MUSLIMS IN MACHINE LEARNING

NEURAL INFORMATION PROCESSING SYSTEMS

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

AH-Translit: A Multi-Domain Dataset and Benchmark for Arabic-to-Hindi Transliteration

Vilal Ali, Mohd Hozaifa Khan, Bassam Adnan

CSE, International Institute of Information Technology - Hyderabad (IIIT-H), India

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop.

Table of contents

01 Introduction & Motivation

02 Related Work

03 Dataset Overview

04 Method

05 Experiments

06 Results and Analysis

07 Future Work

08 Key Takeaway

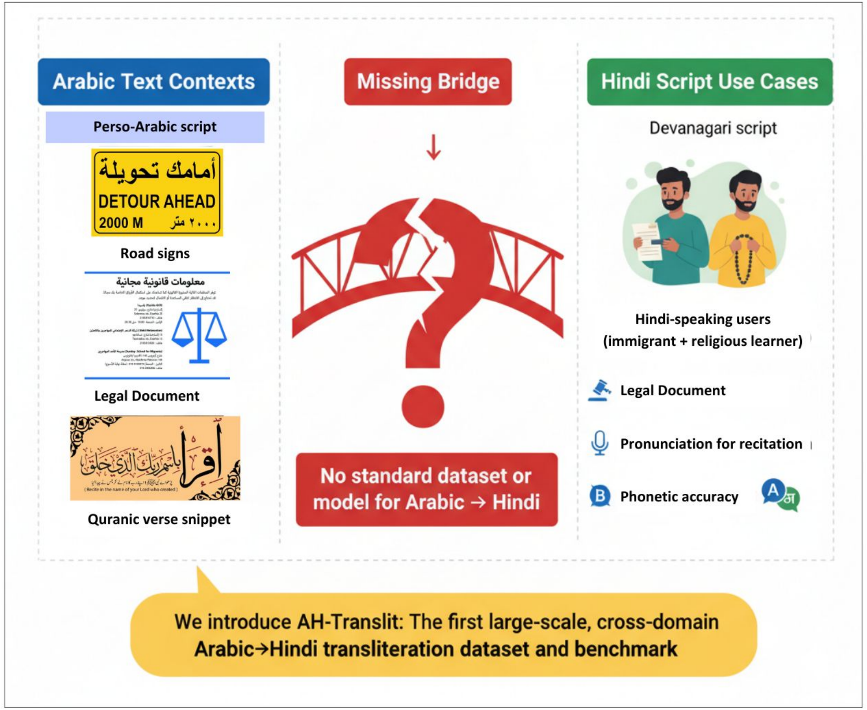
Introduction & Motivation

Problem — Script Barrier:

- Script barrier between Arabic (Perso-Arabic) and Hindi (Devanagari).

Impact:

- Over 8 million Hindi-speaking immigrants in Arab nations face daily challenges navigating documents and signage.
- Millions of South Asian Muslims rely on accurate transliteration for religious texts.



Arabic Text Contexts

Perso-Arabic script

Road signs

Legal Document

Quranic verse snippet

Missing Bridge

No standard dataset or model for Arabic → Hindi

Hindi Script Use Cases

Devanagari script

Hindi-speaking users (immigrant + religious learner)

Legal Document

Pronunciation for recitation

Phonetic accuracy

We introduce AH-Translit: The first large-scale, cross-domain Arabic→Hindi transliteration dataset and benchmark

Introduction & Motivation (Cont.)

Research Gap:

- No large-scale, multi-domain, publicly available Arabic→Hindi transliteration dataset.

Challenges:

- High linguistic diversity across Classical Arabic, MSA, and named entities.
- Significant orthographic ambiguity due to unvocalized Arabic vs. vowel-rich Devanagari.

Introduction & Motivation (Cont.)

Contribution:

- First comprehensive **multi-domain dataset** for Arabic→Hindi transliteration.
- First balanced, **human-verified benchmark** for cross-domain evaluation.
- **Strong baseline models** demonstrating domain generalization behavior.

Related Work: Datasets, Benchmarks & Research Gaps

Aksharantar dataset (Indic transliteration):

- Supports multiple Indic languages
- Focused on Roman ↔ Indic (or among Indic scripts), providing transliteration resources for many users.

Limitations relative to our goal:

- Does not support Arabic script → Indic transliteration
- No cross-domain coverage for Classical Arabic, MSA, named entities, etc.

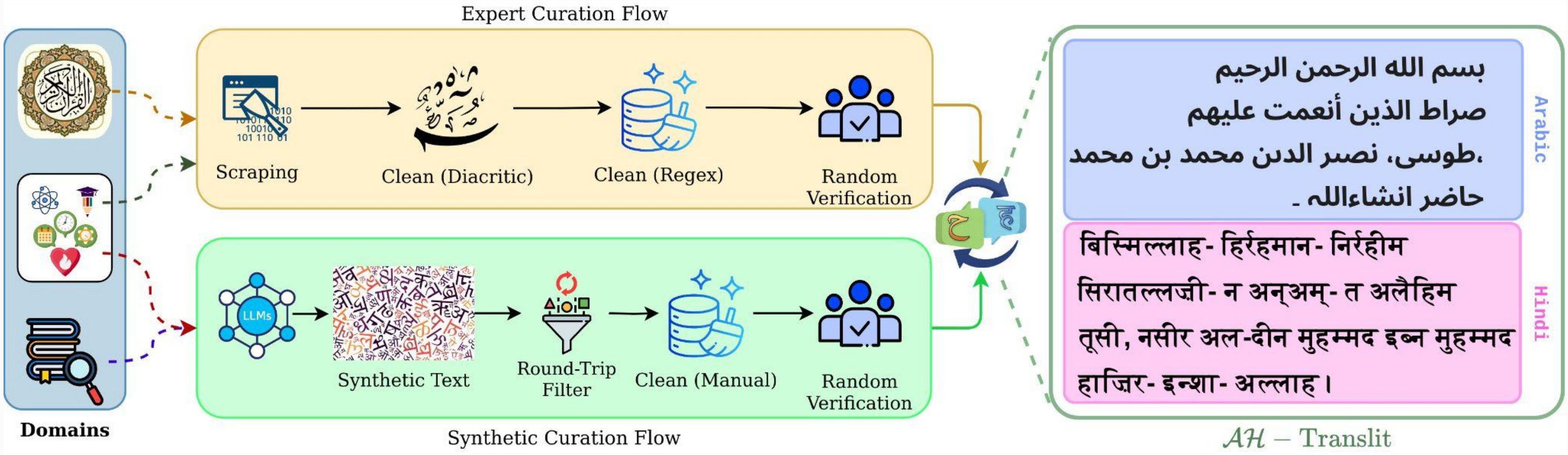
Absence of public Arabic→Indic datasets:

- No benchmark currently allows evaluation of cross-script transliteration from Arabic to any Indic script.
- Existing datasets are either single-domain, small-scale, or not publicly available.

Dataset Overview

- **Three Domains:**
  - Quranic (Classical, vocalized, long syntax)
  - Modern Standard Arabic (MSA)
  - Bibliographic (named entities, dense proper nouns)
- **100K parallel pairs**
- **1.2M Arabic words, 1.5M Hindi words**
- **Balanced 2K-pair benchmark (500 Quranic, 500 MSA, 1000 Biblio)**

Dataset Curation Pipeline



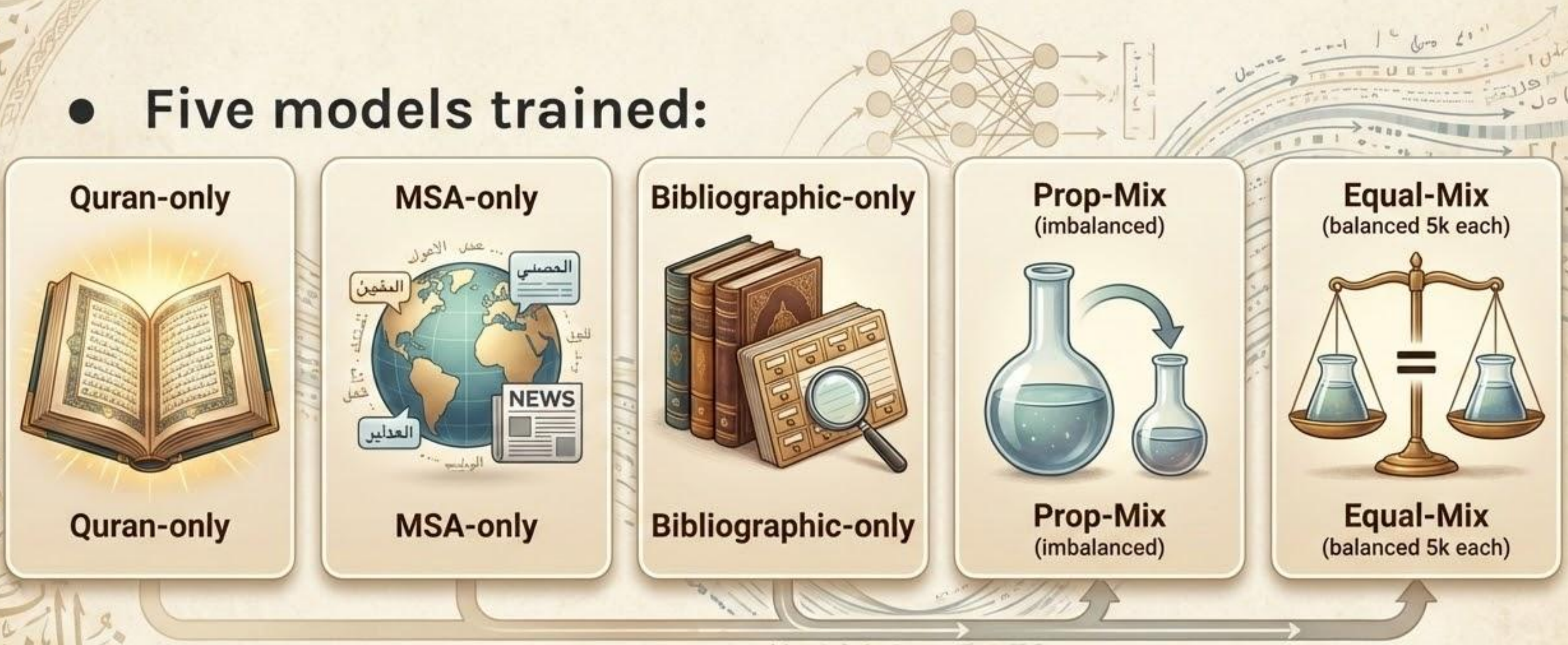
Expert-Curated	Quranic and MSA domains from human experts with direct validation
Synthetic Pipeline	Bibliographic corpus using LLM with round-trip consistency filter

Method: Char-GRU Architecture

- Character-level Seq2Seq GRU encoder–decoder
- 3-layer GRU with Bahdanau attention
- Embedding: 128, Hidden: 256
- Teacher forcing: 0.5

Experiments: Training Setup

- **Five models trained:**



Quran-only

MSA-only

Bibliographic-only

Prop-Mix (imbalanced)

Equal-Mix (balanced 5k each)

Quran-only

MSA-only

Bibliographic-only

Prop-Mix (imbalanced)

Equal-Mix (balanced 5k each)

Quantitative Results

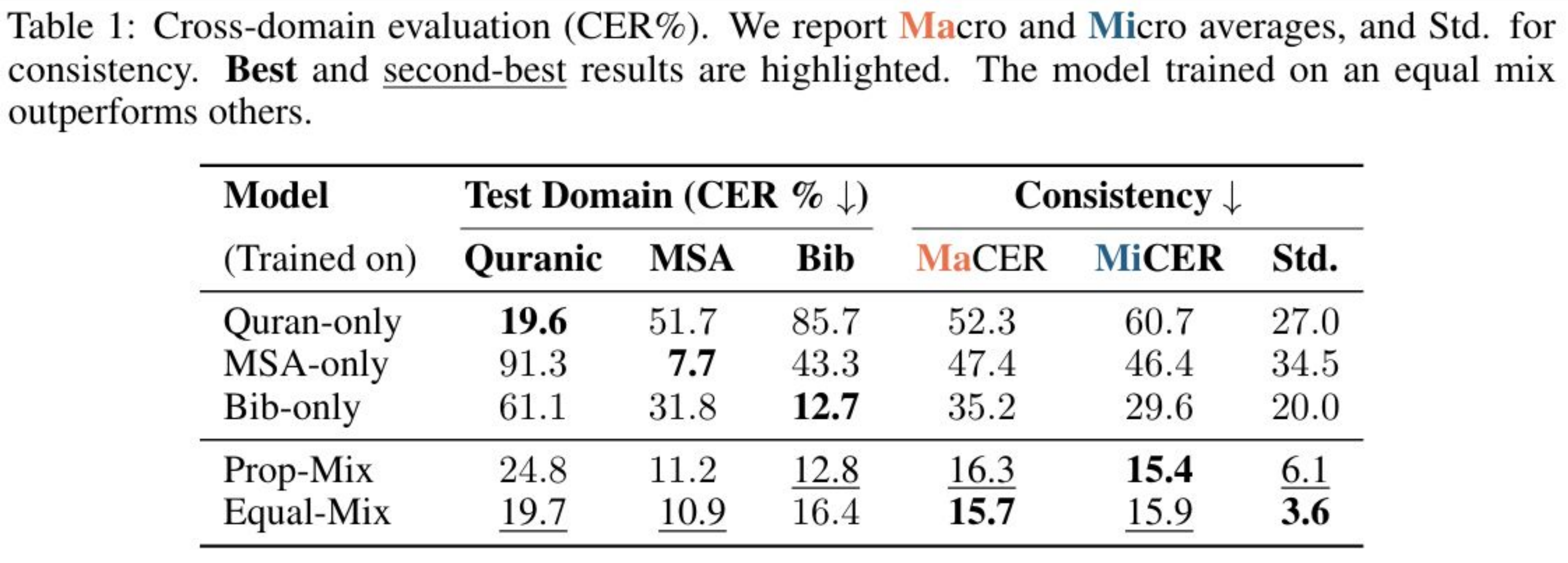


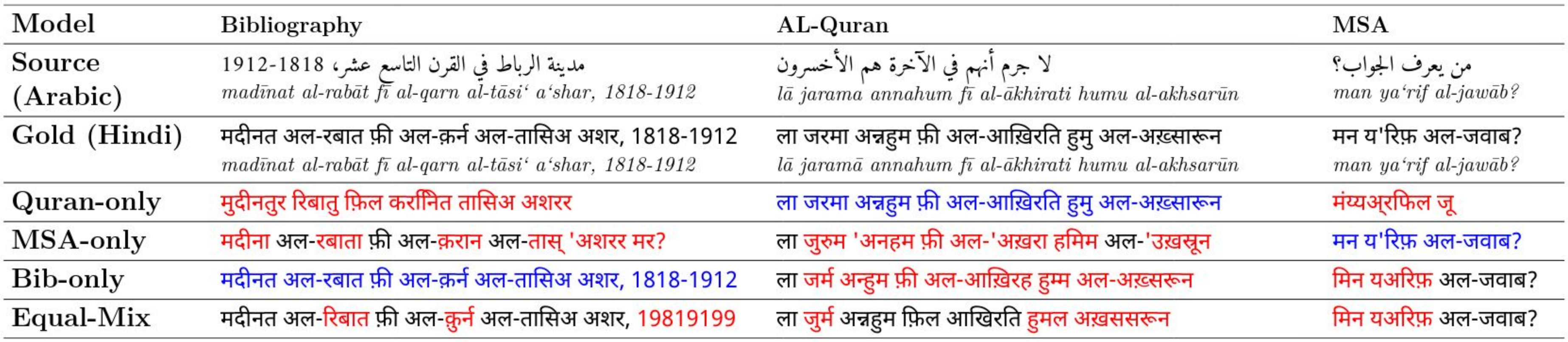
Table 1: Cross-domain evaluation (CER%). We report Macro and Micro averages, and Std. for consistency. **Best** and second-best results are highlighted. The model trained on an equal mix outperforms others.

Model	**Test Domain (CER % ↓)**			**Consistency ↓**		
(Trained on)	**Quranic**	**MSA**	**Bib**	MaCER	**MiCER**	**Std.**
Quran-only	**19.6**	51.7	85.7	52.3	60.7	27.0
MSA-only	91.3	**7.7**	43.3	47.4	46.4	34.5
Bib-only	61.1	31.8	**12.7**	35.2	29.6	20.0
Prop-Mix	24.8	11.2	12.8	16.3	**15.4**	6.1
Equal-Mix	19.7	10.9	16.4	**15.7**	15.9	**3.6**

Quantitative Results (Cont.)

- **Best overall:** Equal-Mix (MaCER 15.7%)
- **Specialist models:** very low in-domain, catastrophic out-of-domain
- **Table showing CER across domains**

Qualitative Analysis



Model	Bibliography	AL-Quran	MSA
Source (Arabic)	مدينة الرباط في القرن التاسع عشر، 1818-1912 *madīnat al-rabāt fī al-qarn al-tāsiʻ aʻshar, 1818-1912*	لا جرم أنهم في الآخرة هم الأخسرون *lā jarama annahum fī al-ākhirati humu al-akhsarūn*	من يعرف الجواب؟ *man yaʻrif al-jawāb?*
Gold (Hindi)	मदीनत अल-रबात फ़ी अल-क़र्न अल-तासिअ अशर, 1818-1912 *madīnat al-rabāt fī al-qarn al-tāsiʻ aʻshar, 1818-1912*	ला जरमा अन्नहुम फ़ी अल-आख़िरति हुमु अल-अख़्सारून *lā jaramā annahum fī al-ākhirati humu al-akhsarūn*	मन य'रिफ़ अल-जवाब? *man yaʻrif al-jawāb?*
Quran-only	मुदीनतुर रिबातु फ़िल करर्नित तासिअ अशरर	ला जरमा अन्नहुम फ़ी अल-आख़िरति हुमु अल-अख़्सारून	मंय्यअ्रफिल जू
MSA-only	मदीना अल-रबाता फ़ी अल-क़रान अल-तास् 'अशरर मर?	ला जुरुम 'अनहम फ़ी अल-'अख़रा हमिम अल-'उख़स्रून	मन य'रिफ़ अल-जवाब?
Bib-only	मदीनत अल-रबात फ़ी अल-क़र्न अल-तासिअ अशर, 1818-1912	ला जर्म अन्हुम फ़ी अल-आख़िरह हुम्म अल-अख़्सरून	मिन यअरिफ़ अल-जवाब?
Equal-Mix	मदीनत अल-रिबात फ़ी अल-कुर्न अल-तासिअ अशर, 19819199	ला जुर्म अन्नहुम फ़िल आखिरति हुमल अख़ससरून	मिन यअरिफ़ अल-जवाब?

Table 2: Cross-domain samples of transliteration models on AH-Translit-Bench

Qualitative Analysis (Cont.)

Specialist models exhibit strong domain overfitting:

- Quran-only: fails on numbers, punctuation, and modern tokens
- MSA-only: mis-transliterates classical/vocalized phonemes
- Bibliographic-only: injects hyphens & segmented patterns into normal text

Cross-domain breakdowns show structural, not random, errors:

- Wrong placement of vowels when unseen in training domain. Script-inconsistent handling of long vowels & gemination. Incorrect segmentation of named entities

Equal-Mix model displays consistent phoneme-level mapping:

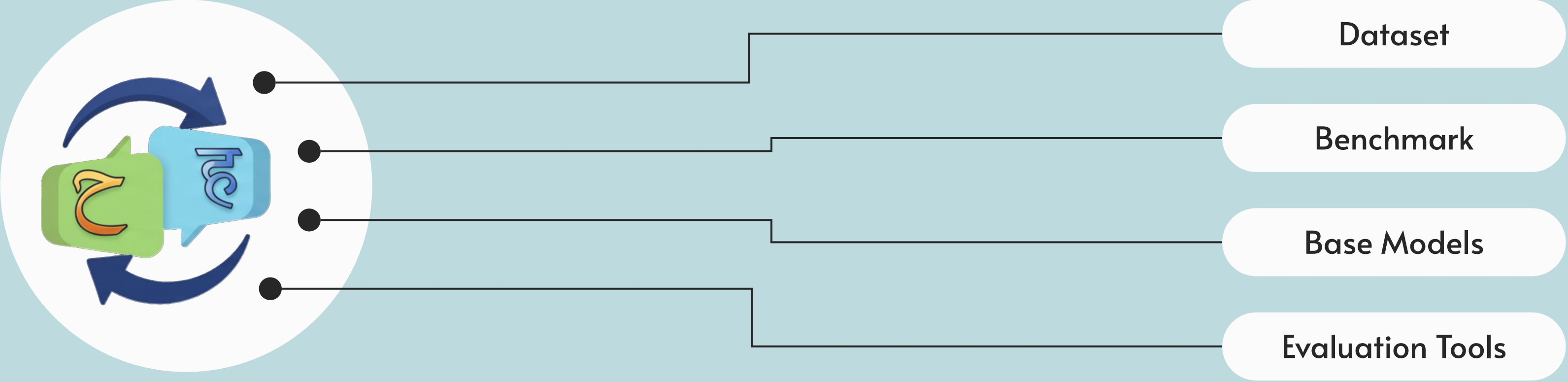
- Preserves classical markers, Handles modern tokens and numerals, Avoids systematic stylistic artifacts seen in domain-specialists

Future Work

- Domain adaptation & curriculum learning
- Leveraging Transformers or lightweight LLMs
- Extending dataset to Urdu, Punjabi, Bengali
- Joint phoneme-aware modeling
- Community evaluation shared task proposal

Key Takeaways

- First multi-domain Arabic→Hindi transliteration dataset (100K sentence pairs)
- Balanced training dramatically improves generalization
- Equal-Mix model: Best macro-CER (15.7%) + highest consistency
- Dataset, benchmark, and evaluation tools released



Dataset

Benchmark

Base Models

Evaluation Tools

Thanks

Do you have any questions?

Email: villa.ali@research.iiit.ac.in
Email: mohd.hozaifa@research.iiit.ac.in
Email: bassam.adnan@research.iiit.ac.in

Actively seeking PhD opportunities. Scan QR code to contact.