LLM-Based Language Similarity

Xinhe Shi, Linchao Zhu
CCAI, Zhejiang University
{xinheshi, zhulinchao}@zju.edu.cn

Abstract

It's noticed that researchers pre-training multilingual models or designing multilingual benchmarks often list the language families of selected languages to prove their large linguistic coverage. However, *does the traditional language family taxonomies based on structural or historical relationships work on Large Language Models (LLMs)*? In this paper, we investigate the LLM-based language similarity, answering which languages are considered similar by LLMs and which languages researchers should choose to maximize linguistic coverage with limited resources. To this end, we propose a dual-methodological framework to visualize LLM-based language families and quantitatively evaluate the similarity within each language pair from the perspective of LLMs, exploring the cross-lingual transferability. Additionally, due to the lack of a dataset meeting our experimental requirements, we construct and make available a Massive Multilingual Multitask Multiple Choice Question (**M4CQ**) dataset, which features content consistency across **31** languages and domain-balanced distribution of **135** tasks. Results show partial alignment with traditional linguistic typology (e.g., Slavic language clustering) alongside novel LLM-specific patterns (e.g., Chinese-Greek affinity) and super-additive transfer effects where fine-tuning on a specific language provides a more effective "bridge" for knowledge transfer than direct fine-tuning on the target language.

1 Introduction

The rapid development of multilingual LLMs has intensified the need for principled language selection strategies. When pre-training multilingual models or designing cross-lingual benchmarks, practitioners often select languages with little explanation or rely on traditional language family taxonomies to demonstrate linguistic diversity (Siddhant et al., 2019; K et al., 2019; Vázquez et al., 2020; Oncevay et al., 2020; Zhang et al., 2020; Kassner et al., 2021; Srinivasan et al., 2021; Chau and Smith, 2021; Sun et al., 2024). However, based on synchronic (focusing on structural features like syntax/morphology) and diachronic (focusing on historical relationships) classifications, these traditional typological frameworks (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 align with how LLMs perceive language similarity. 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 paper, we systematically investigate language similarity perceived by LLMs, exploring the cross-lingual transferability for each language pair. To achieve this, we propose **LLM-based language features** and **quantitative similarity metrics for a language pair**.

Our work provides actionable insights for researchers to: (I) re-examine traditional typology from LLMs' perspective; (II) Select linguistically diverse languages under budget constraints; (III) Identify high-transfer language pairs for efficient cross-lingual adaptation.

Additionally, due to the lack of datasets meeting our experimental requirements (mentioned in Section 2.4), we constructed the **M4CQ** dataset, which has balanced task distribution and strict content consistency across languages, with rigorous proofreading.

2 Methodology

2.1 Method Design Rationale

The primary motivation of this study is to provide multilingual LLM researchers with a sound language selection strategy, enabling them to select a relatively small set of languages (when constrained by budget or computational resources) while maximizing linguistic coverage–specifically by choosing "dissimilar" languages identified by our study. Therefore, we ground our investigation of language similarity in LLM task performance rather than other properties from LLMs (e.g., hidden representations within LLMs), considering that task performance directly reflects LLMs' practical language processing abilities–aligning with our application-oriented objectives.

For instance, in multilingual benchmark design scenarios, designers tend to select dissimilar languages to evaluate the performance of LLMs more comprehensively, that is, although the number of languages covered in the benchmark is limited, we can roughly extrapolate the performance of a model on many uncovered languages once they are similar to one of the covered languages. Hence, the "similar languages" defined in our work are languages on which LLM performance patterns are similar (as operationalized in Section 2.2). For multilingual LLM pre-training scenarios, researchers want their models to perform well on as many languages as possible. Thus, we also explore "similar languages" demonstrating high cross-lingual knowledge transfer efficiency (as formalized in Section 2.3) so that researchers can pre-train with the focus on dissimilar languages. Specifically, if languages A and B exhibit high transfer efficiency, they will be identified as similar in our framework. Consequently, researchers referring to our findings may prioritize only one of them during pre-training, as the resulting model can still perform well on the other language through efficient knowledge transfer. Additionally, to holistically capture model behavioral patterns, we require a multitask benchmark with a comprehensive and balanced task distribution.

2.2 Clustering of LLM-derived Language Features

We derive LLM-based language features from LLMs' 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. 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 N is the number of task categories and $v_{e,j} \in [0, 1]$ represents the LLM'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 (e.g., cosine similarity) reveal LLM-perceived language similarity in task performance distributions.

Clustering these linguistic feature vectors (implementation details discussed in Section 4.1.3) yields **LLM-based language family classification** based on task capability patterns. For resource-constrained multilingual NLP research, whether in model pre-training or benchmark design, researchers should select one representative per cluster at least.

2.3 Quantitative Similarity Metrics for a Language Pair

Besides language family analysis from the perspective of LLM multitask performance patterns, we propose two complementary metrics to quantify similarity between a language pair, specifically, based on **cross-lingual transfer efficiency**. The core hypothesis is: the greater the performance improvement on target language L_e achieved by fine-tuning on language L_f, the higher their similarity in LLM representation space. This approach also requires a multitask dataset comprising parallel versions in multiple languages.

For a language pair (L_f, L_e), the similarity quantification proceeds as:

1. Fine-tuning: 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

3. Baseline acquisition: Evaluate M_e on L_e's N tasks, obtaining baseline accuracies $\{\text{acc}_{e,e,j}\}_{j=1}^{N}$
4. Metric computation: Select similarity metric adaptively based on baseline performance and compute it

Metric Design We propose two complementary metrics to address different baseline performance scenarios:

- **Global Aggregation Metric (Metric-a)** First calculate the averages of cross-lingual transfer performance and baseline performance respectively, then divide and convert to percentages:

$$S_{f,e} = \frac{\frac{1}{N}\sum_{j=1}^{N} \text{acc}_{f,e,j}}{\frac{1}{N}\sum_{j=1}^{N} \text{acc}_{e,e,j}} \times 100\%$$

  Robust against low baseline performance ($\exists j$: $\text{acc}_{e,e,j} < 40\%$) by suppressing extreme value impacts.
- **Task-level Aggregation Metric (Metric-b)** Calculate the task-level similarity scores first, then average them:

$$S_{f,e} = \frac{1}{N}\sum_{j=1}^{N} \underbrace{\frac{\text{acc}_{f,e,j}}{\text{acc}_{e,e,j}}}_{\text{Similarity score on task } j \text{ level}} \times 100\%$$

  More sensitive to differences in multitask transfer patterns. Suitable when there is no low-performance baseline in the denominator ($\forall j : \text{acc}_{e,e,j} \geq 40\%$).

Intuitively, we posit that direct fine-tuning on the target language L_e should yield optimal performance, thus establishing the L_e-fine-tuned model's accuracy ($\text{acc}_{e,e,j}$) as our reference standard. In real-world scenarios, if we don't want to involve L_e in fine-tuning but still hope a good performance on L_e, we may fine-tune on "similar languages" (e.g., L_f) whose resulting performance ($\text{acc}_{f,e,j}$) can approach the reference standard ($\text{acc}_{e,e,j}$). Therefore, our metrics quantify language similarity through this performance ratio and higher values indicate greater similarity (with L_e to itself being 100%).

Metric Selection Rationale The necessity of dual-metric design stems from different baseline conditions. When baseline model M_e exhibits low-performance tasks on L_e, Metric-b might become vulnerable: the denominator $\text{acc}_{e,e,j}$ in task-level ratios may produce extremely large values (e.g., $\text{acc}_{f,e,j}/\text{acc}_{e,e,j} \to \infty$ as $\text{acc}_{e,e,j} \to 0$), causing similarity scores to be skewed by task-level outliers. Metric-a addresses this by averaging task-level accuracies before computing ratios, effectively suppressing outlier dominance. Conversely, when baseline performance remains stable ($\forall j : \text{acc}_{e,e,j} \geq 40\%$), Metric-b's task-level granularity better captures fine-grained cross-lingual transfer patterns.

Higher scores indicate stronger transferability and thus greater LLM-perceived similarity. This adaptive mechanism ensures robustness against diverse baseline distributions.

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.** The base model M_{base} should be pre-trained exclusively on one language to minimize the influence of irrelevant variables as much as possible. Although M_{base}'s performance on L_e will still be impacted by the pre-training language L_p in this situation, the impact manifests as varying degrees of enhancement depending on how similar L_e is to L_p (where higher similarity results in greater enhancement). Therefore, the performance of M_{base} on various languages will naturally further diverge based on the degree of similarity between each language and L_p, which will not hinder the analysis of language similarity. However, even with cross-lingual content consistency, using multiple but partial languages (e.g., L_A and L_B) during pre-training would bias the model's performance towards these specific languages. For instance, M_{base} might show nearly identical accuracy on three evaluation languages: X, which is highly similar to L_A but very dissimilar to L_B; Y, which is highly similar to L_B but very dissimilar to L_A; and Z, which is moderately similar to both L_A and L_B. And X, Y, Z will be considered similar although they might not be. This performance variability, simultaneously influenced by the similarity of multiple language pairs, would make it difficult to achieve reliable clustering results and similarity scores. The same applies to pre-training on all target languages (i.e., the languages present in the multilingual multitask dataset) without cross-lingual content consistency. ***Note***: For our experiments, the ideal model would be one that has been pre-trained on corpora which contain

exact all target languages with completely semantically equivalence across languages. If such a model were available, our experiments could be conducted without introducing any irrelevant variable. 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) **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. (III) **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. (IV) **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. (V) **Moderate Difficulty.** Problems should not be too difficult, since our aim is to measure language capabilities rather than solve highly complex problems. (VI) **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. (VII) **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.

3 The M4CQ Dataset

Given the lack of an existing open-source dataset that meets our requirements (outlined in Section 2.4), we constructed the Massive Multilingual Multitask Multiple Choice Question (**M4CQ**) dataset.

3.1 Dataset Overview

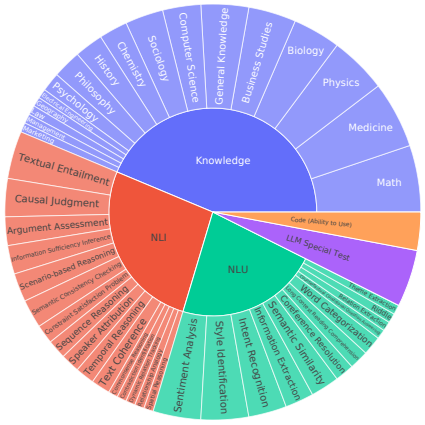

Figure 1: Domain distribution of 135 tasks in M4CQ, where each domain's angular span represents the number of its constituent tasks.

The M4CQ dataset consists of **31 languages**, each containing **135 task categories** in multiple-choice format. Figure 1 shows M4CQ's task distribution, demonstrating a balanced spread across various domains. See Appendix A for detailed information about these 135 tasks. Table 1 provides a complete list of languages covered in M4CQ. Each instance has semantically equivalent versions in all languages (see Figure 2 for examples in M4CQ).

3.2 Construction Methodology

3.2.1 Task Curation

1. Source Selection: We aggregated tasks from 8 established datasets: GLUE (Wang et al., 2018), OpenBookQA (Mihaylov et al., 2018), HellaSwag (Zellers et al., 2019), SuperGLUE (Wang et al., 2019), PiQA (Bisk et al., 2020), MMLU (Hendrycks et al., 2020), Winogrande (Sakaguchi et al., 2021) and BIG-bench (Srivastava et al., 2022). 2. Task Filtering: a) Removed language-biased tasks like code reasoning(biased towards English, but pseudo-code is acceptable), English proverbs, U.S. history questions. b) Excluded low-quality tasks or instances where answer selection of questions is unconvincing. 3. Task Merging: Consolidated tasks of the same category across sources

Figure 2: Six examples from the M4CQ dataset demonstrating cross-lingual content consistency. The first and second choices, though semantically similar, are accurately translated, which also reflects the dataset's quality.

Table 1: Language Codes and Full Names in M4CQ Dataset

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

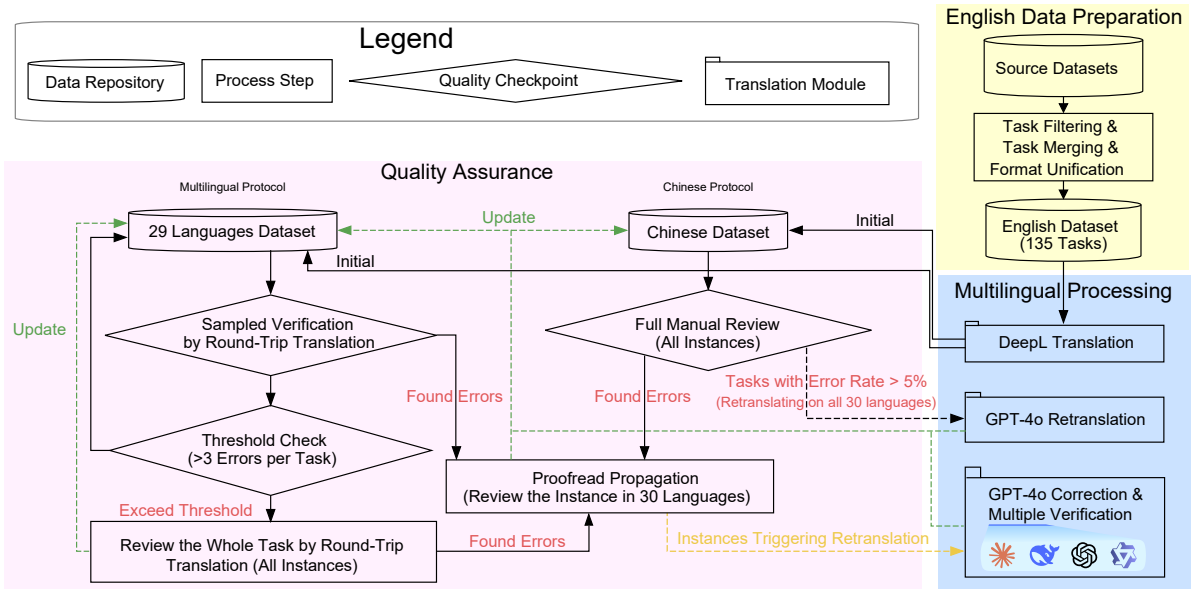

Figure 3: The construction workflow of M4CQ. It begins with the creation and multi-step processing of an English dataset, which is then translated into 30 languages using DeepL, a translation engine standing out in our pre-experiment. To further ensure translation quality, all Chinese instances are manually verified by annotators proficient in both English and Chinese, while the remaining 29 languages undergo rigorous round-trip translation (en→xx→zh→en) checks. Any inconsistencies detected during this process trigger GPT-4o retranslation, followed by multi-verification using Claude 3.5 Sonnet, DeepSeek-R1, GPT-4o, and Qwen2.5-32B to guarantee final accuracy.

(e.g., combining BIG-bench's *presuppositions as nli* with GLUE's *MNLI*).

3.2.2 Format Unification

All tasks were converted to a unified Parquet format with the following schema:

- **`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).

3.2.3 Cross-Lingual Alignment and Quality Control

Translation Pipeline and Validation for Chinese 1. Select Translation Engines: We conducted a preliminary experiment to select suitable translation engines, where DeepL showed great accuracy and stability, contributing to its role as primary engine. Appendix B provides details about the pre-experiment. 2. Initial Translation: All tasks were translated from English to 30 target languages using DeepL at the instance level (question + choices). 3. Full Verification for Chinese: Every instance of the 135 tasks in Chinese was manually reviewed and corrected by native Chinese speakers proficient in English. If a Chinese instance contains errors (e.g., mistranslated technical terms), similar issues may also occur in corresponding instances in other languages. Therefore, for each identified erroneous instance, we perform Proofread Propagation——reviewing its translations in all 29 other languages using Claude 3.5 Sonnet, DeepSeek-R1, GPT-4o, and Qwen2.5-32B in collaboration. 4. Identification of DeepL Limitations: During the verification process above, we observed that DeepL underperformed on 9 tasks (e.g., college computer science, abstract algebra) with error rate >5%. These tasks were re-translated into all 30 languages using GPT-4o and underwent the same manual review process as above.

Sampled Validation for Non-Chinese Languages 1. Round-Trip Translation:

For each non-Chinese target language L_x: a) Randomly sampled 20% instances per task. b) Performed round-trip translation: $L_x \rightarrow$ zh. c) Compared the result with the verified en→zh translation. d) Flagged instances where semantic in-

consistency occurred between en→L_x→zh and en→zh. e) Corrected flagged instances under multi-verification (see Figure 3) and reviewed corresponding instances in other 28 languages.

2. Error Thresholding: If any task in L_x had >3 flagged instances, we conducted full manual review of all instances in that task of language L_x.

Note: Due to the significantly higher availability of high-caliber Chinese–English bilingual experts for us compared to those in other language pairs, we exclusively employed proofreaders proficient in English and Chinese. Therefore, Chinese was treated differently from other target languages in the aforementioned process, and the latter were proofreaded using round-trip translation (the pipeline is effectively equivalent to en→L_x→zh→en, which is more robust compared to en→L_x→en with one more middle language). Although bilinguals for other target languages were not involved, our rigorous proofreading mechanism with multi-verification (see Figure 3) ensured the quality across all target languages.

3.2.4 Workflow Summary

The entire dataset construction process is illustrated in Figure 3. This workflow ensures M4CQ's adherence to experimental requirements while maintaining high quality across all language variants.

3.3 Potential Applications

The M4CQ dataset's unique design enables broad applications beyond our research. Here we highlight some directions. **Multilingual Model Benchmarking**: With balanced task distribution and strict semantic equivalence across languages, M4CQ can serve as a benchmark for evaluating Multilingual Language Models (MLMs), allowing direct comparison of model performance across different languages. **Multitask Learning Optimization**: The 135 task categories spanning diverse domains (Figure 1) enable research on multitask learning optimization, including task weighting strategies and gradient conflict resolution, while supporting analysis of inter-task correlations and their impact on model performance. **Under-Resourced Language Understanding**: M4CQ provides high-quality, translation-verified data for 20+ medium- and low-resource languages (e.g., Ukrainian, Hungarian), enabling research on language-specific pre-training and data augmentation (e.g., back-translation, synthetic data generation) for under-resourced languages.

These applications highlight M4CQ's value as a versatile resource for multilingual NLP research. Its strict cross-lingual alignment and task neutrality uniquely support disentangling linguistic capabilities from task-specific biases.

4 Experiments and Results

See Appendix C for step-by-step experimental instructions.

4.1 Experimental Setup

4.1.1 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.

4.1.2 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.

4.1.3 Clustering Methodology

Given that our clustering objects are 31 vectors of 135 dimensions, high-dimensional data poses the risk of the "curse of dimensionality," where data becomes excessively sparse in the feature space (Bishop, 2006). This sparsity renders traditional distance-based methods (e.g., K-Means (MacQueen, 1967)) ineffective, as distances between samples tend to become uniformly similar

(Hastie, 2009). Additionally, the high-dimensional sparsity also implies that meaningful clustering results necessitate significantly more samples, while we only have feature vectors of 31 languages.

To address these challenges, we performed dimensionality reduction before clustering.

To capture potential nonlinear relationships within data and better preserve global data structures, we employed **UMAP** (Uniform Manifold Approximation and Projection) (Healy and McInnes, 2024) as the dimensionality reduction technique instead of PCA (Principal Component Analysis, which captures only linear structures within data) (F.R.S., 1901; Hotelling, 1933) or t-SNE (t-Distributed Stochastic Neighbor Embedding, which primarily preserves local structures) (van der Maaten and Hinton, 2008).

Given the potential complexity of clusters after dimensionality reduction, **spectral clustering** (Ng et al., 2001) was chosen due to its ability to handle non-convex clusters effectively.

Then two critical parameters required optimization: the target dimensionality for UMAP reduction and the number of clusters for clustering. We systematically explored all combinations of UMAP dimensions (3–29) and cluster counts (5-13), using the **Silhouette Score** (Rousseeuw, 1987) to evaluate both within-cluster cohesion and between-cluster separation. Figure 4 illustrates the top three Silhouette Scores for each cluster count across all tested dimension numbers. Based on the evaluation result, the optimal configuration is determined: reducing the 135-dimensional data to **18** dimensions using UMAP and applying spectral clustering with **6** clusters.

4.1.4 Fine-tuning Settings

See Appendix D for LoRA (Hu et al., 2021) fine-tuning settings.

4.2 Results and Analysis

4.2.1 Experiment 1: Clustering of LLM-derived Language Features

Figure 5 visualizes the clustering result through 2D t-SNE projection, providing a more intuitive perception of LLM-based language family division.

Notably, while some clusters align with established language families, others suggest novel groupings based on LLM performance patterns across languages. For instance:

Consistent Groupings

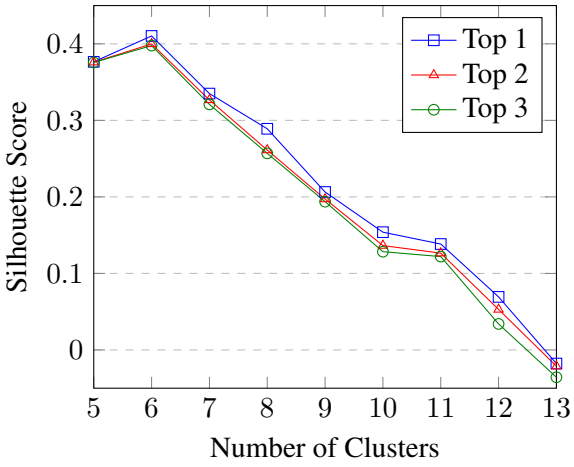

Figure 4: Top 3 Silhouette Scores for each cluster count across all tested dimension numbers. Maximum Silhouette Score occurs when the number of clusters is 6 and the target dimensionality is 18.

- Polish (PL), Czech (CS), and Slovak (SK) cluster together, matching their classification as West Slavic branch.
- Danish (DA), Norwegian (NB), Swedish (SV), and Dutch (NL) cluster together, matching their classification as Germanic group. Danish, Norwegian and Swedish are of the North Germanic branch and Dutch is of the West Germanic branch.
- Lithuanian (LT) and Latvian (LV) cluster together, consistent with their shared Baltic language group.
- Finish (FI) and Hungarian (HU) cluster together, matching their classification as Uralic language family.
- Spanish (ES), Portuguese (PT-PT & PT-BR), and Romanian (RO) cluster together, consistent with their shared Romance language group.

Divergent Groupings

- English (EN) and French (FR) are separated into different clusters, despite their historical and lexical connections.
- Chinese (ZH) and Greek (EL) are grouped together, a pairing not supported by traditional linguistic classification.

The clusters shown in Figure 5 can be interpreted as emergent linguistic families identified by the model, offering a new perspective on traditional classifications.

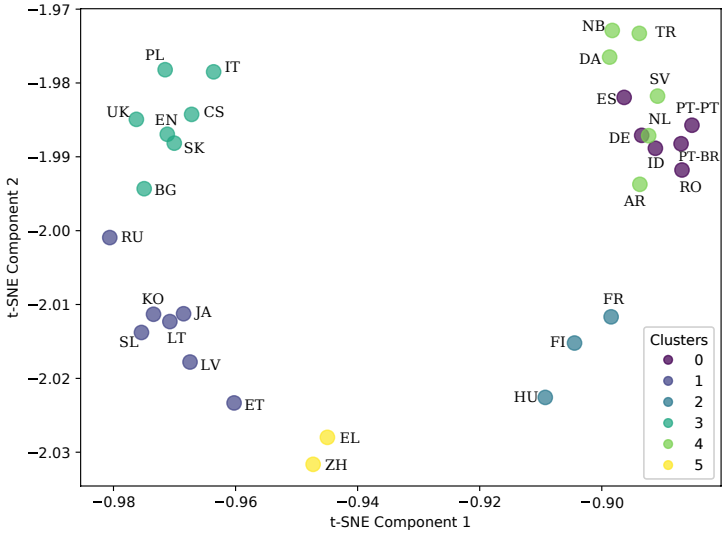

Figure 5: 2D t-SNE visualization of language clusters. Thirty-one languages are clustered based on LLM-derived feature vectors, with each color representing a cluster. For researchers aiming to maximize language coverage under budget constraints, it's recommended to include at least one language from each cluster to ensure broad linguistic diversity.

4.2.2 Experiment 2: Cross-Language Fine-tuning Transferability

To quantify similarity within language pairs, we computed the average similarity score $S_{f,e}$ for each pair of fine-tuning language L_f and evaluation language L_e, using Metric-a defined in Section 2.3. The results are presented in Table 5, Table 6 and Table 7, where higher values indicate stronger cross-lingual transferability, hence higher similarity. See Table 1 for a cross-reference of the language's full name and abbreviation codes.

Given a language of interest, just find the row or column corresponding to the language and observe it to get the correlation between the two languages from the perspective of LLM.

Additionally, there are some counter-intuitive findings. **Super-Additive Transfer**: Many $S_{f,e}$ values exceed 100% (e.g., FR→FI=107.04%, KO→IT=107.52%) when $L_f \neq L_e$, indicating that fine-tuning on a different but similar language may yield better performance than direct fine-tuning on the target language. Additionally, this effect is not necessarily symmetrical (e.g., NB→RU=105.01% vs. RU→NB=91.95%), highlighting directional transfer preferences. **Cluster vs. Transfer Divergence**: The language similarity reflected in Table 5, Table 6 and Table 7 does not strictly align with the clustering results in Figure 5. This discrepancy arises because Figure 5 captures similarity in task performance distributions, while similarity scores reflect cross-lingual knowledge transferability. For researchers who concern on efficient knowledge transfer across languages, Table 5 6 7 might provide more actionable insights.

5 Conclusion

5.1 Key Findings

Our systematic research yields **LLM-based language families** (Figure 5) and **similarity scores within language pairs** (Table 5 6 7) for 31 involved languages, from which two principal findings are concluded:

Discrepant Perspectives on Language Similarity

The clustering results (Figure 5) reveal that LLM-derived language similarity only partially aligns with traditional linguistic typology. There exists consistency (e.g., the North Germanic branch) and discrepancies (e.g., English and French not clustering together, or Chinese and Greek being grouped) between LLM-derived language clusters and traditional linguistic typology. This perspective may prioritize aspects like syntactic complexity, lexical overlap, or task-solving performance over historical or genealogical language relationships.

Super-Additive Transfer Effects

The similarity scores in Table 5, Table 6 and Table 7 demonstrate an unexpected phenomenon where fine-tuning on language L_f yields better performance on language L_e than direct L_e fine-tuning ($S_{f,e} > 100\%$) in some cases. This challenges conventional transfer learning assumptions and suggests that: (I) Fine-tuning on a related but distinct

language may provide a more effective "bridge" for knowledge transfer than direct fine-tuning on the target language. (II) Certain language pairs may share complementary linguistic features in LLM parameter space, enabling enhanced transfer. (III) Task-specific knowledge transfer could bypass surface linguistic differences, leading to unexpected performance enhancements.

5.2 Additional Contribution

We present **M4CQ** (Massive Multilingual Multitask Multiple Choice Question), which comprises **31** languages with **135** task categories. Key features include: a) Full semantic equivalence across languages via professional translation and rigorous proofreading (Figure 3). b) Balanced task distribution (Figure 1) spanning Natural Language Inference (NLI), Natural Language Understanding (NLU), Knowledge, LLM Special Test, and Code (Ability to Use). c) Flexible schema supporting variable-length options and future extension to multiple answers (Section 3.2.2).

This resource addresses gaps in existing multilingual datasets and will be publicly released to facilitate further research into LLM-based language similarity and other area of multilingual NLP research.

5.3 Future Work

a) Extending M4CQ to endangered languages and low-resource dialects. b) Extending M4CQ to multiple-answer questions. c) pre-training a model on all target languages (i.e., all languages present in the multilingual multitask dataset) with completely semantically equivalent corpora across languages, and evaluating the experimental outcomes (as mentioned in Section 2.4). d) Developing theoretical frameworks to explain the super-additive transfer phenomenon.

Limitations

Model Constraints: The requirements for strictly monolingual pre-training limited model selection to architectures that, although classic, are not state-of-the-art, potentially affecting generalizability to latest LLMs. **Language Coverage Limitations**: While M4CQ 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 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-4.0 license with explicit prohibitions against military or surveillance applications. Potential biases in monolingual pre-training corpora remain an unresolved limitation.

References

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.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

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*.

Ethan C. Chau and Noah A. Smith. 2021. Specializing multilingual language models: An empirical study. *CoRR*, abs/2106.09063.

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.

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*.

Karl Pearson F.R.S. 1901. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 1*, 2:559–572.

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*.

Trevor Hastie. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*.

John Healy and Leland McInnes. 2024. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4:82.

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.

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520.

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.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual BERT: an empirical study. *CoRR*, abs/1912.07840.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: investigating knowledge in multilingual pretrained language models. *CoRR*, abs/2102.00894.

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.

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

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.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

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).

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2001. On spectral clustering: analysis and an algorithm. In *Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 849–856, Cambridge, MA, USA. MIT Press.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

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.

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.

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

Peter Rousseeuw. 1987. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65.

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.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Arivazhagan, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2019. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. *CoRR*, abs/1909.00437.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual NLP models. *CoRR*, abs/2110.08875.

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, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Annasaheb Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Santilli, Andreas Stuhlmuller, Andrew M. Dai, Andrew D. La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakacs, Bridget R. Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Ozyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Stephen Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, C'esar Ferri Ram'irez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Tatiana Ramirez, Clara Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Daniel H Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Gonz'alez, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, D. Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth P. Donoway, Ellie Pavlick, Emanuele Rodolà, Emma FC Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fan Xia, Fatemeh Siar, Fernando Mart'inez-Plumed, Francesca Happ'e, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-L'opez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Han Sol Kim, Hannah Rashkin, Hanna Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hubert Wong, Ian Aik-Soon Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, J. Brooker Simon, James Koppel, James Zheng, James Zou, Jan Koco'n, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Narain Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jenni Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Jane W Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jorg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Ochieng' Omondi, Kory Wallace Mathewson, Kristen Chia-

fullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Col'on, Luke Metz, Lutfi Kerem cSenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Madotto Andrea, Maheen Saleem Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, M Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew Leavitt, Matthias Hagen, M'aty'as Schubert, Medina Baitemirova, Melissa Arnaud, Melvin Andrew McElrath, Michael A. Yee, Michael Cohen, Mi Gu, Michael I. Ivanitskiy, Michael Starritt, Michael Strube, Michal Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Monica Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, T MukundVarma, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas S. Roberts, Nicholas Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W. Chang, Peter Eckersley, Phu Mon Htut, Pi-Bei Hwang, P. Milkowski, Piyush S. Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, QING LYU, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ram'on Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib J. Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Sam Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Sameh Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo hwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Rose Biderman, Stephanie C. Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq A. Ali, Tatsuo Hashimoto, Te-Lin Wu, Theo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, T. N. Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler O. Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay V. Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, W Vossen, Xiang Ren, Xiaoyu F Tong, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yang Song, Yasaman Bahri, Ye Ji Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yu Hou, Yushi Bai, Zachary Seid, Zhao Xinran, Zhuoye Zhao, Zi Fu Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu, Sahib Singh, and Uri Shaham. 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.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. 2024. Parrot: Multilingual visual instruction tuning. *CoRR*, abs/2406.02539.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Raúl Vázquez, Alessandro Raganato, Mathias Creutz, and Jörg Tiedemann. 2020. A systematic study of inner-attention-based sentence representations in multilingual neural machine translation. *Computational Linguistics*, 46(2):387–424.

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/. Accessed: [].

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

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

A Task Information of M4CQ

Table 2 shows domains and descriptions of 135 tasks in M4CQ.

Table 2: Task information of M4CQ

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 fomula	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 2 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 2 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.
	winograde	The first half of the sentence involves two people, and the second half digs in to ask which one should be filled in.

Table 2 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 2 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 2 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 2 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 2 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 2 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.

B Specifications of the Pre-experiment on Translation Engine Selection

To select suitable translation engines for dataset construction, we conducted a preliminary evaluation across seven translation systems: DeepL, Google Translation, Bing Translation, Claude 3.5 Sonnet, Qwen2.5-32B, DeepSeek-R1, and GPT-4o. The experiment involved 1,350 instances (10 per task) from the M4CQ dataset, which were then translated into five languages: German, Chinese, Japanese, Russian, and Latvian. Each engine produced 6,750 translation outputs (1,350 instances × 5 languages), totaling 47,250 translations for human evaluation.

A three-tiered classification method was employed to evaluate the translations:

- **Perfect**: Flawless translation with accurate terminology
- **Acceptable with Minor Issues**: Generally acceptable but containing minor flaws (awkward phrasing or non-standard terminology)
- **Erroneous**: Meaning-altering errors or incomprehensible output

As shown in Table 3, DeepL demonstrated superior performance with 97.35% perfect translations and a remarkably low error rate of 0.07%. GPT-4o, DeepSeek-R1, Claude 3.5 Sonnet, and Qwen2.5-32B showed competitive accuracy ranging from 89.54% to 95.96%, though with slightly higher error rates (0.62%–4.40%). Traditional translation services (Bing and Google) exhibited significantly poorer performance, with error rates exceeding 35%.

Based on these findings, DeepL was selected as the primary translation engine due to its exceptional accuracy (high perfect rate) and stability (low error rate). The four language models (GPT-4o, DeepSeek-R1, Claude 3.5 Sonnet, and Qwen2.5-32B) were retained for collaborative error correction during quality control phases. Conventional translation engines (Google and Bing) were excluded from subsequent pipeline stages due to their substantially higher error rates.

C Experimental Instructions

C.1 Experiment 1: Language Feature Clustering

Objective

Explore LLM-based language family classification by performing clustering analysis on language feature vectors.

Procedure

Model Preparation: 1. Select a pre-trained model M_{base}, trained exclusively on a single language L_p (abbreviated from $L_{\text{pre-train}}$). 2. Select a multilingual multitask dataset , where each language variant contains linguistically diverse but *semantically equivalent* task instances, maintaining content consistency. 3. For each language, split the dataset into training and test sets at the task level and keep the instances in train/test sets *semantically equivalent* across all languages. Aggregate the training portions of all tasks to form a unified training set for each language. 4. Fine-tune M_{base} on the training set of L_p, yielding the fine-tuned model M_f (abbreviated from $M_{\text{fine-tune}}$). *Why choose M_f instead of M_{base}*: The performance of M_f is expected to correlate with the evaluation language's similarity to L_p. And compared with M_{base}, M_f exhibits varying degrees of improvement across languages. The improvement is more pronounced when the evaluation language is closer to L_p. This differentiation in performance enhances the effectiveness of subsequent clustering analysis.

Model Evaluation: Evaluate M_f on all language variants of the multitask test sets. Let L_e denote the evaluation language.

Feature Extraction: For each evaluation language L_e, construct a performance vector $\mathbf{v}_e \in \mathbb{R}^N$ (outlined in Section 2.2). These vectors are considered as features of these languages on LLM.

Clustering Analysis: If the number of task categories is large, first apply dimensionality reduction (e.g., PCA) to the set $\{\mathbf{v}_e\}$ to mitigate the curse of dimensionality. Then, perform clustering (e.g., k-means) and interpret the cluster assignments as indicators of language similarity from the LLM's perspective.

C.2 Experiment 2: Cross-Language LoRA Fine-tuning Transferability

Objective

Quantify pairwise language similarity scores through cross-lingual transfer analysis.

Procedure

Base Model: Select a pre-trained model M_{base}, trained exclusively on one language.

Adapter Training: Select a multilingual multitask dataset where each task instance is *semantically equivalent* across all languages, maintaining

Table 3: Translation Engine Performance Comparison (Total Evaluation Number = 6,750 per Engine)

Engine	Perfect (%)	Acceptable with Minor Issues (%)	Erroneous (%)
DeepL	97.35 (6,571)	2.58 (174)	0.07 (5)
GPT-4o	95.96 (6,477)	1.07 (72)	2.98 (201)
DeepSeek-R1	94.79 (6,398)	3.66 (247)	1.56 (105)
Claude 3.5	93.73 (6,327)	5.64 (381)	0.62 (42)
Qwen2.5-32B	89.54 (6,044)	6.06 (409)	4.40 (297)
Bing	34.83 (2,351)	28.40 (1,917)	36.77 (2,482)
Google	30.86 (2,083)	33.97 (2,293)	35.17 (2,374)

content consistency. Following the same data split procedure as in Experiment 1, fine-tune the M_{base} model on the training set of each language L_f to obtain the corresponding fine-tuned model M_f.

Cross-Lingual Evaluation: For each M_f and each L_e, evaluate performance on the L_e language variant of the multitask dataset.

Similarity Metric: Two metrics are designed for different situations (outlined in Section 2.3). If there exists a task in language L_e where M_e's accuracy is less than 40%, choose Metric-a, otherwise Metric-b. Because in the first situation, the average similarity score computed by Metric-b might be dominated by extremely large similarity scores of specific tasks.

D Fine-tuning Settings

The fine-tuning settings are detailed in Table 4.

E Similarity Scores for 31*31 language pairs

The similarity scores obtained in Experiment 2 are listed in Table 5, Table 6 and Table 7.

Table 4: 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 5: Similarity Scores for the First 11*31 Language Pairs (round to two decimal place and omit percentage signs)

L_e \ L_f	AR	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR
AR	100.00	99.46	98.17	98.14	99.59	98.91	97.73	97.71	98.50	99.84	100.14
BG	99.77	100.00	102.65	101.05	99.06	97.78	104.14	100.71	100.80	105.23	103.47
CS	101.09	102.00	100.00	100.26	99.21	99.18	104.33	99.62	100.28	100.00	102.49
DA	99.94	102.97	102.13	100.00	102.32	99.23	97.60	100.48	103.45	102.54	104.63
DE	101.93	97.24	100.76	102.08	100.00	97.13	96.68	100.19	100.64	100.77	101.84
EL	98.88	98.59	98.36	100.14	100.94	100.00	99.28	96.07	100.28	96.63	102.69
EN	103.11	100.52	99.96	102.07	97.72	99.27	100.00	98.98	99.54	102.10	100.65
ES	100.65	100.06	100.61	96.80	96.63	98.97	98.85	100.00	97.45	98.22	97.17
ET	105.10	102.96	99.69	100.43	100.95	101.87	100.85	101.03	100.00	100.68	102.49
FI	104.26	104.75	104.34	103.27	106.16	103.92	98.27	102.84	100.72	100.00	107.04
FR	101.40	101.20	99.92	100.23	96.87	98.75	97.01	99.59	99.39	99.94	100.00
HU	101.43	98.73	101.66	98.70	98.87	101.22	97.68	98.38	98.16	100.22	102.56
ID	101.85	103.88	103.88	103.25	103.05	102.64	100.20	101.56	99.92	103.74	103.90
IT	108.04	104.97	105.60	104.84	102.18	104.42	104.87	103.89	103.24	106.08	105.68
JA	102.41	100.98	102.34	103.86	100.22	103.62	103.52	103.99	105.20	102.30	103.48
KO	101.47	101.62	104.25	105.53	102.21	102.52	103.39	102.10	105.26	104.49	102.91
LT	97.59	96.10	98.44	97.03	96.07	96.27	98.84	98.23	96.67	95.85	99.36
LV	102.24	103.30	102.40	101.79	100.39	101.63	102.59	101.43	99.73	101.42	103.14
NB	96.43	96.54	95.77	99.07	97.22	97.47	93.86	97.34	96.75	100.16	100.61
NL	100.71	100.28	99.84	99.54	98.23	98.87	96.09	98.17	98.08	99.58	100.29
PL	103.88	103.09	103.25	101.92	102.42	101.78	105.99	102.23	102.50	101.21	101.79
PT-BR	103.95	100.63	100.06	100.98	101.29	99.95	100.08	101.01	99.82	101.08	101.44
PT-PT	100.82	97.46	96.53	99.83	97.57	97.35	97.33	98.62	97.54	99.71	99.39
RO	100.65	98.64	98.92	98.61	99.69	99.78	99.74	99.77	100.42	99.70	99.31
RU	104.54	106.37	105.57	105.46	101.64	103.34	103.69	102.94	103.14	104.85	106.02
SK	101.16	102.06	101.35	99.53	99.88	99.44	103.60	99.13	100.11	100.12	101.07
SL	102.30	103.31	102.65	100.77	99.83	100.71	99.90	102.44	97.19	98.87	102.70
SV	104.74	102.74	103.55	102.01	99.06	101.62	101.08	98.96	98.59	102.86	103.79
TR	99.78	98.71	98.13	97.75	98.97	97.69	97.46	97.72	100.52	97.29	99.58
UK	99.50	99.78	98.27	99.26	99.13	96.38	102.71	97.43	100.62	100.25	100.46
ZH	103.97	100.68	100.42	102.46	100.21	100.50	101.01	102.47	102.90	103.78	103.60

Table 6: Similarity Scores for the Middle 10*31 Language Pairs (round to two decimal place and omit percentage signs)

L_e \ L_f	HU	ID	IT	JA	KO	LT	LV	NB	NL	PL
AR	98.39	96.53	97.56	99.37	95.41	97.94	99.40	99.02	97.73	100.91
BG	102.16	98.03	98.71	101.97	99.91	100.33	101.99	100.12	99.88	102.58
CS	101.35	101.03	100.48	99.73	103.07	101.57	101.32	101.22	100.99	99.78
DA	101.71	98.51	98.36	98.82	103.59	105.30	102.35	101.49	101.36	102.58
DE	98.43	99.74	98.49	98.92	101.84	101.64	99.26	101.74	99.07	96.66
EL	100.62	98.18	95.89	99.63	97.06	98.83	97.42	99.92	98.39	100.01
EN	99.57	100.09	100.85	101.85	99.80	100.37	100.66	102.95	99.39	100.41
ES	97.70	98.07	95.04	100.02	100.94	98.76	97.36	98.44	98.30	97.90
ET	103.20	99.15	99.63	103.33	103.88	103.55	100.67	103.11	102.22	102.60
FI	103.21	104.02	101.43	105.89	103.32	105.77	104.68	104.81	103.57	102.58
FR	99.49	100.61	96.52	99.80	99.06	99.73	99.24	99.92	97.32	99.62
HU	100.00	98.76	99.36	100.32	97.70	98.71	101.51	98.56	101.14	99.86
ID	103.15	100.00	99.60	101.61	101.12	104.02	102.74	102.39	101.87	103.27
IT	104.26	104.89	100.00	105.46	107.52	102.95	103.93	106.26	103.57	102.71
JA	104.38	102.78	103.75	100.00	102.28	101.79	102.97	104.02	102.84	104.90
KO	104.12	101.44	104.00	104.15	100.00	105.29	102.97	103.39	102.55	105.28
LT	96.41	98.91	99.24	97.57	97.63	100.00	96.75	100.53	98.42	97.13
LV	102.35	104.03	104.35	102.01	104.44	102.83	100.00	100.60	102.07	102.66
NB	97.74	97.22	96.46	96.89	100.65	98.76	98.90	100.00	95.68	97.77
NL	98.35	98.99	96.63	99.20	99.17	98.02	99.32	100.53	100.00	100.02
PL	104.31	101.18	104.06	101.66	103.22	103.66	102.74	104.58	101.30	100.00
PT-BR	102.53	100.34	97.56	100.52	101.20	101.67	101.29	103.74	101.26	102.32
PT-PT	99.50	97.66	94.89	98.85	100.01	100.63	99.14	102.29	98.00	99.28
RO	98.72	98.11	99.08	99.33	99.12	100.37	96.78	98.81	98.99	99.62
RU	106.18	97.50	102.79	104.00	102.21	103.74	104.33	105.01	103.23	103.78
SK	101.26	101.26	100.00	99.47	105.27	104.23	101.09	102.09	100.62	100.99
SL	101.48	101.26	102.84	99.47	102.83	101.88	99.77	101.27	101.99	99.99
SV	104.51	101.90	99.80	100.30	104.81	102.57	99.77	102.60	100.58	100.96
TR	98.69	97.82	96.76	99.18	97.67	98.13	97.27	98.82	97.10	97.23
UK	99.63	94.40	99.22	99.76	97.59	100.99	97.25	98.96	98.46	100.11
ZH	98.58	100.09	103.29	99.62	101.97	99.17	103.58	102.66	102.37	98.85

Table 7: Similarity Scores for the Last 10*31 Language Pairs (round to two decimal place and omit percentage signs)

L_e \ L_f	PT-BR	PT-PT	RO	RU	SK	SL	SV	TR	UK	ZH
AR	101.06	99.72	97.91	99.18	99.03	98.22	99.14	97.62	99.30	98.89
BG	101.70	101.54	101.99	101.98	101.34	103.21	100.58	99.87	103.65	100.49
CS	98.90	102.56	101.17	100.89	101.18	101.53	100.69	101.25	101.17	104.56
DA	97.55	103.87	101.06	99.00	100.85	103.24	101.62	102.39	95.58	103.44
DE	99.70	102.20	97.64	93.09	98.04	100.51	99.29	100.52	94.15	102.18
EL	96.84	102.74	98.32	96.46	101.03	97.71	98.52	97.50	101.68	98.06
EN	102.88	103.76	98.77	99.83	97.80	97.17	98.57	100.60	96.04	101.90
ES	96.90	99.85	98.08	93.98	97.12	98.85	98.80	98.35	95.00	101.21
ET	97.54	102.92	101.73	93.91	99.48	99.54	100.63	102.10	98.61	105.64
FI	100.88	105.02	102.74	95.52	96.53	104.55	104.05	103.12	98.62	105.26
FR	95.76	99.63	100.61	97.04	99.53	99.35	98.95	100.60	92.79	101.87
HU	97.73	101.95	99.17	94.39	99.30	101.15	97.88	98.30	93.90	100.07
ID	102.88	103.09	100.09	96.52	100.96	104.29	104.38	102.11	97.88	104.64
IT	104.49	104.84	104.09	104.04	103.41	104.23	104.48	101.81	103.11	106.40
JA	101.69	102.16	104.04	100.75	101.41	103.48	102.10	102.49	100.29	102.36
KO	102.86	103.86	102.17	98.04	103.43	103.45	102.77	102.13	99.45	102.94
LT	98.13	98.77	95.79	96.91	96.87	96.55	98.82	95.34	97.76	97.79
LV	104.60	103.27	102.97	99.46	104.84	100.08	102.70	100.85	100.61	104.38
NB	94.65	98.47	96.83	91.95	95.28	98.57	97.28	97.26	91.28	97.65
NL	97.59	100.46	95.47	94.11	93.29	99.60	101.23	101.64	91.72	101.75
PL	102.39	104.55	102.67	100.68	102.47	101.25	101.92	104.07	103.09	104.82
PT-BR	100.00	102.53	100.57	100.24	95.68	99.98	102.25	99.47	95.63	101.50
PT-PT	98.03	100.00	98.16	97.22	92.99	98.48	98.36	96.44	92.47	99.54
RO	99.08	99.71	100.00	94.97	98.59	100.35	100.24	97.77	96.34	101.10
RU	103.59	104.59	102.73	100.00	103.76	105.62	106.41	102.33	99.71	102.12
SK	101.28	100.49	100.24	101.37	100.00	101.44	101.21	101.39	101.41	102.93
SL	102.47	102.74	99.74	98.02	98.87	100.00	102.13	99.26	101.66	100.35
SV	99.00	100.73	101.16	99.97	98.91	102.04	100.00	100.75	98.74	104.74
TR	98.52	101.64	98.33	94.99	95.68	98.06	98.71	100.00	98.08	97.52
UK	100.63	101.29	100.71	102.16	96.38	98.79	100.44	98.31	100.00	98.56
ZH	101.73	105.36	99.56	100.66	101.20	102.94	101.32	102.54	105.02	100.00