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

Sanjeev Kumar, Preethi Jyothi, Pushpak Bhattacharyya

EACL 2026 (findings) Poster

Department of Computer Science & Engineering
IIT Bombay



# Introduction

**Problem Definition**

- ELRLs have <10K parallel sentences
- Zero-shot transfer performs poorly
- Naive multilingual training causes negative transfer
- Millions of speakers remain digitally excluded

**Can structured multilinguality help?**

# Problem with Naive Multilingual Mixing

- Common approach:
  - Many-to-Many (M2M)
  - One-to-Many (O2M)
- Issues:
  - Negative interference
  - Decoder confusion
  - Performance degradation

**Hypothesis**: Decoder distributional stability is the primary bottleneck in ELRL multilingual transfer.

If we: Mix related source languages, by keeping target fixed

Then: Target distribution stabilizes, transfer improves

# Methodology

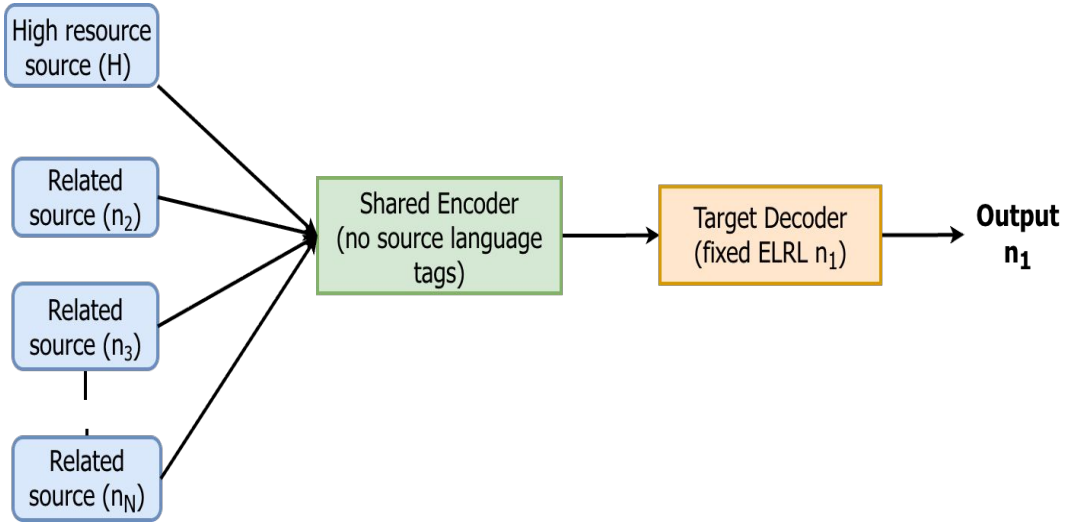
- Training strategies:
  - Zero-shot
  - Supervised Fine-Tuning (SFT)
  - Many-to-Many (M2M)
  - **Source-Mix (SrcMix) – Our proposed approach**
  - Target-Mix (TgtMix)

    **SrcMix mixes related source languages while keeping the target fixed.**

# Source-Mix (SrcMix)

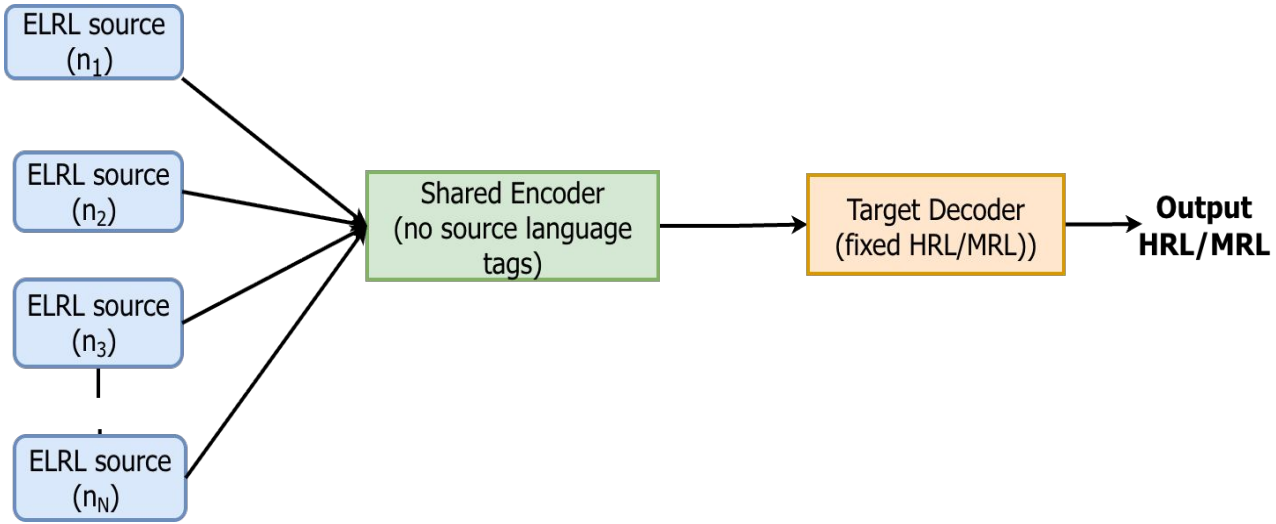
**Key Idea:** Mix multiple linguistically related source languages while keeping the target language fixed

**Only linguistically related source languages are mixed.**



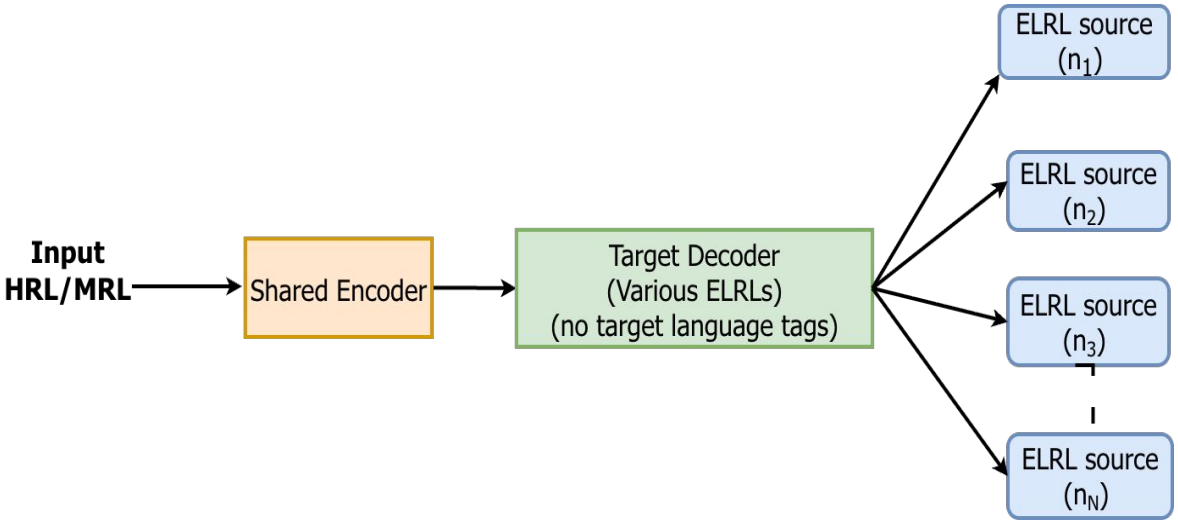
# Source-Mix (SrcMix)

## Reverse Direction (ELRL to HRL/MRL):



# Target-Mix (TgtMix)

## Forward Direction (HRL/MRL to ELRL):



# Introducing New Resources for Angika MT

- Angika (anp): An Indo-Aryan ELRL spoken in India and Nepal
- Written in the Devanagari script and follows SOV word order
- Significant speaker base (15M), but limited MT resources
- We release **first** public dataset for Angika MT

# Experimental Setup

- **Datasets**:
  - NLLB Seed Corpus (6,192 sentences each)
  - FLORES-200 (dev: 997, dev-test: 1,012 samples)
  - Custom-created Angika dataset
- **Language Groups** (14 ELRLs):
  - 4 African: Nigerian Fulfulde, Nuer, Bambara, Tamasheq
  - 4 Romance: Friulian, Ligurian, Limburgish, Sardinian
  - 3 Indic: Angika, Bhojpuri, Magahi
  - 3 Arabic: Dari, Kashmiri_Arab, Southern Pashto

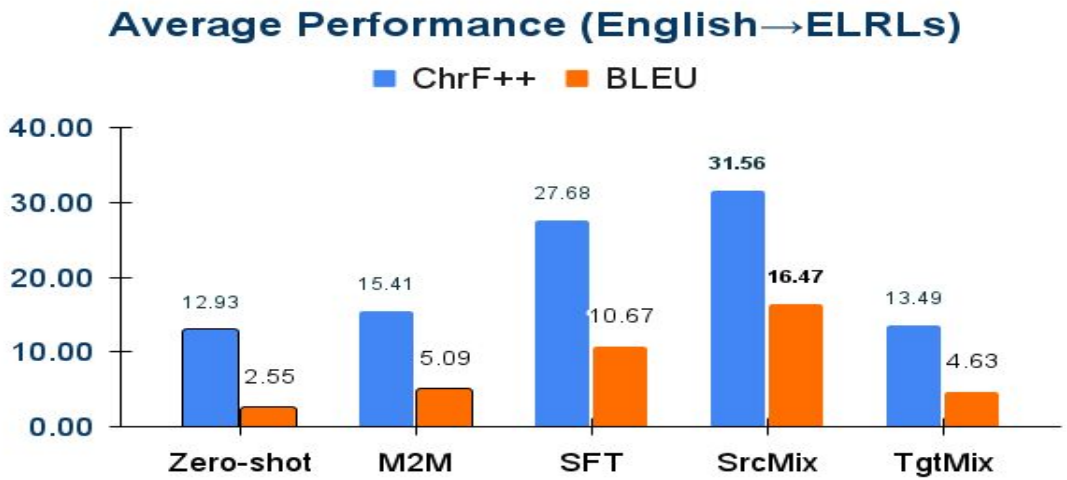
# Experimental Setup

- **Models:**
  - Aya-101 (13B): LLM based on mT5 architecture
  - mT5-large (1.2B): Traditional NMT model
  - **Decoder-only** models such as LLaMa-3.1-8B, and Gemma-7B perform **poorly** in ELRLs.

# Results

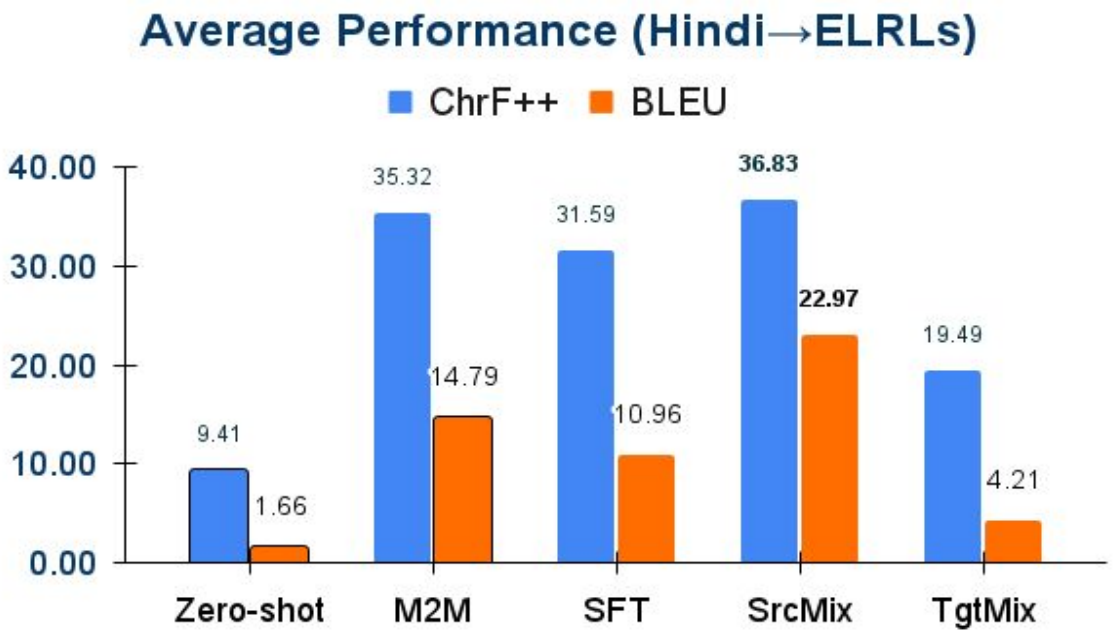
- Key Result:
  - Zero-shot: very low performance
  - M2M: often degrades ELRL translation
  - SFT: strong improvement
  - SrcMix: highest average BLEU & ChrF++

SrcMix improves performance by **+6 BLEU and +4 ChrF++** over SFT



# Results

SrcMix improves performance by **+12 BLEU and +5 ChrF++** over SFT



# Why Does SrcMix Work?

- Related languages share morphology and syntax
- Shared script improves lexical alignment
- Decoder learns from multiple related sources while keeping target language fixed

  **Multilinguality helps only when it is structured**

# Discussion & Conclusions

- Linguistic relatedness enables positive transfer
- Mixing related source languages yields consistent gains
- Naive multilingual training introduces negative transfer in ELRL MT
- Structured multilingual transfer outperforms naive multilinguality
- **Directionality matters:** Source-side mixing (SrcMix) consistently outperforms Target-side mixing (TgtMix)

# Thanks :)

@snjev310

Code

Paper

Dataset

# Key Finding: Asymmetry in Transfer

- Multilingual Transfer is Directionally Asymmetric
  - Source-side mixing helps
  - Target-side mixing hurts
  - Decoder specialization is critical

**Conclusion**: Protecting the decoder from modeling multiple ELRL distributions improves stability and accuracy.

# Ablation Studies

- To understand why SrcMix works, we conducted following ablations.
  - Mixing Direction (SrcMix vs. TgtMix)
    - SrcMix: Multiple sources → Single target
    - TgtMix: Single source → Multiple targets

Result: TgtMix underperforms SFT, SrcMix consistently improves.

TgtMix forces decoder to learn multiple ELRL distributions, causes interference and instability.

Observation: Decoder-side multilinguality is harmful in ELRL settings.

# Ablation Studies: Typology & Script Compatibility

## Does Linguistic Relatedness Affect Gains?

- We analyze gains across: 4 language families, 3 scripts
  - Observations
    - Stronger gains within closely related families (e.g., Indic)
    - Shared scripts show higher ChrF++ improvements
    - Gains reduce when typological distance increases
  - Conclusion
    - Transfer strength correlates with: Morphological similarity, word order similarity, and script overlaps

Structured multilingual transfer is most effective, when languages are typologically and orthographically aligned.

# Ablation Studies: Data Volume & Zero-Shot Controls

## Are Gains Due to More Data?

- We control: training steps, data volume
- Observations:
  - M2M uses same total data but performs worse
  - SrcMix gains persist under controlled training
- Conclusion
  - Improvements are not due to more data
  - Improvements are not due to longer training
  - Gains arise from structured decoder stabilization

In ELRL MT, performance depends more on how languages are mixed than on how much data is mixed.

# Contributions & Limitation of SrcMix

## Contributions

- Simple training-time strategy
- Systematic multilingual directionality study
- New Angika dataset
- Accepted at EACL 2026 (Findings)

## Limitations

- Still requires retraining
- Single model per direction
- Require n-way parallel data.
- Source copy and transliteration of named entity terms.