SrcMix: Mixing of Related Source Languages Benefits Extremely Low-resource Machine Translation

Sanjeev Kumar, Preethi Jyothi, Pushpak Bhattacharyya
Department of Computer Science and Engineering, IIT Bombay, India
{sanjeev, pjyothi}@cse.iitb.ac.in

Abstract

Multilingual models are widely used for machine translation (MT). However, their effectiveness for extremely low-resource languages (ELRLs) depends critically on how related languages are incorporated during fine-tuning. In this work, we study the role of language mixing directionality, linguistic relatedness, and script compatibility in ELRL translation. We propose SrcMix, a simple source-side mixing strategy that combines related ELRLs during fine-tuning while constraining the decoder to a single target language. Compared to its target-side counterpart TgtMix, SrcMix improves performance by 3 ChrF++ and 5 BLEU points in high-resource to ELRL translations, and by 5 ChrF++ and 12 BLEU points in mid-resource to ELRL translations. We also release the first Angika MT dataset[1] and provide a systematic comparison of LLM (Aya-101) and NMT (mT5-Large) models under ELRL settings, highlighting the importance of directional mixing and linguistic compatibility.[2]

1 Introduction

Linguistic diversity continues to decline due to the lack of digital documentation and annotated data, posing a major barrier to the inclusion of ELRLs. Recent efforts have broadened coverage for many ELRLs. For instance, initiatives like NLLB (Costa-Jussà et al., 2022), Masakhane (Nekoto et al., 2020; Adelani et al., 2022), and Samanantar (Ramesh et al., 2022) have contributed substantial monolingual and multilingual resources. In the context of MT, constructing ELRL datasets commonly begins by translating high-quality, standard corpora from a high-resourced language (HRL) into the target ELRL. This approach ensures consistency, immediate usability, and compatibility with evaluation benchmarks such as FLORES-200 (Goyal et al., 2022). Following this approach, *we create the first MT dataset for Angika*, an Indo-Aryan ELRL spoken in eastern India, by translating the NLLB seed corpus (Maillard et al., 2023) and the FLORES-200 devtest sets.

While dataset creation is a critical first step, developing effective MT systems for ELRLs remains challenging. State-of-the-art approaches often rely on large multilingual pretraining (Aharoni et al., 2019; Gu et al., 2018), or two-stage fine-tuning pipelines (Eriguchi et al., 2022), which are computationally expensive and infeasible in ELRLs. Moreover, naively adding multilingual data can degrade performance (Iyer et al., 2024). This raises several important questions: What is the best way to integrate related languages during multilingual training for ELRLs? Does the direction of mixing on the source or target side affect performance? Moreover, how do language groups and script compatibility influence generalization?

To address these questions, we propose SrcMix, a single-stage fine-tuning strategy that mixes related source languages while keeping the ELRL target fixed. By anonymizing source tags, SrcMix encourages decoder specialization and implicit cross-lingual transfer. We compare this with TgtMix (multiple ELRL targets with a fixed source), bilingual fine-tuning, and many-to-many (M2M) training, and zero-shot transfer.

We evaluate these strategies using both NMT and multilingual LLM models. Specifically, we experiment with mT5-Large (Xue et al., 2021) and Aya-101 (Üstün et al., 2024). Using Aya-101 as the multilingual LLM was a deliberate choice. Decoder-only LLMs such as LlaMA-3.1 (8B) (Dubey et al., 2024) and Gemma-7B (Team et al., 2024) performed poorly for our chosen ELRLs with very limited amounts of parallel text (please refer to §A.3). Aya-101, in contrast, performed much better on our ELRLs as also evidenced in prior work on translation for low-resource languages (Xu et al.,

[1] https://huggingface.co/datasets/snjev310/AngikaMT
[2] https://github.com/csalt-research/SrcMix

Findings of the Association for Computational Linguistics: EACL 2026, pages 6306–6323
March 24-29, 2026 ©2026 Association for Computational Linguistics

2024; Köksal et al., 2025).

Our study systematically compares directional multilingual mixing across 14 ELRLs from four geographically distinct language groups and three scripts using both NMT and LLM models. Despite its simplicity, SRCMIX yields consistent gains across typologically diverse ELRLs. This is unlike naive mixing or TGTMIX, which often degrades performance, revealing a key asymmetry in multilingual transfer. These findings underscore the importance of directional mixing and typological compatibility in ELR generalization.

Specifically, our work aims to answer two key questions: (1) How does multilingual training influence generalization to unseen ELRLs? (2) Does source-side or target-side mixing yield better performance? Results show SRCMIX consistently outperforms bilingual fine-tuning and TGTMIX, and remains competitive with or superior to M2M and zero-shot baselines; our reported gains are not from data volume or longer training (as shown in §7 and §A.11). Our contributions are:

1. SRCMIX. We introduce SRCMIX, a simple and scalable source-side multilingual fine-tuning strategy for ELRLs. Unlike prior one-to-many (O2M) or M2M setups, SRCMIX requires no language tags or multi-stage training, and consistently improves translation quality across diverse ELRLs (§3 and §6.1).

2. *First MT corpus for Angika.* We release the first MT dataset for Angika, an Indo-Aryan ELRL, comprising 6,192 training and 1,012 evaluation examples, extending MT coverage to a previously undocumented language (§4).

3. *New ELRL Baselines.* We establish new state-of-the-art baselines for translation into and out of 14 ELRLs, comparing both NMT (mT5-Large) and LLM (Aya-101) models (§5 and §6.1).

4. *Analysis of multilingual mixing strategies.* We provide the first large-scale analysis of multilingual mixing direction (source vs. target), language group, and script similarity. Our results show that multilingual transfer is most effective when mixing is source-side and languages are typologically and orthographically related (§6.2 and §7).

2 Related Work

Multilingual Training Strategies. A common approach to support ELRLs has been to enlarge the training corpus by pooling data from multiple languages in either one-to-many (O2M) or many-to-many (M2M) configurations. In O2M setups, a single source (typically English) is paired with multiple targets, relying on explicit language tags to guide the decoder (Johnson et al., 2017). M2M approaches extend this further, training on all available source–target combinations within a language set (Aharoni et al., 2019; Fan et al., 2021; Costa-Jussà et al., 2022). While these methods enable large-scale multilingual transfer, they often suffer from negative interference, especially when unrelated or typologically distant languages are mixed (Wang et al., 2020; Chang et al., 2024; Kumar et al., 2024). Moreover, ELRLs rarely benefit from such naive concatenation (Iyer et al., 2024) since decoder supervision becomes fragmented across diverse targets, weakening specialization.

Recent work has highlighted this challenge more broadly as *conflict resolution* in massively multilingual LLMs (Zheng et al., 2025; Wu et al., 2024), where competing linguistic patterns interfere with one another. These studies explore architectural or decoding-level mechanisms to resolve such conflicts. By contrast, our approach is deliberately simple: SRCMIX introduces multilinguality only on the *source side*, while keeping the decoder focused on a single ELRL target. Rather than resolving conflicts after they arise, this strategy prevents them, enabling decoder specialization and more effective structural transfer without the drawbacks of O2M or M2M setups.

Pre-trained Models and Data Efficiency for ELRL MT. The rise of multilingual pre-trained language models (PLMs) such as mT5 (Xue et al., 2021), NLLB (Costa-Jussà et al., 2022), and LLaMA (Touvron et al., 2023) has shifted the paradigm from training from scratch to adapting general-purpose models. Parameter-efficient fine-tuning (PEFT) techniques like LoRA (Hu et al., 2022) and adapters (Pfeiffer et al., 2020) further reduced the computational cost of adapting large models to low-resource tasks. Recent instruction-tuned LLMs, such as GPT-3 (Brown et al., 2020) and Aya-101 (Üstün et al., 2024), extend coverage to 100+ languages, supporting zero-shot and few-shot translation (Kojima et al., 2022; Zhu et al., 2023). Nevertheless, these models often perform inconsistently for ELRLs, where training data is scarce and typological diversity is high.

Another line of research investigates how far high-quality but small datasets can go. Tanzer et al. (2023) demonstrated translation for Kalamang using only a grammar book, though later work revealed hidden parallel examples in the resource (Aycock et al., 2024). Similarly, Maillard et al. (2023) and Wu et al. (2024) showed that as few as 100–6K sentences can yield reasonable results in ELRL MT. Participatory projects like Masakhane (Nekoto et al., 2020; Adelani et al., 2022) emphasize the importance of community-driven data curation. These works highlight that curated data and typological alignment matter as much as scale. However, typological differences, such as word order and script, can hinder transfer effectiveness (Östling et al., 2017).

3 Methodology

We investigate how multilingual training strategies can improve translation for ELRLs by exploring the direction and structure of language mixing. Specifically, we compare five fine-tuning strategies: (i) bilingual fine-tuning (Standard FT), (ii) source-side mixing (SRCMIX), (iii) target-side mixing (TGTMIX), (iv) many-to-many training (M2M), and (v) zero-shot transfer. These setups allow us to study the impact of directional mixing and how our approach differs from naive multilingual concatenation and conventional many-to-many systems.

Standard FT. In this setting, we adopt standard bilingual fine-tuning, where a separate model is fine-tuned for each source-to-target language pair using supervised training. The training process employs the LoRA method, as described in Section 5.3. The resulting models are then evaluated individually for each translation direction.

SRCMIX. SRCMIX is a single-stage multilingual training-time strategy tailored for ELRLs. It mixes multiple related source languages targeting a fixed ELRL. Let n_1 denote the target ELRL, and $n_2, n_3, \ldots, n_N$ be related sources within the same family and script. To train $H \rightarrow n_1$, we aggregate training pairs of the form $H \rightarrow n_1, n_2 \rightarrow n_1, n_3 \rightarrow n_1, \ldots, n_N \rightarrow n_1$, where H is a high-resource language (e.g., English or Hindi). For reverse translation, the model is trained on $n_1 \rightarrow H, n_2 \rightarrow H, \ldots, n_N \rightarrow H$.

During training, all source language identities are anonymized, meaning the model is not explicitly informed about the language the input came from. This setup allows the model to leverage training signals from structurally similar but lexically diverse sources while keeping the decoder focused on a single target language distribution. For instance, to train a model for English-to-Angika translation, we construct a mixed training dataset using parallel examples from English→Angika, Magahi→Angika, and Bhojpuri→Angika, with source identity removed. By directly integrating language similarity and target specialization into training, SRCMIX offers a simple and scalable alternative to conventional multilingual fine-tuning, especially for ELRL settings with limited supervision. Unlike naive concatenation or M2M training, which enlarge the dataset while preserving explicit language IDs, SRCMIX removes these cues and compels the model to transfer implicitly.

TGTMIX. TGTMIX reverses the mixing direction: a fixed high-resource source language H is paired with multiple ELRL targets $\{n_1, n_2, \ldots, n_N\}$. Training data includes $H \rightarrow n_1, H \rightarrow n_2, \ldots, H \rightarrow n_N$, with anonymized targets.

This encourages the decoder to generalize across multiple ELRLs. However, unlike source-side mixing, target anonymization forces the decoder to model multiple distributions simultaneously, often leading to instability. We therefore treat TGTMIX as a contrastive baseline that highlights the asymmetry between encoder- and decoder-side transfer. In our setup, TGTMIX is not applied for ELRL→HRL, since only one HRL (English) and one MRL (Hindi) exist as targets. Furthermore, even with language tags, TGTMIX underperforms both SRCMIX and Standard FT (§A.10).

Many-to-Many (M2M) and Zero-Shot. We include a conventional many-to-many (M2M) setup, where the model is trained simultaneously on all available source–target pairs within a language group. This represents the naive concatenation baseline common in large multilingual systems (e.g., NLLB, M2M-100). For comparability, we construct M2M corpora only over languages present in our dataset, avoiding reliance on external predefined tags.

We also evaluate a *zero-shot* setting, without any additional fine-tuning. Instead, we directly evaluate the pretrained LLM models on unseen ELRL translation directions during their training. This evaluation measures the inherent generalization ability of pretrained models to ELRLs without explicit supervision.

4 New MT Corpus for Angika

To extend language coverage, we introduce the first MT dataset for Angika (*ISO 639-2: anp*), an Indo-Aryan ELRL spoken in India (Bihar, Jharkhand, West Bengal) and Nepal, written in Devanagari. The dataset extends the NLLB Seed and FLORES-200 corpora by translating English data into Angika.

To create the Angika MT dataset, we engaged two native Angika speakers who have extensive language expertise and possess proficient typing skills. Each translator independently translated 50 sentences at a time, followed by a peer review process where they verified each other's translations and resolved any discrepancies. Finally, a third native Angika speaker performed a thorough verification of all the translated sentences. This process resulted in the first training, validation, and test sets for Angika MT. The training dataset size is similar to that in the NLLB Seed corpus (6,192). The evaluation sets are derived from translating the dev (997) and devtest (1,012) splits of the FLORES-200 dataset, facilitating meaningful comparisons with established benchmarks. Further details about the Angika language are provided in §A.1, along with our quality control process in §A.2, and information about the compensation provided to translators in §A.2. Since the NLLB Seed corpus lacks Hindi data, we translated English sentences into Hindi using Google Translate, followed by verification by two native Hindi speakers. Incomplete or transliterated outputs from Google Translate were fixed during this process.

5 Experimental Setup

5.1 Selected Languages and Datasets

Our experiments use three datasets: the NLLB Seed corpus, FLORES-200, and a newly created Angika MT dataset. We focus on 14 ELRLs, with Angika data created from scratch and the remaining 13 languages sourced from the NLLB Seed corpus, each containing 6,192 sentence pairs. Our selection includes *four African languages: Nigerian Fulfulde (fuv_Latn), Nuer (nus_Latn), Bambara (bam_Latn), and Tamasheq(taq_Latn), four European languages: Friulian (fur_Latn), Ligurian (lij_Latn), Limburgish (lim_Latn), and Sardinian(srd_Latn), three Indic languages: Angika (anp_Deva), Bhojpuri (bho_Deva), and Magahi (mag_Deva), and three Indo-Iranian languages: Dari (prd_Arab), Kashmiri (kas_Arab), and Southern Pashto (pbt_Arab)*. We use the FLORES-200 devtest sets for evaluation, which contain 1,012 sentence pairs, respectively. We conduct bidirectional translation between each ELRL and English (HRL), and for six Indic and Indo-Iranian languages, we also include Hindi (MRL) as an additional source and target. This set of languages was intentionally chosen to maximize coverage across families, scripts, and geographic regions, thereby enabling us to study how SRCMIX performs under varying degrees of typological relatedness and orthographic similarity.

5.2 Models

We use a multilingual NMT (mT5-large, 1.2B) and a multilingual LLM (Aya-101, 13B), that are most performant on our chosen ELRLs, to evaluate our proposed training strategies. For NMT, we use the multilingual variant of mT5-large. We refer to Aya-101 as a multilingual LLM (in line with prior work (Ghorbanpour et al., 2025; Ermis et al., 2024; Tao et al., 2024; Manchanda et al., 2024)), as it has been trained with general-purpose instructions, supports zero-shot prompting, and offers large-scale multilingual coverage, making it particularly suitable for probing cross-lingual generalization in ELRLs. In both cases, we apply PEFT using LoRA adapters (Hu et al., 2022), which enables efficient training under ELR constraints without updating the full model parameters.

Models such as NLLB (Costa-Jussà et al., 2022), M2M-100 (Fan et al., 2021), or IndicTrans2 (Gala et al., 2023) rely on explicit language tags in their tokenizers, which is infeasible for most of our target languages, thus making direct fine-tuning with these models impractical. In contrast, mT5 and Aya-101 allow flexible adaptation without predefined vocabularies or identifiers, making them particularly well-suited for ELRL experiments.

As we mentioned in the introduction, we also considered LLaMA-3.1 (8B) (Dubey et al., 2024) and Gemma-7B (Team et al., 2024), but excluded them from our main experiments due to consistently poor performance on ELRLs (as shown in §A.3). Our final evaluations, therefore, focus on Aya-101 and mT5-Large, which demonstrate stronger and more consistent performance across diverse ELRLs.

5.3 Training Details

We fine-tune models using parameter-efficient fine-tuning (PEFT) techniques, which allow effective adaptation of LLMs to ELRL tasks with low computational cost. Specifically, we use LoRA, which freezes pretrained model weights and introduces trainable rank-decomposition matrices within transformer layers. We follow prior work in applying LoRA to only the query and value projections in self-attention modules (Alves et al., 2023; Xin et al., 2024; Zhang et al., 2024), using a rank of $r = 16$ and a scaling factor of $\alpha = 32$. See Appendix A.4.

5.4 Evaluation

We evaluate translation quality using BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2017), computed with the standard `sacreBLEU` implementation (Post, 2018). While BLEU remains the most widely reported metric in MT, it is known to underperform in low-resource and morphologically rich settings due to its reliance on exact n-gram overlap (Post, 2018). In contrast, ChrF++, which computes a character-level F-score, has been shown to correlate more strongly with human judgments in these scenarios (Mathur et al., 2020; Freitag et al., 2022), making it particularly suitable for ELRLs, where minor surface-form differences may obscure translation adequacy.

We avoid learned metrics such as BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2022), as their training data do not cover our target ELRLs. The COMET documentation explicitly notes that "results for language pairs containing uncovered languages are unreliable," and prior work shows BLEURT performs inconsistently in ELRLs (Kocmi et al., 2021). For completeness, we report BLEURT scores in §A.5.

6 Results and Analysis

6.1 Results

We report the main results for Aya-101 and mT5-large, which consistently outperform other open-source LLMs such as LLaMA-3.1 and Gemma-7B. The results across each language group for LLaMA-3.1 and Gemma-7B are provided in §A.3, where we observe substantially lower performance, likely due to their limited multilingual coverage (Dubey et al., 2024; Team et al., 2024).

For automatic evaluation, we follow our setup in §5.4 and primarily report BLEU and ChrF++, while BLEURT scores are reported in §A.5. In addition to bilingual fine-tuning and our proposed mixing strategies, we also evaluate Aya-101 in a conventional M2M setup and zero-shot setting. Both setups have lower performance than SRCMIX, confirming that naive concatenation and zero-shot generalization are inadequate for ELRLs MT.

For reference, we include prior baselines from Maillard et al. (2023), trained on the same NLLB Seed corpus, except for Angika (anp_Deva), which has no existing benchmark. Finally, we conduct human evaluation on a subset of Angika translations, focusing on fluency and adequacy, with details in §A.6.

HRL-to-ELRLs Table 1 presents the translation results from a high-resource language (English) into ELRLs across all target languages, evaluated under five training setups: Zero-shot, M2M, Standard FT, SRCMIX, and TGTMIX. We report results for Aya-101 (an LLM model), while results for mT5-Large (an NMT model) are provided in §A.7 (Table 9). Importantly, we establish the first MT benchmark for Angika (anp_Deva), an Indic ELRL for which no prior baselines exist.

Compared to the baseline from Maillard et al. (2023), Aya-101 achieves an average improvement of +3 ChrF++ across all languages. Our proposed SRCMIX strategy further boosts performance, achieving +6.7 ChrF++ over the baseline. Notably, when comparing naive many-to-many (M2M) training to SRCMIX, we observe substantial gains of +16.2 ChrF++ and +11.4 BLEU. These results underscore the limitations of simple multilingual concatenation, which suffers from the well-known *curse of multilinguality* (Conneau et al., 2020; Wang et al., 2020). By contrast, SRCMIX leverages related source-side languages while maintaining a fixed ELRL target, leading to more effective cross-lingual transfer.

MRL-to-ELRLs Table 2 shows results for Hindi (MRL) → ELRL translation using Aya-101, with mT5-Large results in §A.8 (Table 10). No prior baselines exist for these pairs.

Aya-101 achieves strong performance under both M2M and SRCMIX. On average, SRCMIX yields the best results, with much higher BLEU (23.0) than M2M (14.8) or Standard FT (11.0). Although M2M achieves similar (1.5) ChrF++ to SRCMIX, it underperforms in BLEU, indicating that naive multilingual concatenation captures character-level similarity but not word-level fidelity. This ef-

Direction: English → XXX

Language	Baseline	Zero-shot		M2M		Standard FT		SRCMIX		TGTMIX	
Code	ChrF++	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
fuv_Latn	16.6	10.6	1.5	7.4	0.4	21.5	**7.1**	**24.6**	6.4	5.2	1.3
nus_Latn	21.8	4.7	1.3	11.4	0.7	22.9	11.8	**24.6**	**14.2**	18.3	13.5
taq_Latn	15.2	13.4	1.6	11.5	0.9	16.5	3.6	**18.6**	**13.2**	6.6	1.2
bam_Latn	19.9	13.2	1.6	10.5	0.9	28.0	12.8	**27.0**	**18.5**	6.3	2.4
fur_Latn	35.4	20.9	2.8	4.8	0.3	30.4	16.9	**50.2**	**29.9**	19.8	2.4
lij_Latn	**34.1**	24.5	4.3	5.2	0.5	25.5	9.9	25.1	**11.1**	19.1	6.3
lim_Latn	**30.0**	23.3	3.2	6.4	0.4	29.1	5.2	29.6	**9.9**	18.6	1.5
srd_Latn	35.6	27.1	6.0	3.3	0.2	34.1	12.7	**37.5**	**19.1**	27.7	7.9
anp_Deva	–	1.0	0.4	27.2	20.3	42.3	17.9	**44.1**	**20.2**	3.6	3.7
bho_Deva	21.9	12.1	2.4	26.1	**11.9**	31.8	8.5	**33.1**	9.3	5.0	5.7
mag_Deva	27.1	1.9	1.6	46.9	21.6	45.9	20.2	**53.6**	**38.0**	1.9	4.8
kas_Arab	19.2	8.9	3.5	6.9	0.2	24.5	**13.5**	**23.3**	10.3	24.3	7.0
prs_Arab	24.1	13.4	3.1	28.6	9.2	25.5	9.9	**28.3**	**15.3**	16.2	2.4
pbt_Arab	21.9	6.0	2.6	19.6	3.8	11.4	3.3	**22.2**	**15.3**	16.7	4.8
Avg	24.8	13.0	2.6	15.4	5.1	27.8	11.0	**31.6**	**16.5**	13.5	4.6

Table 1: Results for Aya-101 across Zero-shot, M2M, Standard Fine-tuning, SRCMIX, and TGTMIX. Baselines from Maillard et al. (2023) are shown, except for Angika (anp_Deva), which has no prior baseline.

Direction: Hindi → XXX

Language	Baseline	Zero-shot		M2M		Standard FT		SRCMIX		TGTMIX	
Code	ChrF++	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
anp_Deva	–	2.5	0.4	50.7	26.6	44.9	20.5	**52.6**	**28.7**	16.1	2.6
bho_Deva	–	18.8	5.1	**37.8**	12.4	27.7	6.8	35.9	**25.2**	19.9	6.0
mag_Deva	–	19.5	3.5	**47.4**	**21.5**	46.0	20.0	46.0	20.3	33.5	6.9
kas_Arab	–	7.1	0.3	20.8	3.9	24.2	8.7	**32.6**	**24.9**	11.9	2.7
prs_Arab	–	7.2	0.3	30.9	8.8	28.4	7.4	**27.8**	**15.1**	14.9	2.9
pbt_Arab	–	1.5	0.3	24.3	15.6	18.3	2.4	**26.2**	**23.9**	20.7	4.3
Avg	–	9.4	1.7	35.3	14.8	31.6	11.0	**36.8**	**23.0**	19.5	4.2

Table 2: Results for Aya-101 across Zero-shot, M2M, Standard Fine-tuning, SRCMIX, and TGTMIX. No prior baseline exists.

fect is especially visible for Indic ELRLs, where Hindi's typological and geographic proximity provides stronger character-level transfer than English in the HRL→ELRL setting.

For mT5-Large, SRCMIX and Standard FT perform similarly in ChrF++, but SRCMIX provides a slight BLEU advantage. Target mixing consistently underperforms, confirming that merging multiple ELRL targets dilutes decoder specialization. Overall, these findings show that leveraging a mid-resource language like Hindi with SRCMIX provides substantial gains, and that the benefits generalize across both Aya-101 and mT5-Large.

ELRLs-to-HRL. Table 3 presents results for translation from ELRLs→English, comparing Aya-101 and mT5-Large. As discussed in §3, TGTMIX is not applicable here since there is only one target language. Moreover, SRCMIX differs fundamentally from the HRL→ELRL case: instead of mixing diverse HRL sources, it combines multiple ELRL→English pairs (e.g., Angika→English, Magahi→English), offering limited scope for source-side multilinguality. This effectively reduces the task to a many-to-one setting where the target (English) is uniform and the encoder receives weak, low-resource inputs, limiting potential gains.

Interestingly, unlike HRL→ELRL, zero-shot translation from ELRLs into English sometimes matches or outperforms Standard FT setups. For instance, in Friulian (fur_Latn), Ligurian (lij_Latn), and Sardinian (srd_Latn), zero-shot Aya-101 achieves higher ChrF++ scores than both M2M and Standard FT. This trend supports prior findings that translation into HRL is easier because pretrained models possess stronger target-side representations of English (Guzmán et al., 2019; Conneau et al., 2020).

Overall, these results indicate that ELRL→English translation is more forgiving, with even zero-shot inference producing usable outputs for certain languages. Nonetheless, SRCMIX still provides consistent, though modest, improvements in several groups, and Aya-101 remains the strongest performer in this configuration.

Direction: XXX → English

		AYA-101								MT5-Large			
Language	Baseline	Zero-shot		M2M		Standard FT		SrcMix		Standard FT		SrcMix	
Code	ChrF++	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
fuv_Latn	19.8	21.6	5.1	17.8	1.8	26.9	**14.6**	**30.5**	12.9	16.6	2.7	16.4	2.4
nus_Latn	20.1	12.7	2.2	16.7	1.4	26.6	**9.7**	**33.9**	8.0	14.3	1.9	14.6	1.5
taq_Latn	19.4	14.0	2.5	18.0	2.0	17.1	1.5	**19.6**	**2.9**	16.1	17.0	2.0	2.5
bam_Latn	20.7	23.9	7.1	18.1	2.2	18.9	2.9	**35.7**	**15.0**	8.1	0.0	17.4	2.5
fur_Latn	35.6	**38.3**	**34.2**	17.3	1.3	36.2	17.6	32.0	11.9	31.6	12.3	29.0	7.0
lij_Latn	32.1	35.0	17.0	17.6	1.7	**39.6**	**21.3**	31.3	12.3	22.7	12.0	21.6	10.1
lim_Latn	30.7	**33.8**	14.0	17.7	1.9	29.9	14.4	28.4	13.1	27.6	12.5	25.9	**22.6**
srd_Latn	34.0	**46.0**	**23.0**	17.1	1.4	38.4	20.2	33.8	14.5	19.3	4.7	32.0	12.5
anp_Deva	–	21.3	9.7	47.9	**22.4**	47.6	21.9	**48.0**	22.1	29.4	9.7	28.1	9.1
bho_Deva	24.1	29.3	22.5	46.6	21.1	**47.8**	**21.7**	46.2	19.9	33.9	12.8	26.1	8.6
mag_Deva	28.8	53.8	30.5	49.7	25.8	**58.3**	**33.9**	56.0	31.2	39.8	16.1	43.3	14.8
kas_Arab	22.8	21.7	10.1	29.7	8.9	**30.9**	**24.2**	25.9	15.9	23.4	16.8	25.2	10.5
prs_Arab	28.5	**48.4**	**24.7**	32.0	13.8	36.4	17.4	34.2	14.0	29.2	11.0	29.0	10.5
pbt_Arab	24.1	1.4	1.7	31.3	12.5	**31.7**	12.6	30.5	**18.3**	25.1	7.4	23.4	6.3
Avg	26.2	28.7	14.6	27.0	8.4	**34.8**	**16.7**	34.7	14.6	24.1	8.7	25.0	8.6

Table 3: Comparison of Aya-101 and mT5-Large across Zero-shot, M2M, Standard Fine-tuning, and Source Mixing. Baselines are from Maillard et al. (2023), except for Angika (anp_Deva), which has no prior baseline.

6.2 Analysis

What Makes SrcMix Effective. The effectiveness of SrcMix stems from *decoder specialization*. In this setup, the decoder is always trained to generate a single ELRL target while receiving inputs from multiple related source languages without explicit language identifiers. This forces the model to abstract away from surface-level lexical differences and instead capture deeper structural and semantic regularities across related sources. As a result, the decoder produces more fluent and consistent target-language outputs.

By contrast, TgtMix requires the decoder to handle multiple low-resource targets from a fixed source. Even when tags are introduced, as shown in our (§A.10), the decoder struggles to disambiguate diverse target distributions with minimal supervision. This leads to degraded performance due to decoder confusion, more substantial negative interference among typologically distant ELRLs, and inadequate language tags as reliable conditioning signals. A similar weakness is observed in M2M training, where naive concatenation of all source–target pairs results in negative transfer and diluted supervision. Both TgtMix and M2M resemble simple one-to-many and many-to-many setups, where the decoder must generalize across diverse targets or sources, reducing its ability to specialize in any single ELRL. In contrast, SrcMix arrives at the decoder on one ELRL, leveraging source-side diversity while avoiding interference.

To further isolate the cause of performance gains, we conduct two ablations. First, we compare SrcMix with Standard FT under a *fixed training step* (§7), ensuring both models are trained for the same number of updates. The improvements of SrcMix persist under this setting, showing that its advantage is not merely due to longer exposure or data volume, but rather its ability to exploit cross-lingual structural alignment. Second, we control for dataset size (§A.11), confirming that gains are not solely attributable to additional training pairs but arise from the structural signal introduced by related sources.

Model Comparisons (Aya vs. mT5). A cross-model comparison between Aya-101 (LLM) and mT5-Large (NMT) highlights complementary strengths and limitations. On average, Aya-101 achieves substantially higher BLEU scores, indicating better lexical fidelity and word-level accuracy. This is consistent with its larger capacity (13B), instruction tuning, and wider multilingual coverage, which make it better at capturing cross-lingual mappings in ELRL scenarios. In contrast, mT5-Large often produces higher or comparable ChrF++ scores, particularly in morphologically rich Indic and Arabic ELRLs. This suggests that despite its smaller size (1.2B parameters), mT5's encoder–decoder structure is better at preserving character-level correspondences and handling morphological variation.

These trends highlight a trade-off: Aya-101 generalizes better at the word level, while mT5 offers robustness at the subword and character level. Both models benefit consistently from SrcMix. However, Aya-101 gains more from multilingual supervision in HRL→ELRL and ELRL →HRL directions, whereas mT5 shows competitive performance in Hindi→ELRL settings, where typological proximity and script similarity play a stronger role. This divergence suggests that the optimal choice of model may depend on the typological properties of the target ELRL and the available supervision, with LLM-based models excelling at cross-lingual generalization and semantic consistency, while NMT models remain stronger at preserving surface-form fidelity.

Typology Matters. We find that the gains from SrcMix are strongest when source languages share scripts or belong to the same group, underscoring the importance of structural and typological compatibility in enabling effective transfer. In contrast, mixing unrelated languages from different groups having similar scripts leads to a substantial drop in performance (Table 6 in §7). In contrast, using typologically distant or cross-script sources leads to sharp performance drops, even when the dataset size is held constant (§A.11). confirming that the improvements are not simply due to more data, but from meaningful cross-lingual alignment.

Why Naive Mixing Fails in ELRLs. In ELRL settings, naively mixing data from multiple languages often degrades performance instead of improving it. Due to the minimal parallel data available (∼6K sentences per language), multilingual models are susceptible to noise, and mixing unrelated or loosely related languages introduces conflicting lexical and syntactic signals, leading to negative transfer. This issue is more pronounced when languages share scripts, as many-to-many training often results in source copying or mixed-script outputs. Moreover, naive mixing spreads decoder capacity across multiple target languages, preventing effective specialization for the ELRL. In contrast, SrcMix restricts variation to the source side while fixing the target language, enabling controlled transfer from related languages without weakening the target distribution.

7 Ablation Analysis

To perform the ablation study, we selected four distinct languages, each representing a different language group and script, to ensure that the findings generalize across diverse linguistic settings.

Effect of Source Anonymization. To isolate the role of source language anonymization in SrcMix, we conducted an ablation with Aya-101, comparing performance with and without source language tags. Table 4 shows that anonymizing source identities significantly boosts both ChrF++ and BLEU scores, with average gains of +15.3 ChrF++ and +6.5 BLEU. These results confirm that removing source language tags reduces token-level overfitting and helps the decoder generalize across structurally related but lexically distinct inputs, especially beneficial in ELR settings.

Language Code	SrcMix With Lang Tag		SrcMix Without Lang Tag	
	ChrF++	BLEU	ChrF++	BLEU
fur_Latn	23.4	14.9	**50.2**	**30.0**
fuv_Latn	17.2	2.4	**24.6**	**6.4**
bho_Deva	12.7	6.8	**33.0**	**9.3**
prs_Arab	21.5	10.6	**28.3**	**15.1**
Avg	18.7	8.7	**34.0**	**15.2**

Table 4: Effect of source language anonymization in SrcMix with Aya-101 for English→ELRL translation.

Language Code	Standard FT		SrcMix	
	ChrF++	BLEU	ChrF++	BLEU
fuv_Latn	11.3	0.8	**11.4**	**1.2**
fur_Latn	24.3	7.5	**30.2**	**11.6**
bho_Deva	**34.7**	**10.7**	34.6	10.5
prs_Arab	27.2	8.2	**30.6**	**10.3**
Avg	24.4	6.8	**26.7**	**8.4**

Table 5: Comparison of fixed training step performance between Standard fine-tuning and SrcMix with Aya-101 for English→ELRL.

Impact of Training Steps on Performance. To ensure that the improvements of SrcMix are not simply due to longer training, we conducted an ablation where both Standard FT and SrcMix are trained for exactly 10k steps (Table 5). Even under this fixed training budget, SrcMix outperforms Standard FT on average (+2.3 ChrF++, +1.6 BLEU). For languages such as Friulian (fur_Latn) and Dari (prs_Arab), SrcMix shows gains in both metrics, while in Bhojpuri (bho_Deva), performance remains comparable. These results confirm that the gains of SrcMix arise from structural cross-lingual transfer rather than additional training steps, reinforcing its effectiveness for ELRL translation.

Utility of Related Languages in SrcMix. To evaluate whether linguistic proximity plays a critical role in the success of SrcMix, we conducted an ablation contrasting configurations where source languages were selected to be typologically and

Direction: English → XXX

	SrcMix		SrcMix	
Language	Different group		Same group	
Code	ChrF++	BLEU	ChrF++	BLEU
fur_Latn	28.0	9.6	**50.2**	**29.9**
fuv_Latn	16.1	6.0	**24.6**	**6.4**
bho_Deva	**33.7**	**9.9**	33.0	9.3
prs_Arab	25.7	6.7	**28.3**	**15.1**
Avg	25.9	8.1	**34.0**	**15.2**

Table 6: Effect of source-language relatedness on English→ELRL translation with Aya-101. We compare SrcMix performance using within-language group vs. cross-group sources.

orthographically closer ("Same group") versus distant ("Different group"). In the standard SrcMix setup, we mix source languages that share a common script and originate from geographically proximate regions. This encourages the decoder to align on features more likely to transfer effectively across related ELRLs. For comparison, we selected four representative ELRLs and trained SrcMix models with unrelated source languages from different typological profiles and scripts, while keeping the overall data size constant.

As shown in Table 6, using more distant and orthographically diverse sources leads to an average drop of −8.17 ChrF++ and −7.09 BLEU compared to the "Same group" setup. This suggests that typological and orthographic proximity among sources contributes substantially to effective cross-lingual transfer. To further probe this, we ran an additional ablation with typologically distant languages that share the same script (e.g., African and European languages in Latin script), reported in Appendix A.12. This helps clear whether gains arise primarily from genealogical relatedness, script compatibility, or both.

8 Conclusion

In this work, we introduced SrcMix, a simple yet scalable training-time strategy for machine translation in ELRLs. Unlike prior many-to-one or massively multilingual approaches, SrcMix explicitly leverages structural transfer by combining multiple related source languages without relying on explicit language identifiers. This design makes it particularly well-suited for ELRL settings.

Our analysis underscores the importance of multilingual data, directionality of language mixing, and typological compatibility. SrcMix consistently improves translation performance, particularly for languages within the same group and shared script. Through comprehensive experiments on 14 ELRLs spanning four language groups and three scripts, we further compare NMT and LLM model under ELR conditions, offering insights into their strengths and limitations.

Future work can extend our approach to additional languages, explore adaptive training, and further improve cross-lingual transfer in ELRL MT.

Limitations

- Our evaluation focuses on four models: Aya-101, mT5-large, LLaMA-3.1 (8B), and Gemma-7B, but does not explore other multilingual architectures that may perform differently. Moreover, while we use PEFT techniques like LoRA, limited computational resources restrict extensive hyperparameter tuning, which could influence performance in ELRL settings.

- Although our study covers 14 ELRLs from diverse language families, it still represents only a small fraction of the world's underrepresented languages, limiting the generalizability of our findings across all ELRLs.

- In our work, we utilize pretrained language models that were not specifically trained on our languages of interest. As a result, there may be instances where the nuances of these languages are not fully captured, potentially leading to a loss of context in the translations. Additionally, the relatively small size of our datasets, including the Angika MT dataset, constrains the models' ability to learn complex linguistic patterns, affecting translation quality.

- Unlike massively multilingual systems, which train a single model across dozens of targets, SrcMix deliberately trains separate models for each ELRL. This design prioritizes translation quality and avoids the negative transfer often observed in multilingual setups under ELRL conditions.

Acknowledgements

The first author gratefully acknowledges Ph.D. fellowship support from the TCS Research Foundation for conducting research on extremely low-resource Indian languages. The second author gratefully acknowledges financial support from the

Amazon–IITB AI/ML Initiative. The authors thank the anonymous reviewers for their insightful and constructive comments.

References

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, and 26 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Duarte Alves, Nuno Guerreiro, Jo
textasciitilde ao Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2024. Can llms really learn to translate a low-resource language from one grammar book? *arXiv preprint arXiv:2409.19151*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Akiko Eriguchi, Shufang Xie, Tao Qin, and Hany Hassan. 2022. Building multilingual machine translation systems that serve arbitrary XY translations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 600–606, Seattle, United States. Association for Computational Linguistics.

Beyza Ermis, Luiza Pozzobon, Sara Hooker, and Patrick Lewis. 2024. From one to many: Expanding the scope of toxicity mitigation in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15041–15058, Bangkok, Thailand. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Faeze Ghorbanpour, Daryna Dementieva, and Alexandar Fraser. 2025. Can prompting LLMs unlock hate speech detection across languages? a zero-shot and few-shot study. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 413–425, Vienna, Austria. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409, Miami, Florida, USA. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2025. Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions. *Transactions of the Association for Computational Linguistics*, 13:1032–1055.

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. Part-of-speech tagging for extremely low-resource Indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14422–14431, Bangkok, Thailand. Association for Computational Linguistics.

Garry Kuwanto, Eno-Abasi E. Urua, Priscilla Amondi Amuok, Shamsuddeen Hassan Muhammad, Anuoluwapo Aremu, Verrah Otiende, Loice Emma Nanyanga, Teresiah W. Nyoike, Aniefon D. Akpan, Nsima Ab Udouboh, Idongesit Udeme Archibong, Idara Effiong Moses, Ifeoluwatayo A. Ige, Benjamin Ajibade, Olumide Benjamin Awokoya, Idris Abdulmumin, Saminu Mohammad Aliyu, Ruqayya Nasir Iro, Ibrahim Said Ahmad, and 5 others. 2024. Mitigating translationese in low-resource languages: The storyboard approach. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11349–11360, Torino, Italia. ELRA and ICCL.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Jiya Manchanda, Laura Boettcher, Matheus Westphalen, and Jasser Jasser. 2024. The open source advantage in large language models (llms). *arXiv preprint arXiv:2412.12004*.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Robert Östling, Yves Scherrer, Jörg Tiedemann, Gongbo Tang, and Tommi Nieminen. 2017. The Helsinki neural machine translation system. In *Proceedings of the Second Conference on Machine Translation*, pages 338–347, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book. *arXiv preprint arXiv:2309.16575*.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. 2024. Scaling laws with vocabulary: Larger models deserve larger vocabularies. *Advances in Neural Information Processing Systems*, 37:114147–114179.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.

Chunlei Xin, Yaojie Lu, Hongyu Lin, Shuheng Zhou, Huijia Zhu, Weiqiang Wang, Zhongyi Liu, Xianpei Han, and Le Sun. 2024. Beyond full fine-tuning: Harnessing the power of LoRA for multi-task instruction tuning. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2307–2317, Torino, Italia. ELRA and ICCL.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *arXiv preprint arXiv:2410.03115*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. 2024. Personalized lora for human-centered text understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19588–19596.

Tong Zheng, Yan Wen, Huiwen Bao, Junfeng Guo, and Heng Huang. 2025. Asymmetric conflict and synergy in post-training for LLM-based multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18362–18383, Vienna, Austria. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Appendix

A.1 Angika Language

Angika (anp) is an Eastern Indo-Aryan language spoken primarily in the Indian states of Bihar and Jharkhand, with some presence in West Bengal and Nepal. It has approximately 15 million speakers. Historically, it was written in Anga Lipi, later replaced by Kaithi, and is now written in Devanagari. Like Hindi, Angika follows a Subject-Object-Verb (SOV) structure but has distinct grammatical features. For example, in assertive sentences, Hindi uses *Main aam khaata hoon*(I eat mango.), whereas Angika expresses the same as *Hammae aam khaay chiyai*. In negative sentences, Hindi uses *Aap/tum nahin ja rahe hain/ho*(You are not going), while Angika says *Aapanae/tonae nai jaay rahlo cho/chai*. Similarly, interrogative sentences differ, as seen in *Hamlok USA kab jaayenge?* (When will we go to USA) in Hindi and *Hamrasini USA kahia jaibai?* in Angika. Pluralization in Angika is marked by *-sini*, whereas Hindi uses different forms; for instance, *Kitaabon* (Book) in Hindi becomes *Kitaab sini*, *Hamlog* becomes *Hamra sini*, and *Ve log* becomes *Hunka/Okra sini*. Emphasis is also marked differently in Angika—unlike Hindi, which uses *hi*, Angika modifies nouns and pronouns using *-e/ai* and *-i/hi*. For example, *Ram hi Shyam hai.* (Ram is Shyam.) in Hindi is *Rame Shyam chekae* in Angika, and *Maine jise dekha, wah Ram hi tha.* (The one I saw was Ram himself.) translates to *Hammae jekra dekhliyai, u Rame chelai.* Angika also differs in expressing *bhi*, using *-o/au* for nouns and *-u/au* for pronouns; thus, *Main bhi* in Hindi is *Hammu* in Angika, *Train bhi* is *Traino*, *Tum bhi* is *Tahu*. Additionally, Angika does not differentiate gender in verb forms, making it grammatically gender-neutral. For example, *Lata khayegi.* (Lata will eat.) or *Sunil khayega.* (Sunil will eat.) in Hindi both become *Latane khaitai* or *Sunilne khaitai* in Angika. Another key feature is Angika's unique case markers: the subject marker (ne) becomes *-ae ni / -ne as in Ramae ni / Ramnae*, the object marker (*ko*) becomes *-k*, as in *Ramo k*, the instrumental/ablative marker *(se/ke dwara)* becomes *-s / delai*, as in *Ramo se / delai*, the dative marker (ke liye) is *-leli / bastae / l*, as in *Khaay leli / Khaay lae / Khaay bastae*, and the locative marker *(me/par)* is *-m / p*, as in *Jalo m*. While Angika shares some vocabulary and structural similarities with Hindi and other Indo-Aryan languages, it retains distinct linguistic features that set it apart.

A.2 Quality Control

To assess the quality of the translation task of Angika, we abstained from computing automatic inter-annotator agreement (IAA) measures such as Cohen's or Fleiss' kappa (Fleiss, 1971). However, these metrics are designed for categorical data, while our translation data is sequence-based. Since our two annotators worked on alternating segments, direct sentence-level overlap was limited. During the peer-review verification process, annotators marked each other's translations, resulting in an IAA of 85% for training data and 91% for test data. Most disagreements occurred at the word level, influenced by regional language variations and the increasing influence of high-resource languages on Angika. These were resolved during the final verification process. Recent MT dataset creation efforts, such as Guzmán et al. (2019) and

Kuwanto et al. (2024), prioritize reviewer-driven verification rather than traditional IAA metrics like Cohen's or Fleiss' kappa. Our approach aligns with these established practices, ensuring translation quality through careful reviewer validation and agreement checks. We believe this approach aligns with established practices in machine translation evaluation.

Compensation and Ethical Considerations. We adhere to fair and regionally appropriate compensation practices for all language contributors. For Angika, an extremely low-resource language with limited access to professional translators, we offer $0.20 per sentence, reflecting standard rates in the region for linguistic annotation and translation tasks. For Hindi, where initial translations are generated via Google Translate, we compensate human verifiers at $0.10 per sentence for their role in reviewing and correcting the outputs. These rates are consistent with compensation practices in prior research on low-resource language technologies and ensure ethical engagement with speakers and translators while supporting ongoing efforts in language documentation and inclusivity in NLP.

A.3 LLaMA and Gemma Results

We compared LLaMA-3.1 , Gemma-7B, Aya-101, and mT5-Large under the English→ELRL setting. Table 7 shows that, in the Standard FT setup, LLaMA-3.1 and Gemma-7B substantially underperform Aya-101 and mT5-Large across all evaluated languages. The only partial exception is Dari (prs_Arab), where LLaMA-3.1 achieves a +5.08 improvement in ChrF++ over mT5-Large, but at the cost of a –6.35 BLEU drop, highlighting weak word-level accuracy despite some character-level overlap.

When comparing SrcMix to Standard FT, both LLaMA-3.1 and Gemma-7B exhibit modest gains across most languages, consistent with the pattern observed for Aya-101. Bhojpuri remains the only exception where SrcMix underperforms Standard FT across all models, indicating that language-specific factors may limit the benefits of source-side mixing in this case. Despite these relative improvements, the absolute performance of LLaMA-3.1 and Gemma-7B remains far below Aya-101 and mT5-Large, with the gap particularly pronounced in BLEU. This suggests that while decoder-only LLMs may capture some higher-level character correspondences (as reflected in ChrF++), these do not translate into reliable word- or phrase-level matches.

Overall, LLaMA-3.1 and Gemma-7B struggle with ELRL translation. We attribute this to their limited exposure to such languages—or even closely related HRLs—during pretraining, resulting in weak generalization to unseen ELRLs. The model exhibits common issues such as hallucinations, repetitive outputs, and copying the source text, negatively impacting translation quality. While it performs reasonably well on high-resource language pairs, its effectiveness drops sharply in high-resource-to-ELRL translation tasks. This indicates that its multilingual capabilities are limited and not well-suited for scenarios with extreme data scarcity. These findings highlight the limitations of general-purpose decoder-only LLMs without targeted multilingual pretraining or task-specific adaptation, and underscore the need for specialized strategies tailored to ELRL scenarios.

A.4 Experiment setup

The experiment setup consists of the following hyperparameter configurations. We use a batch size of 32. The learning rate is set to 1e-5. The maximum sequence length is 128. For LoRA fine-tuning, we set the LoRA rank to 16 and the LoRA alpha to 32. To access the MT performance, we use Scarebleu's implementation of ChrF++ and BLEU. The model is trained for a total of 5 epochs. All of our experiments were conducted on Nvidia A100 (80G) GPUs. Our Standard FT training time for Aya-101 is approximately 2 hours, while for mT5-large, it takes around 25 minutes. With SrcMix and TgtMix, Aya-101 requires 6 hours of training, whereas mT5-large completes training in 60 minutes. Across all training methods, the inference time for Aya-101 is 25 minutes, while for mT5-large, it is 12 minutes.

A.5 BLEURT Scores

We compared BLEURT, ChrF++, and BLEU across diverse ELRL translation directions (see Table 8). BLEURT produces reasonable trends for Indo-Aryan languages such as Angika and Bhojpuri, likely due to their shared Devanagari script and proximity to Hindi, a language included in BLEURT's pretraining corpus. However, it behaves inconsistently for typologically distant or underrepresented languages like Friulian and Sardinian, where it yields very low or even negative scores despite clear gains in BLEU and ChrF++.

Language Code	Aya-101 Standard FT		mT5-Large Standard FT		LLaMA-3.1 Standard FT		Gemma-7B Standard FT		LLaMA-3.1 SrcMix		Gemma-7B SrcMix	
	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
bho_Deva	**31.8**	8.5	24.5	**15.9**	29.0	3.4	1.9	0.3	27.6	2.8	1.5	0.2
prs_Arab	23.9	6.1	24.6	**14.1**	29.7	3.9	1.3	0.2	**30.1**	4.2	4.0	0.5
fur_Latn	30.4	**17.0**	16.0	4.9	28.1	5.1	11.9	0.3	**36.6**	6.1	20.6	2.3
fuv_Latn	**21.5**	**7.1**	7.9	3.8	11.0	0.8	9.3	0.1	13.6	0.8	12.5	0.7
Avg	**27.0**	**10.0**	18.3	9.7	24.5	3.3	6.1	0.2	**27.0**	3.5	9.7	0.9

Table 7: Comparison of Aya-101, mT5-Large, LLaMA-3.1, and Gemma-7B under Standard FT and SrcMix. Best scores per language in **bold**.

Language Pair	Standard FT			SrcMix		
	BLEURT	ChrF++	BLEU	BLEURT	ChrF++	BLEU
Friulian → Eng	47.2	36.2	17.6	**49.1**	**32.1**	**12.0**
Bhojpuri → Eng	**69.2**	**47.8**	**21.7**	66.8	46.2	19.9
Dari → Eng	48.5	**36.4**	17.4	**49.4**	34.2	**14.0**
Nigerian Fulfulde → Eng	29.1	26.9	**14.6**	**31.7**	**30.5**	12.9
Angika → Hindi	**69.0**	**54.4**	25.1	67.4	49.5	**27.0**
Bhojpuri → Hindi	66.4	**41.0**	14.8	**69.3**	40.0	**17.0**
Dari → Hindi	40.9	22.6	16.9	**42.6**	**30.2**	**18.4**
Southern Pashto → Hindi	**44.7**	**41.5**	**25.1**	42.2	34.2	19.0
Eng → Friulian	4.9	30.4	17.0	**5.2**	**50.2**	**30.0**
Eng → Ligurian	9.0	**25.5**	10.0	**12.0**	25.1	**11.1**
Eng → Limburgish	23.2	29.1	5.2	**23.3**	**29.6**	**10.0**
Eng → Sardinian	-13.5	34.0	12.7	**-10.5**	**37.5**	**19.1**
Eng → Angika	56.7	42.3	17.9	**61.2**	**44.1**	**20.2**
Eng → Bhojpuri	59.1	31.8	8.5	**67.6**	**33.0**	**9.3**

Table 8: Comparison of Aya-101 results, Standard FT vs. SrcMix across multiple language pairs using BLEURT, ChrF++, and BLEU.

This highlights the limitations of learned metrics in ELRL settings, particularly when the evaluation language is absent from the metric's pretraining distribution. As noted by Freitag et al. (2022), BLEURT performs well on some low-resource pairs (e.g., Yakut–Russian) in the WMT22 Metrics Task. However, this success is partly due to BLEURT's exposure to Russian during training, and Yakut's script and typological similarity to Russian. A comparable situation may explain BLEURT's stability for Bhojpuri, Magahi, and Angika languages, which benefit from their closeness to Hindi.

In contrast, BLEURT underperforms for languages like Friulian and Sardinian, which are both typologically and orthographically distant from the models seen during pretraining. As Freitag et al. (2022) caution that learned metrics may not be able to generalize across text styles, scripts, and linguistic domains, an issue that becomes more noticeable in ELRL scenarios involving unseen or underrepresented language families.

A.6 Human Evaluation

While our primary evaluation relied on automatic metrics such as BLEU and ChrF++ for their consistency and reproducibility, we recognize that human evaluation provides deeper insights, particularly for ELRLs. However, due to the limited availability of fluent speakers who can read and write most of these languages and the resources required, conducting a comprehensive human evaluation presented logistical challenges. Nevertheless, we carried out a targeted human evaluation for our Angika MT dataset. Specifically, we randomly selected 302 samples (out of 1,012 test examples) and asked native Angika speakers to rate the outputs generated under the SrcMix setup. Native speakers rated each translation for fluency and adequacy on a 1–5 scale. The average fluency score is 3.32, and the adequacy score is 2.66. Here, the higher fluency scores are compared to adequacy in our human evaluation results. This outcome is not uncommon in ELRL translation scenarios, where models may generate grammatically well-formed sentences that appear fluent but occasionally drift away from the intended meaning. Such discrepancies can arise due to limited parallel data (in our case, only 6K), making it challenging for the model to capture fine-grained semantic nuances while still producing natural-sounding text.

A.7 HRL-to-ELRLs

Direction: English → XXX

Language Code	Standard FT		Source Mixed		Target Mixed	
	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
fuv_Latn	7.9	3.8	**19.4**	**6.2**	3.3	0.0
nus_Latn	12.8	6.9	**26.4**	**16.2**	8.6	5.5
taq_Latn	9.9	1.8	**17.6**	**9.1**	3.8	0.2
bam_Latn	13.7	9.4	**13.8**	**10.1**	7.6	0.9
fur_Latn	**16.0**	**4.9**	**16.0**	4.1	20.5	7.2
lij_Latn	17.9	10.7	**23.4**	**11.0**	13.2	1.8
lim_Latn	26.5	5.6	**27.6**	**7.0**	18.9	4.9
srd_Latn	**22.9**	10.9	22.4	**11.9**	8.9	1.2
anp_Deva	**28.3**	14.3	26.0	**15.7**	15.7	8.1
bho_Deva	**24.5**	**15.9**	22.1	12.9	16.1	8.3
mag_Deva	**23.2**	**30.2**	22.0	19.1	20.8	8.7
kas_Arab	14.1	6.0	**14.2**	**7.3**	16.9	8.9
prs_Arab	24.6	14.1	**26.5**	**18.1**	4.6	0.0
pbt_Arab	15.1	**11.4**	**27.3**	11.3	3.5	0.0
Avg.	18.4	10.4	**21.8**	**11.4**	11.6	4.0

Table 9: Results for MT5-Large across Standard Fine-tuning, SrcMix, and TgtMix.

Table 9 shows the mT5-large results in direction HRL →ELRL. When comparing the same translation direction (Eng→ELRL) across models, we observe that SrcMix improves performance for mT5 compared to Standard fine-tuning. For several languages such as `nus_Latn`, `bho_Deva`, and `prs_Arab`, mT5 in the SrcMix setup achieves higher performance than Aya-101 under Standard fine-tuning. Similarly, for `pbt_Arab`, mT5 achieves better results than Aya-101 in the Standard FT configuration.

However, in terms of overall gains, Aya-101 benefits more from SrcMix than mT5. While mT5 shows improvements mainly at the character level (reflected in ChrF++), its BLEU score gains are relatively smaller. This suggests that, as a smaller model, mT5 is better at recovering character-level patterns but struggles with capturing longer n-gram matches compared to the larger Aya-101. In contrast, Aya-101 demonstrates stronger improvements across both ChrF++ and BLEU, indicating its capacity to leverage structural transfer more effectively under the SrcMix setup.

A.8 MRL-to-ELRLs

Table 10 reports mT5-large results for the MRL →ELRL direction (Hindi→ELRL). Comparing across models, we find that SrcMix improves mT5 mainly in terms of ChrF++, whereas Aya-101 shows larger gains in BLEU. This suggests that mT5 captures character-level patterns but struggles with higher n-gram, while Aya-101 benefits more strongly from SrcMix in word-level matching.

Direction: Hindi → XXX

Language Code	Standard FT		Source Mixed		Target Mixed	
	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
anp_Deva	48.0	18.9	**49.8**	**23.4**	17.3	5.4
bho_Deva	**27.6**	**11.4**	25.5	9.3	19.8	7.6
mag_Deva	41.4	**18.1**	**45.4**	15.4	33.5	6.6
kas_Arab	13.4	6.2	**18.6**	**8.6**	8.1	3.4
prs_Arab	**30.9**	**18.4**	25.1	9.8	10.2	4.0
pbt_Arab	13.9	8.9	**23.8**	**15.9**	17.9	11.2
Avg	29.2	13.6	**31.4**	**13.7**	17.8	6.4

Table 10: Results for Hindi → ELRLs using MT5-Large across Standard FT, Source Mixing, and Target Mixing.

For most Indic ELRLs such as `anp_Deva`, `bho_Deva`, and `mag_Deva`, performance drops with mT5 under SrcMix. One likely reason is that Hindi is both typologically and script-wise very close to these target languages. Since mT5 was pretrained on large amounts of Hindi data, the model may overgeneralize and fail to adequately distinguish between Hindi and these closely related ELRLs. In contrast, when evaluating on Arabic-script languages such as `kas_Arab` and `pbt_Arab`, we observe consistent improvements under SrcMix for both ChrF++ and BLEU. This indicates that script diversity helps mT5 leverage structural transfer more effectively, avoiding overgeneralization.

A.9 ELRLs-to-MRL

Tables 11 present the translation results for ELRL-to-MRL translations across all six languages. TgtMix is not applicable for ELRL-to-HRL and ELRL-to-MRL setups due to the presence of only one target language.

Furthermore, the SrcMix differs in this setting. In HRL/MRL→ELRL translation, SrcMix combines training data from HRLs together with multiple ELRLs as sources, all pointing to a single ELRL target. In contrast, for ELRL→HRL/MRL, SrcMix aggregates training data from multiple ELRL sources but constrains generation to a single HRL/MRL target. As a result, the decoder always specializes on one target distribution, and performance becomes comparable to Standard FT.

A.10 TgtMix with Language Tags

To validate the effectiveness of TgtMix with target language tags, we conducted additional experiments incorporating explicit language tags in the target-side multilingual fine-tuning setup. Specifically, we prepended language identifiers (e.g., ‘<magahi>‘, ‘<angika>‘) to the decoder input to in-

Direction: XXX → Hindi

		AYA-101								MT5-Large			
Language	Baseline	Zero-shot		M2M		Standard FT		SrcMix		Standard FT		SrcMix	
Code	ChrF++	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
anp_Deva	–	41.8	18.1	47.7	24.6	**54.4**	25.1	49.5	**27.0**	49.1	21.7	39.0	10.0
bho_Deva	–	31.8	13.3	40.4	16.7	**41.0**	14.8	40.0	**17.0**	33.6	13.7	29.8	13.6
mag_Deva	–	46.9	21.4	49.7	25.8	**50.9**	**27.8**	50.7	27.3	49.7	26.6	50.0	26.8
kas_Arab	–	3.1	1.1	19.3	4.2	27.6	16.0	**30.5**	**20.9**	16.7	8.9	18.4	12.5
prs_Arab	–	33.7	14.5	21.0	5.9	22.6	16.9	**30.2**	**18.4**	15.0	2.6	12.8	1.1
pbt_Arab	–	23.0	8.4	21.6	6.0	**41.5**	**25.1**	34.2	19.0	23.0	23.8	19.7	8.6
Avg	–	30.0	12.8	33.3	13.9	**39.7**	21.0	39.2	**21.6**	31.2	16.2	28.3	12.1

Table 11: Comparison of Aya-101 & MT5-large results, Standard Fine-Tuning (SFT) vs. Source Mixing (SRCMIX) across multiple language pairs using BLEURT, ChrF++, and BLEU.

Direction: English → XXX

		AYA-101							
Language	Baseline	Standard FT		SrcMix		TgtMix		TgtMix with tag	
Code	ChrF++	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
fuv_Latn	16.6	21.5	**7.1**	**24.6**	6.4	5.2	1.3	6.6	0.2
nus_Latn	21.8	22.8	11.8	**24.6**	**14.2**	18.3	13.5	11.0	0.8
taq_Latn	15.2	16.5	3.6	**18.6**	**13.2**	6.6	1.2	11.3	0.7
bam_Latn	19.9	28.0	12.8	**27.0**	**18.5**	6.3	2.4	11.6	1.2
fur_Latn	35.4	30.4	17.0	**50.2**	**29.9**	19.8	2.4	24.7	7.3
lij_Latn	**34.1**	25.5	9.9	25.1	**11.1**	19.1	6.3	23.1	6.9
lim_Latn	best30.0	29.1	5.2	29.6	**9.9**	18.6	1.5	24.1	5.3
srd_Latn	35.6	34.0	12.7	**37.5**	**19.1**	27.7	7.9	1.2	0.5
anp_Deva	-	42.3	17.9	44.1	20.2	3.6	3.7	**46.2**	**21.7**
bho_Deva	21.9	31.8	8.5	33.0	9.2	5.0	5.7	**36.9**	**12.2**
mag_Deva	27.1	45.9	20.2	**53.7**	**38.0**	1.9	4.8	49.3	23.8
kas_Arab	19.2	24.5	13.5	23.3	**10.3**	**24.3**	7.0	9.3	0.3
prs_Arab	24.1	25.5	10.0	**28.3**	**15.3**	16.2	2.4	28.7	9.1
pbt_Arab	11.4	11.4	3.3	**22.2**	**15.3**	16.7	4.8	21.7	6.7
Avg	24.8	27.8	10.9	**31.6**	**16.5**	13.5	4.6	19.6	9.5

Table 12: Translation results (ChrF++ and BLEU) for 14 ELRLs from English using Aya-101 under three training setups: Standard FT, SrcMix, and TgtMix (with and without language tags). Bold values indicate the best performance per row. Even with language tags, TgtMix lags behind SrcMix, validating our decoder specialization hypothesis.

form the model of the intended target language, following standard multilingual NMT practices (Johnson et al., 2017).

Despite this enhancement, the results show that TgtMix with language tags still underperforms relative to SrcMix across multiple ELRL directions. These findings support our central hypothesis in ELRL settings that target-side multilinguality leads to decoder confusion due to limited supervision across diverse and low-resource targets. Language tags alone cannot resolve this ambiguity, particularly when the model has not seen substantial amounts of each target language during pretraining or fine-tuning.

In contrast, SrcMix trains the decoder exclusively on a single ELRL, allowing it to specialize in generating coherent and fluent translations without interference from other target languages. The consistent performance gap between SrcMix and TgtMix (even with tags) highlights the advantage of source-side mixing in ELRL scenarios. *Takeaway: Even with language tags,* TgtMix *consistently underperforms* SrcMix *across all* ELRL*s, confirming the limitations of decoder-side multilinguality under weak supervision.* Detailed results of this experiment are included in Table 12, and Table 13, where we present BLEU and ChrF++ scores for each direction under both tagged and untagged TgtMix settings for Aya-101.

A.11 Amount of Data vs Multilinguality

Language	Standard FT		SrcMix	
Code	ChrF++	BLEU	ChrF++	BLEU
fur_Latn	30.4	17.0	**48.2**	**23.4**
fuv_Latn	**21.5**	**7.1**	**21.5**	6.8
bho_Deva	31.8	8.5	**35.3**	**10.8**
prs_Arab	25.5	9.9	**26.7**	**15.3**
Avg	27.3	10.6	**32.9**	**14.0**

Table 14: ChrF++ and BLEU for English→ELRL using Aya-101 under Standard FT, SrcMix multilingual.

Direction: Hindi → XXX

Language Code	Baseline ChrF++	AYA-101							
		Standard FT		SrcMix		TgtMix		TgtMix with tag	
		ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
anp_Deva	-	44.9	20.5	**52.6**	**28.7**	16.1	2.6	47.5	27.2
bho_Deva	-	27.7	6.8	**35.9**	**25.2**	19.9	6.0	20.1	12.7
mag_Deva	-	**46.0**	20.1	**46.0**	**20.3**	33.5	6.9	47.2	21.1
kas_Arab	-	24.2	8.7	**32.7**	**24.9**	12.0	2.7	7.4	0.2
prs_Arab	-	**28.4**	7.4	27.8	**15.2**	14.9	2.9	26.6	6.8
pbt_Arab	-	18.3	2.4	**26.2**	**23.7**	20.7	4.3	19.0	3.0
Avg	-	31.6	11.0	**36.8**	**23.0**	19.5	4.2	28.0	11.8

Table 13: Translation results (ChrF++ and BLEU) for 14 ELRLs from Hindi using Aya-101 under three training setups: Standard FT, SrcMix, and TgtMix (with and without language tags).

To isolate the effect of multilinguality from that of data volume, we conducted an ablation where the total number of training samples is fixed across conditions. For example, if the Standard FT baseline used 6k samples, then in SrcMix with four source languages, we sampled 1.5k examples per source–target pair, ensuring the overall dataset size remained constant. To prevent overlap, each language contributed a distinct set of sentences.

The results in Table 14 show that even under a fixed data budget, SrcMix achieves consistent improvements, with an average gain of +5.63 ChrF++ and +3.43 BLEU over Standard FT. Notably, languages such as Friulian (fur_Latn) and Dari (prs_Arab) show substantial jumps in both metrics. These findings confirm that the gains of SrcMix do not stem from using more data, but from leveraging structural cross-lingual signal sharing among related sources.

A.12 Strict language relatedness

To empirically examine whether the requirement of script similarity alone is sufficient without considering whether the family can support transfer to ELRLs, we conducted additional experiments using languages from different groups but with the same script (Latin). Specifically, we explored SrcMix using African and European language groups. Despite sharing the same script, we observed a notable degradation in performance when combining these unrelated groups. These results indicate that script similarity alone is insufficient for effective transfer; language group relatedness is crucial in facilitating successful cross-lingual knowledge transfer. Table 15 presents the results for two languages, Nigerian Fulfude and Friulian, in the SrcMix setting.

Direction	SrcMix with independent languages/scripts		SrcMix with related languages/scripts	
	ChrF++	BLEU	ChrF++	BLEU
Eng→fuv_Latn	**24.6**	**6.4**	17.8	5.3
Eng→fur_Latn	**30.4**	**17.0**	18.2	13.1

Table 15: Ablation study comparing SrcMix with typologically *independent* versus *related* languages/scripts.

Direction	**Standard FT**		**25% Mixing**		**100% Mixing**	
	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
prs_Arab	23.90	6.09	25.68	13.26	**27.80**	**15.14**
fur_Latn	30.42	16.95	48.18	23.36	**50.24**	**29.91**
bho_Deva	31.79	8.46	30.28	8.79	**33.04**	**9.27**
fuv_Latn	21.45	**7.11**	21.52	6.76	**24.59**	6.43
Avg	26.89	9.65	31.42	13.04	**33.92**	**15.19**

Table 16: Ablation on mixing ratio. Even with only 25% of mixed data from related source languages, SrcMix substantially improves over standard fine-tuning, with further gains obtained using full mixing.

A.13 Effect of Mixing Ratio

We conduct a controlled ablation by varying the proportion of mixed data from related source languages while keeping all other training settings fixed.

Table 16 reports results when using only 25% of the available mixed data compared to standard fine-tuning (no mixing) and full (100%) mixing. Even with partial mixing, SrcMix consistently outperforms standard fine-tuning across language pairs, demonstrating that the gains do not rely on aggressive data augmentation. Performance further improves with full mixing, indicating that SrcMix benefits from increased exposure to related source languages while remaining robust under limited mixing.