SilkRoadNLP 2026 • 1st Workshop on NLP and LLMs for Iranian Language Family • March 29, 2026 • Rabat, Morocco

PersianPunc: A Large-Scale Dataset and BERT-Based Approach for Persian Punctuation Restoration

Mohammad Javad Ranjbar Kalahroodi[1], **Heshaam Faili**[1], **Azadeh Shakery**[1,2]
[1]University of Tehran, Iran [2]Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

ABSTRACT

We introduce **PersianPunc**, a 17M-sample dataset for Persian punctuation restoration — the largest of its kind, spanning 6 diverse corpora. Fine-tuning **ParsBERT** achieves 91.33% Macro F1, outperforming GPT-4o (85.96%) while avoiding over-correction and requiring far less compute. Dataset and model are publicly available on Hugging Face.

Why Punctuation Matters

WITHOUT PUNCTUATION:

```
bakhshesh lazem nist e’damesh konid
```

↓ ADD COMMA

WITH PUNCTUATION:

```
bakhshesh, lazem nist e’damesh konid
```

→ Meaning completely reversed by punctuation!

- ASR systems remove punctuation
- Meaning ambiguity increases
- Persian lacks large, clean datasets

PersianPunc Dataset

17M
Sentences
Largest Persian punctuation dataset

6	5
Source Corpora	Punctuation Types

- Formal: Bijankhan-Peykare, Persian Medical QA, Persian Wikipedia | Informal: Telegram Channels, Farsi Stories, Blog Dataset V2
- SHA-256 deduplication prevents data leakage
- Distribution aligned with real-world punctuation usage

Model Architecture

Input Text (No Punctuation)

↓

Fine-tuned ParsBERT
WordPiece Tokenization + Contextualized Embeddings

↓

Classification Layer
5 Classes: EMPTY, COMMA, PERIOD, QUESTION, COLON

↓

Output (With Punctuation)

Training

AdamW • 3 epochs • Dropout 0.1

Efficiency

Lightweight • Real-time ready

Performance Results

OUR MODEL (ParsBERT)

91.33% Macro F1 **61.80%** Exact Match

✓ No Over-Correction

Model	Macro F1	Exact Match	Over-Corr.
Our ParsBERT	91.33%	61.80%	No
GPT-4o	85.96%	50.10%	Yes
GPT-4o-mini	79.54%	38.01%	Yes

Key Advantage: Specialized BERT outperforms LLMs in Macro F1 and avoids Over-Correction (LLMs modify words, delete/replace text beyond adding punctuation). Exact Match (FSM) = % of sentences where predicted punctuation perfectly matches gold standard.

Per-Class Macro F1 Performance

Note: The dataset contains other punctuation (e.g., !), but for comparison with baselines, the model predicts only 5 classes

Key Contributions

1. **Largest Persian punctuation dataset**
   17M sentences across diverse domains
2. **Strong, lightweight baseline**
   Optimized for real-time ASR pipelines
3. **No hallucination or rewriting**
   Preserves lexical integrity
4. **Publicly available**
   Dataset and model on Hugging Face

Access Dataset & Model

Dataset

Model

Workshop
1st Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP 2026)

Conference
19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)

Contact
mohammadjranjbar@ut.ac.ir
hfaili@ut.ac.ir
shakery@ut.ac.ir