

When English Isn't the Best Teacher: Source Language Effects in Cross-Lingual In-Context Learning

Fred Philippy¹, Siwen Guo², Jacques Klein¹, Tegawendé F. Bissyandé¹

¹Snt, University of Luxembourg, Luxembourg

²Luxembourg Institute of Science and Technology, Luxembourg

Correspondence: fred.philippy@uni.lu

Abstract

Cross-lingual transfer in multilingual NLP has been widely explored in supervised fine-tuning contexts, where factors like data availability and linguistic similarity largely determine transfer quality. As the field shifts toward few-shot In-Context Learning (ICL), it is often presumed that insights from fine-tuning carry over unchanged. Yet this assumption has not been rigorously evaluated, leaving open the question of how to choose source languages for cross-lingual ICL. We conduct a broad empirical study of cross-lingual transfer in ICL spanning seven tasks, six models, and a typologically diverse set of languages. We further analyze language confusion, a key obstacle for generative tasks in cross-lingual ICL. Our results show that conventional fine-tuning-based expectations do not consistently apply in the ICL regime and point to alternative heuristics for selecting source languages effectively.

1 Introduction

Large language models (LLMs) have dramatically improved performance across a wide range of NLP tasks, yet their capabilities remain uneven across the world's languages. A core reason is the stark imbalance in training data availability: high-resource languages such as English dominate pretraining corpora, while many others remain severely under-represented. Much of the recent work on multilingual LLMs therefore focuses on cross-lingual transfer, where knowledge gained from high-resource languages is leveraged to improve performance in low-resource ones. Prior research, typically using supervised fine-tuning setups, has shown that cross-lingual transfer is far from uniform: factors such as linguistic similarity, lexical overlap, and writing system strongly influence how well knowledge transfers from a source language to a target one (Pires et al., 2019; K et al., 2020; Muller et al., 2021; Philippy et al., 2023). However, these find-

ings largely reflect the era of task-specific fine-tuning, where models are explicitly trained on labeled data in a source language and then evaluated in a different target language. With the rise of large, instruction-tuned LLMs, the field is increasingly shifting toward In-Context Learning (ICL) approaches that avoid parameter updates altogether (Brown et al., 2020). ICL departs from traditional fine-tuning by allowing models to perform tasks solely through examples and instructions provided at inference time. Instead of relying on parameter modification, LLMs condition on the prompt itself and learn “in context” from natural language demonstrations, task descriptions, or other structured cues. This paradigm has proven surprisingly effective across diverse tasks, and recent work shows that it can also support cross-lingual transfer: models can use demonstrations in one language to perform a task in another (Winata et al., 2022; Tu et al., 2025). Yet, unlike in the fine-tuning setting, it remains unclear which source languages are most effective for ICL for a given target language, and whether the factors known to influence fine-tuning transfer (e.g., linguistic similarity) apply equally in this new paradigm.

In this study, we investigate how source–target language relationships shape cross-lingual transfer in ICL. By systematically examining the effects of typology, writing systems, embedding-based alignment, and resource availability, our goal is to provide clearer guidance on how to select effective source languages for ICL, especially when working with low-resource target languages.

Our analysis reveals that assumptions inherited from fine-tuning do not always hold in ICL. In contrast to established findings in fine-tuning-based cross-lingual transfer, we observe that the target language itself is its most effective source language in only about 24% of cases, and that English, despite its disproportionate presence in pretraining corpora, emerges as the worst source language in

about 16% of our experiments. Moreover, we show that linguistic similarity, which is widely regarded as the strongest predictor of cross-lingual transfer in supervised fine-tuning, plays a far less important role in cross-lingual ICL. In additional experiments focused on language confusion, a phenomenon known to hinder cross-lingual transfer in generative tasks, we likewise uncover substantial discrepancies between source and target languages during ICL.

Altogether, our findings not only call for rethinking cross-lingual transfer in ICL but also point to a promising direction: the least supported languages in LLMs, particularly low-resource non-Latin-script languages, may in fact serve as unexpectedly strong sources from which other languages can benefit.

2 Related Work

Prior research indicates that demonstration language selection and configuration significantly impact cross-lingual ICL performance. Early studies found that cross-lingual prompting often surpasses monolingual baselines, though linguistic proximity is not always a reliable predictor of success (Winata et al., 2022). Recent work explores diverse prompt structures, including mixed-language demonstrations (Kim et al., 2024) and code-switching transitions to English (Yoo et al., 2025), which can boost reasoning in low-resource settings. Conversely, Tu et al. (2025) find that target-only or single high-resource source languages often outperform mixed prompts, suggesting that consistency within the demonstration language may be beneficial.

Research into demonstration selection further highlights the importance of retrieval quality. While alignment-based methods (Tanwar et al., 2023) and multilingual retrievers (Lin et al., 2025) improve results, these studies often restrict sources to English or leave selection unsystematic (Cahyawijaya et al., 2024). Finally, while linguistic similarity has been proposed as a selection predictor (Kaneko et al., 2025), its effectiveness in modern decoder-only LLMs remains an open question.

3 Methodology

3.1 Tasks

To evaluate cross-lingual ICL, we use seven multilingual benchmarks (Table 1). Unlike prior studies that focus on English or target-only demonstration

samples, we evaluate all source-target language combinations.

To manage the quadratic increase in evaluation runs, we selected 18 typologically diverse languages with maximal benchmark overlap. We further limited each benchmark to 1,000 test instances (or the full set, if smaller). Since all benchmarks are fully parallel, using aligned test and in-context examples ensures content consistency across language pairs and isolates cross-lingual transfer dynamics from content-related confounds.

3.2 Models

We conduct experiments with six LLMs, focusing on relatively small models (4B parameters or fewer) to keep the evaluation computationally manageable. To assess robustness with respect to model stochasticity, we repeat all experiments for the three smallest models using two different random seeds, yielding distinct sets of demonstration examples. We then compute the correlation between performance scores across language pairs for the two runs. All correlations are ≥ 0.9 across all tasks and models, indicating high stability. Based on this result, we perform single runs for the two larger models without compromising the reliability of our findings.

More specifically, we use **Llama 3.2 (1B & 3B)**, **Gemma 3 (1B & 4B)** (Team et al., 2025), **Qwen3 (1.7B)** (Yang et al., 2025) and **Phi-4-Mini** (Microsoft et al., 2025).

Following Zhang et al. (2024), who report that adding more than 2–4 demonstrations yields minimal further gains in multilingual scenarios, we adopt a 4-shot configuration for all experiments.

4 Results

Figure 2 presents the average transfer performance across all evaluated language pairs, showing mean z-scores computed over seven tasks and six models.¹ From these results, we identify different high-level observations:

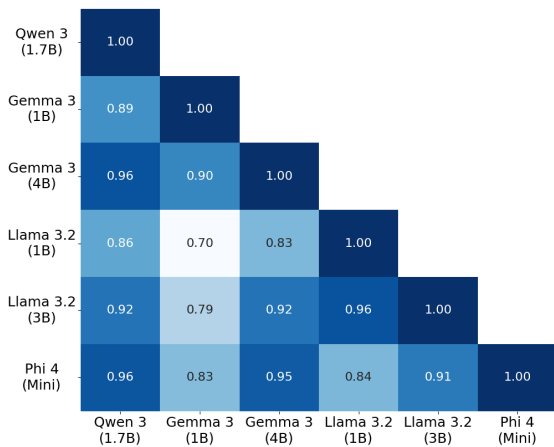
- **Persistent performance disparities across target languages.** As expected, English leads by a wide margin, followed by other high-resource languages such as Spanish, Italian, Indonesian, and German, while lower-resource languages like Swahili, Telugu, and Bengali show substantially lower scores.

¹Z-scores are computed separately for each model–task combination before averaging.

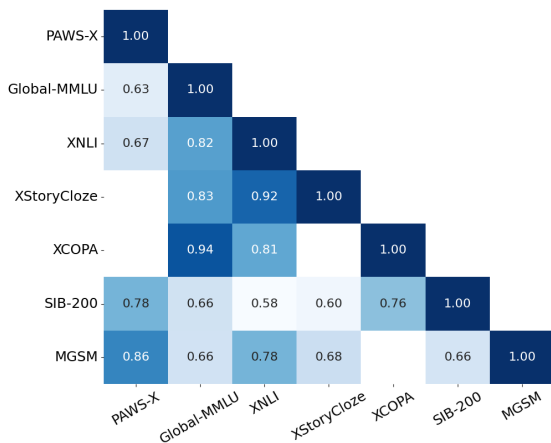
Paper	Dataset	Task	n	$ \mathcal{L} $	$ \mathcal{L}^* $	Metric
Yang et al. (2019)	PAWS-X	Paraphrase Identification	2k	7	7	F1
Lin et al. (2022)	XStoryCloze	Commonsense Reasoning	1.51k	11	9	F1
Singh et al. (2025)	Global-MMLU	Knowledge Comprehension	14k	42	17	Acc.
Conneau et al. (2018)	XNLI	Natural Language Inference	5.01k	15	12	F1
Shi et al. (2022)	MGSM	Mathematical Reasoning	250	11	11	Acc.
Adelani et al. (2024)	SIB-200	Topic Classification	204	206	18	F1
Ponti et al. (2020)	XCOPA	Commonsense Reasoning	500	11	7	F1

Table 1: Multilingual datasets included in our study. Here, n denotes the number of test samples per language, $|\mathcal{L}|$ the full language coverage of each dataset, and $|\mathcal{L}^*|$ the subset of languages we use as source and target languages.

- **The target language is not always its own best source**, contrary to a common assumption. In fact, across all experiments, the target language is the best source in only about 24% of cases.



(a) Across models.



(b) Across tasks.

Figure 1: Correlation matrix of transfer patterns. Pearson correlations are computed over all language-pair transfer scores.

- **Models show highly stable transfer patterns.** Similarities are strongest within model families and among larger models (Figure 1a).
- **Task-level transfer similarities are moderate to high.** Tasks with related reasoning or linguistic structure show more aligned transfer patterns (Figure 1b).

A deeper statistical analysis (§4.1 - §4.4) uncovers several critical findings:

1. **Strong target languages tend to be weak source languages (§4.1).** English, Spanish, German, and Italian perform well as targets but are among the weakest sources, whereas Thai, Telugu, and Bengali show the opposite pattern.
2. **Linguistic similarity between source and target does not predict transfer effectiveness (§4.2).** Unlike prior work on fine-tuning-based transfer, we observe no such relationship in the ICL setting.
3. **Cross-lingual alignment within the model is a stronger indicator of transfer success (§4.3).** This alignment explains far more variance in performance than surface linguistic properties.
4. **Low-resource, non-Latin-script languages make the most effective sources, whereas high-resource Latin-script languages are the least effective (§4.4).** While script and resource level each have independent effects, their interaction amplifies the pattern.

ar	-0.04	-0.13	-0.93	0.59	1.47	0.63	0.58	-0.23	0.48	0.72	0.12	-0.23	0.28	-1.88	-0.97	-0.19	-0.35	0.18	0.62
bn	-0.05	-0.32	-0.54	0.51	1.49	0.72	0.47	-0.43	0.41	0.47	-0.10	-0.25	0.13	-1.71	-0.84	0.22	-0.31	0.23	0.75
de	-0.09	-0.35	-0.88	0.44	1.38	0.48	0.29	-0.40	0.29	0.46	-0.21	-0.40	0.17	-1.79	-1.21	-0.06	-0.34	0.07	0.51
en	-0.19	-0.59	-0.95	0.26	1.36	0.30	0.24	-0.49	0.22	0.44	-0.17	-0.53	0.02	-2.00	-1.24	-0.17	-0.49	0.04	0.41
es	-0.10	-0.28	-0.91	0.19	1.27	0.59	0.30	-0.32	0.24	0.54	-0.31	-0.60	0.19	-1.80	-1.14	0.07	-0.38	0.16	0.49
fr	-0.05	-0.24	-0.81	0.31	1.34	0.48	0.53	-0.48	0.44	0.47	-0.25	-0.52	0.29	-1.89	-1.13	0.18	-0.32	0.20	0.45
hi	0.05	-0.35	-0.92	0.64	1.62	0.71	0.59	-0.21	0.36	0.64	0.15	-0.27	0.31	-2.04	-0.95	-0.02	-0.25	0.23	0.75
id	0.03	-0.46	-1.03	0.58	1.50	0.59	0.66	-0.34	0.64	0.51	0.07	-0.26	0.23	-1.75	-1.17	0.28	-0.31	0.21	0.63
it	-0.05	-0.36	-1.12	0.46	1.66	0.58	0.58	-0.54	0.39	0.78	-0.07	-0.49	0.10	-1.92	-1.18	0.00	-0.48	0.01	0.64
ja	-0.07	-0.35	-0.73	0.20	1.36	0.45	0.41	-0.45	0.40	0.48	0.06	-0.34	0.01	-1.92	-1.13	0.14	-0.30	0.11	0.40
ko	-0.05	-0.33	-0.95	0.49	1.40	0.28	0.42	-0.27	0.32	0.52	-0.06	-0.16	0.12	-2.06	-1.15	0.07	-0.18	0.25	0.44
ru	0.02	-0.31	-0.64	0.49	1.39	0.70	0.47	-0.27	0.31	0.56	-0.07	-0.29	0.35	-1.84	-1.13	0.02	-0.26	0.18	0.62
sw	-0.09	-0.47	-0.88	0.28	1.18	0.29	0.44	-0.43	0.32	0.48	-0.05	-0.27	0.06	-1.38	-1.00	-0.21	-0.54	0.13	0.44
te	0.06	-0.25	-0.66	0.46	1.42	0.62	0.47	-0.25	0.26	0.65	-0.04	-0.29	0.21	-1.76	-0.85	0.36	-0.24	0.28	0.75
th	0.05	-0.23	-0.55	0.48	1.15	0.58	0.42	-0.13	0.47	0.43	-0.12	-0.07	0.41	-1.96	-0.95	0.22	-0.39	0.30	0.83
tr	-0.02	-0.31	-0.98	0.55	1.65	0.65	0.47	-0.39	0.29	0.31	-0.01	-0.35	0.23	-1.99	-1.13	-0.07	-0.06	0.21	0.61
vi	-0.04	-0.34	-1.11	0.43	1.53	0.54	0.51	-0.45	0.34	0.43	0.04	-0.21	0.16	-1.98	-1.27	0.00	-0.31	0.38	0.54
zh	-0.02	-0.26	-0.85	0.41	1.38	0.41	0.38	-0.33	0.25	0.51	-0.04	-0.36	0.24	-1.77	-1.15	0.10	-0.30	0.24	0.70
AVG	-	-0.33	-0.86	0.43	1.42	0.53	0.46	-0.36	0.36	0.52	-0.06	-0.33	0.19	-1.86	-1.09	0.05	-0.32	0.19	0.59
	AVG	ar	bn	de	en	es	fr	hi	id	it	ja	ko	ru	sw	te	th	tr	vi	zh

Figure 2: Cross-lingual transfer matrix showing the mean z-score of model performance across language pairs. Rows represent source languages and columns represent target languages. Scores are normalized per task and model.

4.1 Correlation between Donor and Recipient Capabilities

Inspired by the approaches of Malkin et al. (2022) and Dymkiewicz et al. (2025), we analyze donor and recipient relationships across languages. To compute these scores, we follow the method introduced by Malkin et al. (2022): for each language \mathcal{L} , we define its donor score $\mathcal{D}(\mathcal{L})$ as the average performance it provides to all other target languages (excluding cases where \mathcal{L} is the target). Conversely, the recipient score $\mathcal{R}(\mathcal{L})$ is defined as the average performance that \mathcal{L} receives from all other source languages (excluding cases where \mathcal{L} is the source).

Figure 3 shows that donor and recipient capabilities are strongly negatively related, and we observe a Pearson correlation of -0.932 between source-to-target and target-to-source transfer across all models and tasks.

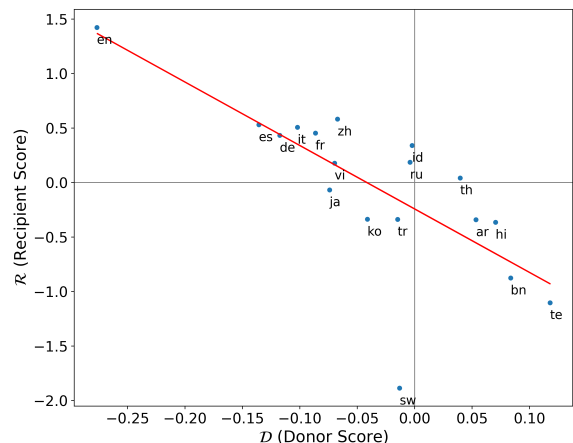


Figure 3: Scatter plot of donor and recipient capabilities; the red line indicates the best-fit regression capturing their inverse relationship.

4.2 Correlation with Linguistic Similarity

We examine whether the well-established relationship between linguistic similarity and transfer performance in fine-tuning scenarios also holds for prompting-based approaches. Using URIEL and

lang2vec representations (Littell et al., 2017), we compute four categories of linguistic similarity (syntactic, genetic, phonological, and featural) and correlate them with performance for each task and model pair individually.

As shown in Figure 4, only a small subset of task and model combinations exhibits low to moderate correlations across these features, and even these effects are inconsistent. For the vast majority of settings, we observe no strong correlation between linguistic similarity and transfer success. These findings suggest that, unlike fine-tuned models, in-context learning does not systematically rely on linguistic proximity, which highlights a fundamental difference in how cross-lingual generalization emerges in the two paradigms.

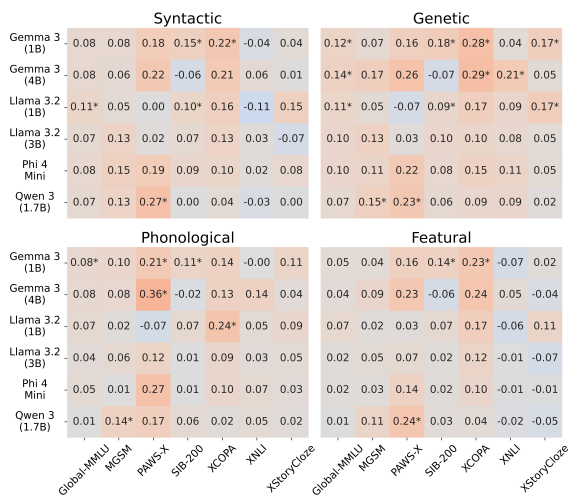


Figure 4: The four heatmaps illustrate the **correlation between linguistic similarity** (syntactic, genetic, phonological, and featural) of source–target language pairs **and cross-lingual transfer performance**. Each heatmap corresponds to one *lang2vec* dimension and reports the Pearson correlation coefficient for every benchmark (column) and model (row) combination. An asterisk (*) indicates statistical significance at $p < 0.05$.

4.3 Correlation with Cross-Lingual Alignment

Beyond surface-level linguistic features, we also analyze whether transfer performance is related to the degree of cross-lingual alignment between the source and target languages within the model’s internal representations. Although cross-lingual alignment is often correlated with underlying linguistic similarity, it captures a fundamentally different notion: it reflects how closely languages are positioned in the model’s learned representation space, which may diverge from typological

similarity due to factors such as training data composition, tokenization, or implicit model biases. To measure this, we use the devtest portion of FLORES-200 (Team et al., 2022), which provides aligned sentences across 200 languages, and for each model we compute the similarity of mean-pooled last-layer hidden states across languages using the Centered Kernel Alignment (CKA) metric (Kornblith et al., 2019). We find that these correlations between cross-lingual alignment and transfer performance are substantially stronger than those obtained from linguistic similarity features, although the strength of the effect varies depending on the task and model (Figure 5). This indicates that, in the context of in-context learning, models rely far more on the cross-lingual structure they have learned during pre-training and the resulting proximity of languages in their representation space than on any surface-level linguistic similarities.

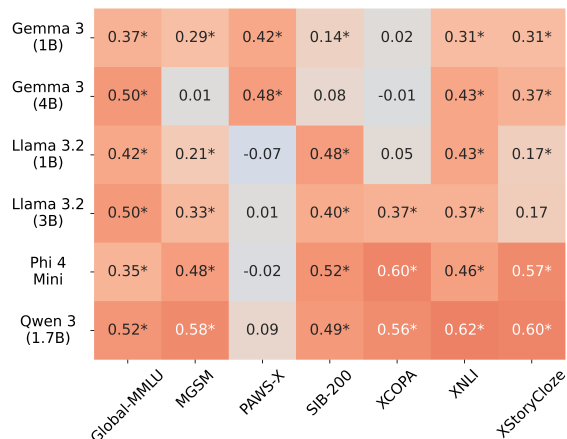


Figure 5: This heatmap shows how strongly a model’s internal alignment (last layer) between a given source–target language pair correlates with its transfer performance for that same pair. Correlation values are shown for each benchmark (column) and each model (row) combination.

4.4 Effect of Script and Resource Availability on Source Language Performance

To assess whether (i) the writing system and (ii) the resource status of the source language influence cross-lingual transfer performance, we perform a set of inferential statistical analyses.

First, we define a binary variable indicating whether the source language uses a Latin-based script. Resource status is likewise modeled as a binary variable, distinguishing between higher- and lower-resource languages based on their rela-

tive representation in the Common Crawl corpus², which we treat as a proxy for the relative availability of a language in typical large-scale pre-training data.

We first fit a two-way ANOVA model with transfer performance as the dependent variable, including script type (Latin vs. non-Latin), binary resource status (high vs. low), and their interaction as predictors. The analysis reveals significant main effects of both *Script* and *Resource Status*. The interaction between script type and resource status is not significant, indicating that the effects of script and resource status on transfer performance are additive rather than interactive (Table 2)³.

Model Term	F	p-value
Script (Latin vs. Non-Latin)	15.29	9.3e-05
Resource Status (High vs. Low)	17.86	2.4e-05
Script \times Resource Status	0.09	0.76

Table 2: **Simplified ANOVA results (script type \times resource level interaction model)** showing F-statistics and p-values for script type, resource level, and their interaction.

Second, to facilitate pairwise comparisons, we conduct post-hoc Tukey HSD tests to assess pairwise differences in transfer performance between all group combinations (Tables 3, 4 & 5).

Averaged across script types, low-resource source languages significantly outperform high-resource source languages (mean difference = 0.1187, $p < .001$), and averaged across resource levels, non-Latin source languages significantly outperform Latin-script source languages (mean difference = 0.1133, $p < .001$). When examining combined script–resource groups, both script type and resource level yield significant differences within each other’s levels. However, the comparison between high-resource non-Latin and low-resource Latin source languages is not significant, indicating that advantages associated with script type and resource availability can partially offset one another.

²Based on CC-MAIN-2025-38, retrieved from <https://commoncrawl.github.io/cc-crawl-statistics/plots/languages>.

³Assumptions were evaluated using residual diagnostics. Q–Q plots indicated approximate normality with mild tail deviations. Homogeneity of variances was assessed using Levene’s test and was satisfied ($p = .59$).

Comparison	meandiff	p-adj	lower	upper
High \rightarrow Low	0.1187	0.0000	0.0767	0.1607

Table 3: Tukey HSD contrast for resource level (High vs. Low). Low-resource languages show significantly higher transfer performance.

Comparison	meandiff	p-adj	lower	upper
NL \rightarrow L	-0.1133	0.0000	-0.1552	-0.0714

Table 4: Tukey HSD contrast for script type (Non-Latin [NL] vs. Latin [L]). The negative value indicates higher performance for NL languages.

Comparison	meandiff	p-adj
High_NL \rightarrow High_L	-0.0807	0.0394
High_NL \rightarrow Low_NL	0.1010	0.0075
High_NL \rightarrow Low_L	0.0068	0.9976
High_L \rightarrow Low_NL	0.1817	0.0000
High_L \rightarrow Low_L	0.0875	0.0284
Low_NL \rightarrow Low_L	-0.0942	0.0206

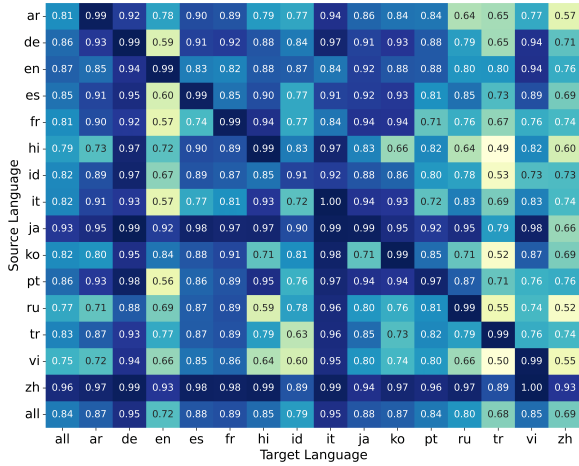
Table 5: Simplified Tukey HSD post-hoc comparison across all Script (Non-Latin [NL] & Latin [L]) and Resource (High & Low) combinations, showing only mean differences and adjusted p-values.

5 Language Confusion in Cross-Lingual In-Context Learning

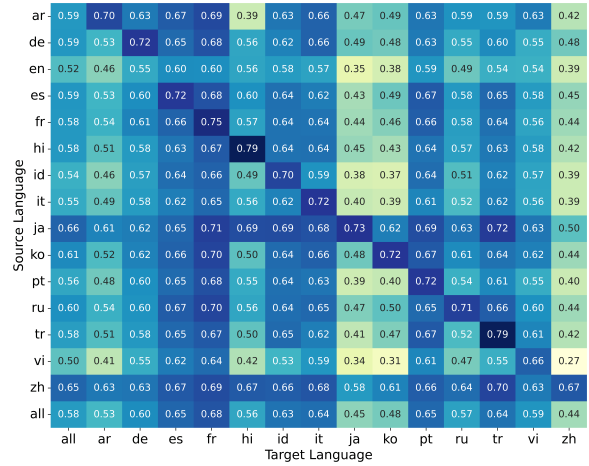
To broaden our analysis beyond classification tasks, we additionally investigate whether source-language variation in ICL produces comparable effects in generative settings.

In our cross-lingual setup, we do not prioritize task-specific performance. Instead, we shift our attention to diagnosing a fundamental obstacle in generative cross-lingual evaluation: *language confusion*. Language confusion occurs when a model’s responses are produced in a language different from the one requested, such as defaulting to English rather than the target language. When a model fails to generate output in the intended language \mathcal{L} , we cannot meaningfully assess cross-lingual transfer quality because no valid output in \mathcal{L} exists to evaluate. In real applications, this failure means the user’s request is not truly addressed, as the response may be partially or entirely incomprehensible. For these reasons, language confusion is arguably a more fundamental metric here than task performance itself.

We therefore evaluate models using the *Lan-*



(a) Monolingual setup.



(b) Cross-lingual setup.

Figure 6: Average Line-level Pass Rates (LPR) across models for different source-target language pairs.

guage Confusion Benchmark (Marchisio et al., 2024), which measures a model’s ability to produce responses in the specified language. The benchmark spans 15 typologically diverse languages that fully overlap with the language set used in Section 4. It includes two settings: (1) **monolingual**, where both the prompt and expected response are in the target language, and (2) **cross-lingual**, where the prompt is given in English but explicitly instructs the model to answer in the target language. Consistent with earlier experiments, we use a 4-shot ICL setup with demonstration examples drawn from the parallel Bactrian-X instruction-tuning dataset (Li et al., 2023), which covers all languages present in the benchmark.

As a metric, we adopt the *Line-level Pass Rate (LPR)* introduced by Marchisio et al. (2024). LPR is defined as the proportion of model responses where every line is correctly identified as being in the user’s intended language⁴.

5.1 Results

We provide the average LPR across all models for each source-target pair in Figures 6a and 6b for the monolingual and cross-lingual settings respectively.

Overall patterns of language confusion.

Unlike the consistent patterns observed in task-specific classification, language confusion shows substantially less consistency across models. As expected, confusion patterns are most similar within

⁴Following Marchisio et al. (2024), we perform line-level Language Identification (LID) using fastText (Joulin et al., 2016) and restrict its use to sequences with more than four words to ensure LID reliability.

model families (Gemma 3 and Llama 3.2), with moderate correlations also observed between Qwen 3 and Llama 3.2, and between Phi 4 and Gemma 3 (Figure 7).

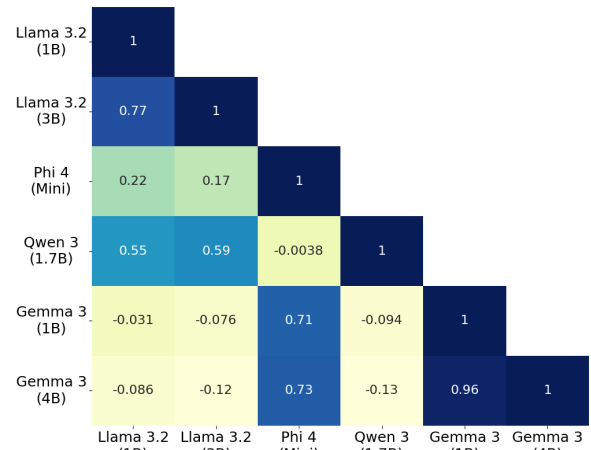


Figure 7: Correlation matrix of language confusion patterns across models. Pearson correlations are computed over all language-pair transfer scores.

Relationship to task transfer performance.

We further compute correlations between task-level transfer performance and language confusion. Overall, we find that transfer performance for some tasks correlates moderately with language confusion in the cross-lingual setting, while other tasks show moderate correlations in the monolingual setting. However, these relationships are not consistent across tasks or models, indicating that **language confusion is not directly linked to unequal transfer performance, but instead represents a distinct challenge.**

Additionally, across all models and both settings,

we find a Pearson correlation of 0.6342 between a language’s tendency to induce confusion when used as a source language and the confusion it experiences as a target language. This source–target symmetry mirrors the correlation previously observed between donor and recipient capabilities of a language in the task transfer experiments (§4.1), albeit with a somewhat weaker magnitude for language confusion.

Linguistic correlates of language confusion.

We investigate correlations between language confusion and two factors: linguistic similarity and the cross-lingual alignment of source and target language embedding spaces. For both the monolingual and cross-lingual settings, no strong correlations were observed for any model tested⁵. Similarly, we found no strong tendency regarding the correlation between the resource level of the source or target language and the resulting language confusion.

However, t-tests comparing language performance based on writing system yielded significant results ($\alpha = 0.05$). Specifically, we found that non-Latin script languages suffer more from language confusion as a target language, but they create less language confusion in other target languages when used as the source language, compared to Latin script languages.

6 Discussion

Our findings suggest that ICL leverages mechanisms distinct from fine-tuning during cross-lingual transfer. The lack of correlation with linguistic similarity indicates that ICL relies less on structural overlap and more on representational properties learned during pretraining. While linguistic similarity guides parameter updates, it is less relevant when models must infer tasks from contextual examples at inference time.

One possible explanation is that source languages differ in the informational constraints they impose during in-context learning. High-resource languages such as English may be highly entropic, activating many overlapping pretraining patterns and encouraging reliance on broad, language-specific heuristics rather than task structure. In contrast, low-resource or typologically atypical languages may act as implicit regularizers, reducing spurious associations and forcing the model to focus more strongly on the abstract input–output mapping illustrated by the examples.

⁵All calculated Pearson correlation values were below 0.3.

The strong effectiveness of low-resource, non-Latin script languages as source languages raises practical and theoretical questions. While this finding suggests that practitioners could improve cross-lingual ICL performance by selecting such languages as sources, it remains unclear whether this strategy is robust or future-proof. As more data becomes available for currently low-resource languages, their role within multilingual models may change, potentially diminishing the very properties that make them effective sources today. Moreover, it is unknown whether there exists a lower bound of resource availability beyond which in-context examples cease to be helpful, because the model no longer reliably understands the language itself. This points to an unresolved trade-off between linguistic distinctiveness and model familiarity.

Finally, our results underscore that source and target languages play fundamentally different roles in cross-lingual ICL, and that strong performance as a target language does not imply effectiveness as a source language. This asymmetry suggests that current evaluation practices, often focused solely on target-side performance, may overlook critical aspects of source-language selection. Future work should aim to better characterize the properties that make a language a good source in ICL, and to disentangle whether these effects stem from script, data imbalance, pretraining dynamics, or more general regularization-like effects induced by atypical language distributions.

7 Conclusion

In this work, we showed that cross-lingual ICL exhibits behaviors that differ markedly from established patterns in fine-tuning-based cross-lingual transfer. In particular, source language effectiveness is not driven by linguistic similarity, and languages that perform poorly as targets are the most effective sources. These findings challenge common assumptions about source language selection and suggest that cross-lingual ICL relies on mechanisms that are still poorly understood. Beyond their immediate practical implications, our results point to the need for a deeper theoretical account of how multilingual models exploit in-context examples across languages, especially in settings involving low-resource and typologically distant languages.

References

- David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. [SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian’s, Malta. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. [LLMs are few-shot in-context low-resource language learners](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Kajetan Dymkiewicz, Ivan Vulic, Helen Yannakoudakis, Eilam Shapira, Roi Reichart, and Anna Korhonen. 2025. [Donors and recipients: On asymmetric transfer across tasks and languages with parameter-efficient fine-tuning](#). *Preprint*, arXiv:2511.13368.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
- Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. [Cross-Lingual Ability of Multilingual BERT: An Empirical Study](#). In *Proc. of the International Conference on Learning Representations*.
- Masahiro Kaneko, Alham Fikri Aji, and Timothy Baldwin. 2025. [Balanced multi-factor in-context learning for multilingual large language models](#). *Preprint*, arXiv:2502.11495.
- Sunyoung Kim, Dayeon Ki, Yireun Kim, and Jin-sik Lee. 2024. [Cross-lingual qa: A key to unlocking in-context cross-lingual performance](#). *Preprint*, arXiv:2305.15233.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMIR.
- Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. [Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation](#). *Preprint*, arXiv:2305.15011.
- Peiqin Lin, Andre Martins, and Hinrich Schuetze. 2025. [XAMPLER: Learning to retrieve cross-lingual in-context examples](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3968–3977, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. [Few-shot learning with multilingual generative language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. [URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.
- Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. [A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.
- Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. [Understanding and mitigating language confusion in LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.
- Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. [Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras](#). *Preprint*, arXiv:2503.01743.

- Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamel Seddah. 2021. [First align, then predict: Understanding the cross-lingual ability of multilingual BERT](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.
- Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. [Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal common-sense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. [Language models are multilingual chain-of-thought reasoners](#). *Preprint*, arXiv:2210.03057.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. [Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.
- Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. [Multilingual LLMs are better cross-lingual in-context learners with alignment](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.
- 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.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Yilei Tu, Andrew Xue, and Freda Shi. 2025. [Blessing of multilinguality: A systematic analysis of multilingual in-context learning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6213–6248, Vienna, Austria. Association for Computational Linguistics.
- Genta Winata, Shijie Wu, Mayank Kulkarni, Tamar Solorio, and Daniel Preotiuc-Pietro. 2022. [Cross-lingual few-shot learning on unseen languages](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.
- 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.
- Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. [PAWS-X: A cross-lingual adversarial dataset for paraphrase identification](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.
- Haneul Yoo, Jiho Jin, Kyunghyun Cho, and Alice Oh. 2025. [Code-switching in-context learning for cross-lingual transfer of large language models](#). *Preprint*, arXiv:2510.05678.
- Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024. [The impact of demonstrations on multilingual in-context learning: A multidimensional analysis](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7342–7371, Bangkok, Thailand. Association for Computational Linguistics.