table 7 continued from previous page

Domain	Task Name	Description
Knowledge - Electrical Engineering	electrical engineering	Questions about electrical engineering.
Knowledge - Business Studies	business ethics	Business-related gap-fill questions (fill in the blanks by selecting words from the context).
	econometrics	Questions about econometrics.
	high school macroeconomics	Macroeconomics questions of high school difficulty.
	high school microeconomics	Microeconomics questions of high school difficulty.
	professional accounting	Questions about accounting.
Knowledge - Management	management	Questions about management.
Knowledge - Marketing	marketing	Questions about marketing.
Knowledge - General Knowledge	global facts	Questions involving some global statistical data.
	kindergarten knowledge	Kindergarten level general knowledge questions.
	factual judgment	Questions about factual judgment.
	realistic interaction problem	Life-oriented problems that simulate real-world interaction
	scientific common sense	Questions about scientific common sense.
NLU (Natural Language Understanding) - Counterfactual Conditional Question Answering	counterfactualQA	Answer a question based on the given text (inconsistent with facts).
NLU - Information Extraction	sentence info extract	Given a sentence, answer a judgment question based on the sentence.
	table info extract	Given a form, answer a question based on the content of the form.
	context info extract	Answer a question based on a passage (containing 2-8 sentences).
NLU - Long Context Reading Comprehension	gre reading comprehension	Reading comprehension questions in GRE test.
	question selection	Given a context and a short answer (usually a number), choose the corresponding question.
NLU - Coreference Resolution	disambiguation qa	Determine what the pronoun in the sentence refers to in the form of a tautological paraphrase.
	winogrande	The first half of the sentence involves two people, and the second half digs in to ask which one should be filled in.

Table 7 continued from previous page

Domain	Task Name	Description
NLU - Sentiment Analysis	movie review attitude	Given an excerpted sentence from a movie review, evaluate whether its sentiment is positive or negative.
	reply attitude	Determine whether the attitude of a reply is supportive, neutral, or opposing.
	sentence emo	Determine whether the sentence reveals a positive, negative, neutral or contradictory emotion.
	suicide risk	Given a text, determine the author's suicide risk.
	character feeling	Given a short context, infer the character's feelings based on the scene.
NLU - Semantic Similarity	concept feature	Given a noun phrase, determine which sentence best characterizes the phrase.
	movie recommendation	Choose a movie that is similar to the four movies specified.
	phrase relatedness	Find the most relevant word or phrase.
NLU - Word Categorization	odd one out	Given a set of words, identify the ones that don't fit together.
	commonality abstraction	Find common ground for two nouns.
NLU - Relation Extraction	character relationship	Given a text (basically three or four sentences), determine the character relationship within it.
NLU - Style Identification	authorship identification	Given a text, determine which of the following texts is of the same author as the given text.
	figure identification	Determine the rhetorical device used in the given sentence.
	irony identification	Determine whether the given sentence is ironic.
	snarks	Given two similar sentences, determine which one is ironic.
	humor identification	Determine whether the given text is (cold) humorous.
NLU - Riddle	riddle sense	Brain teaser questions.

Table 7 continued from previous page

Domain	Task Name	Description
NLU - Intent Recognition	goal	Determine the goal of performing an operation.
	step	Answer the steps needed to achieve the given purpose.
	implicatures	Given a conversation in which speaker1 asks a question and speaker2 doesn't answer directly, determine whether speaker2 means yes or no.
	intent	Determining the intent of a sentence.
NLU - Theme Extraction	story moral	Given a story, extract the point the story is trying to make.
NLI (Natural Language Inference) - Textual Entailment	entailment 1p1h2c	Given 1 premise and 1 hypothesis, determine whether it is an entailment (whether the hypothesis can be derived from the premise). 2 choices: entailment/no-entailment.
	entailment fact 1p1h2c	Given 1 fact and 1 hypothesis, determine whether it is an entailment (whether the hypothesis can be derived from the fact). 2 choices: entailment/no-entailment.
	entailment 1p1h3c	Given two sentences, determine their relationship. 3 choices: entailment/neutral/contradiction.
	entailment 2p1h2c	Given 2 premises and 1 hypothesis, determine whether it is an entailment (whether the hypothesis can be derived from the premises). 2 choices: entailment/no-entailment.
	context entailment 1h2c	Given a premise context (3-6 sentences, usually 3) and 1 hypothesis, determine whether it is an entailment (whether the hypothesis can be derived from the premise context). 2 choices: entailment/no-entailment.
NLI - Relationship Analogy	similarity analogy	Given examples of six types of similarity, determine which similarity a new example belongs to.

Table 7 continued from previous page

Domain	Task Name	Description
NLI - Text Coherence	logical coherence	Given a text of 10 sentences, it is known that the first few sentences were written by a human and then become computer-written. Find out where the shift begins (it's not a matter of language style, it's that the logic back and forth doesn't make sense anymore).
	content coherence	Given the first half of a paragraph, choose the one that best fits as the next sentence.
NLI - Speaker Attribution	movie dialog same or different	Given an unattributed conversation which comes from a movie, pick out two sentences and ask if they are from the same person.
	play dialog same or different	Given an unattributed conversation which comes from a play, pick out two sentences and ask if they are from the same person.
NLI - Dynamic Relation Tracking	tracking shuffled objects	Given the initial pairing relationship and the subsequent rounds of exchanges, ask what is now the paired item/person for the particular object.
NLI - Spatial Reasoning	navigate	Given a sequence of commands to walk back and forth, determine whether it can get back to the original point.
NLI - Temporal Reasoning	temporal computation	Calculate the date based on the given information.
	temporal sequences	Given someone's schedule for part of the day (the time to do Event A is not mentioned), ask when Event A may be done.
NLI - Sequence Reasoning	order of procedures	Given an objective and two operations, determine in what order the two operations should be performed to achieve that purpose.
	logical sequence	Sorting several objects with intrinsic order relationships.

Table 7 continued from previous page

Domain	Task Name	Description
NLI - Scenario-based Reasoning	scenario qa	Given a scene, ask about feelings/reasons/follow-ups, etc.
	scenario hypothesis qa	Given a scenario, assume one more thing that conflicts with what just happened in the scenario, and ask what might have happened.
	fantasy reasoning	Given a description of a scenario that couldn't happen in reality (e.g. hell, demons, etc.), and ask a judgment question based on it
NLI - Commonsense Reasoning	timedial	Given a piece of dialog in which a word indicating duration/time is obscured, choose a reasonable value for the obscured duration/time.
NLI - Causal Judgment	causal judgment	Given a description of a scenario, judge the causal relationship.
	reasonable causation	Given two sentences in which the causal logic is exactly opposite, determine which sentence contains the correct causal relationship.
	cause extraction	Given a short context, answer the reason (LLM needs to infer from the scenario).
	possible cause or effect	Given a premise, choose its possible cause/effect.
NLI - Information Sufficiency Inference	evaluating information essentiality	Assessing Information Importance: Given a question, determine how useful the following two statements are in answering that question.
	not sufficient	Answer a judgment question based on the given short context. When the information in the context is not sufficient enough to judge, select "Either". (The answer to all instances of this task is "Either".)
	whether sufficient	Given a sentence and a question, determine whether the sentence answer the question.

Table 7 continued from previous page

Domain	Task Name	Description
NLI - Argument Assessment	argument logic	Questions about argument logic.
	logical fallacy detection	Determine whether the given causal logical reasoning is correct or not.
	mathematical induction	Mathematical inference questions.
NLI - Contradiction Identification	lie judgment	Given a short context, identify whether what the character has said is true.
NLI - Semantic Consistency Checking	metaphor understanding	Given a sentence that uses metaphorical rhetoric and an explanation of the metaphorical sentence, answer if this explanation conforms to the meaning of the original metaphorical sentence.
	sentence equivalence	Given two sentences, determine if they have the same meaning.
	question equivalence	Given two questions, determine if they have the same meaning.
NLI - Constraint Satisfaction Problem	house number	There is exactly one person living in each house, and the person living in each house has different characteristics in several dimensions. Given a number of hints (relating the positional relationships of the houses of people with different characteristics), answer the number of the house in which a person with a certain characteristic lives.
	logical deduction	Given a paragraph of known conditions (involving the interrelationships of several objects, e.g., location, price, time to accomplish something, age of an antique, etc.), determine which option is correct (each option is a judgment sentence about an object).

Table 7 continued from previous page

Domain	Task Name	Description
LLM Special Test	known unknowns	Factual questions, but some were unknown, testing LLM's ability to answer UNKNOWN.
	hhh alignment	"HHH" stands for 'Helpful, Harmless, and Honest'. Through these tasks, the model can be tested to see if it can be useful, honest, and without negative impacts in real-world applications.
	trolley dilemma	It's a moral question of the Trolley Dilemma type: to do or not to do something.
	ethical question	Test LLM's ability to answer ethically.
	color understanding	Given a color representation in RGB/HCL/hexadecimal/HSL format, ask which is the closest color.
	geometric shapes understanding	Given an SVG path element, answer its shape.
Code (Ability to Use)	longest common subsequence	Given two strings, answer the length of the longest common subsequence.
	bracket match judgement	Given a string with parentheses, center brackets, and curly braces, determine if the left and right brackets are perfectly matched.
	bracket match complement	Given a string with parentheses, center brackets, and braces, complete the string so that the left and right brackets match perfectly.
	symbol interpretation	Use different symbols to refer to specific graphics/specific expressions. Given two symbol strings, determine which option's description matches the first string but does not match the second string.

C Experimental Setup

C.1 Dataset Splitting

For each language variant:

- First 20% instances of every task: Training set (merged into language-level training corpus)
- Remaining 80%: Test set

This split ensures content consistency across languages in both training and test sets.

C.2 Model Selection

Given the requirements of both sufficient context length (indicating newer models) and pre-training on a single language (indicating earlier models), our options were limited. Initially, BERT (Devlin et al., 2018) was considered for its classic transformer architecture, however, its 512-token context window proved insufficient. During this process, it's found that the context length should be no less than 4096. We ultimately adopted `lsg-bert-base-uncased-4096`, a variant of BERT that extends the context length to 4096 tokens through a modified attention mechanism (Local + Sparse + Global attention) (Condevaux and Harispe, 2023) without changing the rest architecture or pre-training weight of $\text{BERT}_{\text{base}}$. Compared to other potential alternative models (Liu et al., 2019; Raffel et al., 2019; Beltagy et al., 2020; Kitaev et al., 2020; Fedus et al., 2021; Guo et al., 2021; Chalkidis et al., 2022), this model, besides meeting our experimental requirements, retains enough similarity to BERT, aligning with our preference for classical architectures.

C.3 Fine-tuning Settings

Table 8 details relevant LoRA fine-tuning configuration.

D Similarity Scores for 30×30 Language Pairs with Three Perspectives

The 30×30 similarity scores obtained with a perspective of LLM-derived language features are listed in Table 9.

The 30×30 similarity scores obtained with a perspective of cross-lingual transferability are listed in Table 10.

The 30×30 similarity scores obtained with an integration perspective are listed in Table 11.

E Complete Validation Results

The complete results are listed in Table 12.

Table 8: Hyperparameters and Settings for Fine-Tuning with LoRA

Parameter	Value/Setting	Description
LoRA Configuration		
r	8	Rank of the low-rank matrices in LoRA.
lora_alpha	32	Scaling factor for LoRA weights.
target_modules	["query", "value"]	Transformer layers to apply LoRA.
lora_dropout	0.1	Dropout probability for LoRA layers.
bias	"none"	Disable bias terms in LoRA.
task_type	"FEATURE_EXTRACTION"	Task type for LoRA fine-tuning.
Training Configuration		
per_device_train_batch_size	4	Batch size per device during training.
num_train_epochs	3	Total number of training epochs.
logging_steps	10	Log metrics every N steps.
save_steps	500	Save model checkpoint every N steps.
save_total_limit	2	Maximum number of checkpoints to save.
eval_strategy	"steps"	Evaluation strategy (evaluate every N steps).
eval_steps	500	Evaluate model every N steps.
metric_for_best_model	"loss"	Metric to determine the best model.
greater_is_better	False	Lower loss indicates better performance.
fp16	True	Enable mixed-precision training.
Data Processing		
max_length	4096	Maximum sequence length for input data.
test_size	0.2	Fraction of data used for validation.

Table 9: Similarity scores for 30×30 language pairs with a perspective of LLM-derived language features.

	AR	BG	CS	DA	DE	EL	ES	ET	FI	FR	HU	ID	IT	JA	KO	LT	LV	NB	NL	PL	PT-BR	PT-PT	RO	RU	SK	SL	SV	TR	UK	ZH
AR	1.0000	0.9801	0.9714	0.9723	0.9704	0.9755	0.9823	0.9826	0.9798	0.9709	0.9832	0.9781	0.9786	0.9705	0.9811	0.9755	0.9857	0.9842	0.9775	0.9773	0.9723	0.9786	0.9866	0.9845	0.9749	0.9818	0.9859	0.9741	0.9689	0.9756
BG	0.9801	1.0000	0.9713	0.9766	0.9762	0.9844	0.9834	0.9821	0.9809	0.9766	0.9874	0.9800	0.9725	0.9795	0.9890	0.9793	0.9829	0.9791	0.9751	0.9753	0.9784	0.9725	0.9793	0.9802	0.9873	0.9885	0.9871	0.9814	0.9685	0.9829
CS	0.9714	0.9713	1.0000	0.9838	0.9870	0.9655	0.9680	0.9748	0.9831	0.9860	0.9715	0.9802	0.9812	0.9780	0.9668	0.9753	0.9752	0.9760	0.9863	0.9831	0.9797	0.9847	0.9720	0.9755	0.9647	0.9741	0.9759	0.9789	0.9557	0.9764
DA	0.9723	0.9766	0.9838	1.0000	0.9859	0.9752	0.9720	0.9724	0.9869	0.9836	0.9768	0.9843	0.9866	0.9699	0.9708	0.9820	0.9818	0.9758	0.9891	0.9872	0.9896	0.9801	0.9768	0.9763	0.9734	0.9789	0.9812	0.9831	0.9597	0.9794
DE	0.9704	0.9762	0.9870	0.9859	1.0000	0.9663	0.9683	0.9700	0.9837	0.9848	0.9720	0.9866	0.9859	0.9716	0.9710	0.9797	0.9776	0.9713	0.9850	0.9865	0.9843	0.9815	0.9716	0.9763	0.9702	0.9759	0.9778	0.9830	0.9582	0.9803
EL	0.9755	0.9844	0.9655	0.9752	0.9663	1.0000	0.9789	0.9793	0.9744	0.9683	0.9809	0.9722	0.9619	0.9751	0.9842	0.9719	0.9765	0.9744	0.9690	0.9694	0.9699	0.9663	0.9713	0.9725	0.9799	0.9859	0.9801	0.9763	0.9615	0.9745
ES	0.9823	0.9834	0.9680	0.9720	0.9683	0.9789	1.0000	0.9828	0.9747	0.9756	0.9893	0.9802	0.9738	0.9734	0.9865	0.9838	0.9790	0.9824	0.9732	0.9715	0.9733	0.9714	0.9812	0.9829	0.9834	0.9866	0.9846	0.9752	0.9746	0.9789
ET	0.9826	0.9821	0.9748	0.9724	0.9700	0.9793	0.9828	1.0000	0.9744	0.9731	0.9839	0.9786	0.9681	0.9808	0.9851	0.9760	0.9791	0.9827	0.9737	0.9730	0.9701	0.9776	0.9763	0.9776	0.9775	0.9847	0.9812	0.9768	0.9673	0.9742
FI	0.9798	0.9809	0.9831	0.9869	0.9837	0.9744	0.9747	0.9744	1.0000	0.9820	0.9757	0.9867	0.9886	0.9746	0.9723	0.9832	0.9801	0.9799	0.9892	0.9883	0.9881	0.9849	0.9810	0.9819	0.9753	0.9787	0.9832	0.9835	0.9648	0.9828
FR	0.9709	0.9766	0.9860	0.9836	0.9848	0.9683	0.9756	0.9731	0.9820	1.0000	0.9781	0.9830	0.9822	0.9787	0.9744	0.9841	0.9712	0.9718	0.9836	0.9816	0.9835	0.9797	0.9736	0.9769	0.9749	0.9774	0.9780	0.9842	0.9684	0.9836
HU	0.9832	0.9874	0.9715	0.9768	0.9720	0.9809	0.9893	0.9839	0.9757	0.9781	1.0000	0.9813	0.9738	0.9750	0.9897	0.9857	0.9832	0.9813	0.9769	0.9733	0.9753	0.9708	0.9805	0.9792	0.9854	0.9896	0.9847	0.9824	0.9743	0.9807
ID	0.9781	0.9800	0.9802	0.9843	0.9866	0.9722	0.9802	0.9786	0.9867	0.9830	0.9813	1.0000	0.9853	0.9775	0.9787	0.9879	0.9765	0.9782	0.9837	0.9848	0.9853	0.9858	0.9741	0.9851	0.9749	0.9769	0.9802	0.9870	0.9672	0.9855
IT	0.9786	0.9725	0.9812	0.9866	0.9859	0.9619	0.9738	0.9681	0.9886	0.9822	0.9738	0.9853	1.0000	0.9648	0.9686	0.9827	0.9788	0.9744	0.9887	0.9887	0.9854	0.9813	0.9831	0.9795	0.9715	0.9745	0.9809	0.9785	0.9618	0.9777
JA	0.9705	0.9795	0.9780	0.9699	0.9716	0.9751	0.9734	0.9808	0.9746	0.9787	0.9750	0.9775	0.9648	1.0000	0.9762	0.9765	0.9704	0.9747	0.9696	0.9709	0.9721	0.9810	0.9625	0.9739	0.9750	0.9767	0.9764	0.9779	0.9663	0.9804
KO	0.9811	0.9890	0.9668	0.9708	0.9710	0.9842	0.9865	0.9851	0.9723	0.9744	0.9897	0.9787	0.9686	0.9762	1.0000	0.9821	0.9786	0.9746	0.9702	0.9722	0.9704	0.9702	0.9739	0.9787	0.9834	0.9880	0.9808	0.9824	0.9687	0.9823
LT	0.9755	0.9793	0.9753	0.9820	0.9797	0.9719	0.9838	0.9760	0.9832	0.9841	0.9857	0.9879	0.9827	0.9765	0.9821	1.0000	0.9740	0.9749	0.9814	0.9812	0.9817	0.9794	0.9742	0.9793	0.9782	0.9804	0.9765	0.9867	0.9704	0.9872
LV	0.9857	0.9829	0.9752	0.9818	0.9776	0.9765	0.9790	0.9791	0.9801	0.9712	0.9832	0.9765	0.9788	0.9704	0.9786	0.9740	1.0000	0.9826	0.9815	0.9807	0.9779	0.9742	0.9850	0.9776	0.9771	0.9883	0.9883	0.9748	0.9594	0.9699
NB	0.9842	0.9791	0.9760	0.9758	0.9713	0.9744	0.9824	0.9827	0.9799	0.9718	0.9813	0.9782	0.9744	0.9747	0.9746	0.9749	0.9826	1.0000	0.9744	0.9731	0.9748	0.9811	0.9821	0.9796	0.9726	0.9788	0.9820	0.9719	0.9648	0.9727
NL	0.9775	0.9751	0.9863	0.9891	0.9850	0.9690	0.9732	0.9737	0.9892	0.9836	0.9769	0.9837	0.9887	0.9696	0.9702	0.9814	0.9815	0.9744	1.0000	0.9896	0.9877	0.9818	0.9791	0.9797	0.9711	0.9792	0.9836	0.9831	0.9611	0.9783
PL	0.9773	0.9753	0.9831	0.9872	0.9865	0.9694	0.9715	0.9730	0.9883	0.9816	0.9733	0.9848	0.9887	0.9709	0.9722	0.9812	0.9807	0.9731	0.9896	1.0000	0.9864	0.9812	0.9789	0.9822	0.9704	0.9787	0.9827	0.9816	0.9631	0.9797
PT-BR	0.9723	0.9784	0.9797	0.9896	0.9843	0.9699	0.9733	0.9701	0.9881	0.9835	0.9753	0.9853	0.9854	0.9721	0.9704	0.9817	0.9779	0.9748	0.9877	0.9864	1.0000	0.9811	0.9764	0.9813	0.9772	0.9750	0.9822	0.9792	0.9651	0.9844
PT-PT	0.9786	0.9725	0.9847	0.9801	0.9815	0.9663	0.9714	0.9776	0.9849	0.9797	0.9708	0.9858	0.9813	0.9810	0.9702	0.9794	0.9742	0.9811	0.9818	0.9812	0.9811	1.0000	0.9718	0.9801	0.9640	0.9686	0.9752	0.9805	0.9602	0.9812
RO	0.9866	0.9793	0.9720	0.9768	0.9716	0.9713	0.9812	0.9763	0.9810	0.9736	0.9805	0.9741	0.9831	0.9625	0.9739	0.9742	0.9850	0.9821	0.9791	0.9789	0.9764	0.9718	1.0000	0.9776	0.9765	0.9807	0.9885	0.9659	0.9647	0.9690
RU	0.9845	0.9802	0.9755	0.9763	0.9763	0.9725	0.9829	0.9776	0.9819	0.9769	0.9792	0.9851	0.9795	0.9739	0.9787	0.9793	0.9776	0.9796	0.9797	0.9822	0.9813	0.9801	0.9776	1.0000	0.9750	0.9781	0.9837	0.9804	0.9790	0.9834
SK	0.9749	0.9873	0.9647	0.9734	0.9702	0.9799	0.9834	0.9775	0.9753	0.9749	0.9854	0.9749	0.9715	0.9750	0.9834	0.9782	0.9771	0.9726	0.9711	0.9704	0.9772	0.9640	0.9765	0.9750	1.0000	0.9862	0.9817	0.9742	0.9731	0.9763
SL	0.9818	0.9885	0.9741	0.9789	0.9759	0.9859	0.9866	0.9847	0.9787	0.9774	0.9896	0.9769	0.9745	0.9767	0.9880	0.9804	0.9883	0.9788	0.9792	0.9787	0.9750	0.9686	0.9807	0.9781	0.9862	1.0000	0.9878	0.9810	0.9685	0.9777
SV	0.9859	0.9871	0.9759	0.9812	0.9778	0.9801	0.9846	0.9812	0.9832	0.9780	0.9847	0.9802	0.9809	0.9764	0.9808	0.9765	0.9883	0.9820	0.9836	0.9827	0.9822	0.9752	0.9885	0.9837	0.9817	0.9878	1.0000	0.9758	0.9697	0.9765
TR	0.9741	0.9814	0.9789	0.9831	0.9830	0.9763	0.9752	0.9768	0.9835	0.9842	0.9824	0.9870	0.9785	0.9779	0.9824	0.9867	0.9748	0.9719	0.9831	0.9816	0.9792	0.9805	0.9659	0.9804	0.9742	0.9810	0.9758	1.0000	0.9682	0.9868
UK	0.9689	0.9685	0.9557	0.9597	0.9582	0.9615	0.9746	0.9673	0.9648	0.9684	0.9743	0.9672	0.9618	0.9663	0.9687	0.9704	0.9594	0.9648	0.9611	0.9631	0.9651	0.9602	0.9647	0.9790	0.9731	0.9685	0.9697	0.9682	1.0000	0.9735
ZH	0.9756	0.9829	0.9764	0.9794	0.9803	0.9745	0.9789	0.9742	0.9828	0.9836	0.9807	0.9855	0.9777	0.9804	0.9823	0.9872	0.9699	0.9727	0.9783	0.9797	0.9844	0.9812	0.9690	0.9834	0.9763	0.9777	0.9765	0.9868	0.9735	1.0000

Table 10: Similarity scores for 30×30 language pairs with a perspective of cross-lingual transferability.

L_f \ L_e	AR	BG	CS	DA	DE	EL	ES	ET	FI	FR	HU	ID	IT	JA	KO	LT	LV	NB	NL	PL	PT-BR	PT-PT	RO	RU	SK	SL	SV	TR	UK	ZH
AR	1.0000	0.9977	1.0109	0.9994	1.0193	0.9888	1.0065	1.0510	1.0426	1.0140	1.0143	1.0185	1.0804	1.0241	1.0147	0.9759	1.0224	0.9643	1.0071	1.0388	1.0395	1.0082	1.0065	1.0454	1.0116	1.0230	1.0474	0.9978	0.9950	1.0397
BG	0.9946	1.0000	1.0200	1.0297	0.9724	0.9859	1.0006	1.0296	1.0475	1.0120	0.9873	1.0388	1.0497	1.0098	1.0162	0.9610	1.0330	0.9654	1.0028	1.0309	1.0063	0.9746	0.9864	1.0637	1.0206	1.0331	1.0274	0.9871	0.9978	1.0068
CS	0.9817	1.0265	1.0000	1.0213	1.0076	0.9836	1.0061	0.9969	1.0434	0.9992	1.0166	1.0388	1.0560	1.0234	1.0425	0.9844	1.0240	0.9577	0.9984	1.0325	1.0006	0.9653	0.9892	1.0557	1.0135	1.0265	1.0355	0.9813	0.9827	1.0042
DA	0.9814	1.0105	1.0026	1.0000	1.0208	1.0014	0.9680	1.0043	1.0327	1.0023	0.9870	1.0325	1.0484	1.0386	1.0553	0.9703	1.0179	0.9907	0.9954	1.0192	1.0098	0.9983	0.9861	1.0546	0.9953	1.0077	1.0201	0.9775	0.9926	1.0246
DE	0.9959	0.9906	0.9921	1.0232	1.0000	1.0094	0.9663	1.0095	1.0616	0.9687	0.9887	1.0305	1.0218	1.0022	1.0221	0.9607	1.0039	0.9722	0.9823	1.0242	1.0129	0.9757	0.9969	1.0164	0.9988	0.9983	0.9906	0.9897	0.9913	1.0021
EL	0.9891	0.9778	0.9918	0.9923	0.9713	1.0000	0.9897	1.0187	1.0392	0.9875	1.0122	1.0264	1.0442	1.0362	1.0252	0.9627	1.0163	0.9747	0.9887	1.0178	0.9995	0.9735	0.9978	1.0334	0.9944	1.0071	1.0162	0.9769	0.9638	1.0050
ES	0.9771	1.0071	0.9962	1.0048	1.0019	0.9607	1.0000	1.0103	1.0284	0.9959	0.9838	1.0156	1.0389	1.0399	1.0210	0.9823	1.0143	0.9734	0.9817	1.0223	1.0101	0.9862	0.9977	1.0294	0.9913	1.0244	0.9896	0.9772	0.9743	1.0247
ET	0.9850	1.0080	1.0028	1.0345	1.0064	1.0028	0.9745	1.0000	1.0072	0.9939	0.9816	0.9992	1.0324	1.0520	1.0526	0.9667	0.9973	0.9675	0.9808	1.0250	0.9982	0.9754	1.0042	1.0314	1.0011	0.9719	0.9859	1.0052	1.0062	1.0290
FI	0.9984	1.0523	1.0000	1.0254	1.0077	0.9663	0.9822	1.0068	1.0000	0.9994	1.0022	1.0374	1.0608	1.0230	1.0449	0.9585	1.0142	1.0016	0.9958	1.0121	1.0108	0.9971	0.9970	1.0485	1.0012	0.9887	1.0286	0.9729	1.0025	1.0378
FR	1.0014	1.0347	1.0249	1.0463	1.0184	1.0269	0.9717	1.0249	1.0704	1.0000	1.0256	1.0390	1.0568	1.0348	1.0291	0.9936	1.0314	1.0061	1.0029	1.0179	1.0144	0.9939	0.9931	1.0602	1.0107	1.0270	1.0379	0.9958	1.0046	1.0360
HU	0.9839	1.0216	1.0135	1.0171	0.9843	1.0062	0.9770	1.0320	1.0321	0.9949	1.0000	1.0315	1.0426	1.0438	1.0412	0.9641	1.0235	0.9774	0.9835	1.0431	1.0253	0.9950	0.9872	1.0618	1.0126	1.0148	1.0451	0.9869	0.9963	0.9858
ID	0.9653	0.9803	1.0103	0.9851	0.9974	0.9818	0.9807	0.9915	1.0402	1.0061	0.9876	1.0000	1.0489	1.0278	1.0144	0.9891	1.0403	0.9722	0.9899	1.0118	1.0034	0.9766	0.9811	0.9750	1.0126	1.0126	1.0190	0.9782	0.9440	1.0009
IT	0.9756	0.9871	1.0048	0.9836	0.9849	0.9589	0.9504	0.9963	1.0143	0.9652	0.9936	0.9960	1.0000	1.0375	1.0400	0.9924	1.0435	0.9646	0.9663	1.0406	0.9756	0.9489	0.9908	1.0279	1.0000	1.0284	0.9980	0.9676	0.9922	1.0329
JA	0.9937	1.0197	0.9973	0.9882	0.9892	0.9963	1.0002	1.0333	1.0589	0.9980	1.0032	1.0161	1.0546	1.0000	1.0415	0.9757	1.0201	0.9689	0.9920	1.0166	1.0052	0.9885	0.9933	1.0400	0.9947	0.9947	1.0030	0.9918	0.9976	0.9962
KO	0.9541	0.9991	1.0307	1.0359	1.0184	0.9706	1.0094	1.0388	1.0332	0.9906	0.9770	1.0112	1.0752	1.0228	1.0000	0.9763	1.0444	1.0065	0.9917	1.0322	1.0120	1.0001	0.9912	1.0221	1.0527	1.0283	1.0481	0.9767	0.9759	1.0197
LT	0.9794	1.0033	1.0157	1.0530	1.0164	0.9883	0.9876	1.0355	1.0577	0.9973	0.9871	1.0402	1.0295	1.0179	1.0529	1.0000	1.0283	0.9876	0.9802	1.0366	1.0167	1.0063	1.0037	1.0374	1.0423	1.0188	1.0257	0.9813	1.0099	0.9917
LV	0.9940	1.0199	1.0132	1.0235	0.9926	0.9742	0.9736	1.0067	1.0468	0.9924	1.0151	1.0274	1.0393	1.0297	1.0297	0.9675	1.0000	0.9890	0.9932	1.0274	1.0129	0.9914	0.9678	1.0433	1.0109	0.9977	0.9977	0.9727	0.9725	1.0358
NB	0.9902	1.0012	1.0122	1.0149	1.0174	0.9992	0.9844	1.0311	1.0481	0.9992	0.9856	1.0239	1.0626	1.0402	1.0339	1.0053	1.0060	1.0000	1.0053	1.0458	1.0374	1.0229	0.9881	1.0501	1.0209	1.0127	1.0260	0.9882	0.9896	1.0266
NL	0.9773	0.9988	1.0099	1.0136	0.9907	0.9839	0.9830	1.0222	1.0357	0.9732	1.0114	1.0187	1.0357	1.0284	1.0255	0.9842	1.0207	0.9568	1.0000	1.0130	1.0126	0.9800	0.9899	1.0323	1.0062	1.0199	1.0058	0.9710	0.9846	1.0237
PL	1.0091	1.0258	0.9978	1.0258	0.9666	1.0001	0.9790	1.0260	1.0258	0.9962	0.9986	1.0327	1.0271	1.0490	1.0528	0.9713	1.0266	0.9777	1.0002	1.0000	1.0232	0.9928	0.9962	1.0378	1.0099	0.9999	1.0096	0.9723	1.0011	0.9885
PT-BR	1.0106	1.0170	0.9890	0.9755	0.9970	0.9684	0.9690	0.9754	1.0088	0.9576	0.9773	1.0288	1.0449	1.0169	1.0286	0.9813	1.0460	0.9465	0.9759	1.0239	1.0000	0.9803	0.9908	1.0359	1.0128	1.0247	0.9900	0.9852	1.0063	1.0173
PT-PT	0.9972	1.0154	1.0256	1.0387	1.0220	1.0274	0.9985	1.0292	1.0502	0.9963	1.0195	1.0309	1.0484	1.0216	1.0386	0.9877	1.0327	0.9847	1.0046	1.0455	1.0253	1.0000	0.9971	1.0459	1.0049	1.0274	1.0073	1.0164	1.0129	1.0536
RO	0.9791	1.0199	1.0117	1.0106	0.9764	0.9832	0.9808	1.0173	1.0274	1.0061	0.9917	1.0009	1.0409	1.0404	1.0217	0.9579	1.0297	0.9683	0.9547	1.0267	1.0057	0.9816	1.0000	1.0273	1.0024	0.9974	1.0116	0.9833	1.0071	0.9956
RU	0.9918	1.0198	1.0089	0.9900	0.9309	0.9646	0.9398	0.9391	0.9552	0.9704	0.9439	0.9652	1.0404	1.0075	0.9804	0.9691	0.9946	0.9195	0.9411	1.0068	1.0024	0.9722	0.9497	1.0000	1.0137	0.9802	0.9997	0.9499	1.0216	1.0066
SK	0.9903	1.0134	1.0118	1.0085	0.9804	1.0103	0.9712	0.9948	0.9653	0.9953	0.9930	1.0096	1.0341	1.0141	1.0343	0.9687	1.0484	0.9528	0.9329	1.0247	0.9568	0.9299	0.9859	1.0376	1.0000	0.9887	0.9891	0.9568	0.9638	1.0120
SL	0.9822	1.0321	1.0153	1.0324	1.0051	0.9771	0.9885	0.9954	1.0455	0.9935	1.0115	1.0429	1.0423	1.0348	1.0345	0.9655	1.0008	0.9857	0.9960	1.0125	0.9998	0.9848	1.0035	1.0562	1.0144	1.0000	1.0204	0.9806	0.9879	1.0294
SV	0.9914	1.0058	1.0069	1.0162	0.9929	0.9852	0.9880	1.0063	1.0405	0.9895	0.9788	1.0438	1.0448	1.0210	1.0277	0.9882	1.0270	0.9728	1.0123	1.0192	1.0225	0.9836	1.0024	1.0641	1.0121	1.0213	1.0000	0.9871	1.0044	1.0132
TR	0.9762	0.9987	1.0125	1.0239	1.0052	0.9750	0.9835	1.0210	1.0312	1.0060	0.9830	1.0211	1.0181	1.0249	1.0213	0.9534	1.0085	0.9726	1.0164	1.0407	0.9947	0.9644	0.9777	1.0233	1.0139	0.9926	1.0075	1.0000	0.9831	1.0254
UK	0.9930	1.0365	1.0117	0.9558	0.9415	1.0168	0.9500	0.9861	0.9862	0.9279	0.9390	0.9788	1.0311	1.0029	0.9945	0.9776	1.0061	0.9128	0.9172	1.0309	0.9563	0.9247	0.9634	0.9971	1.0141	1.0166	0.9874	0.9808	1.0000	1.0502
ZH	0.9889	1.0049	1.0456	1.0344	1.0218	0.9806	1.0121	1.0564	1.0526	1.0187	1.0007	1.0464	1.0640	1.0236	1.0294	0.9779	1.0438	0.9765	1.0175	1.0482	1.0150	0.9954	1.0110	1.0212	1.0293	1.0035	1.0474	0.9752	0.9856	1.0000

Table 11: Similarity scores for 30×30 language pairs with an integration perspective.

L_f \ L_e	AR	BG	CS	DA	DE	EL	ES	ET	FI	FR	HU	ID	IT	JA	KO	LT	LV	NB	NL	PL	PT-BR	PT-PT	RO	RU	SK	SL	SV	TR	UK	ZH
AR	0.5203	0.2787	0.2074	0.1931	0.2104	0.2031	0.3353	0.5010	0.4207	0.2076	0.3766	0.3187	0.5168	0.2218	0.3489	0.1687	0.4432	0.1976	0.2769	0.3671	0.2825	0.2938	0.3900	0.5143	0.2553	0.3874	0.5467	0.2104	0.1468	0.3401
BG	0.2688	0.5203	0.2246	0.3285	0.1644	0.2825	0.3279	0.4152	0.4566	0.2795	0.3185	0.4131	0.3094	0.3104	0.4634	0.1532	0.4409	0.1659	0.2353	0.3115	0.2864	0.1396	0.2345	0.4985	0.4582	0.5313	0.4848	0.2573	0.1467	0.3439
CS	0.1455	0.2381	0.5203	0.4110	0.3992	0.0933	0.1544	0.2167	0.4812	0.3531	0.2215	0.4164	0.4925	0.3320	0.1938	0.1893	0.2921	0.1225	0.3524	0.4414	0.2838	0.2048	0.1680	0.3812	0.1216	0.2819	0.3345	0.2145	0.0000	0.2548
DA	0.1530	0.2746	0.3401	0.5203	0.4399	0.2323	0.1208	0.2063	0.5030	0.3367	0.2112	0.4610	0.5649	0.2398	0.2903	0.2038	0.3695	0.2109	0.3711	0.4511	0.4433	0.2814	0.2088	0.3934	0.1965	0.2962	0.3687	0.2385	0.0432	0.3573
DE	0.1642	0.2147	0.3339	0.4500	0.5203	0.1379	0.0910	0.1868	0.5618	0.2188	0.1670	0.4893	0.4440	0.1913	0.2258	0.1548	0.2692	0.1249	0.2741	0.4627	0.3856	0.2184	0.1804	0.2869	0.1684	0.2324	0.2317	0.2825	0.0262	0.2955
EL	0.2039	0.2513	0.1041	0.2085	0.0835	0.5203	0.2405	0.3362	0.3186	0.1264	0.3370	0.2533	0.1092	0.3223	0.4316	0.1091	0.2900	0.1559	0.1358	0.1940	0.1655	0.0870	0.1790	0.2725	0.2653	0.3833	0.3394	0.1779	0.0400	0.2331
ES	0.2300	0.3519	0.1380	0.2013	0.1514	0.1496	0.5203	0.3558	0.2965	0.2234	0.3216	0.3390	0.3072	0.3022	0.4488	0.2632	0.3190	0.2177	0.1623	0.2327	0.2306	0.1557	0.2920	0.4267	0.2933	0.4641	0.2986	0.1690	0.1565	0.3491
ET	0.2615	0.3384	0.2318	0.2744	0.1809	0.2858	0.2252	0.5203	0.2385	0.1903	0.2613	0.2660	0.1992	0.4700	0.5535	0.1476	0.2667	0.1990	0.1645	0.2621	0.1659	0.1843	0.2537	0.3491	0.2591	0.2305	0.2509	0.2633	0.1461	0.2899
FI	0.2775	0.4729	0.3215	0.4724	0.3583	0.1349	0.1780	0.2373	0.5203	0.3068	0.2413	0.5197	0.6556	0.2805	0.2957	0.1694	0.3332	0.2894	0.3748	0.4358	0.4276	0.3316	0.2873	0.4798	0.2339	0.2346	0.4292	0.2246	0.1103	0.4559
FR	0.1817	0.3436	0.4581	0.5024	0.4134	0.1932	0.1583	0.2632	0.5585	0.5203	0.3402	0.4641	0.5139	0.3772	0.2926	0.3092	0.2481	0.2026	0.3386	0.3668	0.3804	0.2618	0.1935	0.4217	0.2527	0.3343	0.3760	0.3185	0.1576	0.4634
HU	0.2638	0.4650	0.2147	0.2968	0.1572	0.3165	0.2910	0.4529	0.3222	0.2477	0.5203	0.4102	0.3171	0.3404	0.5886	0.2074	0.4098	0.2232	0.2015	0.3083	0.2963	0.1670	0.2485	0.4719	0.3990	0.4654	0.5161	0.2668	0.2091	0.2459
ID	0.1583	0.2211	0.3223	0.2784	0.3517	0.1537	0.2237	0.2423	0.5315	0.3432	0.2583	0.5203	0.5422	0.3372	0.3143	0.3312	0.3576	0.1802	0.2911	0.3880	0.3608	0.2586	0.1695	0.2466	0.2586	0.2850	0.3499	0.2759	0.0485	0.3540
IT	0.1936	0.1678	0.3165	0.2948	0.2936	0.0382	0.0915	0.1391	0.4496	0.1871	0.1975	0.3314	0.5203	0.1530	0.2216	0.2898	0.4065	0.1302	0.2380	0.5687	0.2514	0.1247	0.2878	0.3691	0.1852	0.2923	0.2892	0.1682	0.0654	0.3566
JA	0.1612	0.3422	0.2537	0.1437	0.1634	0.2179	0.2077	0.4068	0.3721	0.2633	0.2347	0.3030	0.1738	0.5203	0.3552	0.1762	0.2124	0.1434	0.1482	0.2126	0.2045	0.2575	0.0742	0.3121	0.2125	0.2320	0.2518	0.2362	0.1213	0.2774
KO	0.1412	0.3868	0.1762	0.2508	0.2181	0.2219	0.4008	0.4987	0.2695	0.1957	0.2943	0.3045	0.2830	0.3036	0.5203	0.2260	0.4057	0.2383	0.1543	0.2651	0.1964	0.1702	0.1918	0.3385	0.5231	0.5024	0.4579	0.2293	0.1103	0.3828
LT	0.1780	0.2878	0.2720	0.4969	0.3350	0.1653	0.2833	0.3362	0.5370	0.3234	0.3007	0.5531	0.4251	0.2942	0.4984	0.5203	0.2843	0.1932	0.2332	0.4246	0.3644	0.2985	0.2268	0.3960	0.3927	0.3526	0.3157	0.2859	0.1921	0.3346
LV	0.3286	0.3927	0.2638	0.3892	0.2360	0.1721	0.1909	0.2963	0.4404	0.1664	0.3787	0.3216	0.3934	0.2314	0.3602	0.1346	0.5203	0.2765	0.2798	0.3855	0.2997	0.1957	0.2167	0.3844	0.2833	0.3728	0.3726	0.1542	0.0301	0.2361
NB	0.2969	0.2787	0.2713	0.2765	0.2200	0.2177	0.2574	0.4302	0.4408	0.1875	0.2513	0.3373	0.3768	0.3257	0.3080	0.2387	0.3382	0.5203	0.2325	0.3118	0.3199	0.3771	0.2673	0.4423	0.2459	0.3110	0.4013	0.1643	0.0939	0.2599
NL	0.1894	0.2248	0.3997	0.4530	0.3071	0.1273	0.1653	0.2648	0.5547	0.2267	0.2810	0.3998	0.5472	0.2164	0.2203	0.2470	0.3755	0.1105	0.5203	0.4579	0.4308	0.2365	0.2426	0.3860	0.1935	0.3396	0.3494	0.2150	0.0523	0.3378
PL	0.2807	0.2982	0.3134	0.4790	0.2235	0.1614	0.1407	0.2645	0.4959	0.2909	0.2030	0.4703	0.5086	0.2790	0.3110	0.2006	0.3829	0.1521	0.3994	0.5203	0.4569	0.2743	0.2607	0.4470	0.1921	0.2704	0.3516	0.2078	0.0881	0.2445
PT-BR	0.2182	0.3193	0.2462	0.2866	0.3243	0.1061	0.1332	0.1217	0.4187	0.1676	0.1700	0.4618	0.5285	0.2306	0.2293	0.2403	0.3987	0.0866	0.2723	0.4597	0.5203	0.2306	0.2173	0.4246	0.2892	0.2906	0.2752	0.2289	0.1187	0.4037
PT-PT	0.2599	0.2320	0.4404	0.4143	0.3789	0.1642	0.1818	0.3430	0.5405	0.2694	0.2166	0.4789	0.4686	0.3702	0.2453	0.2391	0.2984	0.2462	0.3231	0.4554	0.3845	0.5203	0.1823	0.4374	0.1032	0.1998	0.2484	0.3460	0.0611	0.4829
RO	0.2761	0.3410	0.2175	0.2786	0.1365	0.1483	0.2341	0.2900	0.3913	0.2247	0.2638	0.2188	0.4727	0.1177	0.2667	0.1126	0.4613	0.1969	0.1319	0.3559	0.2588	0.1487	0.5203	0.3377	0.2507	0.2848	0.4367	0.0964	0.1143	0.1480
RU	0.3064	0.3533	0.2562	0.2141	0.0499	0.1170	0.0986	0.0772	0.1500	0.1648	0.0984	0.2077	0.4090	0.2324	0.2093	0.1787	0.2409	0.0214	0.0913	0.3361	0.3090	0.1950	0.1087	0.5203	0.2621	0.2034	0.3278	0.1232	0.3415	0.3505
SK	0.2002	0.4279	0.1195	0.2280	0.1324	0.3171	0.2181	0.2408	0.1387	0.2130	0.3203	0.2509	0.2577	0.2630	0.4541	0.1694	0.3918	0.0910	0.0415	0.2214	0.1272	0.0191	0.2045	0.3240	0.5203	0.3114	0.2670	0.1098	0.1196	0.2755
SL	0.2437	0.5265	0.2543	0.3733	0.2511	0.2615	0.3148	0.3223	0.4106	0.2362	0.4503	0.3717	0.3273	0.3459	0.5295	0.1752	0.3864	0.2268	0.2640	0.3094	0.2260	0.1254	0.3056	0.4324	0.4172	0.5203	0.4652	0.2314	0.1298	0.3452
SV	0.3191	0.3934	0.2565	0.3555	0.2384	0.2377	0.2924	0.3207	0.4736	0.2304	0.2575	0.4317	0.4480	0.3021	0.3889	0.2109	0.5013	0.2125	0.3738	0.3867	0.3911	0.1861	0.3962	0.5711	0.3477	0.4691	0.5203	0.2012	0.1731	0.2811
TR	0.1569	0.2976	0.3121	0.4098	0.3396	0.1726	0.1854	0.3081	0.4428	0.3576	0.2528	0.4570	0.3233	0.3352	0.3894	0.1693	0.2466	0.1302	0.3830	0.4465	0.2591	0.1722	0.0887	0.3672	0.2524	0.2722	0.2563	0.5203	0.1186	0.4712
UK	0.1431	0.2136	0.0000	0.0232	0.0096	0.0816	0.0946	0.1147	0.0903	0.0259	0.0655	0.1026	0.0974	0.1288	0.1429	0.1283	0.0470	0.0000	0.0032	0.1178	0.0552	0.0072	0.0613	0.2647	0.2376	0.1795	0.1410	0.1148	0.5203	0.3291
ZH	0.2040	0.3370	0.3701	0.3885	0.3607	0.1714	0.3096	0.3585	0.5097	0.3981	0.2959	0.5369	0.4490	0.3685	0.4176	0.2758	0.2515	0.1454	0.3190	0.4372	0.3949	0.2832	0.1755	0.4051	0.3235	0.2687	0.3768	0.2610	0.1744	0.5203

Llama-3B																						
		AR	CS	DE	EL	ES	FR	ID	IT	JA	KO	LT	NL	PL	PT	RO	RU	SV	TR	UK	ZH
Traditional Language Similarity	Acc.	-	0.383	0.406	-	**0.448**	0.446	-	0.439	-	-	-	0.416	0.392	0.447	0.418	0.413	-	-	**0.402**	-
	Selected L_f	-	PL	NL	-	IT, FR, RO, PT	IT, ES, RO, PT	-	FR, ES, RO, PT	-	-	-	DE	CS	FR,IT, ES, RO	FR, IT, ES, PT	UK	-	-	RU	-
LLM-Based Language Similarity (feature)	Acc.	0.352	0.403	**0.453**	0.363	0.443	**0.447**	**0.455**	**0.471**	0.388	0.385	**0.353**	**0.444**	**0.425**	0.449	**0.450**	**0.430**	0.410	**0.401**	**0.402**	0.411
	Selected L_f	RO	DE	CS	KO	KO	CS	LT	NL	PT	ES	ID	PL	NL	ID	SV	ID	RO	ID	RU	LT
LLM-Based Language Similarity (transferability)	Acc.	0.358	0.392	0.438	0.365	0.439	0.430	0.424	0.438	0.373	**0.394**	0.333	0.430	0.399	0.448	0.420	0.418	0.391	0.387	**0.402**	**0.422**
	Selected L_f	PT	ZH	PT	PT	ZH	ZH	ZH	AR	PL	LT	FR	ZH	ZH	AR	ZH	SV	KO	PT	RU	PT
LLM-Based Language Similarity (integration)	Acc.	**0.359**	**0.410**	0.426	**0.372**	0.443	0.430	**0.455**	**0.471**	**0.393**	**0.394**	**0.353**	**0.444**	0.413	**0.460**	**0.450**	0.418	**0.419**	0.387	**0.402**	**0.422**
	Selected L_f	SV	FR	FR	SV	KO	ZH	LT	NL	FR	LT	ID	PL	IT	LT	SV	SV	AR	PT	RU	PT
Reference: Direct Fine-tuning	Acc.	0.358	0.403	0.439	0.374	0.444	0.435	0.438	0.453	0.368	0.369	0.358	0.447	0.410	0.437	0.416	0.429	0.418	0.377	0.396	0.390

Qwen-4B																					
		AR	CS	DE	EL	ES	FR	ID	IT	JA	KO	LT	NL	PL	PT	RO	RU	SV	TR	UK	ZH
Traditional Language Similarity	Acc.	-	0.585	0.617	-	**0.660**	0.647	-	0.657	-	-	-	0.614	0.591	0.650	0.611	**0.633**	-	-	**0.600**	-
	Selected L_f	-	PL	NL	-	IT, FR, RO, PT	IT, ES, RO, PT	-	FR, ES, RO, PT	-	-	-	DE	CS	FR,IT, ES, RO	FR, IT, ES, PT	UK	-	-	RU	-
LLM-Based Language Similarity (feature)	Acc.	0.550	0.600	0.630	0.549	0.653	**0.659**	0.632	**0.659**	0.594	**0.591**	**0.534**	**0.627**	0.605	0.653	**0.620**	0.609	0.604	0.547	**0.600**	0.649
	Selected L_f	RO	DE	CS	KO	KO	CS	LT	NL	PT	ES	ID	PL	NL	ID	SV	ID	RO	ID	RU	LT
LLM-Based Language Similarity (transferability)	Acc.	0.570	**0.610**	0.629	0.554	0.645	0.657	**0.633**	0.649	0.598	0.584	0.517	0.626	0.608	0.663	0.620	0.627	0.605	**0.557**	**0.600**	**0.657**
	Selected L_f	PT	ZH	PT	PT	ZH	ZH	ZH	AR	PL	LT	FR	ZH	ZH	AR	ZH	SV	KO	PT	RU	PT
LLM-Based Language Similarity (integration)	Acc.	**0.573**	0.603	**0.639**	**0.557**	0.653	0.657	0.632	**0.659**	**0.613**	0.584	**0.534**	**0.627**	**0.613**	**0.664**	**0.620**	0.627	**0.614**	**0.557**	**0.600**	**0.657**
	Selected L_f	SV	FR	FR	SV	KO	ZH	LT	NL	FR	LT	ID	PL	IT	LT	SV	SV	AR	PT	RU	PT
Reference: Direct Fine-tuning	Acc.	0.568	0.597	0.641	0.539	0.661	0.664	0.627	0.656	0.602	0.593	0.533	0.629	0.588	0.659	0.624	0.619	0.594	0.552	0.600	0.663

Gemma-4B																					
		AR	CS	DE	EL	ES	FR	ID	IT	JA	KO	LT	NL	PL	PT	RO	RU	SV	TR	UK	ZH
Traditional Language Similarity	Acc.	-	0.446	0.482	-	0.506	0.496	-	0.493	-	-	-	0.494	0.463	0.492	0.472	0.454	-	-	**0.453**	-
	Selected L_f	-	PL	NL	-	IT, FR, RO, PT	IT, ES, RO, PT	-	FR, ES, RO, PT	-	-	-	DE	CS	FR,IT, ES, RO	FR, IT, ES, PT	UK	-	-	RU	-
LLM-Based Language Similarity (feature)	Acc.	0.443	0.464	0.491	0.412	0.479	0.489	0.411	**0.505**	0.438	**0.443**	**0.440**	0.482	0.464	0.498	0.455	**0.479**	0.470	0.443	**0.453**	0.392
	Selected L_f	RO	DE	CS	KO	KO	CS	LT	NL	PT	ES	ID	PL	NL	ID	SV	ID	RO	ID	RU	LT
LLM-Based Language Similarity (transferability)	Acc.	0.443	**0.473**	0.488	0.446	**0.514**	**0.496**	**0.482**	0.497	0.429	0.392	0.420	**0.505**	0.478	**0.511**	**0.481**	0.463	0.439	**0.459**	**0.453**	**0.486**
	Selected L_f	PT	ZH	PT	PT	ZH	ZH	ZH	AR	PL	LT	FR	ZH	ZH	AR	ZH	SV	KO	PT	RU	PT
LLM-Based Language Similarity (integration)	Acc.	**0.450**	0.459	**0.501**	**0.460**	0.479	**0.496**	0.411	**0.505**	**0.453**	0.392	**0.440**	0.482	**0.483**	0.442	0.455	0.463	**0.485**	**0.459**	**0.453**	**0.486**
	Selected L_f	SV	FR	FR	SV	KO	ZH	LT	NL	FR	LT	ID	PL	IT	LT	SV	SV	AR	PT	RU	PT
Reference: Direct Fine-tuning	Acc.	0.450	0.484	0.493	0.455	0.502	0.485	0.486	0.486	0.460	0.422	0.377	0.483	0.467	0.505	0.488	0.456	0.455	0.437	0.429	0.478

Table 12: The validation results of four fine-tuning language selection strategies, conducted on the Global MMLU dataset using Llama-3B, Qwen-4B, and Gemma-4B. Each column represents a target language. The reported values are average task accuracies (Acc.) on the target language after fine-tuning the model with the selected language. "-" means there is no candidate language for fine-tuning within these 20 languages. **The integrated similarity measure** (in green) **performs**, providing **comparable performance with direct fine-tuning** (in gray).