INTRODUCTION AND MOTIVATION

- Persian language exhibits **digraphia**: Cyrillic in Tajikistan vs. Arabic in Iran/Afghanistan, creating a digital content barrier.
- Automatic transliteration is a low-cost bridge, but faces challenges: Arabic→Cyrillic requires restoring unwritten short vowels; reverse needs positional shaping and grapheme ambiguity.
- Prior work limited by scarce parallel data and narrow-domain evaluation. **Goal:** first systematic benchmark across six model families on a unified multi-domain corpus.
- **Main contributions:**
  1. New corpus of 328,253 sentence pairs + stratified 40k subset (FAIR).
  2. Controlled comparison of rule-based, LSTM, char-level Transformer, G2P Transformer, pretrained multilingual (mBART, mT5+LoRA, ByT5).
  3. Evidence that byte-level models dominate; subword models fail.

RELATED WORK

- Early SMT-based transliteration (Davis, 2012); theoretical foundations via intermediate Latin transcription (Grashchenko, 2003).
- Neural G2P Transformers achieved chrF++ 59 (FarsiTajik) and 74 (reverse), establishing baselines.
- Datasets: ParsText (2,813 pairs), Shahnameh-derived corpus; ParsTranslit (SOTA: 87.9 / 92.3 chrF++ on full 328k).
- Our prior resources: TajPersLexon (40k pairs), char-level Transformer with CER 0.3182.
- Persian NLP ecosystem: DadmaTools V2, APARSIN benchmark. **Gap:** no controlled byte-level vs. subword comparison on unified multi-domain data.

DATA: MULTI-DOMAIN PARALLEL CORPUS

Sources: official/news, terminology, classical poetry (Shahnameh, Masnavi, Khayyam), crowdsourcing, TajPersParallelLexicalCorpus.
Cleaning: removal of artifacts, space unification, script filtering, deduplication → **328,253 pairs**. Stratified subsample: **40,000 pairs** (seed 42).

Category	Full (328k)	Subset (40k)
Poetic fragments (PF)	47.2%	47.1%
Rumi's Masnavi (RM)	11.9%	12.0%
Unique Tajik words (UT)	11.6%	11.5%
Shahnameh (SH)	7.6%	7.4%
Prose fragments (PR)	7.3%	7.4%
Proper names–Persons	6.1%	6.2%
General vocabulary	4.4%	4.3%
Proper names–Locations	2.9%	3.0%
Other	1.0%	1.1%

Table: Domain distribution preserved

Length statistics (mean chars): Tajik 33.4, Farsi 28.0; subset statistically equivalent.
Character frequencies (top 5, 40k subset):

Tajik	Freq.	Farsi	Freq.
space	190,188	space	212,338
	167,061	alef	113,481
	94,200	re	78,884
	89,231	ye	77,101
	78,567	nun	73,665

MODEL ARCHITECTURES

1. **Rule-based** 70 rules (TajFar), 47 (FarTaj), parameter-free baseline.
2. **LSTM + attention** 2-layer biLSTM encoder (256), attention decoder, AdamW, dropout 0.1.
3. **CharTransformer** 4 layers, emb 256, 8 heads, sinusoidal pos, trained from scratch.
4. **G2P Transformer** CharTransformer + label smoothing 0.1, weight decay 10^{-4}, max len 512.
5. **Pretrained multilingual**:
   - mBART-large-50 (610M) fine-tuned.
   - mT5-small + LoRA (rank 8, α=32).
   - ByT5-small (byte-level, no subwords).

Neural models use character/byte vocabularies except mT5 (subword).

EXPERIMENTAL SETUP

- 40k pairs: 80% train, 10% val, 10% test (stratified), 3 seeds (42, 43, 44).
- Metrics: chrF++ (primary), BLEU, CER, TER, WER, exact match. Bootstrap CIs (95%), Wilcoxon and t-tests ($p < 0.05$).
- Hardware: NVIDIA A100 40GB, PyTorch 2.0, Transformers 4.30.

PRIMARY RESULTS (chrF++ / BLEU / CER)

Farsi → Tajik

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	**80.1 ± 0.2**	**56.6 ± 0.3**	**0.09 ± 0.00**

Tajik → Farsi

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	**87.4 ± 0.1**	**73.6 ± 0.2**	**0.05 ± 0.00**

Key observations: ByT5 dominates all metrics; G2P Transformer outperforms mBART in TajikFarsi; LSTM asymmetric; mT5 fails.

chrF++ COMPARISON

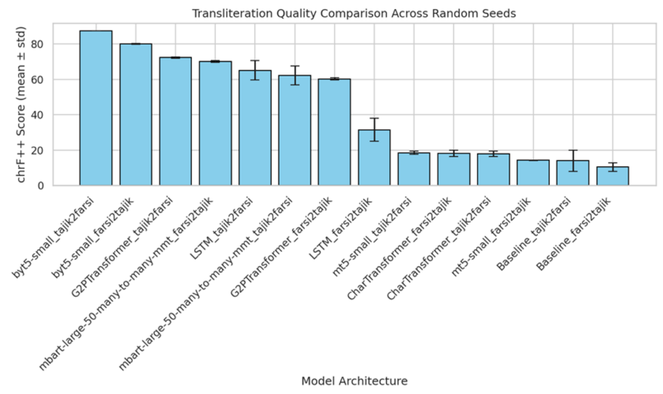

Figure: Mean chrF++ across models and directions

ERROR ANALYSIS (TER / WER)

Farsi→Tajik

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**

Tajik→Farsi

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**

Most common error: omission of *ezafe* marker (short vowel), e.g., cheshm-e mast → cheshm mast.

COMPUTATIONAL EFFICIENCY

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

*CPU RAM. ByT5 best quality–cost trade-off; G2P fastest/lowest memory.

QUALITY VS. TRAINING TIME

Figure: Pareto front: ByT5 and G2P Transformer

PER-CATEGORY PERFORMANCE

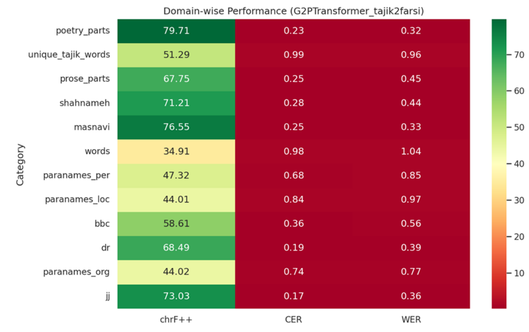

Figure: chrF++ per category (G2P Transformer, TajikFarsi)

Hardest: unique Tajik words, proper names; easiest: poetry.

LEARNING DYNAMICS

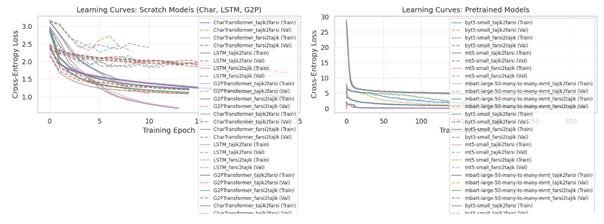

Figure: Learning curves: from scratch vs. pretrained

QUALITATIVE EXAMPLES

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
суханронони..."	سخنران..."	سخنران..."	суханронони..."

Figure: ByT5 transliteration examples (TajikFarsi)

COMPARISON WITH SOTA

ByT5 (40k) vs. ParsTranslit (328k): chrF++ 87.4 vs. 92.28 (TajFar), 80.1 vs. 87.9 (FarTaj). Our model uses only 12% of data, confirming byte-level efficiency. All pairwise differences significant.

LIMITATIONS

- Experiments on 40k subset; full data may change rankings.
- No external test set.
- Default hyperparameters for pretrained models.
- Automatic metrics; no human evaluation.
- Potential copyright issues in some sources.
- Single researcher; independent replication needed.

CONCLUSION

- ByT5 achieves SOTA: chrF++ 87.4 / 80.1 with moderate cost.
- G2P Transformer efficient and competitive (72.3 TajikFarsi).
- Subword tokenization (mT5) fails; character/byte-level necessary.
- Open FAIR benchmark released: code, data, models.
- **Recommendation:** ByT5 for maximum quality; G2P Transformer for resource-constrained deployment.
- Future: scale to full corpus, larger ByT5, external test set, extend to other Iranian languages.