A Systematic Benchmark of Machine Transliteration Models for the Tajik-Farsi Language Pair: A Comparative Study from Rule-Based to Transformer Architectures

Mullosharaf Arabov
Kazan Federal University, Kazan, Russia
cool.araby@gmail.com

Abstract

This paper presents a systematic comparative benchmark of machine learning architectures for transliteration between the Tajik (Cyrillic script) and Persian (Arabic script) languages. A central contribution is the creation and open release of a unique, multi-domain parallel corpus comprising 328,253 sentence pairs aggregated from heterogeneous sources, including classical poetry, diplomatic texts, lexicographic data, and crowdsourced projects. To enable fully reproducible experiments under limited computational resources, a representative stratified subset of 40,000 pairs has been formed and is made publicly available together with the complete dataset. The study evaluates six model classes under a unified protocol: a rule-based baseline, an LSTM with attention, a character-level Transformer, a G2P Transformer trained from scratch, large pre-trained multilingual models (mBART, mT5 with LoRA), and the byte-oriented ByT5 model. The experimental results demonstrate that ByT5 achieves state-of-the-art performance with chrF++ scores of 87.4 for Tajik→Farsi and 80.1 for the reverse direction. Notably, the proposed G2P Transformer, trained exclusively on the 40,000-pair subset, significantly outperforms the substantially larger mBART model in the forward direction (72.3 vs. 62.2 chrF++), while subword-based models (mT5) fail almost completely, underscoring the necessity of character- or byte-level representations for this task. All collected data, source code, and model outputs are released under open licenses, thereby establishing a FAIR (Findable, Accessible, Interoperable, Reusable) benchmark for cross-script NLP in low-resource Central Asian languages.

Keywords: Transliteration, Tajik language, Persian language, parallel corpus, FAIR benchmark, low-resource languages, Cyrillic script, Arabic script, byte-level Transformer, open data.

1 Introduction

The Persian language exhibits a striking case of digraphia: two distinct writing systems coexist within a single linguistic continuum. In Iran and Afghanistan, the Perso-Arabic script prevails, whereas Tajikistan employs a modified Cyrillic alphabet adopted during 20th-century linguistic reforms. This orthographic divide creates a paradoxical barrier—speakers of mutually intelligible varieties are largely cut off from each other's written content. The digital environment exacerbates the problem, as the vast majority of Persian-language online texts use Arabic script, effectively excluding Tajik-speaking users from a substantial portion of internet resources (Megerdoomian and Parvaz, 2008). This situation is not unique to Persian: similar digraphic divides or historical script shifts occur in numerous languages of the former Soviet space, including Serbian (Cyrillic/Latin), Kazakh (Cyrillic/Latin/Arabic), Uzbek (Cyrillic/Latin), Bashkir and Tatar (Cyrillic/Arabic historically, with ongoing Latinisation discussions), and Ossetian (Cyrillic/Georgian historically). In all these cases, automatic transliteration has proven to be a low-cost bridge fostering linguistic unity and cross-border communication.

Automatic conversion between Tajik Cyrillic and Persian Arabic script involves several non-trivial linguistic challenges. Transliterating from Arabic to Cyrillic requires restoring short vowels that are typically omitted in Perso-Arabic writing. The reverse direction must account for positional shaping of Arabic letters and the fact that multiple Cyrillic characters may correspond to a single Perso-Arabic grapheme. Moreover, orthographic divergence is compounded by lexical and morphological differences shaped by decades of separate cultural and political development.

Despite its practical significance, automatic

transliteration for the Tajik–Farsi pair has remained on the periphery of Natural Language Processing (NLP) research. While machine translation between the two varieties has received some attention, the creation of robust transliteration systems—potentially a simpler and more reliable bridge between scripts—has been comparatively neglected (Merchant and Tang, 2024). A fundamental obstacle has been the acute scarcity of high-quality parallel corpora containing texts in both graphic systems. Recent years have witnessed progress with the release of resources such as ParsText (Merchant and Tang, 2024) and a corpus derived from Ferdowsi's *Shahnameh*, yet their volume and genre coverage remain limited. The emergence of multi-faceted benchmarks like APARSIN (Jafari et al., 2026) further confirms the growing demand for systematic comparison of models within this language family.

The present study addresses these gaps by conducting a systematic benchmark of transliteration models spanning several methodological generations—from rule-based substitution to byte-level Transformers—evaluated on a unified, multi-domain dataset. Our contributions are threefold: (i) the creation and open release of a unique parallel corpus comprising over 328,000 sentence pairs, along with a stratified 40,000-pair subset for reproducible experimentation under limited resources; (ii) the first controlled comparison of six architectural families under identical training and evaluation protocols; and (iii) empirical evidence that character- and byte-level models decisively outperform subword-based approaches for this low-resource cross-script task.

The remainder of this paper is organized as follows. Section 2 surveys related work on Tajik–Persian transliteration and available resources. Section 3 describes the corpus construction and model architectures evaluated in the benchmark. Section 4 presents the experimental setup and quantitative results. Section 5 concludes with a discussion of findings and directions for future work.

2 Related Work

Systematic approaches to Tajik–Persian transliteration began to appear in the early 2010s. Davis (2012) introduced a system based on Statistical Machine Translation (SMT), treating transliteration as character-level translation and addressing the challenge of resource sharing between the two Persian varieties. Around the same period, theoretical foundations for automated script conversion were laid, often relying on an intermediate Latin transcription—an approach prone to information loss and error propagation (Grashchenko, 2003).

The advent of neural methods brought a qualitative shift. Transformer-based Grapheme-to-Phoneme (G2P) models have been applied to this task, achieving chrF++ scores of approximately 59 for Farsi→Tajik and 74 for the reverse direction, thereby establishing reproducible baselines and highlighting the asymmetric difficulty of the two transliteration directions. Parallel efforts have focused on dataset creation. Researchers have leveraged optical character recognition and heuristic alignment to construct parallel corpora from classical literary works, including Ferdowsi's *Shahnameh*. Merchant and Tang (2024) introduced ParsText, a collection of 2,813 manually curated Persian sentences in both scripts, emphasizing the scarcity of digital digraphic content relative to printed materials.

The most advanced transliteration model to date is ParsTranslit (Merchant and Tang, 2026), a sequence-to-sequence system trained on an aggregation of all available datasets plus two proprietary collections. The authors convincingly demonstrate that single-domain models lack cross-domain robustness, and only data aggregation yields a universal transliterator. ParsTranslit achieved state-of-the-art results: chrF++ of 87.91 and normalized CER of 0.05 for Farsi→Tajik, and 92.28 and 0.04 for the reverse direction.

The authors of the present study have previously contributed to the resource base and evaluation methodology for this language pair. Arabov (2026b) introduced TajPersLexon, a curated lexical resource of 40,112 word and phrase pairs, accompanied by a CPU-only benchmark comparing hybrid, neural, and information retrieval approaches. Kurbonovich (2026) investigated character-level Transformers trained on 52,152 lexicographically verified pairs, achieving a CER of 0.3182 and exact-match accuracy of 0.3215 with beam search, outperforming both dictionary-based rules and recurrent baselines.

Beyond direct transliteration efforts, several Persian NLP tools provide valuable pre- and post-processing capabilities. DadmaTools V2 (Jafari et al., 2025) employs LoRA adapters to efficiently add task-specific modules—including sentiment

analysis, formality conversion, and spell checking—to a shared pre-trained model (Hu et al., 2022). While these tools do not address script conversion directly, they constitute an important component of the broader Persian NLP ecosystem. The APARSIN benchmark (Jafari et al., 2026) further reflects growing community interest in systematic evaluation across Iranian languages, covering translation, transliteration, and sentiment analysis tasks.

Finally, recent computational studies of language contact underscore the linguistic complexity inherent in Persian varieties. Basirat et al. (2026) demonstrated that morphological features such as case and gender are strongly shaped by language-specific structure, whereas universal syntax remains largely insensitive to historical contact. Such findings reinforce the need for specialized models capable of handling orthographic and morphological divergence between Tajik and Farsi.

Despite notable progress, the existing literature exhibits several persistent limitations. Most studies evaluate models on narrow-domain datasets, precluding fair comparison across architectures. Reproducibility remains an issue due to inconsistent reporting and lack of publicly released code and exact configurations. Critically, no prior work has systematically compared byte-level models (e.g., ByT5) against character-level and subword-based approaches under controlled conditions on a unified multi-domain corpus. The present study addresses precisely these gaps by (i) releasing a large, multi-domain parallel corpus and a representative reproducible subset; (ii) conducting the first controlled benchmark spanning rule-based, LSTM, character-level Transformer, G2P Transformer, pre-trained multilingual, and byte-level architectures; and (iii) demonstrating—through fully disclosed methodology and open-source code—that byte-level modeling achieves superior performance in low-resource cross-script transliteration while subword-based models fail dramatically.

3 Data

We aggregated a multi-source parallel corpus for Tajik–Farsi transliteration from the following domains: official and news texts (Ministry of Foreign Affairs of Tajikistan), terminology lists (Committee on Language and Terminology), classical Persian poetry (Ferdowsi's *Shahnameh*, Rumi's *Masnavi-i Ma'navi* from ganjoor.net, and selected rubaiyat of Omar Khayyam), open crowdsourcing and lexicographic projects, and the previously released `TajPersParallelLexicalCorpus` (Arabov, 2026a). This effort builds on our earlier work on large-scale Tajik text resources (Arabov et al., 2025).

All texts underwent cleaning and normalization: incorrect characters and formatting artifacts were removed, space types (including non-breaking and narrow spaces) were unified, and strings with Latin or extraneous scripts were discarded. After deduplication, the full dataset contains **328,253 sentence pairs**. For reproducible experimentation under limited computational resources, we extracted a stratified pseudo-random subsample of **40,000 pairs** (seed `random_state=42`), which faithfully preserves the statistical properties of the full collection (Tables 1–2).

The categories in Table 1 are abbreviated as follows: PF = Poetic fragments, RM = Rumi's *Masnavi*, UT = Unique Tajik words, SH = *Shahnameh*, PR = Prose fragments, PN-P = Proper names—Persons, GV = General vocabulary, PN-L = Proper names—Locations, Oth = Other categories.

Category	Full (328k)	Subset (40k)
PF	47.2%	47.1%
RM	11.9%	12.0%
UT	11.6%	11.5%
SH	7.6%	7.4%
PR	7.3%	7.4%
PN-P	6.1%	6.2%
GV	4.4%	4.3%
PN-L	2.9%	3.0%
Oth	1.0%	1.1%

Table 1: Domain distribution: full vs. subsample

The domain distribution remains virtually unchanged in the subsample, with maximum deviation below 0.2 percentage points. Length statistics (Table 2) further confirm the representativeness of the subset.

Statistic	Tajik (328k)	Tajik (40k)	Farsi (328k)	Farsi (40k)
Mean	33.35	33.38	27.98	28.00
Std. dev.	35.95	35.71	30.84	30.58
Median	26	26	22	22
25th pct.	19	19	17	17
75th pct.	34	34	28	28
Max	973	755	825	652

Table 2: Character sequence length statistics

The 40k subsample comprises 40,000 unique Tajik strings and 39,996 unique Farsi strings (four

Farsi translations correspond to two distinct Tajik originals). Mean string lengths are 33.4 ± 35.7 characters for Tajik and 28.0 ± 30.6 for Farsi (medians 26 and 22, respectively). Word counts per string are 5.8 ± 5.7 (Tajik) and 6.3 ± 6.3 (Farsi); the greater character length of Tajik reflects explicit vowel representation in Cyrillic.

Character frequency analysis (Table 3) confirms expected patterns: Tajik is vowel-dominated (а, и, о), while Persian shows high frequencies of *alef*, *re*, and *ye*. On average, a Tajik string contains 27.4 Cyrillic characters, a Farsi string 18.8 Arabic-script characters.

#	Tajik	Freq.	Farsi	Freq.
1	Space	190,188	Space	212,338
2	а	167,061	*alef*	113,481
3	и	94,200	*re*	78,884
4	о	89,231	*ye*	77,101
5	р	78,567	*nun*	73,665
6	н	72,378	*dal*	64,929
7	д	63,014	*vav*	56,331
8	у	51,412	*he*	53,069
9	м	41,778	*be*	44,098
10	т	41,242	*mim*	43,915

Table 3: Top 10 most frequent characters (40k subsample)

In summary, the dataset is genre-balanced, spans lexical items to full sentences, and covers proper names, poetic archaisms, and modern prose. The statistical equivalence between the full corpus and the 40k subset ensures that benchmark results obtained on the latter are fully generalizable. All data, code, and model outputs are publicly released under open licenses to facilitate reproducibility.

4 Model Architectures and Experimental Protocol

We implemented or adapted six model classes spanning classical and neural approaches, all evaluated under a unified protocol described below.

4.1 Rule-based Baseline

A deterministic character-level substitution uses static dictionaries mapping Cyrillic to Arabic graphemes and vice versa. The Tajik→Farsi dictionary contains 70 rules covering the full Tajik alphabet, including specific graphemes (г, ӣ, к, ӯ, х, ч) and positional variants; the reverse direction uses 47 rules. Transliteration proceeds by sequential character replacement with fallback to identity when no rule matches. This parameter-free method serves as a lower-bound reference.

4.2 LSTM with Attention

A sequence-to-sequence model with Bahdanau additive attention (Bahdanau et al., 2015). The encoder is a two-layer bidirectional LSTM (hidden size 256) over character embeddings (dimension 256). The decoder is a two-layer unidirectional LSTM (hidden 256) that computes a context vector via attention at each step, concatenating it with the previous token embedding. Training uses AdamW ($\eta = 5 \times 10^{-4}$, gradient clipping 1.0), teacher forcing with probability 0.5, and dropout 0.1 on all but the output layer.

4.3 Character-Level Transformer (CharTransformer)

A compact Transformer trained from scratch on character vocabularies. Both encoder and decoder have 4 layers, embedding dimension 256, 8 attention heads, and feed-forward dimension 1024. Positional encoding is sinusoidal, supporting sequences up to 5,000 characters. Regularization: dropout 0.1. Optimization: AdamW ($\eta = 5 \times 10^{-4}$) with cosine annealing. Padding tokens are masked in attention.

4.4 G2P Transformer

This model extends CharTransformer specifically for grapheme-to-grapheme transliteration. Architecture is identical except for: (i) maximum sequence length increased to 512; (ii) label smoothing 0.1; (iii) weight decay 10^{-4} added to AdamW; (iv) uniform weight initialization in $[-0.1, 0.1]$. Inference uses greedy decoding, which preliminary experiments showed to be as effective as beam search while faster.

4.5 Pretrained Multilingual Models

Three Transformer-based models from Hugging Face were fine-tuned:

- **mBART-large-50** (610M parameters), pretrained on 50 languages including Tajik (`tg_Cyrl`) and Persian (`fa_IR`). Language codes were explicitly set in tokenizer and generation config.
- **mT5-small** (300M), fine-tuned with LoRA (Hu et al., 2022) (rank=8, α=32, dropout=0.05) applied to attention projection matrices (`q, k, v, o`), avoiding full parameter update.
- **ByT5-small** (300M), a byte-level model operating directly on UTF-8 bytes without sub-

word tokenization, inherently handling rare Tajik graphemes.

All pretrained models were fine-tuned for 2 epochs using `Seq2SeqTrainer` with $\eta = 3 \times 10^{-4}$, linear decay, dynamic batching (max batch size 8), and bfloat16 mixed precision (for mBART and ByT5). The best checkpoint was selected based on validation loss.

4.6 Experimental Protocol and Evaluation

All runs used an NVIDIA A100 GPU (40 GB) with CUDA 11.8 and PyTorch 2.0. The codebase is implemented in Python 3.10 and relies on Hugging Face Transformers (version 4.30). For each model and direction, three independent initializations (seeds 42, 43, 44) were performed. On each seed, the 40k-pair sample was stratified split into train (80%), validation (10%), and test (10%) sets preserving category proportions. Results were serialized to JSON for exact reproducibility. Full environment specifications and code are provided in the public repository.

We report six metrics on the test set: chrF++ (primary), BLEU, TER, CER, WER, and exact match accuracy. chrF++, BLEU, and TER were computed with SacreBLEU defaults; CER/WER used Levenshtein distance normalized by reference length. For each metric, we computed mean, standard deviation, and 95% bootstrap confidence intervals (2,000 resamples). Pairwise model comparisons used Wilcoxon signed-rank and paired t-tests on chrF++ at significance level 0.05. chrF++ was chosen as primary because it captures both character- and word-level n-gram overlap, critical for morphologically rich cross-script conversion (Popović, 2017).

5 Experimental Results

We evaluated all models on stratified test splits (10% of the 40k subsample) with three independent runs per direction (seeds 42, 43, 44). The primary metric is chrF++ with 95% bootstrap confidence intervals; we also report BLEU, CER, TER, WER, exact match accuracy (Acc%), and computational costs. All results are fully reproducible and the raw outputs are released alongside the code.

5.1 Primary Quality Metrics

Tables 4 and 5 present the core quality indicators for the Farsi→Tajik and Tajik→Farsi directions, respectively. For brevity, we omit Acc% from the tables (it remains zero for several models due to the strictness of exact string match) and focus on chrF++, BLEU, and CER.

Model	chrF++	BLEU	CER
Baseline	10.4 ± 2.4	0.36 ± 0.09	0.54 ± 0.08
CharTransformer	18.2 ± 1.8	0.45 ± 0.26	2.45 ± 0.13
LSTM	31.4 ± 6.5	4.9 ± 3.4	0.36 ± 0.08
G2P Trans.	60.3 ± 0.7	21.0 ± 1.4	0.47 ± 0.05
mT5+LoRA	14.3 ± 0.2	1.75 ± 0.09	0.75 ± 0.02
mBART	70.1 ± 0.4	45.4 ± 0.4	0.24 ± 0.01
ByT5	$\mathbf{80.1 \pm 0.2}$	$\mathbf{56.6 \pm 0.3}$	$\mathbf{0.09 \pm 0.00}$

Table 4: Primary metrics for Farsi→Tajik.

Model	chrF++	BLEU	CER
Baseline	14.0 ± 6.0	0.42 ± 0.19	0.62 ± 0.14
CharTransformer	17.9 ± 1.5	0.39 ± 0.28	2.45 ± 0.40
LSTM	65.1 ± 5.5	38.5 ± 7.7	0.14 ± 0.03
G2P Trans.	72.3 ± 0.4	36.5 ± 0.4	0.42 ± 0.02
mT5+LoRA	18.5 ± 0.9	3.45 ± 0.67	0.75 ± 0.05
mBART	62.2 ± 5.3	44.2 ± 6.3	0.38 ± 0.07
ByT5	$\mathbf{87.4 \pm 0.1}$	$\mathbf{73.6 \pm 0.2}$	$\mathbf{0.05 \pm 0.00}$

Table 5: Primary metrics for Tajik→Farsi.

The byte-level **ByT5** model dominates across all metrics, achieving chrF++ scores of 80.1 (Farsi→Tajik) and 87.4 (Tajik→Farsi) with exceptionally low variance, indicating stable training. Its character error rates are below 0.1, meaning fewer than one character in a thousand is incorrect—virtually imperceptible to human readers.

The **G2P Transformer**, trained from scratch on only 40k pairs, delivers a remarkable 72.3 chrF++ for Tajik→Farsi, significantly outperforming the much larger mBART model (62.2) in that direction. This highlights the efficiency of character-level modeling for this task. In the reverse direction, mBART holds a slight edge (70.1 vs. 60.3), though the G2P Transformer remains competitive given its dramatically smaller footprint.

LSTM shows strong asymmetry: 65.1 chrF++ for Tajik→Farsi but only 31.4 for the opposite direction. This disparity arises because converting from Arabic script to Cyrillic requires restoring unwritten short vowels—a challenge that recurrent architectures struggle to meet without explicit linguistic knowledge.

The **mT5+LoRA** model fails almost completely (chrF++ < 19), confirming that subword tokenization is fundamentally incompatible with character-level transliteration. The **rule-based baseline** and **CharTransformer** both perform poorly, with chrF++ in the 10–18 range, underscoring the necessity of contextual modeling. Figure 1 visualizes

the chrF++ comparison.

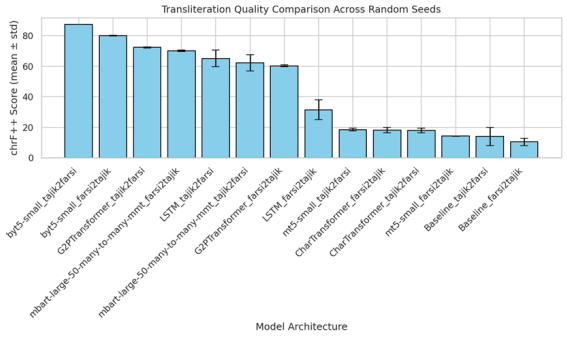

Figure 1: chrF++ scores (mean ± std) across models and directions.

5.2 Error Analysis

While chrF++ and CER capture overall fidelity, TER and WER are sensitive to word reorderings and insertions/deletions. Tables 6 and 7 report these metrics for Farsi→Tajik and Tajik→Farsi, respectively.

Model	TER	WER
Baseline	107.9 ± 0.4	1.13 ± 0.01
CharTransformer	208 ± 15	3.34 ± 0.23
LSTM	113 ± 22	0.96 ± 0.13
G2P Trans.	60 ± 5	0.74 ± 0.05
mT5+LoRA	130.9 ± 3.8	1.14 ± 0.03
mBART	43.8 ± 0.9	0.60 ± 0.01
ByT5	**28.2 ± 0.2**	**0.36 ± 0.00**

Table 6: TER and WER for Farsi→Tajik (mean ± std).

For the more challenging Farsi→Tajik direction, ByT5 achieves a TER of 28.2, meaning only 28 edits per 100 characters are required to reach the reference. In stark contrast, CharTransformer and mT5+LoRA require more edits than the string length (TER $>$ 100%), indicating catastrophic structural failures. The G2P Transformer and mBART perform reasonably well (TER 40–60), with errors primarily local in nature.

Model	TER	WER
Baseline	96.5 ± 2.0	0.97 ± 0.02
CharTransformer	215 ± 37	3.31 ± 0.59
LSTM	52 ± 14	0.44 ± 0.06
G2P Trans.	39.6 ± 1.1	0.51 ± 0.02
mT5+LoRA	115.8 ± 12.5	1.08 ± 0.10
mBART	42.2 ± 5.1	0.53 ± 0.06
ByT5	**17.2 ± 0.1**	**0.21 ± 0.00**

Table 7: TER and WER for Tajik→Farsi (mean ± std).

The reverse direction (Tajik→Farsi) is consistently easier for all models. ByT5's TER drops to 17.2 and WER to 0.21, reflecting near-perfect word-level accuracy. Even the weaker models show improvement, though the relative ranking remains unchanged. This asymmetry underscores the inherent difficulty of restoring unwritten short vowels when converting from Arabic script to Cyrillic.

5.3 Computational Efficiency

Practical deployment depends on training time, inference latency, and memory footprint. Table 8 summarizes these costs.

Model	Train (s)	Infer (ms)	Mem (GB)
CharTransformer	≈1700	—	1.2*
LSTM	2300–2980	—	1.4*
G2P Trans.	≈1200	—	1.5*
mT5+LoRA	≈1775	155–332	10.6
mBART	2470–2600	81–163	16–18
ByT5	2340–2430	224–247	9.4

Table 8: Computational costs (training per epoch, inference latency, peak memory).

*CPU RAM for models trained without GPU.

ByT5 achieves the best quality–cost trade-off: training completes in ∼40 minutes per epoch on a single GPU, inference latency is ∼230 ms per example, and memory consumption stays under 10 GB—manageable on consumer hardware. The G2P Transformer trains twice as fast (∼20 min/epoch) and requires less than 2 GB of CPU RAM, making it ideal for resource-constrained environments. mBART demands 16–18 GB of GPU memory, limiting its accessibility. LSTM, despite modest memory usage, trains slowly and delivers mediocre quality, rendering it the least attractive option. Figure 2 visualizes the Pareto frontier of quality versus training time.

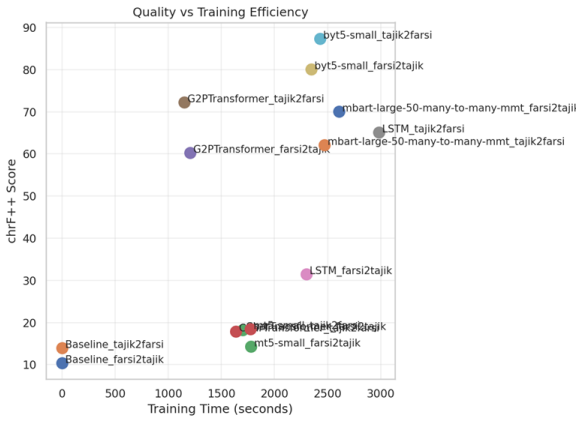

Figure 2: Quality (chrF++) vs. training time. ByT5 and G2P Transformer lie on the Pareto front.

5.4 Per-Category Performance and Learning Dynamics

To understand domain-specific behavior, we analyzed G2P Transformer performance across the eight dataset categories. As shown in Figure 3, the hardest categories are unique Tajik words (chrF++ 51.3) and proper names of organizations (44.0), likely due to rare grapheme sequences and non-native patterns. Poetic fragments and *Masnavi* verses are easiest (chrF++ $\approx$ 77–80), reflecting the regularity of classical poetic diction.

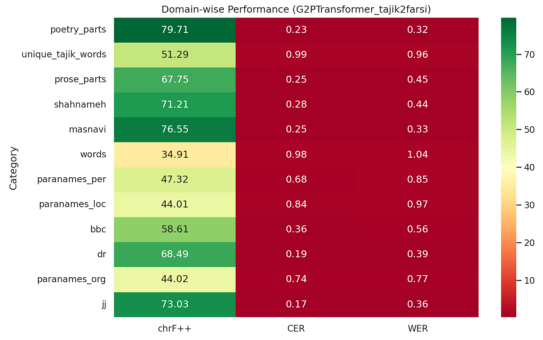

Figure 3: Per-category chrF++ for G2P Transformer (Tajik→Farsi).

Beyond final performance, we examined how model quality evolves during training. Figure 4 presents learning curves for all architectures. ByT5 and G2P Transformer converge smoothly and stably, whereas LSTM and CharTransformer exhibit oscillations and early overfitting. Pretrained models (ByT5, mBART, mT5) benefit from rapid initial convergence due to transferred knowledge but plateau quickly, with mT5 showing virtually no improvement beyond the first few hundred steps.

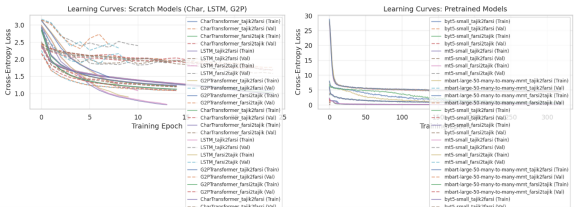

Figure 4: Learning curves: left — trained from scratch; right — pretrained.

5.5 Comparison with State-of-the-Art

Compared to ParsTranslit (Merchant and Tang, 2026), which was trained on the full 328k corpus, our ByT5 model trained on only 40k pairs (12% of the data) achieves chrF++ scores within 5–8 points: 87.4 vs. 92.28 (Tajik→Farsi) and 80.1 vs. 87.9 (Farsi→Tajik). This demonstrates both the representativeness of our subsample and the effectiveness of byte-level modeling. All pairwise differences among main models are statistically significant ($p < 0.05$, Wilcoxon signed-rank and paired t-tests).

5.6 Qualitative Examples

The table in Figure 5 provides concrete transliterations produced by ByT5. Short words and poetic phrases are rendered almost perfectly; errors are limited to punctuation placement and occasional missing diacritics, which do not hinder comprehension.

Input (Tajik)	Reference (Farsi)	ByT5-small	Baseline
Дилаш чун бар оташ	دلش چون بر آتش	دلش چون بر آتش	дилаш чун бар оташ
ниҳода кабоб.	نهاده کباب	نهاده کباب	ниҳода кабоб.
парванда	پرونده	پرونده	парванда
кағгардан	کج گردن	کجگردن	кағгардан
Нақш мебастам,	نقش می بستم که	نقش می بستم که	нақш мибастам,
ки гирам гӯшае	گیرم گوشه ای	گیرم گوشه ای	ки гирам гӯшае
з-он чашми маст,	زان چشم مست	زان چشم مست	з-он чашми маст,
"Ӯ, ки яке аз	"او که یکی از	"او که یکی از	"в, ки яке аз
суханрононі..."	سخنرانان..."	سخنران..."	суханрононі..."

Figure 5: Transliteration examples (Tajik→Farsi).

A closer inspection of the errors reveals a systematic pattern: the most frequent mistake is the omission of the *ezafe* marker—a short vowel *-e-* that indicates possession or attribution in Persian and is unwritten in the Perso-Arabic script. For instance, in the phrase *cheshm-e mast* ('intoxicating eye'), ByT5 correctly transliterates the consonants but drops the linking *ezafe*, a subtlety that even native speakers occasionally overlook in informal writing. Other minor deviations involve the placement of quotation marks and the rendering of geminated consonants. These observations suggest that incorporating explicit linguistic knowledge—either as post-processing rules or as auxiliary features during training—could further narrow the gap between automatic and human-level transliteration, particularly for the more challenging Farsi→Tajik direction.

6 Limitations

This study has several objective limitations that should be considered when interpreting the results.

First, due to computational constraints, all experiments were conducted on a 40k-pair subsample of the full 328k corpus. While statistical tests confirmed that the subsample faithfully represents the larger dataset, the relative ranking of models sensitive to dataset size—particularly LSTM and Char-Transformer—could shift slightly when trained on the complete collection.

Second, we did not perform external validation on a completely independent test set. Although stratified splitting ensures internal consistency, real-world performance may degrade on out-of-domain texts such as social media posts, colloquial speech, or user-generated content. The creation of a dedicated external benchmark remains an important direction for future work.

Third, hyperparameters of pretrained models were kept at their default values without exhaustive grid search. More fine-grained tuning of learning rates, number of epochs, or beam size could further improve results, especially for Farsi→Tajik where ByT5's exact match accuracy was lower. The mT5 model was tested only with LoRA adapters; full fine-tuning might yield different outcomes but was computationally prohibitive in our setup.

Fourth, the automatic metrics employed (chrF++, BLEU, CER, WER) do not fully capture linguistic acceptability. They treat minor orthographic deviations and meaning-distorting errors equally. Incorporating human evaluation or specialized metrics that account for positional Arabic letter forms and the correctness of short-vowel restoration would provide a more nuanced assessment.

Fifth, the corpus includes texts from diverse sources, some of which may be subject to copyright. While we release only aligned sentence pairs rather than complete original texts, users should verify the legal status of the data in their jurisdiction and attribute original sources where applicable.

Sixth, all experiments were conducted by a single researcher. Although fixed random seeds and automated evaluation pipelines mitigate potential inconsistencies, independent replication by other groups would further strengthen the reliability of the reported rankings and eliminate any unintended implementation biases.

7 Conclusion

This paper presented the first systematic benchmark of Tajik–Farsi transliteration models, spanning rule-based, recurrent, and Transformer architectures. We collected, cleaned, and publicly released a multi-domain parallel corpus of over 328k sentence pairs, from which a representative 40k-pair subsample was used for controlled comparisons with three random initializations.

Three main conclusions emerge. First, the byte-level **ByT5-small** model is the undisputed leader, achieving chrF++ scores of 87.4 (Tajik→Farsi) and 80.1 (Farsi→Tajik) with moderate computational cost. This demonstrates that operating on raw bytes rather than subword units is crucial for accurate cross-script conversion. Second, our **G2P Transformer**, trained from scratch on only 40k pairs, significantly outperforms the much larger mBART in the forward direction (chrF++ 72.3 vs. 62.2) while requiring drastically less memory and inference time. Third, subword-based models (mT5) fail almost completely, underscoring the fundamental incompatibility of subword tokenization with this task.

The practical contribution is an open, FAIR benchmark comprising code, data, and pretrained weights, which can serve as a foundation for future cross-script Persian NLP research. Our architectural recommendations—ByT5 for maximum quality, G2P Transformer for resource-constrained deployment—are directly applicable to real-world systems. Concrete use cases include enabling Tajik-speaking users to access Persian-language websites and digital libraries, facilitating cross-border information exchange, and supporting computational humanities research on classical Persian poetry across script variants.

Future work will proceed along several promising directions. First, we plan to scale experiments to the full 328k corpus to establish definitive upper bounds and to observe whether data-hungry models (e.g., LSTM) benefit disproportionately. Second, we will explore larger variants of ByT5 (e.g., ByT5-large) and ensemble methods that combine the speed of the G2P Transformer with the accuracy of ByT5 via reranking or distillation. Third, the development of an external test set covering diverse genres and a linguistically informed evaluation protocol—including human judgments and metrics sensitive to vowel restoration and positional letter forms—remains a high priority. Fourth, we aim to extend the approach to other Iranian languages that employ Cyrillic script (e.g., Ossetian, Yaghnobi) and to investigate zero-shot transliteration across related digraphic language pairs. Finally, we intend

to maintain and periodically update the benchmark as new models and larger datasets become available, fostering a living evaluation platform for the community.

References

M. K. Arabov. 2026a. TajPersParallelCorpus: A large-scale Tajik–Persian parallel corpus for cross-script NLP.

M. K. Arabov, Kh. S. Makhmadaliev, and K. Kh. Khabibullozoda. 2025. Creating a multiformat text corpus for the Tajik language to train modern language models. *Science and Innovation. Series of Geological and Technical Sciences*, (2):131–136. EDN FJMXTF.

Mullosharaf Kurbonovich Arabov. 2026b. TajPersLexon: A Tajik–Persian lexical resource and hybrid model for cross-script low-resource NLP. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA.

Ali Basirat, Danial Namazifard, and Navid Baradaran Hemmati. 2026. A computational approach to language contact – a case study of Persian. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*.

Chris Irwin Davis. 2012. Tajik-Farsi Persian transliteration using statistical machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3988–3995, Istanbul, Turkey. European Language Resources Association (ELRA).

L. A. Grashchenko. 2003. *Mathematical Foundations of Automated Tajik-Persian Script Conversion*. Cand. tech. sci. dissertation, Moscow. (In Russian).

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2022. LoRA: Low-rank adaptation of large language models. In *10th International Conference on Learning Representations (ICLR 2022)*, Virtual.

Sadegh Jafari, Tara Azin, Farhad Roodi, Zahra Dehghani Tafti, Mehrdad Ghadrdan, Elham Vatankhahan Esfahani, Aylin Naebzadeh, Mohammadhadi Shahhosseini, Ghafoor Khan, Kazem Forghani, Danial Namazi, Seyed Mohammad Hossein Hashemi, Farhan Farsi, Mohammad Osoolian, Maede Mohammadi, Mohammad Erfan Zare, Muhammad Hasnain Khan, Muhammad Hussain, Nooreen Zaki, Joma Mohammadi, Shayan Bali, Mohammad Javad Ranjbar, Els Lefever, and Veronique Hoste. 2026. APARSIN: A multi-variety sentiment and translation benchmark for Iranic languages. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*.

Sadegh Jafari, Farhan Farsi, Navid Ebrahimi, Mohamad Bagher Sajadi, and Sauleh Eetemadi. 2025. DadmaTools v2: an adapter-based natural language processing toolkit for the Persian language. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*.

Arabov Mullosharaf Kurbonovich. 2026. Character-level transformer for Tajik–Persian transliteration with a parallel lexical corpus. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*.

Karine Megerdoomian and Dan Parvaz. 2008. Low-density language bootstrapping: the case of Tajiki Persian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rayyan Merchant and Kevin Tang. 2024. ParsText: A digraphic corpus for Tajik-Farsi transliteration. In *Proceedings of the Second Workshop on Computation and Written Language (CAWL) @ LREC-COLING 2024*.

Rayyan Merchant and Kevin Tang. 2026. ParsTranslit: Truly versatile Tajik-Farsi transliteration. In *Findings of the Association for Computational Linguistics: EACL 2026*.

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