$\mathcal{AH}$-*Translit*: A Multi-Domain Dataset and Benchmark for Arabic-to-Hindi Transliteration

Vilal Ali **Mohd Hozaifa Khan** **Bassam Adnan**
CSE, IIIT Hyderabad, India
{vilal.ali, mohd.hozaifa, bassam.adnan}@research.iiit.ac.in

Abstract

The lack of public data for Arabic-to-Hindi transliteration has hindered the development of systems that can handle the languages' diverse linguistic styles. To address this, we introduce ***$\mathcal{AH}$-Translit***, a multi-domain dataset of 100K parallel pairs with over 1.2M Arabic and 1.5M Hindi words. We also present ***$\mathcal{AH}$-Translit*-Bench**[1], a balanced, human-verified benchmark for fair evaluations across diverse linguistic domains. Our analysis reveals that domain-specific models, while strong in-domain, generalize poorly. We show that a single model, trained on a balanced mixture, achieves higher performance consistency across all domains. This approach establishes a strong baseline with a Macro-averaged Character Error Rate (MaCER) of *15.7%*. We release the benchmark and an evaluation package for reproducible, cross-domain assessment.

1 Introduction

The script barrier between Arabic (Perso-Arabic) and Hindi (Devanagari) poses a significant machine transliteration challenge with wide-ranging implications across secular and religious domains. This problem directly impacts two major user groups: the 8 million [1, 2] Hindi-speaking immigrants in Arab nations who face daily obstacles in navigating official documents and public signage [3], and millions of non-Arabic-speaking South Asian Muslims who rely on phonetic transliteration for religious observances, such as reciting the Quran and Duas [4]. Developing effective transliteration systems is essential for improving information access and language learning for these large communities [5, 6].

However, progress in Arabic-to-Hindi transliteration is stalled by a foundational resource gap: the lack of a large-scale, multi-domain, public benchmark dataset. Such a resource is vital for training models, conducting standardized evaluations, and enabling comparative analysis[7]. This scarcity is particularly acute given the orthographic and phonetic ambiguities between unvocalized Arabic and Devanagari [8–11]. Moreover, Arabic exhibits extensive domain diversity, with significant lexical and phonetic shifts between Classical Arabic and Modern Standard Arabic (MSA). Without a comprehensive dataset, effectively learning these complex rules and generalizing across domains remains a prohibitive challenge.

To address this resource gap, we introduce $\mathcal{AH}$-*Translit*, with the following core contributions:

- **A Large-Scale, Multi-Domain Dataset**, $\mathcal{AH}$-*Translit*, the first public dataset for Arabic to Hindi Transliteration, containing 100K parallel sentence pairs across three distinct domains: Quranic, Modern Standard Arabic (MSA), and Bibliographic.
- ***$\mathcal{AH}$-Translit* Bench**, a manually verified benchmark that enables reliable, high-quality evaluation across domains.

[1] The benchmark data is available at: AH_TB Data

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

- **Comprehensive Baseline Analysis:** We provide strong baselines and demonstrate the dataset's effectiveness.

2 Related Work

Machine Transliteration The field of machine transliteration [12, 6, 13] has evolved from rule-based, language-specific systems [14, 12, 15, 16] to data-driven, multilingual frameworks. Data-driven approaches, spanning statistical [15, 17–21], Seq2Seq [20, 22, 23], and Transformer models [20, 24–29], have enabled powerful multilingual systems such as IndicXlit [7]. However, even these models are constrained by the lack of parallel data. Prior works have explored Arabic [24, 22, 23] and Hindi [30, 31, 9, 7] transliteration in isolation, but parallel *Arabic-to-Hindi* coverage remains absent. Our work addresses this gap by providing the *foundational data* necessary for existing models to succeed on this neglected task.

Benchmark Datasets Progress in machine transliteration is driven by the creation of benchmark datasets[32, 33, 24, 34, 35], from seminal efforts like the NEWS Shared Tasks [21] and the Dakshina dataset [36] to large-scale corpora for Indic languages like Aksharantar [7]. To overcome data scarcity for our target pair, we employ Large Language Models (LLMs) to generate a synthetic corpus, a common practice in low-resource scenarios[37, 24]. However, this approach presents a significant quality challenge: LLM-generated data can contain phonetic or orthographic errors. Generic filtering methods are often insufficient for the phonemic accuracy crucial for transliteration. We implement a rigorous *round-trip consistency filter*, a validation pipeline designed to ensure the quality of our synthetic data.

3 The $\mathcal{AH}$-*Translit* Dataset

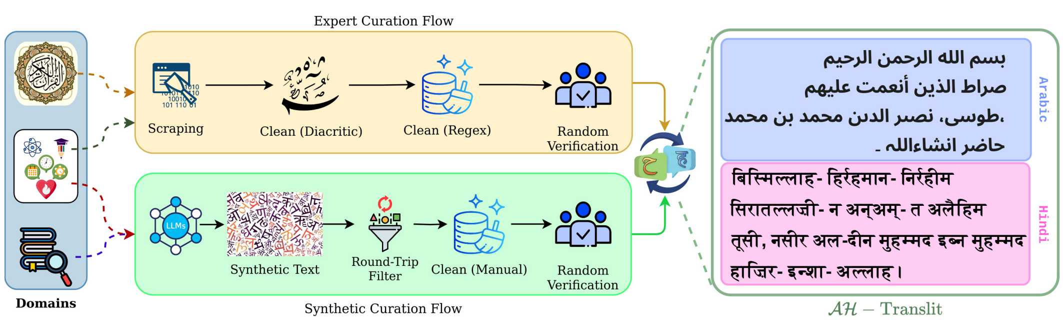

Figure 1: **The $\mathcal{AH}$-*Translit*** curation process uses two pipelines. The top pipeline processes Human-expert data from the Quranic and MSA domains via direct validation. The bottom pipeline processes the Bibliographic corpus synthetically, using LLM for parallel data generation and our round-trip consistency filter for validation.

3.1 Dataset Curation

We construct $\mathcal{AH}$-***Translit*** from three complementary sources to ensure high linguistic quality and broad domain coverage. First, we incorporate two expert-curated corpora as a gold-standard foundation: (1) a **Quranic** domain [38], providing formal, phonetically precise text from fully vocalized Classical Arabic with *Tajwid*[2], and (2) a **Contemporary MSA** domain that combines Modern Standard Arabic (MSA) phrases from an expert-compiled book [39] with synthetically generated scientific and educational text. Second, to address the challenge of transliterating named entities, we add a **Bibliographic** domain [40] rich in proper nouns. Together, these three domains cover classical, contemporary, and entity-rich language.

[2]Tajwid is the set of rules governing how Quranic words should be pronounced during recitation

Synthetic Data Generation Pipeline The Bibliographic domain data was originally an Arabic-to-Roman transliteration corpus [40], requiring us to generate Hindi transliterations synthetically. We applied the same procedure to the synthetic MSA portions. We used Gemini 2.5 Pro for this task due to its strengths in multilingual, non-Latin script handling and superior tokenization efficiency for Perso-Arabic and Devanagari, ensuring practical feasibility for this large-scale data generation task [41]. We found that a robust filtering mechanism was necessary to ensure quality. To this end, we developed a *round-trip consistency filter*. This method validates a generated Hindi transliteration (Hi_{llm}) of an Arabic source sentence (Ar_{src}) by transliterating it back to Arabic ($Ar_{rt} = \text{LLM}(Hi_{llm})$). We retain the (Ar_{src}, Hi_{llm}) pair only if the Character Error Rate (CER) between the original and the round-trip text, $\text{CER}(Ar_{src}, Ar_{rt})$, is below an empirically set threshold of 9%. This automated pipeline is scalable and effective in filtering inaccurate synthetic transliterations.

3.2 Dataset Statistics and Benchmark Design

The final curated $\mathcal{AH}$-*Translit* dataset contains 100K high-quality, parallel Arabic-to-Hindi sentence pairs — over 1.2M Arabic words and a vocabulary of $\approx$ 4K unique tokens. Each domain contributes a distinct linguistic profile: Quranic verses are long and syntactically layered (avg. 41.65 words; σ=27.68), MSA contributes short, everyday phrases (avg. 10.63 words; σ=6.44) and the Bibliographic domain (avg. 10.69 words; σ=9.43) is dense with named entities — a focused stress test for transliteration systems. Together, these domains ensure coverage of varied linguistic styles, laying the groundwork for a challenging cross-domain evaluation. Additional statistics appear in Appendix B.

Given these distinct domain characteristics, a simple proportional sampling would be inadequate for robust evaluation. We therefore construct ***$\mathcal{AH}$-Translit*-Bench**, a curated 2,000-pair test set designed for balanced, cross-domain assessment. It comprises 500 pairs each from the Quranic and MSA domains, and 1,000 from the Bibliographic domain. This stratified composition ensures models are evaluated on both the syntactic complexity of religious texts and the high named-entity density of bibliographic entries.

Ethics. All sources for this work are in the public domain. Our use of LLMs for data generation adheres to their respective terms of use. To mitigate potential inaccuracies, all synthetically generated pairs were rigorously filtered. We will release the dataset for research purposes only.

4 Experiments

Models. We establish our baseline with a standard character-level sequence-to-sequence architecture [42]: a 3-layer Gated Recurrent Unit (GRU) with attention [43] **Char-GRU**. This model choice allows us to measure the utility of our dataset for training and testing robust transliteration systems from scratch. Further details in Appendix A

Training. We investigate the impact of data composition by training five Char-GRU models across different corpora. The `Quran-only`, `MSA-only`, and `Bib-only` models are trained exclusively on their respective corpora, containing 2,000, 4,500, and 92,221 pairs. A fourth model, `Equal-Mix`, is trained on a balanced 15K-pair corpus constructed by sampling 5,000 pairs from each domain to mitigate inherent data imbalance. A fifth model, `Prop-Mix`, is trained on an imbalanced mixture of 40K-pairs in a 7:1:1 ratio (Bib : Quran : MSA) to serve as a baseline against our balanced-mixture approach. All models are trained on a single NVIDIA RTX 2080 GPU (16 GB VRAM). Further training details appear in Appendix A.

Evaluation. Each of the four trained models is evaluated against the three partitions of the $\mathcal{AH}$-*Translit*-Bench. This cross-domain evaluation allows us to measure not only in-domain proficiency but also the ability of models to generalise to unseen linguistic styles. Since transliteration requires character-level phonetic precision, we report Character Error Rate (CER) [44] as our primary metric. CER is also suitable for capturing critical sub-word errors that alter pronunciation, such as misplaced diacritics (e.g., the Hindi *nuqta*).

5 Results and Analysis

Performance Analysis The cross-domain evaluation results in Table 1 reveal a sharp contrast between specialist and generalist models. While models trained on single domains achieve low error

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

Model	**Test Domain (CER % ↓)**			**Consistency ↓**		
(Trained on)	**Quranic**	**MSA**	**Bib**	**Ma**CER	**Mi**CER	**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	<u>12.8</u>	<u>16.3</u>	**15.4**	<u>6.1</u>
Equal-Mix	<u>19.7</u>	<u>10.9</u>	16.4	**15.7**	<u>15.9</u>	**3.6**

rates on matched data, the `MSA-only` model reaches 7.7% CER on MSA; they are fundamentally brittle. Their performance degrades significantly out-of-domain, with the `MSA-only` model's CER reaching 91.3% on Quranic text. This extreme variance, captured by high Standard Deviations (20.0 to 34.5), demonstrates that specialist models overfit to their source domain and are unreliable for real-world use.

In contrast, the `Equal-Mix` model proves to be a generalist. While a close second on each domain, its primary advantage is its performance consistency. Its Standard Deviation of only 3.6 is an order of magnitude lower than any specialist, indicating that its performance is stable across diverse linguistic styles. This reliability translates to superior overall performance, achieving a state-of-the-art Macro-average CER of 15.7%. This confirms that a balanced training mixture is vital for developing a single system for diverse domain data.

Table 2: Cross-domain samples of transliteration models on $\mathcal{AH}$-*Translit*-Bench

Model	Bibliography	AL-Quran	MSA
Source (Arabic)	مدينة الرباط في القرن التاسع عشر، 1912-1818 *madīnat al-rabāt fī al-qarn al-tāsi‘ a‘shar, 1818-1912*	لا جرم أنهم في الآخرة هم الأخسرون *lā jarama annahum fī al-ākhirati humu al-akhsarūn*	من يعرف الجواب؟ *man ya‘rif al-jawāb?*
Gold (Hindi)	मदीनत अल-रबात फ़ी अल-क़र्न अल-तासिअ अशर, 1818-1912 *madīnat al-rabāt fī al-qarn al-tāsi‘ a‘shar, 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 ya‘rif al-jawāb?*
Quran-only	मुदीनतुर रिबातु फ़िल करर्निंत 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-'उख़सून	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-क़ुर्न अल-तासिअ अशर, 19819199	ला जुर्म अन्नहुम फ़िल आखिरति हुमल अख़ससरून	मिन यअरिफ़ अल-जवाब?

Qualitative Analysis Our analysis reveals how specialist models overfit to superficial patterns in their source domains. As shown in Table 2, the `Bib-only` model, trained on data rich with hyphens (-), incorrectly inserts them into classical Quranic transliterations (Col-2). Conversely, because numbers and special characters are absent in the Quranic domain, the `Quran-only` model fails to transliterate them when they appear in other contexts (Col-1). The `Equal-Mix` model avoids these overfitting errors, demonstrating a more generalizable understanding of linguistic rules. This affirms both the effectiveness of the balanced training strategy and the linguistic diversity of the benchmark.

6 Conclusion and Future Work

In this work, we introduced ***$\mathcal{AH}$-Translit***, the first large-scale, multi-domain dataset for Arabic-to-Hindi transliteration covering diverse linguistic styles (Classical to Bibliographic). Using our benchmark, ***$\mathcal{AH}$-Translit*-Bench**, we conducted rigorous cross-domain analysis. Our key finding is that while domain-specific models are brittle, a generalist model trained on a balanced data mixture achieves excellent performance consistency. This approach delivers state-of-the-art performance, showing that reliability across diverse inputs is as vital as peak accuracy for real-world applications. *Limitations:* The raw training corpus is skewed, which may not be ideal for all training scenarios. While our balanced benchmark mitigates this for evaluation, future work could explore its direct use. Furthermore, while effective, the round-trip consistency filter may still permit subtle errors. *Future Work:* Future research could explore advanced domain adaptation techniques or curriculum learning strategies to reduce performance variance across diverse linguistic styles. We release our benchmark to promote such research.

References

[1] Wikipedia contributors. Indian diaspora in the Middle East — Wikipedia, the free encyclopedia, 2025. URL https://en.wikipedia.org/wiki/Indian_diaspora_in_the_Middle_East. [Online; accessed 6-September-2025].

[2] Press Trust of India. More than 66 per cent of 1.34 crore nris live in gulf countries: Rti reply. https://economictimes.indiatimes.com/nri/latest-updates/more-than-66-per-cent-of-1-34-crore-nris-live-in-gulf-countries-rti-reply/articleshow/102177304.cms, July 2023. Citing a Right to Information (RTI) reply from the Ministry of External Affairs, Government of India, with data as of March 2022.

[3] Mohammad Shariq. Use of arabic by the 'others' in saudi arabia: A sociolinguistic study of communication needs leading to interlanguage development. *Migration Letters*, 20(S7):1334–1343, 2023. doi: 10.59670/ml.v20iS7.5036.

[4] Yunan Harahap, Talha Ansari, et al. The impact of transliteration method in the tradition of reading and understanding the quran at madrasah aliyah swasta amaliyah. *Proceeding of The Annual International Conference on Islamic Education*, 2(1):1–13, 2023.

[5] Jong-Hoon Lee and K-S. Choi. A comparison of different machine transliteration models. *Journal of Artificial Intelligence Research*, 27:165–193, 2006.

[6] A'la Syauqi and Aji Prasetya Wibawa. Advances in machine transliteration methods, limitations, challenges, applications and future directions. *Natural Language Processing Journal*, 11:100158, 2025. ISSN 2949-7191. doi: https://doi.org/10.1016/j.nlp.2025.100158. URL https://www.sciencedirect.com/science/article/pii/S2949719125000342.

[7] Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Khapra. Aksharantar: Open Indic-language transliteration datasets and models for the next billion users. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 40–57, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.4. URL https://aclanthology.org/2023.findings-emnlp.4/.

[8] Satish. Phonetic transliteration of arabic script to devanagari. Website, 2014. Retrieved from https://satish.com.in/20140418.

[9] Riyaz A. Bhat, Irshad A. Bhat, Naman Jain, and Dipti Misra Sharma. A house united: Bridging the script and lexical barrier between Hindi and Urdu. In Yuji Matsumoto and Rashmi Prasad, editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 397–408, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1039/.

[10] Assia Boumaraf, Sonia Bekal, and Joël Macoir. The orthographic ambiguity of the arabic graphic system: Evidence from a case of central agraphia affecting the two routes of spelling. *Behavioural Neurology*, 2022:11 pages, 11 2022. doi: 10.1155/2022/8078607.

[11] N.Y. Habash. *Introduction to Arabic Natural Language Processing*. Synthesis digital library of engineering and computer science. Morgan & Claypool Publishers, 2010. ISBN 9781598297959. URL https://books.google.co.in/books?id=kRIHCnC74BoC.

[12] Palakpreet Kaur and Kamal Deep Garg. Machine transliteration for indian languages: Survey. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 462–467, 2022. doi: 10.1109/PDGC56933.2022.10053165.

[13] Sreelekha S. Statistical vs rule based machine translation; A case study on indian language perspective. *CoRR*, abs/1708.04559, 2017. URL http://arxiv.org/abs/1708.04559.

[14] Kevin Knight and Jonathan Graehl. Machine transliteration. *Computational Linguistics*, 24(4):599–612, 1998. URL https://aclanthology.org/J98-4003/.

[15] Yomal De Mel, Kasun Wickramasinghe, Nisansa de Silva, and Surangika Ranathunga. Sinhala transliteration: A comparative analysis between rule-based and Seq2Seq approaches. In Ruvan Weerasinghe, Isuri Anuradha, and Deshan Sumanathilaka, editors, *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 166–173, Abu Dhabi, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.indonlp-1.19/.

[16] Yu-Chun Wang and Richard Tzong-Han Tsai. English-korean named entity transliteration using statistical substring-based and rule-based approaches. In *NEWS@IJCNLP*, 2011. URL `https://api.semanticscholar.org/CorpusID:16125561`.

[17] Nam X. Cao, Nhut M. Pham, and Quan H. Vu. Comparative analysis of transliteration techniques based on statistical machine translation and joint-sequence model. In *Proceedings of the 1st Symposium on Information and Communication Technology*, SoICT '10, page 59–63, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450301053. doi: 10.1145/1852611.1852624. URL `https://doi.org/10.1145/1852611.1852624`.

[18] Soumyadeep Kundu, Sayantan Paul, and Santanu Pal. A deep learning based approach to transliteration. In Nancy Chen, Rafael E. Banchs, Xiangyu Duan, Min Zhang, and Haizhou Li, editors, *Proceedings of the Seventh Named Entities Workshop*, pages 79–83, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2411. URL `https://aclanthology.org/W18-2411/`.

[19] Gia H. Ngo, Minh Nguyen, and Nancy F. Chen. Phonology-augmented statistical framework for machine transliteration using limited linguistic resources. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):199–211, 2019. doi: 10.1109/TASLP.2018.2875269.

[20] Hemanta Baruah, Sanasam Ranbir Singh, and Priyankoo Sarmah. Transliteration characteristics in romanized assamese language social media text and machine transliteration. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(2), February 2024. ISSN 2375-4699. doi: 10.1145/3639565. URL `https://doi.org/10.1145/3639565`.

[21] Haizhou Li, A Kumaran, Vladimir Pervouchine, and Min Zhang. Report of NEWS 2009 machine transliteration shared task. In Haizhou Li and A Kumaran, editors, *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 1–18, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL `https://aclanthology.org/W09-3501/`.

[22] Bashar Talafha, Analle Abuammar, and Mahmoud Al-Ayyoub. Atar: Attention-based lstm for arabizi transliteration. *International Journal of Electrical and Computer Engineering*, 11:2327–2334, 06 2021. doi: 10.11591/ijece.v11i3.pp2327-2334.

[23] Mohamed Seghir Hadj Ameur, Farid Meziane, and Ahmed Guessoum. Arabic machine transliteration using an attention-based encoder-decoder model. *Procedia Computer Science*, 117:287–297, 2017. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2017.10.120. URL `https://www.sciencedirect.com/science/article/pii/S1877050917321774`. Arabic Computational Linguistics.

[24] Soufiane Hajbi, Omayma Amezian, Mouhssine Ziyad, Issame El Kaime, Redouan Korchiyne, and Younes Chihab. Transformer-based model for moroccan arabizi-to-arabic transliteration using a semi-automatic annotated dataset. *International Journal of Information Management Data Insights*, 5(2): 100351, 2025. ISSN 2667-0968. doi: https://doi.org/10.1016/j.jjimei.2025.100351. URL `https://www.sciencedirect.com/science/article/pii/S2667096825000333`.

[25] MohammadAli SadraeiJavaheri, Ehsaneddin Asgari, and Hamid R. Rabiee. Transformers for bridging persian dialects: Transliteration model for tajiki and iranian scripts. In *International Conference on Language Resources and Evaluation*, 2024. URL `https://api.semanticscholar.org/CorpusID:269804499`.

[26] Umer Butt, Stalin Varanasi, and Günter Neumann. Low-resource transliteration for Roman-Urdu and Urdu using transformer-based models. In Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jonathan Washington, Nathaniel Oco, and Xiaobing Zhao, editors, *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 144–153, Albuquerque, New Mexico, U.S.A., May 2025. Association for Computational Linguistics. ISBN 979-8-89176-230-5. doi: 10.18653/v1/2025.loresmt-1.13. URL `https://aclanthology.org/2025.loresmt-1.13/`.

[27] Rohith Gowtham Kodali, Durga Prasad Manukonda, and Maharajan Pannakkaran. byteSizedLLM@DravidianLangTech 2025: Abusive Tamil and Malayalam text targeting women on social media using XLM-RoBERTa and attention-BiLSTM. In Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Sajeetha Thavareesan, Elizabeth Sherly, Saranya Rajiakodi, Balasubramanian Palani, Malliga Subramanian, Subalalitha Cn, and Dhivya Chinnappa, editors, *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 80–85, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-228-2. doi: 10.18653/v1/2025.dravidianlangtech-1.14. URL `https://aclanthology.org/2025.dravidianlangtech-1.14/`.

[28] Davor Lauc, Attapol Rutherford, and Weerin Wongwarawipatr. Ayutthayaalpha: A thai-latin script transliteration transformer, 2024. URL https://arxiv.org/abs/2412.03877.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[30] Dwija Parikh and Thamar Solorio. Normalization and back-transliteration for code-switched data. In Thamar Solorio, Shuguang Chen, Alan W. Black, Mona Diab, Sunayana Sitaram, Victor Soto, Emre Yilmaz, and Anirudh Srinivasan, editors, *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 119–124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.calcs-1.15. URL https://aclanthology.org/2021.calcs-1.15/.

[31] Saqib Maqbool, Nisar Ahmed, Gulshan Saleem, and Muhammad Munawar. An efficient hindi-urdu transliteration system. *Oral Science International*, 27:4549–4553, 07 2015.

[32] Md Fahim, Fariha Tanjim Shifat, Fabiha Haider, Deeparghya Dutta Barua, MD Sakib Ul Rahman Sourove, Md Farhan Ishmam, and Md Farhad Alam Bhuiyan. BanglaTLit: A benchmark dataset for back-transliteration of Romanized Bangla. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14656–14672, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.859. URL https://aclanthology.org/2024.findings-emnlp.859/.

[33] Hemanta Baruah, Sanasam Ranbir Singh, and Priyankoo Sarmah. Assamesebacktranslit: Back transliteration of romanized assamese social media text. In *International Conference on Language Resources and Evaluation*, 2024. URL https://api.semanticscholar.org/CorpusID:269803924.

[34] Mohamed Seghir Hadj Ameur, F. Meziane, and Ahmed Guessoum. Anetac: Arabic named entity transliteration and classification dataset. *ArXiv*, abs/1907.03110, 2019. URL https://api.semanticscholar.org/CorpusID:195833431.

[35] Mohammed Furqan, Raahid Bin Khaja, and Rayyan Habeeb. Erupd - english to roman urdu parallel dataset. *ArXiv*, abs/2412.17562, 2024. doi: 10.48550/arXiv.2412.17562.

[36] Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. Processing South Asian languages written in the Latin script: the Dakshina dataset. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.294/.

[37] Mihai Nadăş, Laura Dioşan, and Andreea Tomescu. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615–134633, 2025. ISSN 2169-3536. doi: 10.1109/access.2025.3589503. URL http://dx.doi.org/10.1109/ACCESS.2025.3589503.

[38] HindiQuran.in. Quran in hindi, 2025. URL https://hindiquran.in/.

[39] A.M.K. SIDDIQUI. JOUDAH INFO-LINGO TECH PVT LTD, 2014. URL https://jilt.in/.

[40] Fadhl Eryani and Nizar Habash. Automatic Romanization of Arabic bibliographic records. In Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb, editors, *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 213–218, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wanlp-1.23/.

[41] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261. Preprint.

[42] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[43] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2015. URL https://arxiv.org/abs/1409.0473.

[44] Andrew C. Morris, Viktoria Maier, and Phil D. Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, 2004. URL https://api.semanticscholar.org/CorpusID:18880375.

A Experimental and Implementation Details

This section consolidates the technical details required for full reproducibility of our experiments.

A.1 Training Hyperparameters

All Char-GRU models were trained using the hyperparameters documented in Table 3. The number of epochs was adjusted based on the dataset size to prevent overfitting on smaller corpora while ensuring convergence on larger ones.

Table 3: Training hyperparameters for all Char-GRU models.

Group	**Hyperparameter**	**Value**
Model Architecture	Embedding Dimension	128
	Hidden Size (GRU)	256
	Number of Layers	3
Optimization	Optimizer	AdamW
	Learning Rate	3×10^{-4}
	Weight Decay	1×10^{-6}
	Batch Size	16
	Gradient Clip Norm	1.0
Regularization	Encoder Dropout	0.2
	Decoder Dropout	0.2
	Teacher Forcing Ratio	0.5
Training Duration	Epochs (`Quran/MSA-only`)	30
	Epochs (`Bib-only/Equal-Mix`)	10

A.2 Round-Trip Consistency Filter

The synthetic portion of our dataset was curated using a round-trip consistency filter. The process is as follows:

1. An Arabic source sentence (Ar_{src}) is transliterated to Hindi (Hi_{llm}) using Gemini 2.5 Pro.
2. The generated Hindi (Hi_{llm}) is then transliterated back to Arabic (Ar_{rt}) using the same model.
3. The Character Error Rate (CER) is calculated between the original and the round-trip Arabic text: CER(Ar_{src}, Ar_{rt}).
4. The pair (Ar_{src}, Hi_{llm}) is permanently kept only if the CER is below a 9% threshold. This threshold was determined empirically by manually reviewing random 1000 samples, as it offered the best balance between filtering out clear orthographic errors while retaining correct transliterations.

Note: A manual audit of 1,000 sampled pairs indicated that a 9% round-trip CER threshold achieves a practical quality–coverage balance, filtering obvious noise while preserving 92% of high-quality pairs.

B Dataset Details

This section provides a detailed breakdown of the dataset statistics for this work.

B.1 Detailed Dataset Statistics

Table 4 provides the complete statistics for the training and benchmark splits of the $\mathcal{AH}$*-Translit* dataset.

Table 4: Detailed statistics for the $\mathcal{AH}$-*Translit* dataset.

Domain	Split	Pairs	Word Count		Avg. Word Len		Avg. Char Len	
			Arabic	Hindi	Ar[a]	Hi[b]	Ar	Hi
Quranic	Train	2000	83,300	91,400	41.65	56.28	89.97	120.78
	Bench.[a]	500	20,825	22,850	41.65	57.27	88.62	122.77
MSA	Train	4500	47,835	55,120	10.63	14.00	21.51	28.16
	Bench.	500	5315	6125	10.63	14.01	21.53	28.00
Bibliographic	Train	92,221	1,010,715	1,319,377	10.69	13.95	24.16	30.76
	Bench.	1000	10,690	12,150	10.69	13.95	24.16	29.79

[a] Bench.: Benchmark set.
[b] Ar: Arabic. Hi: Hindi.

C Extended Results and Analysis

This section provides the complete, unabridged evaluation results to support the analysis in the main paper.

C.1 Complete Evaluation Results

Table 5 presents the complete evaluation metrics. Following prior transliteration evaluation, we also report Word Correctness (100–WER), which can be negative in cases of high insertion/deletion rates.

Table 5: **Complete evaluation results**. We report Character Error Rate (CER, lower is better) and Word Error Rate (WER). Word Correctness (100–WER) may be negative when insertion/deletion errors exceed the total word count; we retain it to allow direct comparability with prior transliteration work. All metrics are reported as percentages (%).

Model	Test Domain	Char Acc	CER	WER	Word Corr.
`MSA-only`	Quranic	8.70	91.30	123.46	−23.46
	MSA	92.32	7.68	26.58	73.42
	Bib.	56.70	43.30	98.24	1.76
	Mi-Avg	**53.61**	**46.40**	**86.63**	**13.37**
`Quran-only`	Quranic	80.36	19.64	57.40	42.60
	MSA	48.31	51.69	110.51	−10.51
	Bib.	14.27	85.73	176.18	−76.18
	Mi-Avg	**39.30**	**60.70**	**130.07**	**−30.07**
`Bib-only`	Quranic	38.91	61.09	106.68	−6.68
	MSA	68.24	31.76	86.06	13.94
	Bib.	87.26	12.74	31.64	68.36
	Mi-Avg	**70.42**	**29.58**	**63.01**	**36.00**
`Prop-Mix`	Quranic	75.14	24.86	63.39	36.61
	MSA	88.78	11.22	38.59	61.41
	Bib.	87.11	12.89	34.12	65.88
	Mi-Avg	**83.67**	**16.33**	**45.36**	**54.63**
`Equal-Mix`	Quranic	80.27	19.73	56.08	43.92
	MSA	89.07	10.93	38.49	61.51
	Bib.	83.63	16.37	44.81	55.19
	Mi-Avg	**84.15**	**15.85**	**46.05**	**53.95**