SilkRoadNLP 2026

The First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP 2026)

Proceedings of the Workshop

March 29, 2026

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-371-5

Preface

It is our pleasure to welcome you to the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP 2026), held as a half-day workshop on March 29, 2026, at EACL 2026 in Rabat, Morocco.

As the first ACL-affiliated workshop dedicated to the Iranian/Iranic linguistic family, we are excited to bring together linguists (both computational and non-computational), AI researchers, and community experts to foster open collaboration and promote culturally informed, inclusive, and ethical approaches to multilingual Natural Language Processing (NLP) and Large Language Models (LLMs).

We received 18 submissions, out of which 14 papers were accepted. Of these, 6 will be presented as oral presentations, and 8 as posters, resulting in an acceptance rate of 33% for the former and 44% for the latter. We regret that the Internet blackout in Iran—which began on January 8th, 2026—prevented many in Iran from submitting their research to us and caused profound worry for Iranians in the diaspora who were unable to contact their family and friends. In future editions of SilkRoadNLP, we hope to welcome our Iranian colleagues who were unable to join us.

Despite these difficulties, contributions to SilkRoadNLP covered a broad range of Iranic languages, demonstrating the community's desire and need for a space of its own within the broader NLP community. Although the most represented language is—as expected—(Iranian) Persian, our program covers lesser-represented languages and varieties such as: Dari (Afghan Persian), Dezfuli, Esfahani, Hazaragi, Kabuli, Kalhori (Kurdish), Khorasani, Luri Bakhtiari, Pashto, Semnani, Shirazi, Shughni, Tajiki (Tajik Persian), Tonekaboni, Yazdi, and Zoroastrian Dari. Beyond just linguistic diversity, this list also showcases the orthographic diversity of the Iranian language family, as only four possess a formal, standardized orthography (Iranian Persian, Dari, Tajiki, and Pashto), two use the Tajik-Cyrillic script (Tajik and Shughni), and one is exclusively orally transmitted (Zoroastrian Dari).

In terms of topics, the papers at SilkRoadNLP focus on key areas such as: dataset curation for low-resource languages, culturally-aware sentiment analysis, automatic speech recognition, and the reasoning and comprehension capabilities of LLMs. In lieu of an invited speaker or panel, we end our program with a community discussion on the future of Iranic NLP.

We are very thankful to our program committee, authors, and the EACL Workshop Chairs for their contributions which have allowed us to develop a robust inaugural workshop. We look forward to stimulating presentations and discussions at SilkRoadNLP 2026, and hope this workshop inspires continued efforts in Iranic NLP.

SilkRoadNLP 2026 Organizers

Organizing Committee

Organizers

Karine Megerdoomian, Zoorna Institute, USA
Rayyan Merchant, Zoorna Institute, USA
Ehsaneddin Asgari, Qatar Computing Research Institute, Qatar
Ali Salehi, University of Buffalo, USA

Program Committee

Program Committee

Hiwa Asadpour, Goethe University Frankfurt and University of Saarland, Germany
Hadi Asghari, TU Berlin, Germany
Zahra Bahmani, Sharif University of Technology, Iran
Ali Emami, Emory University, USA
Heshaam Faili, University of Tehran, Iran
Masood Ghayoomi, Institute for Humanities and Cultural Studies, Iran
Nizar Habash, New York University Abu Dhabi, UAE
Hossein Hassani, University of Kurdistan Hewlêr, Iraq
Carina Jahani, Uppsala University, Finland
Amir Hossein Kargaran, Ludwig Maximilian University of Munich, Germany
Mohsehn Mahdavi Mazdeh, University of Arizona, USA
Yury Makarov, University of Cambridge, UK
Masoud Makrehchi, Ontario Tech University, Canada
Corey Miller, Rev, USA
Seyed AbolGhasem Mirroshandel, University of Guilan, Iran
Behrang Mohit, Apple, USA
Salar Mohtaj, German Research Center for Artificial Intelligence (DFKI), Germany
Isar Nejadgholi, National Research Council Canada, Canada
Muhtasham Oblokulov, Independent Researcher, Germany
Mohammad Taher Pilehvar, Cardiff University, UK
Mohammad Ali Sadraei Javaheri, Part AI Research Center, Iran
Mehrnoush Shamsfard, Shahid Beheshti University, Iran
Yadollah Yaghoobzadeh, University of Tehran, Iran

Table of Contents

Program

Sunday, March 29, 2026

09:00 - 09:15 *Opening & Welcome*

09:15 - 10:15 *Session 1: Low-Resource Languages & Dialectal Diversity*

TajPersLexon: A Tajik–Persian Lexical Resource and Hybrid Model for Cross-Script Low-Resource NLP
Mullosharaf Kurbonovich Arabov

APARSIN: A Multi-Variety Sentiment and Translation Benchmark for Iranic Languages
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

Shughni Machine Translation Enhanced by Donor Languages
Dmitry Novokshanov, Innokentiy S. Humonen and Ilya Makarov

10:15 - 10:55 *Coffee Break & Poster Session*

A Computational Approach to Language Contact – A Case Study of Persian
Ali Basirat, Danial Namazifard and Navid Baradaran Hemmati

Benchmarking Offensive Language Detection in Persian and Pashto
Zahra Bokaei, Bonnie Webber and Walid Magdy

Do Large Language Models Understand Double Mismatches? Evidence from Farsi
Maryam Mohammadi

One Language, Three of Its Voices: Evaluating Multilingual LLMs Across Persian, Dari, and Tajiki on Translation and Understanding Tasks
Noor Mairukh Khan Arnob and Abu Bakar Siddique Mahi

Online Polarization Detection in Persian (Farsi) Social Media
Saeedeh Davoudi and Nazli Goharian

ParsCORE: The Persian Corpus of Online Registers
Alireza Razzaghi, Erik Henriksson and Veronika Laipalla

Sunday, March 29, 2026 (continued)

PersianPunc: A Large-Scale Dataset and BERT-Based Approach for Persian Punctuation Restoration
Mohammad Javad Ranjbar Kalahroodi, Heshaam Faili and Azadeh Shakery

Segmentation Strategy Matters: Benchmarking Whisper on Persian YouTube Content
Reihaneh Iranmanesh, Rojin Ziaei and Joe Garman

10:55 - 11:55 *Session 2: Understanding, Reasoning & Generation*

Unmasking the Factual-Conceptual Gap in Persian Language Models
Alireza Sakhaeirad, Ali Ma'manpoosh and Arshia Hemmat

PMWP: A Benchmark for Math Word Problem Solving in Persian
Marzieh Abdolmaleki, Mehrnoush Shamsfard, Veronique Hoste and Els Lefever

Multi-modal Neural Machine Translation for Low-Resource Classical Persian Poetry: A Culture-Aware Evaluation
Soheila Ansari, Mounir Boukadoum and Fatiha Sadat

11:55 - 12:25 *Community Discussion: Future of Iranic NLP*

12:25 - 12:30 *Closing Remarks*

Unmasking the Factual-Conceptual Gap in Persian Language Models

Alireza Sakhaeirad*
EPFL

Ali Ma'manpoosh
University of Isfahan

Arshia Hemmat
University of Oxford

Abstract

While emerging Persian NLP benchmarks have expanded into pragmatics and politeness, they rarely distinguish between memorized cultural facts and the ability to reason about implicit social norms. We introduce DIVANBENCH, a diagnostic benchmark focused on superstitions and customs, arbitrary, context-dependent rules that resist simple logical deduction. Through 315 questions across three task types (factual retrieval, paired scenario verification, and situational reasoning), we evaluate seven Persian LLMs and reveal three critical failures: most models exhibit severe acquiescence bias, correctly identifying appropriate behaviors but failing to reject clear violations; continuous Persian pretraining amplifies this bias rather than improving reasoning, often degrading the model's ability to discern contradictions; and all models show a 21% performance gap between retrieving factual knowledge and applying it in scenarios. These findings demonstrate that cultural competence requires more than scaling monolingual data, as current models learn to mimic cultural patterns without internalizing the underlying schemas.[1]

1 Introduction

If you offer a Persian guest food, they will refuse. If you offer again, they will refuse. Only on the third offer, the "real" one, will they accept. This three-iteration ritual, called *taarof*, is immediately obvious to any Iranian child raised in the culture. But can a language model, trained on billions of Persian tokens, distinguish genuine *taarof* from a cultural violation? Our results suggest: not really.

The rapid advancement of large language models (LLMs) has sparked significant interest in their multilingual capabilities, including Persian language processing. Current Persian LLM benchmarks primarily focus on factual knowledge retrieval, translation quality, and linguistic tasks such as sentiment analysis and named entity recognition (Khashabi et al., 2020). However, these evaluations critically overlook a dimension central to Persian communication: the ability to reason about implicit cultural concepts, particularly superstitions and customs, that govern appropriate behavior in context-dependent social situations.

Figure 1: Sample Persian cultural concepts from the benchmark, spanning superstitions, traditions, and taboos.

Persian culture embeds meaning across multiple historical layers: ancient Zoroastrian beliefs (*Chaharshanbe Suri* fire-jumping, *esfand* burning for evil eye protection), Islamic traditions (*nazri* vow offerings, *ghorbani* sacrifice), complex social etiquette systems (*taarof*, *jang-e hesab*), rich folk cosmology featuring supernatural beings (*jinn*, *div*, *pari*, *bakhtak*), widespread superstitious practices (whistling at night attracts evil), and elaborate life-cycle ceremonies with highly specific ritualistic requirements. These concepts are not facts to be retrieved from a knowledge base; they are cultural

*Correspondence to: alireza.sakhaeirad@epfl.ch

[1]Dataset publicly available huggingface.co/datasets/divanbench/divanbench

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 1–12
March 29, 2026

Category	Example Concepts
Social Etiquette	*Taarof*, *Jang-e Hesab*, Doorway Deference, Three-Times Rule, *Pishkesh*, *Shirini*
Nowruz Traditions	*Haft Sin*, *Chaharshanbe Suri*, *Haji Firuz*, *Samanu Pazan*, *Khaneh Tekani*, *Eidi*
Supernatural Beings	*Jinn*, *Div*, *Pari*, *Bakhtak*, *Hamzad*, *Al*
Apotropaic Rituals	*Nazar* amulet, *Esfand* burning, Salt circle, *Bismillah* invocations
Wedding Ceremonies	*Sofreh Aghd*, Knife Dance, *Kuzeh Shekani*, Honey ritual
Taboos	Whistling at night, Sweeping at night, Stepping on bread, Shoe taboos, Nail cutting restrictions
Divination/Omens	*Fal-e Hafez*, *Fal-Gush*, Dream interpretation, Itchy palms, Ear ringing

Table 1: Distribution of cultural concepts across categories. Concepts span ancient Zoroastrian beliefs, Islamic traditions, social etiquette systems, folk cosmology, life-cycle ceremonies, taboos, and divination practices.

schemas (Shore, 1996; Strauss and Quinn, 1997) requiring understanding of context-dependent social dynamics, implicit power relations, and culturally-specific rules of appropriateness that are obvious to cultural insiders but opaque to pattern-matchers.

1.1 Motivation: Why Superstitions and Customs?

Superstitions and customs provide particularly demanding testbeds for cultural reasoning because they often lack logical justification-, unlike social etiquette, which has pragmatic benefits, superstitions require acceptance of culturally-transmitted beliefs without empirical grounding. Furthermore, they exhibit context-dependence: the same action (such as sweeping) is neutral by day but taboo at night. These practices involve implicit rules that are typically corrected through social feedback rather than explicit instruction, and Persian superstitions notably reflect historical syncretism, layering Zoroastrian, Islamic, and folk beliefs in complex ways. Recent work on Korean superstitions (Kim and Lee, 2025) demonstrated that culture-bound phenomena reveal limitations invisible in standard benchmarks. We extend this insight to Persian, where superstitions permeate daily life and provide rich signal for distinguishing genuine cultural understanding from keyword-matching.

1.2 Research Questions, Contributions, and Key Findings

This work (illustrated in Fig. 1) investigates whether Persian-capable Large Language Models (LLMs) internalize *cultural schemas*, the implicit social logic mapping context to action, or merely reproduce surface-level *cultural facts* and stereotypes. The investigation focuses on three core dimensions: the propensity for models to exhibit acquiescence bias through pattern-matching, the impact of continuous Persian pretraining on reasoning versus fluency, and the capacity for factual knowledge to be operationalized into scenario-based application. By distinguishing between fluent cultural expression and robust cultural reasoning, this work provides a framework for evaluating cultural competence in low-resource linguistic environments.

Central to this evaluation is the introduction of DIVANBENCH, a diagnostic framework covering 81 concepts across superstitions, customs, and social etiquette, instantiated through 315 questions. The benchmark adopts a three-level architecture designed to isolate distinct layers of cultural competence: Factual MCQ for baseline knowledge retrieval, Binary Belief Verification for testing discernment through paired positive/negative scenarios, and Scenario-Based MCQ for evaluating multi-step inference in complex social contexts. By utilizing a paired design, contrasting behavior that aligns with Persian customs against plausible but culturally inappropriate alternatives, this methodology effectively quantifies acquiescence bias. Furthermore, the study isolates the effects of monolingual adaptation by comparing a base model, Llama 3.1-8B, its Persian-adapted variant Dorna2-8B, revealing how specialized pretraining affects the transition from surface fluency to logical consistency.

Our evaluation of 7 state-of-the-art models yields the following key findings:

- **The acquiescence Trap:** 6 of 7 models exhibit significant asymmetry in reasoning; while they identify appropriate behaviors with 84%–92% accuracy, they fail to reject cultural violations at rates between 19% and 48%, suggesting a reliance on pattern-matching over logic.
- **The Persian Pretraining Paradox:** Continuous pretraining can inadvertently degrade critical reasoning. Llama 3.1-8B's rejection accuracy of 73% dropped to 30% after Persian-specific adaptation (Dorna2), representing a 43-percentage-point decline.
- **The Factual–Conceptual Gap:** Performance

decreases by an average of 21% when transitioning from factual retrieval to scenario-based reasoning, demonstrating that memorized cultural facts do not reliably translate into functional cultural schemas.

- **Distinct Learning Targets:** The results provide empirical evidence that cultural facts and schemas behave as independent objectives, suggesting that cultural competence requires training mechanisms beyond standard data scaling or monolingual adaptation.

The remainder of this paper is structured as follows: Section 2 reviews related work on cultural evaluation and Persian NLP; Section 3 describes our benchmark design, cultural concept coverage, and data collection methodology. Section 4 details our experimental setup including model selection, evaluation protocol, and experimental design. Subsection 4.2 presents our four main findings on acquiescence bias, pretraining effects, factual-conceptual gaps, and model scaling. Subsection 4.3 analyzes the mechanisms behind these findings and their implications for multilingual NLP. Finally, Section 5 concludes and discusses future directions for cultural reasoning research. Appendix A gives more insight into the data and presents more examples, and appendix B adds some additional detail on system prompt variations, experiments, and results.

2 Related Work

2.1 Persian LLM Evaluation

Most Persian NLP benchmarks have traditionally emphasized linguistic competence and factual or domain knowledge. Representative resources include PARSINLU for broad NLU coverage (Khashabi et al., 2020), FARSTAIL for Persian natural language inference (Amirkhani et al., 2021), and PERSIANMEDQA for bilingual medical question answering (Ranjbar Kalahroodi et al., 2025). Recent efforts broaden evaluation toward general knowledge and multimodal educational assessment, including the Khayyam Challenge (PersianMMLU) (Ghahroodi et al., 2024) and MEENA (PersianMMMU) (Ghahroodi et al., 2025). In parallel, a newer line of work targets cultural and pragmatic competence more directly: PERCUL uses story-driven scenarios to probe cultural understanding in Persian (Moosavi Monazzah et al., 2025), TAAROFBENCH evaluates the Persian politeness system and its nuanced social dynamics (Gohari Sadr et al., 2025), and ELAB benchmarks Persian-relevant alignment and safety dimensions (Pourbahman et al., 2025). Together, these works show a shift from "language + facts" evaluation toward culturally grounded pragmatics, though the latter remains less standardized and comparatively underexplored.

2.2 Cultural Competence in LLMs

Beyond Persian, substantial evidence shows that LLMs often fail to capture culture-specific, context-dependent norms and may default to dominant viewpoints or produce plausible-but-shallow cultural explanations (Zhang et al., 2025; Dai et al., 2025; Durmus et al., 2024; Hossain and Afli, 2025). Cross-cultural evaluations similarly report performance disparities between high-resource and low-resource cultures and uneven representation of global perspectives (Cao et al., 2023; Durmus et al., 2023). Related research in social/pragmatic reasoning and perspective-taking further indicates weaknesses on tasks that require implicit modeling of beliefs and intentions (Sap et al., 2019, 2022). Particularly relevant are cultural testbeds based on superstitions (e.g., NUNCHI-BENCH), where models may handle surface facts but struggle with situational reasoning (Kim and Lee, 2025).

Methodologically, evaluation can be confounded by systematic response biases: LLMs exhibit position bias in multiple-choice settings and related tendencies such as acquiescence bias/"yes-saying" (Zheng et al., 2023; Wang et al., 2023; Wallace et al., 2019; Si et al., 2023). Prior work mitigates these effects via prompt design and calibration (Ko et al., 2023; Si et al., 2023), while alternative evaluation designs aim to measure bias explicitly. Finally, cognitive theories distinguish explicit cultural facts from implicit cultural schemas learned through repeated, socially situated experience (Shore, 1996; Strauss and Quinn, 1997). This distinction aligns with arguments that text-only training may be insufficient for grounded understanding (Bisk et al., 2020) and echoes the classic observations that robust common sense reasoning is difficult to recover from text statistics alone (Levesque et al., 2012).

2.3 Positioning Our Work

Our work uniquely combines: (1) comprehensive coverage of Persian superstitions and customs (more than 80 concepts across 7 domains); (2) explicit bias measurement through paired pos-

itive/negative design; (3) direct comparison isolating pretraining effects (Llama3.1 (Dubey et al., 2024) vs. Dorna2 (PartAI Team, 2024)); and (4) theoretical grounding distinguishing cultural facts from schemas. While prior work addresses individual aspects (Persian-related factual retrieval tests, cultural evaluation, low-resource pretraining), we provide integrated analysis revealing systematic limitations in current approaches to cultural competence in Persian LLMs.

3 Benchmark Design and Data Collection

We introduce DIVANBENCH, a benchmark for evaluating cultural reasoning in Persian LLMs. Our design philosophy centers on distinguishing genuine cultural understanding from pattern-matching to plausible-sounding Persian text.

3.1 Design Principles

We evaluate models across three complementary tasks isolating different aspects of cultural competence:

Factual Multiple-Choice Questions The first data type consists of 100 Factual MCQs designed to test whether models can accurately retrieve fundamental information regarding Persian culture, history, geography, and traditions. These questions serve as a critical baseline control, identifying whether a model possesses the raw cultural knowledge necessary for more complex tasks. Success in this category indicates a robust factual foundation, while failure suggests a lack of exposure to Persian-specific data during the model's pre-training phase.

Binary Belief Verification The second task type comprises 162 statements categorized into matched pairs for 81 distinct cultural concepts. Each pair includes a "positive" scenario indicating a person following the Persian customs and a "negative" scenario where the subject acts against those customs in a way that remains plausible in a non-Persian context. This paired design is specifically intended to measure acquiescence bias and discernment; models that rely on simple pattern matching often show an asymmetric performance, erroneously accepting culturally inappropriate behaviors while correctly identifying appropriate ones.

Scenario-Based Multiple-Choice Questions The final data type includes 53 Scenario-Based MCQs that challenge models to apply cultural reasoning to complex, novel social situations. These scenarios require intricate inferences about social hierarchies, contextual appropriateness, and interpersonal nuances that are not explicitly taught. By focusing on concepts typically acquired through lived experience and social interaction, this task isolates the model's ability to move beyond rote memorization and demonstrate authentic cultural competence in dynamic settings.

Example: Scenario-Based MCQ

Arman has a sore throat and a high fever. His mother brings him soup and tea but strictly warns him not to eat the fresh melon sitting on the counter. Arman argues that the fruit is full of vitamins and will help him recover. Why does the mother forbid Arman from eating the melon while he has a cold?

(A) Melon is *Sard* (Cold); adding cold fruit while sick will freeze the lungs and prolong recovery

(B) Melon is *Garm* (Hot); eating it during a fever will ignite the blood, spreading infection faster

(C) Folk belief holds melon aroma attracts night-illness spirits, preventing medicine from working

(D) Hygiene precaution: high sugar content feeds bacteria in the throat, causing permanent voice loss

Expected: (A) — requires understanding the classical Persian food classification system (*Sard-Garm*) and traditional Persian medicine, which lacks modern medical evidence.

Author-Generated Content and Quality Control All questions were drafted and refined through multiple rounds of author review based on cultural knowledge acquired via lived experience in Iranian society. This process ensures authenticity by reflecting real social practices rather than stereotypes, focusing on "insider knowledge" that is intuitive to cultural members but requires active reasoning from outsiders. We prioritized implicit concepts typically learned through social interaction, rather than explicit instruction. By ensuring unambiguity for insiders while maintaining plausible alternatives for outsiders, the dataset enables a diagnostic evaluation of cultural competence that distinguishes genuine understanding from simple pattern

matching.

3.2 Cultural Concept Coverage

Our benchmark spans seven major cultural domains, summarized in Table 1 and detailed in Appendix A.1. These domains encompass superstitions and omens such as divination practices (*Fal-e Hafez*, *Fal-Gush*), dream interpretation, and bodily omens like itchy palms or ear ringing, alongside apotropaic rituals like burning *esfand* to ward off the evil eye, wearing *nazar* amulets, and *Bismillah* invocations. Additionally, the dataset covers taboos governing daily life, including restrictions on whistling or sweeping at night, and prohibitions against stepping on bread.

4 Experiments

4.1 Setup

Model Selection We evaluate 7 models from the Open Persian LLM Leaderboard (OpenPersian Team, 2024) with similar parameter counts to ensure fair comparison. All models are in the 7–12B parameter range, representing state-of-the-art performance for Persian language tasks: Aya-8B (Cohere For AI, 2024), Dorna2-8B (PartAI Team, 2024), Gemma2-9B and Gemma3-12B (Gemini Team, 2023), Llama3.1-8B (Dubey et al., 2024), and Qwen2-7B and Qwen2.5-7B (Qwen Team, 2024). The critical comparison is Llama3.1-8B (base) versus Dorna2-8B (Persian-adapted through continuous pretraining), which isolates the effect of Persian-focused pretraining while controlling for all other modeling choices. (For detailed description about models see App. B.1)

Evaluation Protocol Following best practices for reproducible LLM evaluation (Chang et al., 2024), we use temperature of 0.1 and top-p sampling set to 0.9 with fixed random seeds. For answer extraction, we employ GPT-4o-mini as a systematic extraction agent (Zheng et al., 2023), which parses model outputs to identify selected options (A/B/C/D) for multiple-choice questions and yes/no responses for binary questions. To account for prompt sensitivity (Si et al., 2023), each question is evaluated with 5 diverse system prompts that vary in phrasing while maintaining semantic equivalence. All prompts instruct models to respond as Iranian cultural insiders to ensure fair evaluation conditions. We report mean accuracy and standard deviation across prompt variations. To see exact system prompts, see B.2.

We compute three complementary metrics: (1) **Accuracy** measuring percentage of correct responses per task type; (2) **Acquiescence bias** quantifying the difference between acceptance rates for positive scenarios and rejection rates for negative scenarios in binary tasks, where high positive bias indicates pattern-matching rather than reasoning; and (3) **Factual-Conceptual Gap** measuring performance difference between factual retrieval and scenario-based reasoning. These metrics provide multi-dimensional assessment of cultural competence beyond standard accuracy reporting.

Experimental Design We conduct three complementary experiments to systematically evaluate cultural reasoning in Persian LLMs:

1. **Acquiescence bias Measurement (Experiment 1).** We compare model performance on paired positive and negative binary scenarios testing identical cultural concepts. This paired design directly measures acquiescence bias as $B = \text{Acc}_{\text{pos}} - \text{Acc}_{\text{neg}}$, where Acc_{pos} is accuracy on appropriate behavior and Acc_{neg} is accuracy on violations. Models with high positive bias demonstrate pattern-matching to cultural keywords rather than genuine reasoning.

2. **Pretraining Effects (Experiment 2).** We conduct a controlled comparison between Llama3.1-8B (base model M_{base}) and Dorna2-8B (Persian-adapted model M_{adapted}). This matched-pair design isolates the effect of Persian pretraining by comparing performance changes $\Delta = \text{Acc}(M_{\text{adapted}}) - \text{Acc}(M_{\text{base}})$ across all task types, revealing whether continuous pretraining improves cultural reasoning or reinforces surface-level pattern-matching.

3. **Knowledge Transfer (Experiment 3).** We analyze the factual-conceptual gap as $G = \text{Acc}_{\text{factual}} - \text{Acc}_{\text{scenario}}$, where $\text{Acc}_{\text{factual}}$ measures retrieval and $\text{Acc}_{\text{scenario}}$ measures cultarul reasoning. Large consistent gaps suggest cultural facts and schemas engage different cognitive mechanisms.

4.2 Results

We present results across all evaluation dimensions, revealing systematic patterns in how Persian language models handle cultural reasoning. Table 2

Model	Factual MCQ (Knowledge)	Scenario MCQ (Reasoning)	Gap (Fact - Scenario)	Acquiescence bias (Accept - Reject)
Gemma3-12B	**87.0** ± 1.0	**66.0** ± 1.0	21.0	+61.0
Gemma2-9B	83.0 ± 0.0	65.0 ± 3.0	18.0	+41.0
Llama3.1-8B	80.0 ± 2.0	53.0 ± 5.0	27.0	**-10.0**
Qwen2.5-7B	75.0 ± 1.0	60.0 ± 2.0	**15.0**	+26.0
Qwen2-7B	73.0 ± 3.0	48.0 ± 5.0	25.0	+65.0
Dorna2-8B	73.0 ± 2.0	54.0 ± 3.0	19.0	+61.0
Aya-8B	66.0 ± 2.0	42.0 ± 2.0	24.0	+19.0
Average	**76.7**	**55.4**	**21.3**	**+43.3**

Table 2: Core performance metrics across all models. **Gap** measures difficulty transferring factual knowledge to scenario reasoning. **Acquiescence bias** (Accept - Reject) reveals pattern-matching: positive values (red) indicate models accept appropriate behavior readily but fail to reject violations. Only Llama3.1-8B shows skeptical bias (blue). Percentages shown; standard deviations computed across 5 prompt variations. Models sorted by Factual MCQ performance.

Task	Llama3.1-8B	Dorna2-8B	Change
Reject False	0.73	0.30	-43%
Accept True	0.63	0.91	+28%
Factual MCQ	0.80	0.73	-7%
Scenario MCQ	0.53	0.54	+1%

Table 3: Direct comparison of Llama3.1-8B (base) and Dorna2-8B (Persian-tuned). Persian pretraining dramatically increased acquiescence while destroying critical reasoning.

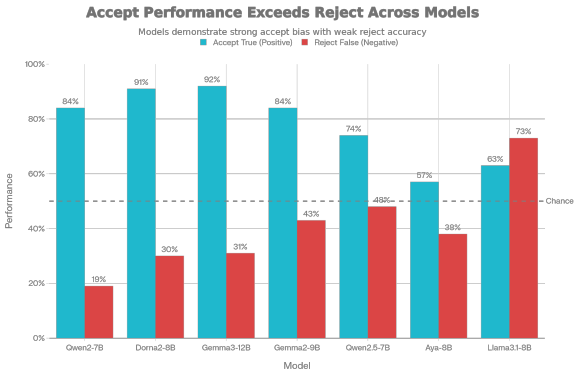


Figure 2: acquiescence bias across models. Most models accept true cultural statements readily (Blue) but fail to reject false ones (Red), indicating pattern-matching rather than reasoning.

shows complete performance across all tasks (See table 5 for detailed results of binary tests).

Finding 1: The acquiescence Trap Most models exhibit severe acquiescence bias on cultural questions, accepting plausible statements far more readily than rejecting violations. This asymmetry reveals that models are pattern-matching to culturally-themed Persian text rather than reasoning about appropriateness. Extreme cases include Qwen2-7B (84% acceptance vs. 19% rejection, bias = +65%), Gemma3-12B (92% acceptance vs. 31% rejection, bias = +61%), and Dorna2-8B (91% acceptance vs. 30% rejection, bias = +61%). These models correctly identify culturally appropriate behavior when presented positively but fail to recognize violations of the same concepts when presented negatively. for instance, accepting both "Standing up when elders arrive" (correct code of respect) and "Remaining seated for comfort when elders arrive" (violates code of respect). The only exception is Llama3.1-8B, showing opposite bias (73% rejection vs. 63% acceptance, bias = -10%), suggesting an skeptical tone in this model.

Finding 2: The Persian Pretraining Paradox Comparing Llama3.1-8B (base) with Dorna2-8B (Persian-tuned) in table 3 reveals counterintuitive effects of continuous pretraining on cultural reasoning. Rejection accuracy collapsed dramatically from 73% to 30% (-43 percentage points), while acceptance accuracy soared from 63% to 91% (+28pp). Notably, factual knowledge slightly declined (80% to 73%) and scenario reasoning remained essentially flat (53% to 54%), indicating that continuous Persian pretraining taught the model to recognize cultural patterns without improving reasoning about cultural logic. This suggests that more Persian data reinforced surface-level pattern-matching over deep cultural understanding. Dorna2 became more “culturally compliant” but less discerning, accepting any plausible-sounding cultural scenario regardless of correctness. This finding challenges the common assumption in low-resource NLP that monolingual pretraining universally improves understanding.

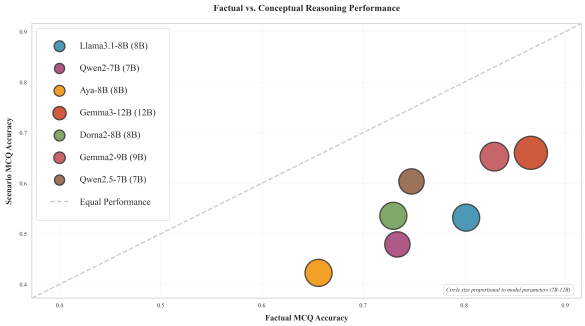


Figure 3: Factual Knowledge vs. Cultural Reasoning Gap. Gemma3-12B (largest bubble, top-right) achieves highest factual accuracy but shows inconsistent scenario reasoning. All models fall below the diagonal, indicating systematic difficulty in transferring factual knowledge to cultural schema application. Bubble size represents model parameters.

Finding 3: The Factual-Conceptual Gap All models show consistent performance degradation from factual to scenario-based tasks, with an average gap of 21 percentage points. Models can retrieve facts about Persian culture ("What is *taarof*?") but struggle to apply cultural logic to novel scenarios ("In this situation, would *taarof* be appropriate?"). This gap demonstrates that knowing that a concept exists differs fundamentally from knowing when it appropriately applies. The gap persists even for the best-performing model (Gemma3-12B shows -21%), suggesting this is a fundamental limitation rather than a simple capacity issue. This evidence supports the cognitive anthropology distinction between cultural facts (retrievable from text) and cultural schemas (requiring embodied social learning).

Finding 4: Size Does Not Guarantee Cultural Intelligence Gemma3-12B, the only model exceeding 10B parameters, shows contradictory performance patterns: it achieves best factual retrieval (87%, 21% above average) yet worst acquiescence bias (+61%, tied with Dorna2). The larger model excels at memorizing cultural facts while showing extreme acquiescence bias, accepting 92% of positive statements but rejecting only 31% of negative ones. This suggests that model size improves fact memorization without improving cultural discernment, and larger models may amplify pattern-matching behaviors learned from training data.

4.3 Analysis

Why Persian Cultural Concepts Are Hard Persian cultural concepts exhibit three critical characteristics that resist pattern-matching. First, they encode implicit social contracts never stated in text: *taarof* requires tracking iteration counts, modeling power asymmetries, and distinguishing ritual from sincerity through multi-turn interaction. Second, they involve context-dependent inversion where identical actions have opposite meanings: paying for a meal is appropriate among peers after negotiation but offensive when done secretly with elders. Third, they conflate factual and normative knowledge: "Is burning *esfand* effective?" (factual, no) versus "Is burning *esfand* appropriate when moving?" (normative, yes). Models trained on factual text systematically conflate these categories.

The Acquiescence Trap Mechanism The severe acquiescence bias across six models stems from distributional patterns in Persian training data. Models learn that cultural keywords like *taarof*, *nazri*, and *esfand* co-occur with positive contexts: respectful news coverage, affirmative social media posts, and celebratory descriptions. When encountering these keywords, models trigger acceptance regardless of scenario details. This explains why they accept both correct *taarof* (refuse twice, accept third time) and violations (accept immediately). Without reasoning about iteration logic, they pattern-match to cultural keywords instead.

Why Persian Pretraining Degraded Reasoning Continuous Persian pretraining on Dorna2 collapsed rejection accuracy by 43 percentage points while boosting acceptance by 28 points. This asymmetry reveals surface pattern reinforcement: Persian corpora contain overwhelmingly affirmative mentions of traditions, creating distributional bias toward accepting culturally-themed statements. Meanwhile, Llama3.1's skeptical stance (73 percent rejection) likely reflects instruction tuning for critical evaluation. Persian pretraining overrode this skepticism, producing cultural fluency without reasoning. This challenges the assumption that monolingual pretraining improves understanding, it may amplify fluency while degrading discernment.

Factual-Conceptual Gap as Schema Evidence The consistent 21 percent performance drop from factual to scenario tasks operationalizes the distinction between cultural facts (retrievable knowledge like "Nowruz marks New Year") and cultural schemas (situational rules like "Would serving *halva* be appropriate here?"). Facts require memorization; schemas require conditional reasoning

about context, relationships, and implicit norms. Current training successfully instills facts but fails at schemas, suggesting that adding more factual data cannot close this gap, different learning mechanisms are needed.

5 Conclusion

We introduced DIVANBENCH, a benchmark evaluating cultural reasoning in Persian LLMs through 81 concepts and 315 questions across three task types. Evaluation of seven models reveals four key findings: most models exhibit severe acquiescence bias, accepting appropriate behavior while failing to reject violations; continuous Persian pretraining amplified this bias rather than improving reasoning; a consistent 21 percent gap between factual and scenario performance demonstrates that cultural facts and schemas engage different mechanisms; and model size improves memorization but not reasoning. These findings challenge assumptions that scaling monolingual data improves cultural competence and suggest that low-resource languages require more advanced methods, rather than simply training the model on more tokens of the target language. We release our benchmark publicly to enable systematic measurement of cultural understanding, providing a template for evaluation in other languages. Persian culture's synthesis of Zoroastrian, Islamic, and folk traditions makes it a demanding testbed for genuine schema reasoning versus surface pattern-matching. Our work establishes a foundation for building language models that truly comprehend the cultures they serve.

Limitations

Model size and scaling regime. Our study evaluates a narrow band of *small-to-mid sized* open models (7–12B parameters), selected to enable controlled comparisons across systems with broadly similar capacity and inference cost. This design improves comparability, but it limits what we can conclude about how cultural conceptual reasoning behaves at larger scales (e.g., 30B–70B+), where instruction tuning, longer context windows, or different training mixtures may change both *bias profiles* and *reasoning strategies*. In particular, our observation that models can improve factual recall while still exhibiting accept-over-reject asymmetry may not extrapolate monotonically to substantially larger model families; scaling could either attenuate or amplify the "acquiescence trap," depending on how the model is trained and aligned. Additionally, by keeping model size relatively constant, we cannot disentangle whether certain failures are fundamentally capacity-limited versus primarily data- and objective-driven.

Hand-curated dataset and author priors. DivanBench is authored manually based on lived cultural knowledge, which increases scenario realism and insider validity but introduces several curation constraints. First, manual writing inevitably reflects the authors' judgments about what counts as the "canonical" interpretation of a concept, which may underrepresent regional, socioeconomic, or generational variation. Second, even with iterative review, scenario phrasing can unintentionally leak cues (e.g., overly salient keywords, unnatural dialogue) that models may exploit via superficial heuristics; especially in paired positive/negative formats. Third, manual coverage is bounded: although we span many (more than 80) concepts, the space of culturally meaningful edge cases is far larger, and some concepts may be overrepresented by the scenarios that are easiest to write unambiguously. Finally, because expert annotation is expensive and time-consuming, we do not incorporate large-scale crowdsourced validation or inter-annotator agreement; thus, while items are intended to be unambiguous to cultural insiders, residual ambiguity cannot be ruled out.

References

Hossein Amirkhani, Mohammad AzariJafari, Zohreh Pourjafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, and Azadeh Amirak. 2021. Farstail: A persian natural language inference dataset. *Preprint*, arXiv:2009.08820.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, and 1 others. 2020. Experience grounds language. *Proceedings of EMNLP*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, and 1 others. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Cohere For AI. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias in llms via word association test. *Preprint*, arXiv:2505.18562.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, and 1 others. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, and 1 others. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Omid Ghahroodi, Arshia Hemmat, Marzia Nouri, Seyed Mohammad Hadi Hosseini, Doratossadat Dastgheib, Mohammad Vali Sanian, Alireza Sahebi, Reihaneh Zohrabi, Mohammad Hossein Rohban, Ehsaneddin Asgari, and Mahdieh Soleymani Baghshah. 2025. Meena (persianmmmu): Multimodal-multilingual educational exams for n-level assessment. *Preprint*, arXiv:2508.17290.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *Preprint*, arXiv:2404.06644.

Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoomian, Laleh Seyyed-Kalantari, and Ali Emami. 2025. We politely insist: Your llm must learn the persian art of taarof. *Preprint*, arXiv:2509.01035.

Shehenaz Hossain and Haithem Afli. 2025. Craft: An explanation-based framework for evaluating cultural reasoning in multilingual language models. *Preprint*, arXiv:2510.14014.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Graham Neubig, and 1 others. 2020. Parsinlu: A suite of language understanding challenges for persian. *arXiv preprint arXiv:2012.06154*.

Kyuhee Kim and Sangah Lee. 2025. Nunchi-bench: Benchmarking language models on cultural reasoning with a focus on Korean superstition. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15328–15342, Vienna, Austria. Association for Computational Linguistics.

Yoonjoo Ko, Hyungjoo Lee, Taehwan Lee, Jaewook Jung, Jinhyeok Park, and Jaihoon Choi. 2023. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. *arXiv preprint arXiv:2310.00492*.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of KR*.

Erfan Moosavi Monazzah, Vahid Rahimzadeh, Yadollah Yaghoobzadeh, Azadeh Shakery, and Mohammad Taher Pilehvar. 2025. Percul: A story-driven cultural evaluation of llms in persian. *Preprint*, arXiv:2502.07459.

OpenPersian Team. 2024. Openpersian leaderboard: Benchmarking persian language models. `https://huggingface.co/spaces/OpenPersian/leaderboard`.

PartAI Team. 2024. Dorna: Persian-optimized language models. `https://huggingface.co/PartAI/Dorna-Llama3-8B-Instruct`.

Zahra Pourbahman, Fatemeh Rajabi, Mohammadhossein Sadeghi, Omid Ghahroodi, Somaye Bakhshaei, Arash Amini, Reza Kazemi, and Mahdieh Soleymani Baghshah. 2025. Elab: Extensive llm alignment benchmark in persian language. *Preprint*, arXiv:2504.12553.

Qwen Team. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Mohammad Javad Ranjbar Kalahroodi, Amirhossein Sheikholselami, Sepehr Karimi, Sepideh Ranjbar Kalahroodi, Heshaam Faili, and Azadeh Shakery. 2025. Persianmedqa: Evaluating large language models on a persian-english bilingual medical question answering benchmark. *Preprint*, arXiv:2506.00250.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 4463–4473.

Bradd Shore. 1996. *Culture in Mind: Cognition, Culture, and the Problem of Meaning*. Oxford University Press.

Chenglei Si, Zihan Gan, Zhengyuan Yang, Shuohang Wang, Jingfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. In *Proceedings of ICLR*.

Claudia Strauss and Naomi Quinn. 1997. *A Cognitive Theory of Cultural Meaning*. Cambridge University Press.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. In *Transactions of the Association for Computational Linguistics*, volume 7, pages 387–401.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Zhifang Sui, and 1 others. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of EMNLP*, pages 9840–9855.

Jinghao Zhang, Sihang Jiang, Shiwei Guo, Shisong Chen, Yanghua Xiao, Hongwei Feng, Jiaqing Liang, Minggui HE, Shimin Tao, and Hongxia Ma. 2025. Culturescope: A dimensional lens for probing cultural understanding in llms. *Preprint*, arXiv:2509.16188.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Dataset Details

A.1 Detailed Cultural Concept Lists

Table 4 provides comprehensive concept coverage across all seven cultural domains.

A.2 Additional Examples

Example: Factual MCQ

What fruit is used as a symbol on the Yalda night table?
(A) Orange (B) Apple (C) Pomegranate (D) Pear
Expected: (C) Pomegranate

Example: Binary Belief Verification

Positive: "During a casual family gathering, an elderly aunt walks into the room. Everyone younger immediately stands up as a respectful acknowledgment of her presence. Did they act according to Persian tradition?"
Expected: Yes (This follows the rule to stand up when elders enter)

Negative: "During a family gathering, an elderly aunt enters the room. Since everyone is already comfortable, no one stands up and they simply greet her warmly from their seats. Did they act according to Persian tradition?"
Expected: No (This violates the rule to stand up when elders enter)

B Appendix

B.1 Model Descriptions

Aya-8B Multilingual model from CohereForAI with Persian support. Trained on diverse multilingual data including Persian web text and translated datasets. Instruction-tuned for helpfulness.

Dorna2-8B Persian-optimized variant of Llama3.1-8B created through continuous pretraining on Persian corpora including news, social media, literature, and web text. Maintained Llama3.1's instruction-tuning while adapting vocabulary and linguistic patterns to Persian.

Gemma2-9B and Gemma3-12B Google's open-source multilingual foundation models (second and third generations). Trained on diverse web data including Persian content. Gemma3 represents scaled-up version with 12B parameters.

Category	Example Concepts
Social Etiquette	*Taarof* (ritual politeness), *Jang-e Hesab* (payment battles), Doorway Deference (older enters first), Three-Times Rule (refuse twice), *Pishkesh* (gift-giving protocol), *Shirini* (sweet bringing obligation), Guest-host dynamics, Shoe removal customs, Greeting hierarchies
Superstitions & Omens	*Fal-e Hafez* (Hafez divination), *Fal-Gush* (eavesdropping for omens), Dream interpretation, Itchy palms (money coming), Ear ringing (someone talking about you), Twitching eye (bad omen), Sneezing omens, Bird flight patterns, Broken mirrors, Spilled salt
Apotropaic Rituals	*Esfand* burning (wild rue fumigation), *Nazar* amulet (evil eye protection), Salt circle protection, *Bismillah* invocations, Garlic hanging, Iron deterrence, Seven knots ritual, Mirror placement, Knife under pillow, Holy verses
Supernatural Beings	*Jinn* (fire spirits), *Div* (demons), *Pari* (fairies), *Bakhtak* (sleep paralysis demon), *Hamzad* (personal spirit double), *Al* (child-stealing demon), *Ghoul*, Evil eye personification, Ancestor spirits
Nowruz Traditions	*Haft Sin* (seven S's table), *Chaharshanbe Suri* (fire-jumping), *Haji Firuz* (blackface herald), *Samanu Pazan* (wheat pudding stirring), *Khaneh Tekani* (spring cleaning), *Eidi* (new year gift), *Sizdah Bedar* (13th day picnic), *Sabzeh* growing, *Goldfish* symbolism, New clothes tradition, Elder visitation order, Fire-jumping prayers, *Senjed* (jujube significance), Mirror watching, Egg decoration
Wedding Ceremonies	*Sofreh Aghd* (marriage spread), Honey ritual, Knife dance, *Kuzeh Shekani* (pot breaking), Sugar cone rubbing, Mirror & candelabra, Needle-sewing ritual, *Aghd* contract, *Shirini Khoran* (dessert ceremony), Witness requirements, Gold coin showering, Henna night
Taboos	Whistling at night (attracts *jinn*), Sweeping at night (sweeps away prosperity), Stepping on bread (brings poverty), Cutting nails at night (invites demons), Giving knives as gifts (cuts relationships), Opening umbrellas indoors, Haircut on Wednesdays, Shoe upside-down, Passing over children (stunts growth), Trimming nails over water, Direct compliments (causes evil eye), Singing while eating

Table 4: Comprehensive cultural concept coverage. Each concept has paired positive/negative binary scenarios plus representation in factual or scenario MCQs.

Model	Accept True (Positive)	Reject False (Negative)	Bias
Qwen2-7B	84.0 ± 4.0	19.0 ± 1.0	+65.0
Gemma3-12B	92.0 ± 3.0	31.0 ± 3.0	+61.0
Dorna2-8B	91.0 ± 5.0	30.0 ± 5.0	+61.0
Gemma2-9B	84.0 ± 2.0	43.0 ± 7.0	+41.0
Qwen2.5-7B	74.0 ± 7.0	48.0 ± 6.0	+26.0
Aya-8B	57.0 ± 4.0	38.0 ± 5.0	+19.0
Llama3.1-8B	63.0 ± 7.0	73.0 ± 5.0	-10.0

Table 5: Binary verification task performance sorted by bias magnitude. Models show severe asymmetry: high acceptance of appropriate behavior but low rejection of violations, indicating surface-level pattern-matching.

Llama3.1-8B Meta's base model serving as foundation for Dorna2. Multilingual capabilities including Persian from pretraining on web-scale data. Instruction-tuned for general helpfulness and reasoning.

Qwen2-7B and Qwen2.5-7B Alibaba's multilingual models (second generation and improved variant). Strong performance on Asian languages including Persian. Qwen2.5 incorporates improved training procedures and data quality.

B.2 System Prompt Variations

We use 5 diverse system prompts to ensure robust evaluation and they are shared among all tasks. Examples for binary verification tasks:

System Prompt 1

"You are a helpful assistant who is an expert in Iranian culture, folklore, traditions, and customs."

System Prompt 2

"You are a knowledgeable cultural expert specializing in Persian traditions, superstitions, and folk beliefs. Answer questions based on traditional Iranian cultural knowledge."

System Prompt 3

"You are an assistant with deep knowledge of Iranian heritage, including traditional customs, superstitions, proverbs, and cultural practices passed down through generations."

System Prompt 4

"As an expert in Iranian folklore and cultural traditions, provide accurate answers about Persian customs, beliefs, superstitions, and traditional practices."

System Prompt 5

"You are a cultural historian specializing in Iranian and Persian traditions. Draw upon your extensive knowledge of Iranian superstitions, customs, ceremonies, and folk beliefs to answer questions."

B.3 Binary Classification Task complementary Results

Exact numbers for figure 2 are given in table 5.

Benchmarking Offensive Language Detection in Persian and Pashto

Zahra Bokaei
School of Informatics
University of Edinburgh
zahra.bokaei@ed.ac.uk

Bonnie Webber
School of Informatics
University of Edinburgh
bonnie.webber@ed.ac.uk

Walid Magdy
School of Informatics
University of Edinburgh
wmagdy@ed.ac.uk

Abstract

Offensive language detection and target identification are essential for maintaining respectful online environments. While these tasks have been widely studied for English, comparatively less attention has been given to other language, including Persian and Pashto, and the effectiveness of recent large language models for these languages remains underexplored. To address this gap, we created a comprehensive benchmark of diverse modeling approaches in Persian and Pashto. Our evaluation covers zero-shot, fine-tuned, and cross-lingual transfer settings, analyzing when detection succeeds or fails across different model approaches. This study provides one of the first systematic analyses of offensive language detection and cross-lingual transfer between these languages.

1 Introduction

With the widespread use of online platforms, offensive language has become a persistent challenge in digital communication, with harmful consequences for individuals and communities (Singh and Li, 2021). Effectively moderating such platforms therefore requires reliable detection of offensive content, as well as identifying its type and intended target to support appropriate responses. Recently, research has increasingly explored offensive language detection beyond English, including work on Iranian languages such as Persian (Ataei et al., 2022; Safayani et al., 2024; Kebriaei et al., 2024) and, to a lesser extent, Pashto (Haq et al., 2023). However, compared to the extensive literature on English, systematic evaluation for these languages remain limited (Ataei et al., 2022), with Pashto in particular receiving substantially less attention.

Persian and Pashto are closely related languages. Persian is spoken primarily in Iran, Afghanistan, and Tajikistan, with approximately 110 million speakers worldwide, while Pashto is spoken mainly in Afghanistan and Pakistan, with an estimated 45–55 million speakers (UNESCO Silk Roads Programme, b,a). Despite their linguistic and cultural proximity, cross-lingual studies of offensive language detection for these languages remain scarce.

Recent advances in large language models (LLMs) have led to substantial gains across many NLP tasks, including harmful content detection. While such models have been widely evaluated for high-resource languages, their effectiveness for Persian and Pashto—particularly for offensive language detection and target identification—remains underexplored (Bokaei et al., 2025). In this paper, we benchmark recent instruction-tuned and multilingual models for offensive language detection in Persian and Pashto across zero-shot, fine-tuned, and cross-lingual transfer settings. We additionally evaluate target identification for individual- and group-directed offenses and analyze cross-lingual transfer between the two languages. Finally, we conduct error analysis to identify systematic model strengths and weaknesses. To the best of our knowledge, this work is the first to jointly study offensive language detection, target identification, and cross-lingual transfer between Persian and Pashto using recent LLMs, with a comprehensive error analysis.

In this study, we address the following research questions (RQs):

RQ1. How do different models perform on offensive language detection in Persian and Pashto?

RQ2. When does Transfer Learning help detect offensive language between Persian and Pashto?

RQ3. How does target type affect offensive language detection in Persian?

To address these research questions, we use two publicly available datasets: Pars-OFF for Persian (Ataei et al., 2022), which contains over 10,000 instances annotated for offensiveness and target type, and POLD for Pashto (Haq et al., 2023), comprising 34,400 instances labeled as offensive or non-offensive. Importantly, target annotations are available only in Pars-OFF, and not in POLD. Our find-

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 13–23

ings reveal consistent patterns across models and languages. *We show that all LLMs reliably detect highly explicit, profanity-based abuse, but systematically struggle with implicit and context-dependent offense.* Fine-tuning substantially improves recall by enabling models to capture language-specific realizations of offensiveness missed in zero-shot settings. Target identification reveals structural limitations, particularly for proxy and mixed targeting, where grammatical form diverges from semantic target. Transfer experiments show a clear asymmetry: Persian-to-Pashto transfer is stronger than the reverse, indicating that cross-lingual effectiveness depends on different factors, such as how offense is encoded in each language.

2 Related Work

This section briefly reviews prior work on offensive language detection and cross-lingual transfer, with a focus on Persian and Pashto. For Persian, multiple datasets benchmark traditional and transformer-based models on social media text (Kebriaei et al., 2024; Safayani et al., 2024; Mozafari et al., 2024), while resources such as Pars-HaO and Pars-OFF provide multi-level and target-aware annotations (Sheykhlan et al., 2023; Ataei et al., 2022). In contrast, Pashto remains severely under-resourced: POLD is currently the only publicly available dataset, and Pashto-specific BERT models outperform multilingual alternatives such as XLM-R on this dataset (Haq et al., 2023). Some prior work explores TL for offensive language detection. Pamungkas and Patti (2019) demonstrates improved robustness through multilingual and cross-domain transfer, El-Alami et al. (2022) shows effective English-to-Arabic transfer with BERT-based models, and Zhou et al. (2023) highlights the limitations of zero-shot transfer due to cultural and contextual mismatch, motivating few-shot and multilingual approaches. Despite this progress, existing work largely focuses on earlier models and does not provide a systematic benchmark of recent LLMs for Persian and Pashto. Moreover, cross-lingual TL between these two closely related Iranian languages with detailed error analysis remain under-explored. Our work addresses these gaps.

3 Dataset

We conduct our experiments on two publicly available datasets for offensive language detection in Persian and Pashto: Pars-OFF (Ataei et al., 2022) for Persian and POLD (Haq et al., 2023) for Pashto. Pars-OFF, a Persian dataset from Twitter data, consists of 10,563 instances, of which 7,381 are labeled as non-offensive and 3,182 as offensive. For the offensive instances, it also includes 2,612 cases targeting individuals, 1,280 targeting groups, and 553 labeled as other targets. For Pashto, POLD Twitter dataset contains 34,400 Twitter instances. Among these, 12,400 instances are labeled as offensive and 22,000 as non-offensive. POLD does not include target annotations.

	Pars-OFF			POLD		
Model	P	R	F1	P	R	F1
GPT-4o	**88**	77	82	**87**	77	82
LLaMA-3-Instruct	83	72	77	78	67	72
Gemma-3-Instruct	74	76	75	75	76	75
Dorna2-Instruct	80	78	79	68	59	60
Mistral-Instruct	70	66	68	71	57	63
Aya-Expanse	69	75	72	59	57	58
LLaMA-3 (FT)	85	**87**	**86**	79	79	79
Gemma-3 (FT)	**88**	82	85	82	**80**	81
ParsBERT	84	82	83	83	72	77
XLM-R	86	84	85	84	82	**83**

Table 1: Offensive language detection performance of LLMs on Pars-OFF and Pashto. FT = fine-tuned

4 Experimental Setup

Motivated by recent benchmark studies demonstrating strong performance of contemporary LLMs on safety- and norm-related tasks (Pourbahman et al., 2025), as well as rapid advances represented by models such as LLaMA 3 and Gemma 3 (Guo and Sarker, 2025), we select a diverse set of recent instruction-tuned and multilingual models for our experiments. We evaluate instruction-tuned LLMs in zero-shot settings: LLaMA-3-Instruct, Gemma-3-Instruct, Mistral-8B-Instruct, Aya-Expanse-8B, GPT-4o, and Dorna2-LLaMA-3-8B. In addition we deployed LLaMA-3-Instruct, Gemma-3-Instruct, ParsBERT (monolingual Persian Model), and XLM-R in fine-tuning and cross-lingual TL settings. Table 5 in the Appendix lists all models used in our benchmarking. For offensive detection on both dataset and target identification on Pars-OFF, we adopt a fixed 80%/10%/10% train/validation/test split across all experiments.

Zero-shot Experiments: We iteratively tested multiple prompts and selected the one with the highest validation performance. The final zero-shot prompt is shown in Figure 1 and 2 in the Appendix.

Fine-tune Experiments: All models were fine-

Model	Group			Individua			Other		
	P	R	F_1	P	R	F_1	P	R	F_1
GPT-40	**83**	81	**82**	81	79	80	58	**75**	**65**
LLaMA 3-Instruct	75	61	67	75	58	65	53	51	52
Gemma 3- Instruct	80	62	70	71	71	71	62	48	54
Dorna 2- Instruct	78	64	70	75	75	75	65	48	55
Llama-3 (FT)	80	**82**	81	71	69	70	59	57	58
Gemma-3 (FT)	**91**	**87**	**89**	82	78	80	**67**	63	**65**
ParsBERT	68	79	73	69	65	67	45	70	55
XLM-R	75	79	77	80	76	78	58	62	60

Table 2: Target identification performance on Persian across LLMs

tuned for 10 epochs on each dataset's training set. Final test results are reported using the checkpoint that achieved the highest validation F1 score.

TL Experiment: LLaMA-3-Instruct, Gemma-3-Instruct, ParsBERT, and XLM-R were fine-tuned for 10 epochs on the source-language training split, selecting the checkpoint with the highest validation F1, and evaluated on the target-language test set.

5 Results

Offensive Detection: Table 1 presents offensive language detection performance across different LLMs. Across both Persian and Pashto, fine-tuned models consistently outperform zero-shot models in F1. On Persian, zero-shot instruction-tuned LLMs achieve F1 scores ranging from 68 to 82, while fine-tuning raises performance to 83–86 F1, with LLaMA-3-Instruct reaching the highest score (F1 = 86). On Pashto, zero-shot performance is notably lower (F1 = 57–72), whereas fine-tuning yields consistent improvements, increasing F1 to 77–83, with XLM-RoBERTa achieving the best result (F1 = 83). These gains are primarily driven by recall improvements: for example, LLaMA-3-Instruct recall increases from 72 to 87 on Persian and from 67 to 79 on Pashto after fine-tuning, while precision remains comparatively stable. In contrast, zero-shot models tend to exhibit higher precision than recall, particularly for Pashto (e.g., GPT-4o precision = 87 vs. recall = 77), indicating conservative predictions that miss offensive instances.

Target Identification: Table 2 reports performance for identifying offensive targets. Performance is highest for group targets across models (F1 = 67–89), followed by individual targets (F1 = 60–80), while other targets are consistently the most challenging (F1 = 52–65). Fine-tuning leads to clear improvements over zero-shot inference for all target types, particularly through recall gains. Gemma-3 fine-tuned achieves the strongest overall performance on group targets (P = 91, R = 87, F1 = 89), whereas GPT-4o obtains the highest F1 on individual targets (P = 75, R = 79, F1 = 80). In contrast, zero-shot models exhibit notably lower recall for individual and other targets indicating difficulty in detecting targets that are implicit or structurally ambiguous.

Model	TL from Pashto			TL from Persian		
	P	R	F_1	P	R	F_1
LLaMA 3	75	86	81	79	89	83
Gemma 3	84	86	85	86	88	87
ParsBERT	57	71	63	82	78	70
XLM-R	79	77	70	79	77	78

Table 3: TL results between Pashto and Persian.

TL Experiments: We evaluate TL between Pashto and Persian using LLaMA-3, Gemma-3, ParsBERT, and XLM-RoBERTa, and compare transfer performance against in-language fine-tuning. The results are shown in Table 3. Overall, TL is more effective from Persian to Pashto than the reverse. When transferring from Pashto to Persian, F1 scores range from 63 to 85: Gemma-3 achieves the best transfer result (F1 = 85), closely matching its in-language Persian performance (F1 = 85), while LLaMA-3 attains F1 = 81, remaining below its Persian in-language score (F1 = 86). In contrast, ParsBERT and XLM-R show larger drops when transferring from Pashto to Persian. Transfer from Persian to Pashto yields stronger gains, with F1 scores between 70 and 87. Gemma-3 again performs best (F1 = 87), exceeding its in-language Pashto result (F1 = 81), while LLaMA-3 also improves under transfer (F1 = 83 vs. 79 in-language). XLM-R maintains stable performance (F1 = 78 vs. 83 in-language), whereas ParsBERT shows limited transferability in both directions (F1 = 63–70). Across models, transfer from Persian to Pashto approaches or surpasses in-language baselines for LLaMA-3 and Gemma-3, while transfer from Pashto to Persian exhibits larger performance drops, particularly for ParsBERT and XLM-R.

6 Discussion

To address RQ1 and RQ2, we analyze offensive language detection from three perspectives: (i)instances detected by all models, (ii) missed by all models, (iii) and detected only after fine-tuning. This highlights which aspects of offensiveness are universally captured, consistently overlooked, or learned through supervision in Persian and Pashto.

6.1 RQ1: Model Performance

Cases detected by all models in both languages: represent a high-signal subset in which offensiveness is realized through explicit lexical cues and direct insult constructions. Across Persian and Pashto, these instances are consistently identified by both zero-shot and fine-tuned models, resulting in near-perfect agreement. This pattern indicates that when offensive intent is overt and targets are explicitly realized, detection is largely architecture- and training-independent. The instances in this category are characterized by extreme lexical explicitness, direct insult speech acts, and clear grammatical realization of targets. They contain dense concentrations of common profanities, sexual and kin-based insults, animalization, and dehumanizing expressions, often stacked within short spans or chained across clauses. Offensive intent is enacted directly rather than implied or reported, and emotional intensity is high. Although some examples reference culturally or politically specific entities, correct detection does not depend on such knowledge, as surface-level abusive markers dominate interpretation. Representative examples are provided in Table 6 in Appendix.

Cases missed by all models: In both languages, these instances lack explicit offensive markers and direct target attachment, relying instead on implicit or discursive forms of hostility. All models converge on non-offensive predictions in such cases, indicating a shared limitation in handling implicit, context-dependent offense. Offensiveness is conveyed implicitly through ideological positioning, moral judgment, evaluative political discourse, slogans, or reflective commentary. These texts frequently resemble legitimate argumentation or social critique, employ neutral or formal vocabulary, and exhibit low emotional arousal. In Pashto, indirect encoding through idioms, proverbs, symbolic group references, and religious or historical framing is particularly prominent. Without explicit surface cues, both zero-shot and fine-tuned models consistently predict non-offensive labels. Examples of these false negatives are reported in Table 7 in Appendix.

Cases detected only after fine-tuning illustrate the contribution of supervised learning. These instances are not reliably captured by zero-shot models but are correctly identified after exposure to in-domain training data. This subset demonstrates that fine-tuning improves sensitivity to non-explicit and discourse-driven expressions of offensiveness. This category includes instances in which offense is implicit, discourse-driven, or culturally embedded rather than lexically explicit. Zero-shot models fail to identify these cases, while fine-tuned models succeed in several cases. The texts express hostility through indirect accusation, moral condemnation, sarcasm, or rhetorical critique, often involving abstract or diffuse targets and limited emotional intensity. Correct predictions after fine-tuning suggest improved alignment with language-specific realizations of implicit offense present in the training data. Illustrative examples are presented in Table 8 in the Appendix. Overall, the results for RQ1 show a consistent pattern across Persian and Pashto: explicit abuse is robustly detected by all models, while implicit offense remains challenging, with fine-tuning partially mitigating this limitation.

6.2 RQ2: TL Effectiveness

We analyze the results of the highest-performing experiment (Gemma 3) for both languages. A clear asymmetry is noticed; transfer from Persian to Pashto is consistently stronger than from Pashto to Persian. In both directions, transfer succeeds for explicit, culturally shared insults; however, Persian-trained models generalize more robustly to Pashto than the other way around. reflecting differences in discourse explicitness and transferable norms. Detailed observation reveals that Persian to Pashto transfer performs particularly well on Pashto offenses that rely on direct profanity, kin-based honor insults, animalization, and moral or religious de-legitimization—patterns that are dominant and highly productive in Persian offensive discourse. Table 9 in the Appendix contains examples that are correctly detected after TL, closely mirroring Persian dehumanization, and moral exclusion. Similarly, politically moralized attacks grounded in shared sociocultural narratives—betrayal, hypocrisy, and foreign dependency —transfer robustly. Religious delegitimization and honor-based insults further strengthen TL, relying on shared Islamic moral frameworks and family-centered notions of shame.

In contrast, Pashto to Persian TL succeeds primarily when Persian offenses adopt Pashto's dominant style of overt, literal abuse with minimal discourse embedding. TL fails when Persian offensiveness is expressed through analytical argumentation, irony, or reflective critique. Pashto to Persian transfer also breaks only when Pashto offense is

implicit, idiomatic, or locally pragmatic rather than lexically explicit. Overall, the observed asymmetry indicates that Persian provides a richer and more generalizable offensive signal for Pashto, while Pashto offers a narrower subset of transferable patterns for Persian. This suggests that TL strength depends not only on how explicitly social norms and offense are expressed, but also on stronger in-language source performance, which enables more transferable supervision

6.3 RQ3: Target Effects

Cases detected by all models exhibit a strong alignment between surface form and semantic target. Offense is explicitly attached to the target through clear grammatical realization: pluralized group nouns, explicit individual reference, or unambiguous syntactic binding between insult and target. In these cases, the addressed and targeted entities coincide, minimizing the need for discourse inference; leading both zero-shot LLMs and FT classifiers converge on the correct target label.

Cases missed by all models show that shared failures arise when there is a mismatch between surface address and semantic target scope. A dominant pattern is proxy or representative targeting, where an individual addressee is used to insult a broader group or institution (e.g., political factions, media organizations, national or gender groups). Models consistently prioritize grammatical address over pragmatic generalization, labeling these cases as individual-targeted even when the semantic intent is clearly collective. Additional failure modes include implicit or abstract targets, metonymic references, proverb-like expressions, and ideological condemnation without explicit pluralization. In such cases, offense is group-directed at the semantic level but realized through individual-directed grammar, leading all models to converge on the same error. Table 1 in the Appendix presents some instances of target identification.

7 Conclusion

We benchmark offensive language detection, target identification, and cross-lingual transfer learning for Persian and Pashto using different LLMs. Results show that performance is driven by how offense is linguistically realized. Across both languages, models reliably detect explicit, profanity-rich abuse but consistently fail on implicit and context-dependent offenses. Fine-tuning improves recall by capturing language-specific realizations missed in zero-shot settings, reducing errors on implicit offense. Target identification further exposes shared structural limitations. Finally, transfer experiments show that Persian-to-Pashto transfer is consistently stronger than the reverse, indicating that effectiveness depends not only on linguistic proximity but also on how explicitly offensive patterns are encoded in the source language and on the overall strength of source-language models.

Limitations

This study has several limitations. First, our experiments are limited to two publicly available Twitter-based datasets—Pars-OFF (Persian) and POLD (Pashto). While standard benchmarks, they may not capture the full diversity of offensive language across platforms, domains, or registers, limiting the generalizability of our findings. Second, Pashto remains severely under-resourced compared to Persian, both in terms of dataset availability and pretrained models. This imbalance likely affects absolute performance and contributes to the observed asymmetry in TL. In particular, the weaker in-language performance for Pashto constrains how much transferable supervision Pashto-trained models can provide for Persian.

Third, our target identification analysis is limited to the target taxonomy provided by Pars-OFF (individual, group, other). The “other” category aggregates heterogeneous and often implicit target types, which may partly explain the consistently lower performance observed for this class. More fine-grained or discourse-aware target annotations could yield additional insights. Future work could focus on providing such resources for more in-depth analysis. While target labels are available for Persian (Pars-OFF), the absence of such annotations for Pashto (POLD) limits direct cross-lingual target analysis. One promising direction is to apply a Persian-trained target identification model to Pashto instances already classified as offensive, generating initial target predictions rather than final labels. These predictions could then be validated by human annotators, reducing annotation effort. Given the stronger Persian-to-Pashto transfer observed for explicit and culturally shared offensive patterns, this approach may achieve sufficient precision on a subset of Pashto data to serve as an effective pre-annotation step.

Ethics Statement

This study analyzes publicly available offensive datasets and does not involve collecting new user data. All datasets were obtained from prior peer-reviewed work or shared tasks that follow established ethical guidelines. Because offensive datasets may contain harmful or toxic language, examples shown in this paper are minimized and presented only when necessary for scientific transparency. No personally identifiable information is included in our datasets or model outputs. All experiments were conducted using anonymized text. Models trained in this work are not intended for deployment without further evaluation, fairness review, and context-specific calibration.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Taha Shangipour Ataei, Kamyar Darvishi, Soroush Javdan, Amin Pourdabiri, Behrouz Minaei-Bidgoli, and Mohammad Taher Pilehvar. 2022. Pars-off: a benchmark for offensive language detection on farsi social media. *IEEE Transactions on Affective Computing*, 14(4):2787–2795.

Zahra Bokaei, Walid Magdy, and Bonnie Webber. 2025. Culture matters in toxic language detection in Persian. *arXiv preprint arXiv:2506.03458*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The LLaMa 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Fatima-zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6048–6056.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: transformer-based model for Persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

Yuting Guo and Abeed Sarker. 2025. Benchmarking open-source large language models on healthcare text classification tasks. *arXiv preprint arXiv:2503.15169*.

Ijazul Haq, Weidong Qiu, Jie Guo, and Peng Tang. 2023. Pashto offensive language detection: a benchmark dataset and monolingual Pashto bert. *PeerJ Computer Science*, 9:e1617.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Emad Kebriaei, Ali Homayouni, Roghayeh Faraji, Armita Razavi, Azadeh Shakery, Heshaam Faili, and Yadollah Yaghoobzadeh. 2024. Persian offensive language detection. *Machine Learning*, 113(7):4359–4379.

Marzieh Mozafari, Khouloud Mnassri, Reza Farahbakhsh, and Noel Crespi. 2024. Offensive language detection in low resource languages: A use case of Persian language. *PLoS one*, 19(6):e0304166.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, pages 363–370.

PartAI. 2024. Partai/dorna2-llama3.1-8b-instruct. `https://huggingface.co/PartAI/Dorna2-Llama3.1-8B-Instruct`. Accessed: 2025-12-19.

Zahra Pourbahman, Fatemeh Rajabi, Mohammadhossein Sadeghi, Omid Ghahroodi, Somayeh Bakhshaei, Arash Amini, Reza Kazemi, and Mahdieh Soleymani Baghshah. 2025. Elab: Extensive LLMs alignment benchmark in persian language. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 458–470.

Mehran Safayani, Amir Sartipi, Amir Hossein Ahmadi, Parniyan Jalali, Amir Hossein Mansouri, Mohammad

Bisheh-Niasar, and Zahra Pourbahman. 2024. Opsd: an offensive Persian social media dataset and its baseline evaluations. *arXiv preprint arXiv:2404.05540*.

Mohammad Karami Sheykhlan, Jana Shafi, Saeed Kosari, and Saleh Kheiri Abdoljabbar. 2023. Parshao: Hate and offensive language detection on Persian tweets using machine learning and deep learning. *Authorea Preprints*.

Sumer Singh and Sheng Li. 2021. Exploiting auxiliary data for offensive language detection with bidirectional transformers. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 1–5.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

UNESCO Silk Roads Programme. a. Pashto. https://en.unesco.org/silkroad/silk-road-themes/languages-and-endanger-languages/pashto.

UNESCO Silk Roads Programme. b. Persian. https://en.unesco.org/silkroad/silk-road-themes/languages-and-endanger-languages/persian.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for Chinese offensive language detection. *arXiv preprint arXiv:2303.17927*.

Appendix

[1]

[1]This appendix contains offensive examples used solely for research purposes.

Sentence	Explanation	Gold Label	Predicted Label
Persian: @USER کیر فلا تو کس مادرت … مادر کصه … خارکصه … سسسسسگ ننتو از Damn you, fuck your mother's pussy... pussy mother... sister-fucker... ssssdog... your mother	The insult is directly bound to the target through second-person address and kin-based constructions ("مادرت"، "ننت"). There is no separation between who is spoken to and who is insulted.	individual	individual
خاک تو سرت سایبری… شما بسیجیا اصلا فیلم و سریال میفهمید چیه؟ "Shame on you, cyber troll... Do you Basijis even understand what film and series are?"	The addressee is a single user. But "سایبری" and "بسیجیا" clearly refer to an institutional / ideological group. The speaker uses one person as a stand-in for the group.	Group	Individual
حزبلیای سلطنت‌طلب "Royalist Hezbollah types."	This is a compressed ideological label. No insult word is needed; the condemnation is implicit. The phrase targets a political category, not a person. Without explicit plural insult syntax, models struggle with group target recognition.	Group	Other

Table 4: Target Identification In Persian.

Model	#Params	Reference
LLaMA-3-Instruct	8B	(Dubey et al., 2024)
Dorna2LLaMA-3	8B	(PartAI, 2024)
Gemma-3-Instruct	4B	(Team et al., 2025)
Mistral-Instruct	8B	(Jiang et al., 2023)
Aya-Expanse	8B	(Dang et al., 2024)
ParsBERT	162M	(Farahani et al., 2021)
XLM-RoBERTa	270M	(Conneau et al., 2020)
GPT-4o	–	(Achiam et al., 2023)

Table 5: Models used in this study.

Sentence	Explanation	Gold Label	Predicted Label
Persian: قبر پدر و مادرتون سگ برینه اگر مجاهدین را فالو میکنید خائن های کثیف \ May a dog shit on your parents' graves.	Here we see chained abusive elements: curse construction ("قبر … سگ برینه"), moral condemnation ("خائن"), and profanity ("کثیف"). Targets are explicit and plural, and the emotional intensity is high.	1	1
سید رضی لاشی تر از تو و امثال تو هست؟ مادرسگ کفتار امیدوارم داغ عزیزات یکی یکی رو دلت بمونه بچه لاشی! \ Is Seyyed Razi more of a lowlife than you and people like you? You dog's mother, you hyena — I hope the pain of losing your loved ones one by one stays in your heart. You filthy bastard	This example shows dense insult stacking across clauses: animalization ("کفتار"), kin-based insult ("مادرسگ"), and explicit malediction. The target is directly addressed and repeatedly reinforced. The offense is extreme, emotionally charged, and fully explicit		
Pashto: «شرم نلری غلام خنخیر» Shameless pig's slave.	This is a maximally direct insult with explicit dehumanization ("خنخیر" 'pig') and clear second-person address. The offensive intent is enacted directly, not implied. Emotional intensity is high		
خوان استشهادي ته جنت الفردوس غواړم او تالب خناورو ته جهنم غواړم "I wish the highest paradise (Jannat al-Firdaws) for the young martyr, and I wish hell for the Taliban beasts."	Here we see explicit animalization ("خناورو") paired with extreme moral polarization and direct malediction. Targets are clearly realized as groups, emotional intensity is high, and abusive intent is unambiguous.		

Table 6: Offensive Instances Correctly Predicted by All Models.

Sentence	Explanation	Gold Label	Predicted Label
Persian: «@USER حتما بفرستش، البته حرفت و تفاوت فرهنگی در پوشش و همینطور قوانین پوشش سلطنتی کاملا درسته، ولی بیماری پز دادن و خودنمایی در شبکه‌های اجتماعی هم در بین ما زیاد شده.» Definitely send it. Of course, what you said about differences in clothing culture, as well as royal dress codes, is completely correct—but the obsession with showing off and self-display on social media has also become widespread among us.	This instance is framed as reasonable social commentary and cultural critique. It contains no profanity, no insult lexicon, and no direct target attachment. Any potential offensiveness emerges implicitly through moral judgment (“بیماری پز دادن و خودنمایی”) rather than abusive language. Emotional arousal is low, vocabulary is neutral, and the text resembles legitimate argumentation.	1	0
«امروز داشتم فکر میکردم به این گزاره که: “چقدر بیچاره ست کسی که جرات نکنه به حضرت زهرا بگه؛ مادر!” (منظورم خودمم به خودتون نگیرید)...» Today I was thinking about this statement: ‘How unfortunate someone is who doesn’t dare to call Lady Fatima “mother”’... (I mean myself—don’t take it personally).	This example conveys ideological and religious positioning through reflective commentary rather than attack. While the statement implicitly encodes moral hierarchy and evaluative judgment, it lacks explicit hostility, insults, or target realization. The self-directed disclaimer further reduces perceived aggression.		
Pashto: 😍 ده لوڼې سره کرکټ نه کېږي You can’t play cricket with a headscarf (veil).	This is a Pashto proverb-like construction (“you can’t play cricket with a tail”), which functions metaphorically rather than literally. Any exclusionary or derogatory meaning is *indirect* and culturally encoded, relying on shared idiomatic knowledge. There is no explicit insult, no profanity, and no clear target attachment. Surface vocabulary is neutral, emotional arousal is low, and the emoji further softens the tone.		
د زمري اتحاد له ګیدړ سره.” The alliance of a lion with a jackal	This Pashto expression (“the alliance of the lion with the fox”) uses symbolic animal imagery to convey moral or political judgment indirectly. While it may encode criticism, opportunism, or betrayal, there is no explicit insult marker or target realization.		

Table 7: Offensive Instances Incorrectly Predicted by All Models.

Task:

Determine the type of target of the offense.

Target labels:
• Individual – the offense targets a specific person or clearly identifiable individual.
• Group – the offense targets a group of people defined by nationality, religion, ethnicity, gender, political affiliation, profession, or similar collective identity.
• Other – the offense targets an institution, organization, abstract entity, ideology, media, country, or the target is implicit, symbolic, or not clearly identifiable.

Instructions:
• Base your decision on the semantic intent, not only grammatical form.
• If an individual is used as a representative to insult a broader group, choose Group.
• If the target is abstract, symbolic, or indirect, choose Other.
• Choose the single best label.

Output format:
Answer with exactly one label on a single line:
Individual
Group
Other

Post:
"<TEXT>"

Figure 1: Prompt used for Target Identification

Sentence	Explanation	Gold Label	Predicted Label
Persian: @USER ازین کاسبا بزودی در ایران هم قبل انتخابات خیلی خواهیم دید. بخصوص براندازها و چپهای خریداری شده توسط سپاه برای زدن نامزدهای میانه و اصلاح‌طلب.» We will soon see a lot of these opportunists in Iran as well, especially before the elections—particularly the regime-change activists and leftists bought by the IRGC to attack moderate and reformist candidates.”	This instance expresses indirect accusation and delegitimization through political labeling (“خریداری شده”، “برای زدن”). There is no direct insult word, but the discourse clearly encodes hostility toward a group. Targets are collective and ideologically defined, not directly addressed. Zero-shot models often classify this as neutral political commentary. Supervised models benefit from exposure to in-domain political discourse where accusations of manipulation or treason function as implicit offense.	1	0
« باز قالیباف کار می‌کرد ... شماها که یا آدم می‌کشید و یا عکس یادگاری می‌گیرید، کاری نمی‌کنید که نقد کنیم یه موی قالیباف تو تَن شماها نیست...!» At least Ghalibaf was doing something... You people either kill others or just take souvenir photos—you don’t actually do anything we could even criticize... You don’t have a single hair of Ghalibaf’s worth in you!	This example is offensive in an indirect way. It does not use insults or swear words. Instead, it attacks the target by comparing two sides: one is shown as imperfect but working, while the other is described as morally worse (e.g., killing people or only taking photos). The criticism comes from this comparison, not from explicit abuse. Zero-shot models usually miss this kind of offense because there are no obvious offensive words. Fine-tuned models perform better because they have seen similar contrastive political criticism during training and learn that such comparisons often function as implicit attacks in Persian discourse.		
Pashto: د پوست مادرن نسل په ذهن کې یو لوی مشکل غرب محوره تفکر دی دوی خپله مطالعه نه کوي ټول مالومات له سوشل میډیا انټرن» One major problem in the mindset of the post-modern generation is West-centric thinking. They don’t study on their own; all their information comes from social media and the internet.	This sentence expresses implicit offense through ideological judgment, not through insults. The speaker criticizes a group (“پوستِ مادرن نسل”) by portraying them as intellectually shallow, dependent on Western thinking, and lacking independent study. There is no profanity, no direct insult, and no emotionally charged language. The hostility is conveyed indirectly through delegitimization of values and competence.		
او د یو میلیون تنو شهیدانو په وینو راغلی نظام به د څو تنو د شخصی هوس او ناپوهی ښکار سی And the system that was established with the blood of a million martyrs will fall victim to the personal whims and ignorance of a few people.	This sentence expresses moral condemnation through ideological framing. There is no profanity and no direct insult. The offense comes from accusing a group of betraying a system built on sacrifice. Zero-shot models often miss this because it looks like historical or political commentary. Fine-tuned models can detect it because they learn that this kind of framing is commonly used to delegitimize political actors in Pashto discourse.		

Table 8: Offensive Instances Predicted Correctly Only After Fine-Tuning LLMs.

Task:
Given a social media post written in Persian, determine whether it contains offensive language.

Definition:
Offensive language includes insults, abuse, humiliation, dehumanization, threats, or moral/religious delegitimization directed at an individual or a group.
Offense may be explicit (e.g., profanity, direct insults) or implicit (e.g., indirect attacks, moral condemnation, exclusionary rhetoric).

Instructions:
• Consider both literal wording and implied meaning.
• Do not require profanity for a post to be offensive.
• Criticism, disagreement, or political discussion is NOT offensive unless it attacks, humiliates, or degrades a target.
• If the intent is ambiguous, choose the label that best reflects the overall tone and implied hostility.

Output format:
Answer with exactly one label on a single line:
Offensive
Non-offensive

Post:
"<TEXT>"

Figure 2: Prompt used for Offensive Language Detection

Sentence	Explanation	Gold Label	Predicted Label
Pashto «ستا د خور درکوس غیږ واکاړم «ستا بڼخه وغیم ... ستا ناموس وغیم» I'll rip open your sister's cunt." "I'll fuck your wife... I'll fuck your honor	Family-cantered insults (mother / sister / wife / honor) are highly productive in Persian as well. The insult schema (attacking honor via female kin) is structurally identical across languages. Models trained on Persian already learn this pattern, so they generalize easily to Pashto.	1	1
«مرتده خدای مو تباه کره» «دا ټول د شیطان ابلیس ګروه دي» «خوارج خبیثان کلاب النار» "Heretic, may God destroy you." "They are all the group of Satan, Iblis." "Filthy Kharijites, dogs of hell."	Persian offensive discourse also uses these moral exclusion, and Islamic judgment. Shared Islamic vocabulary and moral framing enable strong transfer. These are explicit moral labels, not subtle pragmatics.	1	1
هره ونه میوه نیسي دغي وني بیا ټول خره نیولي دینه Every tree bears fruit; this tree, though, has only attracted donkey	Meaning is metaphorical and idiomatic, not literal. It requires cultural grounding in Pashto proverb logic. Persian models lack exposure to this implicit evaluative style.	1	0
وروره بعضي چې ګمراه شي بیا لاره په روښانه ورځ هم نه شي موندلی کیدی شي کوم تکلیف به لري او یا پیسي به اخلي په دغه بې ایمانه کار Brother, if some people go astray, they won't be able to find their way even in broad daylight. They might have some trouble or they might get paid for this dishonest work.	No profanity or insult terms appear. The offense is insinuation (corruption, moral failure) expressed as reflective commentary.	1	0

Table 9: TL Performance on Offensive Instances

Do Large Language Models Understand Double Mismatches? Evidence from Farsi

Maryam Mohammadi
Bielefeld University
maryam.mohammadi@uni-bielefeld.de

Abstract

Large language models (LLMs) are increasingly used for communication in many languages, therefore, understanding their limitations with respect to culture-specific pragmatics is important. While LLMs perform well on statistically frequent structures, their shortcomings are most evident in rare pragmatic phenomena. This study investigates whether LLMs can generate a (rare) complex honorific mismatch in Farsi. The pattern arises at two levels: (i) a plural pronoun disagrees with a singular referent for the sake of honorification, and (ii) the related components violate the Polite Plural Generalization due to intimacy implication. This *double mismatch* pattern is attested in everyday speech, though it is statistically sparse. We tested GPT-4 across multiple scenarios. The results reveal that the model successfully employs the first mismatch to indicate honorific, but fails to adopt the second mismatch that simultaneously conveys intimacy. The model thus deviates from human-like behavior at the syntax–pragmatics interface. These findings suggest that, while machine models demonstrate partial success in generating honorifics, they rely primarily on statistical patterns and lack the deeper pragmatic understanding necessary for contextual competence.

1 Introduction

In many languages, plural pronouns are used to express politeness when addressing a single honorific individual (Brown and Levinson, 1987), resulting in the (first) mismatch at the syntax–semantics interface between the plural form and its singular referent (e.g., using *Sie* 'you.2.PL' instead of *du* 'you.2.SG' in German to address a single individual politely). Comrie (1975) investigates the behavior of honorific pronouns and shows that while in some languages all related components (e.g., possessives and verb conjugations) symmetrically agree with polite plurals, in other languages the honorific pronoun controls plural agreement only on certain components. Notably, when the target components exhibit plural agreement with polite pronouns, they tend to do so consistently.

Such mismatch patterns are well studied across languages and are frequently used by speakers to systematically signal politeness. Consequently, large language models can readily learn these patterns, adopt their pragmatic functions, and deploy them in appropriate contexts (Noh et al., 2024). Farsi, as a positive face–saving language, employs many types of politeness (see Gohari Sadr et al., 2025, for the study on *Taarof*). Interestingly, Farsi exhibits a second type of mismatch that goes beyond politeness and is used to express intimacy toward an honorific addressee. This double–mismatch is statistically less frequent than the purely polite form, however, it is pragmatically effective in appropriate contexts.

This study investigates number (dis)agreement patterns in Farsi, focusing on their interaction at the syntax–pragmatic interface. The pronoun *šoma* serves both as the second-person plural and as the honorific second-person singular. In line with the *Polite Plural Generalization* (Wechsler, 2011), Farsi typically requires the polite plural controller to determine plural number agreement on all components marked for person and number features, i.e., verbs and related pronouns. Our data show that speakers employ singular agreement with the honorific plural pronoun in order to convey intimacy alongside politeness, but this pragmatic emerges only under specific settings. The present study explores how LLMs process double mismatches in Farsi. Our initial results reveal that LLMs successfully employ the first mismatch, exhibiting plural pronouns in honorific contexts. However, they fail to capture the second mismatch, expressing intimacy with the honorific addressee.

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 24–28

2 Polite Plural

Plural pronouns are frequently used to express politeness when addressing a single honorific individual (Brown and Levinson, 1987). According to the Polite Plural Generalization (Wechsler, 2011), a polite plural pronoun functions as an agreement controller, imposing plural number on any related components marked for person and number features. In many languages, second-person plural pronouns employed honorifically for a single addressee systematically trigger plural agreement on all eligible targets (Wechsler, 2011).

Comrie (1975) examines the syntax–semantics behavior of polite/honorific pronouns, showing that while in some languages (e.g., Croatian) all agreement targets (e.g., verbs, possessives, reflexives pronouns) uniformly agree with polite plurals, in others (e.g., French) the honorific pronoun triggers plural agreement only on certain targets (e.g., verbal predicates but not adjectival predicates). This contrast motivates a typology distinguishing *uniform* and *mixed* agreement systems. Generally, languages tend to adhere to their (mis)matching patterns consistently, whether uniform or mixed. In other words, while such patterns vary across languages, they are fixed within individual languages.

There are two main approaches in the literature; in the first syntactic approach, honorificity is encoded on a functional projection (e.g., CP) in the clause–periphery, via a feature. This projection is responsible for licensing honorificity on all 2.person components in the clause, via binding on 2.person pronouns (Portner et al., 2019; Alok, 2021). In the second semantic–pragmatic approach, the expressive honorific content on the lexical entries of the two items is comparable in semantics/pragmatics (Potts, 2007; McCready, 2019). That is, if a speaker uses an item which honors the addressee, and also an item that dishonors the addressee within the same sentence, the combination would be conceptually strange and language would block the combination.

We will see that the double mismatch in Farsi does indeed occur, pragmatically conveying both honorificity and intimacy simultaneously (see also Puškar-Gallien, 2019, for discussions on verb–politeness conflicts).

3 Core Data

In this section, we present data from Farsi; particularly the colloquial form form spoken in Tehran, Iran. Farsi is an SOV language and exhibits subject–verb agreement in person and number. Like many other languages, Farsi employs polite plural: the pronoun *šoma* functions both as the canonical second person plural form and as an honorific singular form. In line with the Polite Plural Generalization (Wechsler, 2011), Farsi typically requires the polite plural controller to determine plural number agreement on all components marked for number feature, including verbs and related pronouns (e.g., possessives and reflexives).

As illustrated in example (1), the speaker addresses her grandmother in polite plural pronoun *šoma* you.2.PL (the $\rightsquigarrow$ indicates the added implication). The polite pronoun controls the number agreement on the verb and the possessive. Note that in Farsi, subject (A1) and object (A2) pronouns have the same form.

(1) **Context:** A is talking with her grandma.

A1: *šoma* *čâe-tun* *ro* *xord-id?*
you.2.PL tea-2.PL.POSS ACC eat-2.PL

'Did you have your tea?' $\rightsquigarrow$ Honorific implication

A2: *man* *lebâse* *šoma* *ro* *behe-tun* *dâdam.*
I clothes you.2.PL ACC to-2.PL gave

'I gave your clothes to you.' $\rightsquigarrow$ Honorific implication

Colloquial data shows that number disagreement can be strategically employed to convey an *Honorific* yet *Intimacy* relationship with the addressee. In the same context (1), alternative (A1′) features a plural subject pronoun, while the corresponding possessive and the verb appear in singular form, resulting in a number mismatch with the plural pronoun controller (Ferguson, 1991).

A1′: *šoma* *čâe-t* *ro* *xord-i?*
you.2.PL tea-2.SG.POSS ACC eat-2.SG

'Did you have your tea?' $\rightsquigarrow$ Honorific and Intimacy.

Crucially, the pattern is pragmatically infelicitous in the absence of intimacy between the interlocutors, for instance, when addressing a socially superior individual (e.g., in a manager–employee relationship). This deliberate mismatch reflects a nuanced role of discourse on syntactic structures.

By contrast, in (A1″), the use of plural object possessive pronouns, agreeing with the plural subject, alongside a singular verb results in infelicity. This suggests that, the number feature must be symmetrically underspecified across all target components in the sentence.

A1″:# *šoma* *čâe-tun* *ro* *xord-i?*
you.2.PL tea-2.PL.POSS ACC eat-2.SG

It is worth to note that since Farsi is a subject-drop language, the subject pronoun *šoma* can optionally be omitted. While removing *šoma* from (A1′) eliminates the honorific implication, omitting it in (A1″) does not resolve the infelicity caused by the number mismatch, though the form still includes a plural pronoun, in clitic form, and a singular verb.

The same pattern holds when *šoma* appears as an object pronoun. Object pronouns in Farsi can occur in either free form or clitic form that attaches to objects or verbs (Rasekh Mahand, 2014). In example (2), using the free object pronoun *šoma* with number disagreement with the reflexive in (A1) is pragmatically felicitous, however, in (A2), the object clitic form *-tun* cannot achieve the same effect and results in infelicity. Thus, this underspecification is possible only with the free pronouns *šoma*, but not with the clitic counterpart *-tun*.

(2) **Context:** A is talking with her grandma.

A1: *qorse* *šoma* *ro* *emruz* *bara-t* *mixram.*
tablet you.2.PL ACC today for-2.SG buy

'I will buy your tablet today.'

A2:# *qorse-tun* *ro* *emruz* *bara-t* *mixaram.*
tablet-your.PL ACC today for-2.SG buy

4 LLMs and Honorific Mismatches

Recent advances in large language models have prompted linguists to explore how these artificial systems handle the complex structures characteristic of natural languages. While some theoretical linguists remain skeptical about the relevance of LLMs for studying natural language, their utility extends beyond serving as sophisticated computational artifacts. They can function as valuable tools for testing linguistic hypotheses and refining theories of human linguistic competence.

Given the complexity of double mismatch pattern in Farsi, the present study aims to evaluate whether LLMs can recognize and adopt these less frequent yet highly functional forms. Although honorific (mis)matches have been extensively investigated in English and other European languages, the behavior of LLMs with respect to Farsi remains largely unexplored (see Gohari Sadr et al., 2025, among all). In this paper, we aim to control the performance of LLMs on Farsi double mismatches.

It is worth acknowledging that this study primarily focused on the linguistic analysis of double mismatch phenomena. However, we extended our investigation as a pilot study on LLM behavior in double mismatches. Here, we report only the preliminary results of this first attempt. Regarding the intriguing initial observations, in which the tested model failed to capture the second mismatch in intimacy, a more comprehensive evaluation is reserved for our future work. Our next steps will involve a strengthened experimental design, testing more contexts across multiple LLMs, including Persian models (e.g., ParsBERT[1]), and comparing model performance with a larger baseline of native speakers. Although our current investigation is limited in both the number of trials and the number of models tested, the findings underscore the importance of examining how LLMs process complex pragmatic patterns.

We tested GPT-4 on the double mismatch pattern using six carefully designed contexts. Each context involved two interlocutors with an age difference (e.g., grandchild–grandparent, niece–elder uncle) to trigger honorific marking, as well as a friendly relation to trigger intimacy. The experimental settings, illustrated in example (3), positioned the model as the speaker, who asks a question from the honorific addressee. In all trials, we included a possessive or reflexive pronoun that, in addition to the verb, should agree with the subject. If the LLM only adopts intimacy, all referring forms should appear in the singular, as in (A1). If it adopts the single mismatch (honorific), pronouns and verbs should agree with the polite plural, as in (A2). If it adopts the double mismatch (honorific + intimacy), the agreement components should appear in the singular form, regardless of the polite plural pronoun, as in (A3). Note that the prompts were originally presented in Farsi, they are provided here in English for brevity.

(3) **Prompt:** You have a close relationship with your grandma. You are sitting with your grandparents in the living-room. Address your grandma and ask if she has already taken her tablets.

A1: *to* *qorshâ-t* *ro* *xord–i?*
you.2.SG tablets-2.SG ACC ate-2.SG

A2: *šoma* *qorshâ-tun* *ro* *xord–id?*
you.2.PL tablets-2.PL ACC ate-2.PL

A3: *šoma* *qorshâ-t* *ro* *xord–i?*
you.2.PL tablets-2.SG ACC ate-2.SG

'Did you have your tablets?'

[1]Thanks to the EACL reviewer for the helpful suggestion.

Across all six contexts, GPT consistently adopted the intimacy interpretation over politeness and produced singular forms, as illustrated in (A1). We then tested the same contexts with the explicit close-relationship cue replaced by an implicit formulation. For instance, in example 3, we replaced '*You have a close relationship with your grandma*' with '*You grew up with your grandma*'. This time, the model constantly produced the honorific mismatch pattern, as in (A2), ignoring the intimacy.[2]

Notably, we validated the trial contexts with four native Farsi speakers. The results indicate that they readily interpreted both experimental settings, those with explicit and implicit cues, and predominantly produced the double mismatch forms shown in (A3), while occasionally producing the other two forms.

Although the number of trials is admittedly limited, the results underscore the importance of investigating complex pragmatic patterns in LLMs. Double mismatches require the language user, whether human or machine, to consider two layers of social relations, encompassing both honorific marking and intimacy. While the honorific and intimacy distinctions are each frequent and readily captured by LLMs in isolation, their combination at the second layer appears less accessible, likely due to its relative infrequency.

It is worth noting that we observe double mismatch patterns in two conditions: (i) when the addressee is older than the speaker, as shown in (3), and (ii) when the addressee is (significantly) younger than the speaker. Following the politeness principles (Brown and Levinson, 1987), in the second condition the speaker typically does not use the honorific plural.

However, in our investigation, we observed several instances of the second condition in which double mismatches were employed. For example, some parents use the polite plural when addressing their children to emphasize the child's developing character, signaling that, despite their young age, they should behave respectfully. In this regard, some parents use the double mismatch to maintain intimacy with their children.

As illustrated in example (4), the mother addresses her three-year-old son with the polite plural *šoma* to practically teach respectful behavior, while using the singular verb to preserve intimacy.

(4) **Context:** A mother is talking to her three-year-old son.
A: šoma pofak mixor–i?
you.2.PL crunchy eat-2.SG

'Do you want a crunchy?' ⇝ Honorific and Intimacy.

We evaluated this condition using GPT-4 across four scenarios. In every case, the model consistently used the second-person singular pronoun '*to*' (corresponding to pattern (A1) in example (3) above), failing to produce the double mismatch option. Notably, our pilot native informants also showed less frequent use of double mismatches in this setting. This can be explained by the fact that using honorifics for children reflects emerging psychological and behavioral practices among younger parents and is less familiar to older parents. Consequently, the double mismatch in this context is currently less prevalent. We expect it to be used more frequent among younger speakers and, consequently, to increase in the near future by (at least human) speakers.

5 Discussion

This study examines syntax–pragmatic mismatches in Farsi, focusing on the interaction between the polite plural and its number-agreement targets. The data show that the number feature is strategically underspecified to signal intimacy alongside politeness. Specifically, the polite plural pronoun *šoma* mismatches with its singular referent due to honorific implication, while other components, such as the verb and associated pronouns, remain in the singular form to convey intimacy. Crucially, this underspecification operates symmetrically across all relevant components. Thus, although the polite plural appears to mismatch its targets, Farsi nonetheless exhibits *uniform* (dis)agreement in number at the syntax–pragmatics interface.

We tested GPT-4 on double mismatch constructions to investigate how a representative large language model handles such complex pragmatic function. The initial results suggest that, although the model exhibits partially human-like behavior in processing honorifics, it fails to produce the double mismatch form and instead appears to rely primarily on statistical cues in the training data.

Although the model was prompted with only a small number of trials, it consistently failed

[2] As an anonymous reviewer rightly noted, changing the explicit close-relationship cue to an implicit one may make it unclear what exactly the model is sensitive to (honorifics, kinship, register, or lexical triggers). We will consider this in our future work.

to generate double mismatches and instead produced the plural honorific in a single-mismatch form. This outcome highlights the importance of linguistic patterns that, while statistically sparse, are pragmatically functional. Such patterns are often overshadowed by highly frequent constructions in LLMs, which can lead to their omission in downstream LLM-based frameworks or applications.

Given the increasing role of LLMs in everyday communication, it is crucial that these models are trained to recognize and produce complex pragmatic patterns, ensuring that low-frequency yet functionally significant forms are not lost in future language technologies. Such rare sociopragmatic constructions are precisely where LLMs often struggle, and our data provide clear evidence of this limitation. Understanding these shortcomings in handling culture-specific pragmatic phenomena is therefore essential. We emphasize that the statistical rarity of certain syntactic exceptions encoding complex pragmatic meanings warrants further investigation, as such research may inform improvements in LLM architectures and training strategies.

Limitations

This study was conducted under a tight timeline as a pilot extension of a broader linguistic investigation of mismatch patterns at the syntax–pragmatics interface. As such, our novel human-based findings were not systematically tested in large language models. Consequently, the results should be interpreted as preliminary rather than robust. Our investigation is limited in several respects. First, the number of prompt contexts was small. Second, we tested only a single model (GPT-4), which restricts the generalizability of the findings. Third, no statistical analysis was conducted, and the evaluation is therefore underpowered.

While the findings underscore the importance of examining LLMs with rare pragmatic patterns, the limitations prevent strong empirical claims regarding model behavior. Furthermore, the present study focuses exclusively on the generation of double mismatch patterns. The perception and interpretation of such mismatches by LLMs were not examined and remain for future research. As this is an ongoing project, we plan to address these limitations by expanding the range of models and contexts tested, incorporating systematic statistical analyses, and investigating both production and perception of mismatch patterns in future works.

Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): CRC 1646/1 2024 – 512393437, Project INF.

References

Deepak Alok. 2021. The morphosyntax of magahi addressee agreement. *Syntax*, 24:263–296.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge, UK.

Bernard Comrie. 1975. Polite plurals and predicate agreement. *Language*, 51(2):406–418.

Charles A. Ferguson. 1991. Individual and social in language change: Diachronic changes in politeness agreement in forms of address. In Robert L. Cooper and Bernard Spolsky, editors, *The Influence of Language on Culture and Thought: Essays in Honor of Joshua A. Fishman's Sixty-Fifth Birthday*, pages 183–198. De Gruyter Mouton, Berlin and Boston.

Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoomian, Laleh Seyyed-Kalantari, and Ali Emami. 2025. We politely insist: Your llm must learn the persian art of taarof. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1819–1838. Association for Computational Linguistics.

Elin McCready. 2019. *The Semantics and Pragmatics of Honorification*. Oxford University Press, Oxford.

Kangsan Noh, Sanghoun Song, and Eunjeong Oh. 2024. How language models understand honorific mismatches in korean. *Language Research*, 60(3):303–322.

Paul Portner, Miok Pak, and Raffaella Zanuttini. 2019. The speaker–addressee relation at the syntax–semantics interface. *Language*, 95(1):1–36.

Christopher Potts. 2007. The expressive dimension. *Theoretical Linguistics*, 33(2):165–198.

Zrinka Puškar-Gallien. 2019. Resolving polite conflicts in predicate agreement. *Glossa: A Journal of General Linguistics*, 4(1):33.

Mohammad Rasekh Mahand. 2014. Persian clitics: Doubling and agreement. *Modern Language Journal*, 24:16–33.

Stephen Wechsler. 2011. Mixed agreement, the person feature, and the index/concord distinction. *Natural Language & Linguistic Theory*, 29(3):999–1031.

TajPersLexon: A Tajik–Persian Lexical Resource and Hybrid Model for Cross-Script Low-Resource NLP

Mullosharaf Kurbonovich Arabov
Department of Data Analysis and Technological Programming
Institute of Computational Mathematics and Information Technologies
Kazan (Volga Region) Federal University
420008, Kazan, Russia
MKArabov@kpfu.ru

Abstract

This work introduces **TajPersLexon**, a curated Tajik–Persian parallel lexical resource of **40,112** word and short-phrase pairs for cross-script lexical retrieval, transliteration, and alignment in low-resource settings. We conduct a comprehensive **CPU-only benchmark** comparing three methodological families: (i) a lightweight hybrid pipeline, (ii) neural sequence-to-sequence models, and (iii) retrieval methods. Our evaluation establishes that the task is essentially solvable, with neural and retrieval baselines achieving 98-99% top-1 accuracy. Crucially, we demonstrate that while large multilingual sentence transformers fail on this exact lexical matching, our interpretable hybrid model offers a favorable accuracy-efficiency trade-off for practical applications, achieving 96.4% accuracy in an OCR post-correction task. All experiments use fixed random seeds for full reproducibility. The dataset, code, and models will be publicly released.

1 Introduction

Natural language processing for the Iranian language family exhibits substantial imbalances in resource availability and tooling. While Persian (in its Iranian and Dari standards) has been the focus of numerous computational efforts, related varieties such as Tajik remain comparatively under-resourced. Tajik is primarily written in Cyrillic script, whereas Persian varieties commonly employ the Perso-Arabic script; this digraphic landscape creates a cross-script challenge that complicates lexical alignment, transliteration, retrieval and downstream applications such as machine translation and optical character recognition (OCR) (Megerdoomian and Parvaz, 2008; Merchant and Tang, 2024). Furthermore, existing textual and lexicographic materials for Tajik are often heterogeneous in format and not directly aligned with Persian resources, limiting their immediate usefulness for cross-script computational methods (Shukurov et al., 1969; Ghiyosiddin, 1987–1989; Nazarzoda et al., 2008; taj).

Prior research has explored transliteration and mapping across Tajik and Persian scripts using statistical machine translation techniques (Davis, 2012), rule-based systems and neural approaches including transformer models (SadraeiJavaheri et al., 2024; Merchant et al., 2025). At the same time, a maturing toolkit ecosystem (e.g. fairseq) and well-established evaluation measures (e.g. the chrF family) support experimental rigour (Ott et al., 2019; Popović, 2017). However, many modern neural approaches depend on substantial pretraining and GPU resources, which reduces accessibility and reproducibility in computationally constrained environments; this motivates research that prioritises lightweight, interpretable and reproducible pipelines (Arabov et al., 2025; Arabov and Sedykh, 2025).

To address these gaps we introduce **TajPersLexon**, a curated Tajik–Persian parallel lexical resource of approximately **40,112** word and short-phrase pairs annotated with part-of-speech information and illustrative examples. Building on this resource, we conduct a comprehensive evaluation of multiple methodological families: (i) a compact CPU-only hybrid pipeline combining joint subword tokenisation (SentencePiece), subword-aware distributional embeddings (FastText and Word2Vec), and a ranking model fusing embedding similarity, edit-distance retrieval and rule-based transliteration; (ii) modern sequence-to-sequence models (LSTM and Transformer architectures); and (iii) retrieval-based methods (BM25). Our design goals prioritise reproducibility, interpretability and accessibility: while establishing strong neural baselines, we particularly focus on lightweight approaches usable without GPU infrastructure, providing practical solutions for low-resource scenarios. **To our knowledge,**

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 29–37

this is the largest publicly available machine-readable Tajik-Persian lexicon and the first comprehensive benchmark for this cross-script task.

Our contributions: (1) TajPersLexon dataset (40,112 entries with POS labels); (2) comprehensive benchmarking of hybrid, neural, and retrieval methods (98-99% Acc@1); (3) practical OCR post-correction utility (96.4% accuracy); (4) fully reproducible CPU-only setup. All resources will be publicly released.

2 Related work

We review five relevant areas: lexicographic resources, transliteration methods, tokenization techniques, evaluation tools, and low-resource methodologies, positioning TajPersLexon within this landscape.

Lexicographic and corpus resources. Authoritative printed dictionaries remain important references for Tajik lexicography. Historical and large-scale dictionaries such as Shukurov et al.'s dictionary (Shukurov et al., 1969), Ghiyosiddin's comprehensive dictionary (Ghiyosiddin, 1987–1989) and Nazarzoda et al.'s explanatory dictionary (Nazarzoda et al., 2008) provide rich descriptions, but are rarely distributed in machine-readable parallel form suitable for computational experiments. Digital corpora and national collections (e.g. the Tajik National Corpus) supply raw text but often lack aligned bilingual lexical pairs (taj). Recent efforts to assemble multiformat corpora for Tajik attempt to close this gap (Arabov et al., 2025); TajPersLexon aims to complement these resources by providing an explicitly parallel, POS-annotated lexicon with usage examples.

Transliteration and cross-script mapping. Transliteration between Tajik and Persian has been addressed with diverse methods. Early approaches treated transliteration as a sequence mapping problem using SMT-style models (Davis, 2012). More recently, neural seq2seq and transformer-based models have been applied to transliteration and dialect bridging, achieving strong results when sufficient parallel data and compute are available (SadraeiJavaheri et al., 2024; Merchant et al., 2025). Corpora such as ParsText support these efforts by providing digraphic data (Merchant and Tang, 2024). Neural methods are powerful but frequently depend on large pretrained models and GPU resources; this motivates complementary lightweight and hybrid approaches that are accessible in constrained environments.

Tokenisation, embeddings and hybrid methods. Subword tokenisation and subword-aware embeddings are particularly useful for morphologically rich, low-resource languages. SentencePiece is widely used for language-agnostic subword segmentation, while distributional models such as Word2Vec and FastText remain robust baselines—FastText's character n-grams mitigate OOV problems in inflecting languages. Combining distributional similarity with string-based measures (e.g., edit distance) and rule-based transliteration exploits complementary strengths: embeddings capture distributional semantics, string methods capture orthographic correspondences. Prior comparative studies show that such hybrid strategies improve robustness in low-resource, cross-script tasks (Arabov and Sedykh, 2025; Mohtaj et al., 2018); our hybrid ranking follows this rationale and tunes component weights on held-out data.

Tooling and evaluation. A mature tooling ecosystem (e.g. fairseq) facilitates reproducible sequence modelling experiments (Ott et al., 2019). For transliteration and related tasks, character-level metrics (chrF, CER) alongside retrieval metrics (Accuracy@k, MRR) provide complementary perspectives on performance (Popović, 2017). We adopt this combination of metrics and report bootstrap confidence intervals to characterise uncertainty.

Low-resource methodology and accessibility. There is an active research strand on bootstrapping methodologies and practical workflows for low-resource languages, covering corpus preparation, preprocessing and lightweight modelling strategies (Megerdoomian and Parvaz, 2008; Arabov et al., 2025; Arabov and Sedykh, 2025). Language-specific preprocessing tools (e.g. Parsivar for Persian) illustrate the gains from tailored pipelines (Mohtaj et al., 2018). Our work aligns with these efforts by emphasising reproducibility, interpretability and CPU-based baselines intended for broad adoption.

Summary and differentiation. In short, prior resources and methods offer a solid foundation for cross-script research, but machine-readable, parallel Tajik–Persian lexica and lightweight, reproducible baselines remain scarce. TajPersLexon addresses this gap by combining lexicographic curation with a compact hybrid modelling framework (subword tokenisation, distributional embeddings and string-based measures) and by releasing data

and code to support follow-up research.

3 Dataset

Sources. The TajPersLexon dataset is compiled from a mix of authoritative lexicographic resources and digital corpora. Primary sources include printed Tajik dictionaries and the Tajik National Corpus (TNC) (Shukurov et al., 1969; Ghiyosiddin, 1987–1989; taj). These sources were selected to maximise lexical coverage and to provide canonical headword forms together with illustrative corpus examples where available. Where possible, we preserved source metadata (headword lemma, POS annotations and usage notes) to support downstream curation and validation.

3.1 Curation Pipeline and Normalization

Dataset construction followed a semi-automated, reproducible pipeline:

1. **Extraction:** Candidate Tajik–Persian correspondences were extracted from digitized dictionary renditions and aligned corpus segments.

2. **Normalization:** Unicode NFC normalization was applied. For Tajik Cyrillic, we standardized orthographic variants (e.g., russian → russian where appropriate) and regularized the use of **russian** in diphthongs. For Persian Perso-Arabic, we unified common aleph (ٱ) and hamza variations where semantically neutral, following standard Persian text-processing conventions.

3. **Deduplication:** Exact duplicates were removed, followed by fuzzy matching (Levenshtein distance < 2) for near-identical forms.

4. **Enrichment:** Entries were aligned with part-of-speech labels and illustrative examples where available.

5. **Quality control:** A stratified random sample of 5% (2,006 pairs) was manually reviewed by a native Tajik speaker, correcting systematic OCR, tokenization, and script-conversion errors. The pre-correction error rate in the sample was 3.2%, reduced to 0.4% after manual review.

Multi-word expressions (MWEs) such as compound nouns (russian – arabic) and light-verb constructions (russian – arabic) are included as atomic units without compositional decomposition, reflecting dictionary-lookup usage. **Morphological variants** (inflected forms) are preserved as separate entries rather than lemmatized, maintaining surface-form diversity important for retrieval tasks.

Statistic	Value
Records (N)	40,112
Tajik types	40,112
Persian types	37,546
Distinct queried forms (`_queried_word`)	1,630
Avg. examples per record	0.52
Avg. example length (chars)	83.5
Curation sample reviewed	2,006 (5%)
Pre-correction error rate	3.2%
Post-correction error rate	0.4%

Table 1: High-level statistics and curation quality metrics for TajPersLexon.

Part-of-speech annotation. POS labels are derived from a combination of sources: where available, we retain POS metadata from the original lexicographic sources; for entries lacking explicit tags we apply a lightweight rule- and lexicon-based assignment heuristic that leverages dictionary headword information and simple morphological cues. A random subset of automatically-assigned labels was manually inspected during curation and corrected where necessary. The final POS inventory is reported in Table 2.

Record format. Each dataset record is stored as a single JSON object in a newline-delimited file (JSONL) with the canonical fields `tajik` (Cyrillic), `persian` (Perso-Arabic), `part_of_speech` (Tajik POS labels) and `examples` (array of illustrative sentences, if available). A typical record:

```
{"tajik":"russian",        "persian":"",
"part_of_speech":"",
"examples":["russian                    .
- c"]}
```

High-level statistics. The cleaned dataset contains **40,112** records. Table 1 summarises key corpus-level statistics and curation quality metrics.

Table 2 reports the part-of-speech distribution. As typical for dictionary-derived lexica, open-class categories dominate.

Observations. Several points are notable: (1) **Asymmetry in token coverage:** Fewer unique Persian forms (37,546) than Tajik (40,112) reflects normalisation choices and genuine one-to-many Tajik→Persian correspondences. (2) **Low example density:** 0.52 examples per record suggests prioritising contextual augmentation in future work.

Part of speech (Tajik label)	Count
Noun	21,987
Adjective	14,375
Adverb	1,458
Verb	1,302
Proper noun	398
Interjection	227
Numeral	133
Conjunction / particle	85
Pronoun	35
Functional morpheme / particle	29
Preposition	23
Postposition	9
Unclassified / other	52

Table 2: Distribution of parts of speech in TajPersLexon.

(3) **POS skew:** Nouns and adjectives dominate; verbs and closed-class words are under-represented, which may affect downstream systems requiring robust morphological coverage. (4) **Canonical query forms:** The field _queried_word contains 1,630 distinct normalised lookup forms, enabling both surface-form evaluation and lemma-level analysis.

Splits and partitioning. For experiments we use deterministic train/dev/test splits (80/10/10) stratified by POS, generated with fixed seed `seed=42`. After removing five malformed records from the test partition, the final evaluation set contains **4,011** Tajik–Persian pairs (used in Section **??**).

Reproducibility and release. All preprocessing, normalisation and curation scripts are versioned. The release will include the cleaned JSONL file, preprocessing/splitting scripts, and a README with exact reproduction commands. The dataset aggregates material under fair-use principles for non-commercial research, with all original sources credited.

The following sections describe the methodological families evaluated on TajPersLexon: lightweight hybrid models, neural sequence-to-sequence baselines, retrieval-based methods, and multilingual sentence encoders.

4 Methodology

We design and evaluate multiple methodological families for Tajik–Persian cross-script lexical retrieval, ranging from lightweight symbolic methods to modern neural architectures. All experiments enforce strict CPU-only constraints to ensure reproducibility and accessibility in low-resource settings, with fixed random seeds and deterministic execution.

Task definition. We formulate the task as *cross-script lexical retrieval*: given a Tajik query word in Cyrillic script, retrieve the corresponding Persian lexical form in Perso-Arabic script from a fixed candidate set. This subsumes dictionary-lookup and transliteration scenarios under closed-vocabulary retrieval evaluation.

Data splits and reproducibility. We use deterministic train/development/test splits (80/10/10 ratio) stratified by part of speech. Generated with fixed seed `seed=42`, the splits yield approximately 32,090 training, 4,011 development, and 4,011 test instances (after removing 5 malformed records, final test $N = 4,011$). All experiments run on CPU with single-threaded execution to ensure deterministic reproduction.

Hybrid and Neural Models

Hybrid ranking model integrates complementary signals from multiple sources. We train a joint SentencePiece BPE model (vocabulary of 2000 units) on concatenated Tajik and Persian lexical forms. For distributional embeddings, we train FastText (with character n-grams $n \in [3, 6]$) and Word2Vec skip-gram models (200-dimensional vectors, window size 5, minimum frequency 2, 10 epochs), obtaining word vectors by averaging constituent subword embeddings. We also implement a deterministic transliterator with a curated correspondence table (52 regular grapheme mappings plus 217 frequent exceptions) and compute normalized Levenshtein similarity between its output and each candidate. These signals are combined via linear fusion: for each query–candidate pair we compute

$$S = \alpha S_{FastText} + \beta S_{Word2Vec} + \gamma S_{edit} + \delta S_{rule}, \quad (1)$$

where weights $\alpha, \beta, \gamma, \delta$ are tuned on the development set to maximise Mean Reciprocal Rank (MRR).

Neural sequence-to-sequence baselines include a bidirectional LSTM encoder (256 hidden units) with decoder and Bahdanau attention, trained for 10 epochs with teacher forcing and cross-entropy loss; and a compact transformer with 2 encoder/decoder layers, 4 attention heads, 128-dimensional embeddings, trained for 15 epochs. Both models operate at character level and generate Persian transliterations via greedy decoding, providing direct comparison to transliteration-focused approaches.

Retrieval, Phonetic and Transfer Learning Methods

Retrieval and phonetic approaches encompass traditional information-retrieval ranking (BM25 with parameters $k_1 = 1.5$, $b = 0.75$,

indexing Tajik queries against Persian candidates) and Soundex-based phonetic matching with script-specific mappings for Cyrillic and Perso-Arabic.

Multilingual sentence encoders provide transfer-learning baselines using four pre-trained models: SentenceTransformer: `paraphrase-multilingual-MiniLM-L12-v2` (50 languages), FastMultilingualST: `distiluse-base-multilingual-cased-v2` (50+ languages), PowerfulMultilingualST: `paraphrase-xlm-r-multilingual-v1` (100+ languages), and MultilingualSimilarityST: `stsb-xlm-r-multilingual` (semantic-similarity tuned). These models offer strong cross-lingual representations but require loading substantial pre-trained weights (120–550MB).

Evaluation Framework

Evaluation regimes employ two complementary setups: the primary regime uses a candidate pool of 1,000 distractors (gold + 1,000), yielding the main results in Table 3; the stress regime uses 3,000 distractors with variable query subset sizes for diagnostic analysis.

Metrics and statistical analysis include retrieval metrics (Accuracy@1/5/10, Mean Reciprocal Rank), transliteration metrics (Character Error Rate, chrF for sequence-to-sequence outputs), statistical uncertainty via bootstrap confidence intervals (1,000 iterations), and efficiency metrics (training/evaluation times, memory footprint).

Implementation details cover the software stack: SentencePiece for tokenisation; Gensim for Word2Vec/FastText; PyTorch for neural models; and the sentence-transformers library for multilingual encoders. Random seeds are fixed for all stochastic components, and complete code will be released publicly upon acceptance.

5 Experimental Setup

Data splits. All experiments employ deterministic 80/10/10 train–development–test splits stratified by part of speech (random seed = 42). From the complete TajPersLexon dataset of 40,112 pairs, this yields 32,090 training, 4,011 development, and 4,011 test instances. After post-split validation and removal of five malformed records, the final evaluation set contains **4,011** Tajik–Persian pairs. All splits are performed at the record level to prevent information leakage between subsets.

Model configurations. We implement multiple methodological families under strict CPU-only constraints: Hybrid model components: SentencePiece BPE with a joint vocabulary of 2000 subword units trained on concatenated Tajik–Persian forms; Embeddings using FastText (character n-grams $n \in [3, 6]$) and Word2Vec skip-gram models (200 dimensions, window=5, $\min_count = 2, 10epochs); Fusionweightsinitialisedata\alpha = 0.4$ (FastText), $\beta = 0.3$ (Word2Vec), $\gamma = 0.2$ (edit distance), $\delta = 0.1$ (rule-based), then tuned on the development set to maximise MRR. Neural sequence-to-sequence models: LSTM with attention using a bidirectional encoder (256 hidden units), Bahdanau attention decoder, trained for 10 epochs with scheduled teacher forcing (ratio decay from 1.0 to 0.5); Transformer with 2 encoder/decoder layers, 4 attention heads, 128-dimensional embeddings, trained for 15 epochs with learning rate warmup (1,000 steps) and cosine decay. Retrieval and phonetic baselines: BM25 with parameters $k_1 = 1.5$, $b = 0.75$, using Tajik query against Persian candidate text; Phonetic similarity using a custom Soundex implementation with script-specific phonetic mappings for Cyrillic and Perso-Arabic. Multilingual sentence encoders: Four pre-trained multilingual models (as suggested by reviewers): SentenceTransformer: `paraphrase-multilingual-MiniLM-L12-v2` (117MB); FastMultilingualST: `distiluse-base-multilingual-cased-v2` (470MB); PowerfulMultilingualST: `paraphrase-xlm-r-multilingual-v1` (1.1GB); MultilingualSimilarityST: `stsb-xlm-r-multilingual` (1.1GB).

Evaluation regimes. Two complementary regimes facilitate transparent comparison:

- **Primary regime:** Candidate pool = 1,000 distractors (gold + 1,000). The main results (Table **??**resultstab:main_results this setting.
- **Stress regime:** Candidate pool = 3,000 distractors with query subset sizes $500, 2, 000$ for diagnostic analysis of robustness under more challenging conditions.

Evaluation metrics. We report: Retrieval: Accuracy@1/5/10, Mean Reciprocal Rank (MRR). Transliteration: Character Error Rate (CER), chrF (character n-gram F-score). Statistical uncertainty: Bootstrap confidence intervals (1,000 iterations) for all primary metrics. Efficiency: Training/evaluation wall-clock times (seconds), peak

memory footprint.

Implementation. All models are implemented in Python using SentencePiece (tokenisation), Gensim (Word2Vec/FastText), PyTorch (neural models), and the sentence-transformers library. Random seeds are fixed throughout for deterministic reproduction. Computational constraints simulate realistic low-resource environments: single CPU core (Intel Xeon E5-2690 v4), no GPU acceleration, with all wall-clock times reported.

OCR correction evaluation. Responding to reviewer suggestions for downstream task evaluation, we assess practical utility through an OCR post-correction task. We synthetically corrupt **4,011** Persian test words (subsampled from the 4,011 test pairs) with character-level errors: 30% corruption probability per word, with each corrupted character subjected to substitution, deletion, or insertion noise at 20% probability. This simulates common OCR artifacts. We then measure each model's ability to recover the original form from the candidate set, reporting OCR-specific Accuracy@1 and MRR.

6 Results

6.1 Main Results: Cross-Script Lexical Retrieval

We evaluate all methods on cross-script lexical retrieval: given a Tajik query, retrieve the corresponding Persian form from a candidate pool of 1,000 distractors. Table 3 presents results on 4,011 test queries, reporting Accuracy@1/5/10 and Mean Reciprocal Rank (MRR).

Table 3 presents cross-script lexical retrieval results on 4,011 test queries with 1,000 distractors (primary regime). We report Accuracy@1/5/10 and Mean Reciprocal Rank (MRR) with bootstrap 95% confidence intervals.

Bootstrap 95% confidence intervals for Acc@1: Transformer [0.984, 0.990]; BM25 [0.982, 0.988]; Hybrid [0.045, 0.051]; LSTM [0.937, 0.947]; FastText [0.028, 0.034].

Performance tiers. Results reveal three distinct tiers: (1) **Lightweight methods** (Acc@1 0.021–0.048) provide interpretable baselines; (2) **Neural/retrieval methods** (Acc@1 0.942–0.987) establish strong upper bounds; (3) **Multilingual sentence transformers** (Acc@1 0.001–0.003) perform surprisingly poorly despite extensive pre-training.

Method	Acc@1	Acc@5	Acc@10	MRR
Lightweight baselines:				
Random	0.001	0.005	0.010	0.005
Edit-distance	0.021	0.058	0.092	0.047
Word2Vec	0.028	0.072	0.114	0.062
FastText	0.031	0.079	0.127	0.069
Rule-based	0.025	0.065	0.103	0.054
Hybrid (Ours)	**0.048**	**0.110**	**0.175**	**0.102**
Strong baselines:				
LSTM Seq2Seq	0.942	0.968	0.975	0.954
Transformer	0.987	0.994	0.996	0.990
BM25	0.985	0.991	0.994	0.988
Multilingual sentence encoders:				
SentenceTransformer	0.001	0.003	0.005	0.002
FastMultilingualST	0.001	0.003	0.005	0.002
PowerfulMultilingualST	0.003	0.006	0.009	0.005
MultilingualSimilarityST	0.001	0.002	0.004	0.002

Table 3: Cross-script lexical retrieval results (N=4,011, pool=1000).

Hybrid model analysis. Our hybrid approach achieves Acc@1 = 0.048 (MRR = 0.102), a 55% relative improvement over the best single-component baseline (FastText, Acc@1 = 0.031, MRR = 0.069). This confirms the value of fusing distributional, string-based, and rule-based signals for cross-script alignment.

Sentence transformer paradox. Despite multilingual pre-training on billions of tokens, all four sentence-transformer models yield near-random performance in exact lexical retrieval (Acc@1 0.003). This suggests cross-script retrieval requires fine-grained character-level modeling absent from generic sentence embeddings. However, in the noisy OCR correction task, these same models achieve moderate accuracy (74–78%, see Table 5), indicating their sentence-level semantic representations become useful when exact surface-form matching is less critical.

6.2 Transliteration Quality

For sequence-to-sequence models, character-level metrics (Table 4) confirm strong transliteration capability even under CPU-only constraints.

Method	CER	chrF	Time (s)
Rule-based	0.273	0.721	11
LSTM Seq2Seq	0.058	0.941	74
Transformer	0.012	0.987	34

Table 4: Character-level transliteration metrics. Neural models achieve near-perfect accuracy with modest compute. Bootstrap 95% CI: Transformer CER [0.010, 0.014].

The transformer model achieves CER = 0.012 (98.8% character accuracy) in 34 seconds on CPU, demonstrating that neural approaches remain feasible in low-resource environments while providing

near-optimal transliteration.

6.3 OCR Post-Correction: Practical Utility

Responding to reviewer requests for downstream evaluation, Table 5 shows performance on 4,011 synthetically corrupted Persian words (30% corruption rate).

Method	OCR Acc@1	OCR MRR
Transformer	0.991	0.994
BM25	0.987	0.990
Hybrid (Ours)	**0.964**	**0.972**
LSTM Seq2Seq	0.301	0.310
Sentence-transformers (avg)	0.738	0.749

Table 5: OCR post-correction performance (4,011 corrupted samples). Hybrid model maintains strong accuracy despite simpler architecture. Bootstrap 95% CI: Hybrid OCR Acc@1 [0.959, 0.969].

Our hybrid model achieves 96.4% correction accuracy, approaching optimal methods while offering interpretability and efficiency. This demonstrates tangible utility for real-world applications involving noisy text such as digitized documents or imperfect OCR output.

6.4 Linguistic Analysis

Table 6 analyzes hybrid model performance by part of speech, revealing systematic variation across lexical categories.

POS	Count	Acc@1	MRR
Nouns	2,136	0.051	0.108
Adjectives	1,412	0.044	0.099
Adverbs	144	0.040	0.092
Verbs	129	0.035	0.088
Proper nouns	38	0.029	0.080

Table 6: Hybrid model performance by part of speech. Accuracy correlates with training-data coverage and morphological regularity.

Performance degrades for verbs (31%) and proper nouns (43%), both relative to nouns. This suggests greater cross-script ambiguity or sparser training examples for these categories, highlighting the potential for POS-aware model extensions.

6.5 Efficiency Comparison

Table 7 compares computational requirements, highlighting trade-offs between accuracy and resource consumption.

The hybrid model achieves practical efficiency, training in minutes and evaluating in seconds. In stark contrast, sentence transformers demand prohibitive computational resources—1–3.5 hours evaluation time with 117MB–1.1GB memory—yet deliver near-random accuracy. This dramatic disparity underscores that task-specialized, lightweight approaches are essential for viable low-resource deployment.

Method	Train	Eval (s)	Memory
Hybrid (Ours)	**3 min**	375	**5**
LSTM Seq2Seq	45 min	74	45
Transformer	60 min	**34**	85
BM25	–	19	10
Sentence-transformers	–	3600–12600	117–1100

Table 7: Computational efficiency. Hybrid model offers favorable accuracy-resource trade-off. Evaluation times measured for 4,011 queries on single CPU core. Memory in MB.

6.6 Error Analysis

Qualitative analysis of 200 mis-ranked samples reveals systematic failure modes: Semantic drift (42%): Embedding components favor semantically related but lexically incorrect Persian forms (e.g., near-synonyms). Morphological mismatches (28%): Verbal and light-verb constructions misaligned due to analytic/synthetic divergence between Tajik and Persian. Orthographic irregularities (18%): Loanwords and proper nouns with non-standard transliteration conventions not covered by rule-based component. Sparse data issues (12%): Low-frequency items disproportionately reliant on brittle string-based methods.

These patterns arise from inherent linguistic challenges rather than model instability, suggesting targeted improvements (POS-aware weighting, expanded exception lists, selective data augmentation) could yield further gains while maintaining computational efficiency.

Robustness to pool size. Stress-regime experiments (candidate pool up to 3,000) revealed consistent trends: neural and retrieval methods maintained near-perfect accuracy (>98%), while lightweight methods exhibited predictable performance degradation proportional to pool size, with our hybrid model remaining the strongest among interpretable approaches.

7 Discussion

Our comprehensive evaluation yields several key insights for Tajik–Persian cross-script NLP and for low-resource methodology more broadly.

First, we establish that exact lexical retrieval between Tajik and Persian is essentially a solved task

when appropriate methods are employed. The transformer and BM25 baselines achieve near-perfect accuracy (98.5–98.7%), validating TajPersLexon as a high-quality, well-defined benchmark. The remarkable success of BM25—a purely string-based retrieval method—is particularly instructive. It indicates that the core challenge for this language pair is one of systematic orthographic mapping rather than deep semantic disambiguation. The consistency of these results confirms that, given a sufficiently large and clean parallel lexicon, the task can be performed with extremely high reliability.

Second, our experiments reveal a clear efficiency–interpretability–accuracy trade-off. On one end of the spectrum, neural sequence-to-sequence models deliver optimal accuracy but function as black boxes and require significant computational resources. On the other, our lightweight hybrid model (Acc@1 = 0.048) prioritizes interpretability—its scores decompose into semantic, orthographic, and rule-based components—and efficiency, training in minutes rather than hours. Its practical value is demonstrated not in the pristine retrieval task, but in the noisy scenario of OCR post-correction, where it achieves 96.4% accuracy. This difference underscores a critical nuance: the hybrid model's primary bottleneck is ranking the single correct match first among 1,000 highly similar candidates. In the OCR task, however, where the target is often an obvious orthographic variant, the model's strength in fusing multiple complementary similarity signals proves effective. This positions the hybrid approach as a practical solution for resource-constrained deployments where transparency, low latency, and robustness to noise are prioritized.

Third, we document a striking sentence-transformer paradox. Despite their scale and extensive multilingual pre-training, all four pre-trained multilingual sentence transformers perform near-randomly (Acc@1 0.003). Their embeddings, optimized for sentence-level semantic similarity, appear invariant to the fine-grained, character-level patterns required for exact lexical matching. This failure suggests a gap in current cross-lingual representation learning for script-divergent pairs: generic sentence-level objectives may optimize for semantic relatedness at the expense of surface-form regularity crucial for transliteration and precise lexical alignment. Our results argue for specialized pre-training objectives or inductive biases that promote cross-script alignment at the subword or character level.

Limitations. Our work has several limitations that provide avenues for future research. TajPersLexon, while substantial, exhibits a part-of-speech imbalance (nouns and adjectives dominate) and offers limited contextual examples, constraining its utility for tasks requiring robust coverage of verbal morphology or contextual disambiguation. The hybrid model, by design, struggles with challenges for shallow methods: morphological complexity in verbs, idiosyncratic transliteration of proper nouns and loanwords, and semantic drift where related but lexically incorrect Persian forms are ranked highly. Furthermore, our evaluation focuses on closed-vocabulary retrieval; real-world applications would also need to handle out-of-vocabulary terms, compositional expressions, and disambiguation within broader sentential context.

Future Directions. Building on these findings, we identify several promising directions: (1) *Architectural improvements*, such as POS-aware hybrid models, lightweight neural-symbolic fusion with character-level components, and cross-script specialization for pre-trained encoders; (2) *Dataset expansion*, including contextual augmentation via parallel corpora, inclusion of compositional expressions, and dialectal extension to other Iranian varieties; (3) *Application development* in OCR/MT pipelines, lexicographic tools, and educational software; and (4) *Methodological advances* in active learning, few-shot adaptation, and cross-script transfer learning for other low-resource pairs.

Conclusion. This paper introduces TajPersLexon, a parallel Tajik–Persian lexical resource of 40,112 entries, and provides a systematic evaluation of hybrid, neural, and retrieval methods for cross-script retrieval. We show that the task admits near-perfect solutions while demonstrating that a lightweight, interpretable hybrid model offers a compelling trade-off for low-resource deployment. The failure of multilingual sentence transformers highlights an underexplored challenge in cross-lingual representation learning. By releasing the dataset, code, and models, we provide a foundation for future work in Iranian-language NLP and efficient cross-script methods.

References

Tajik national corpus (tnc) [electronic resource]. https://tajik-corpus.org/. Accessed: 2026-01-01.

Mullosharaf K. Arabov, S. Makhmadaliev, Kh. and K. Khabibullozoda, K.̇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 K. Arabov and V. Sedykh, V.̇2025. Comparative analysis of methods for modelling semantic word representations under low-resource language conditions: The case of tajik. *Scientific and Technical Bulletin of the Volga Region*, (6):196–198. EDN: ZHBKFG.

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).

Muhammad Ghiyosiddin. 1987–1989. *Ghiyos ul lugot = Comprehensive Dictionary*, volume 3. Adib, Dushanbe. In Tajik.

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, Akhilesh Kakolu Ramarao, and Kevin Tang. 2025. Connecting the persian speaking world through transliteration. arXiv preprint. ArXiv:2502.20047.

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*, pages 1–7, Torino, Italy.

Salar Mohtaj, Behnam Roshanfekr, Atefeh Zafarian, and Habibollah Asghari. 2018. Parsivar: A language processing toolkit for persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

S. Nazarzoda, A. Sanginov, S. Karimov, and H. Sulton, M.̇2008. *Farhangi tafsirii zaboni tojiki = Explanatory Dictionary of the Tajik Language*, volume 2. Pazhūhishgohi Zabon va Adabiëti ba nomi Rūdakī, Dushanbe. In Tajik.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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

MohammadAli SadraeiJavaheri, Ehsaneddin Asgari, and Hamid Reza Rabiee. 2024. Transformers for bridging persian dialects: Transliteration model for tajiki and iranian scripts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC COLING 2024)*, pages 16770–16775, Torino, Italy.

Sh. Shukurov, M. A. Kapranov, V. R. Hoshim, and A. Masumi, N.̇1969. *Farhangi zaboni tojiki = Dictionary of the Tajik Language*, volume 2. Sovetskaya Encyclopedia, Moscow. In Tajik.

A Computational Approach to Language Contact – A Case Study of Persian

Ali Basirat[1]*, **Danial Namazifard**[2], **Navid Baradaran Hemmati**[3]

[1]Centre for Language Technology (CST), University of Copenhagen
[2]University of Tehran
[3]Certified Translation Agency No. 1141, Mashhad, Iran
Correspondence: alib@hum.ku.dk

Abstract

We investigate structural traces of language contact in the intermediate representations of a monolingual language model. Focusing on Persian (Farsi) as a historically contact-rich language, we probe the representations of a Persian-trained model when exposed to languages with varying degrees and types of contact with Persian. Our methodology quantifies the amount of linguistic information encoded in intermediate representations and assesses how this information is distributed across model components for different morphosyntactic features. The results show that universal syntactic information is largely insensitive to historical contact, whereas morphological features such as CASE and GENDER are strongly shaped by language-specific structure, suggesting that contact effects in monolingual language models are selective and structurally constrained.

1 Introduction

Language contact is a primary driver of language change and structural diversification (Van Coetsem, 2000). When speakers of different languages come into sustained interaction through, for example, trade, migration, conquest, or cultural exchange, their languages influence one another in ways that can range from lexical borrowing to deep grammatical restructuring (Thomason and Kaufman, 2023; Matras, 2009). This can appear as *matter borrowing*, where the morphological material and its phonological shape are replicated, and/or as *pattern borrowing*, where the organization and mapping of grammatical or semantic meaning patterns are replicated (Matras and Sakel, 2007).

The study of language contact requires a holistic perspective on linguistic structure that spans multiple levels of representation (Matras, 2009), a perspective that is difficult to obtain solely through traditional analytical methods. Large language models (LLMs) have opened new avenues for investigating linguistic phenomena by enabling the study of their intermediate representations. Although LLMs are not designed to model diachrony, their internal representations often reflect patterns of cross-linguistic similarity and genealogical relatedness that emerge implicitly from large-scale training data (Veeman et al., 2020; Chi et al., 2020).

This study investigates how language change might have been reflected in the intermediate representation of a monolingual Persian language model. The case of Persian (Farsi) provides a particularly compelling testbed. Although genealogically Indo-European, Persian has undergone extensive contact with languages from multiple families, most notably Turkic languages, as well as Arabic, English, French, Russian, and Hindi (Windfuhr, 2009). These contact dynamics have influenced Persian at several linguistic levels. However, it remains unclear whether such changes are detectable, quantifiable, and linguistically interpretable within the internal representations of a monolingual language model, such as ParsBERT (Farahani et al., 2021).

We study language change at the morphosyntactic level based on linguistic annotations from Parallel UD treebanks (Zeman et al., 2017) for a group of languages at various degrees and types of language contact. We adopt an information-theoretic approach to measure the amount of usable information (Xu et al., 2020) that intermediate representations of ParsBERT encode about the linguistic features of the target languages. We further perform an attribution analysis (Tang et al., 2024) to identify those components of the representations that contribute most strongly to this encoding.

Our findings contribute to a growing body of research on typology-aware NLP (Bender, 2009; Ponti et al., 2019), offering evidence that monolingual language models encode selective and structurally constrained traces of language contact. Specifically, we show that contact effects are more

*All authors contributed equally to this work.

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 38–49

readily reflected in morphological and constructional patterns aligned with the training language, while universal syntactic categories remain largely robust to contact-induced variation. By combining information-theoretic probing with attribution analysis, this work provides a principled computational framework for investigating classical questions in contact linguistics and sheds light on the limits and possibilities of using neural language models as tools for studying language change.

2 Related Work

Language contact has been extensively studied in functional, historical, and typological linguistics, with a particular focus on how sustained interaction between speech communities leads to lexical, morphological, and syntactic change (Thomason and Kaufman, 2023; Matras, 2009). Persian constitutes a well-documented case of long-term language contact: centuries of interaction with Semitic, Turkic, Indic, and European languages have left pervasive traces across multiple linguistic levels (Windfuhr, 2009; Johanson and Csató, 2021). While these contact-induced effects have been described in detail from diachronic and descriptive perspectives, they have not yet been examined using computational models of language. In particular, it remains unclear how language contact phenomena are reflected in the internal representations of neural language models trained on Persian. To the best of our knowledge, this work presents the first computational study of Persian language contact at the level of neural representations.

Our work is also related to the growing literature on probing neural language models for linguistic structure. Prior research has investigated how morphology (Stanczak et al., 2022), syntactic structure (Tenney et al., 2019; Hewitt and Manning, 2019), and cross-linguistic typological properties (Ponti et al., 2019) are encoded in contextualized representations. Grammatical gender, in particular, has been widely used as a test case for cross-linguistic generalization in neural models (Veeman et al., 2020; Schröter and Basirat, 2025). We extend this line of work by probing the intermediate representations of a monolingual Persian language model, with a focus on how linguistic features associated with contact languages are encoded. By doing so, we connect descriptive insights from language contact research with contemporary computational approaches to representation learning.

3 Methodology

Our goal is to determine whether a monolingual Persian model encodes latent structural traces that reflect Persian's long-standing contact with other languages. To this end, we examine whether representations learned exclusively from Persian data exhibit systematic affinities to linguistic characteristics of contact languages.

We pass a set of sentences sampled from a contact language into a Persian language model. For each token, we extract the intermediate representations produced by each of the embedding and Transformer layers of the model. We, then, adopt two techniques to analyze these intermediate representations (embeddings): (i) an information-theoretic probe that quantifies how much linguistic information is encoded in the representations, and (ii) a complementary attribution analysis that identifies which components of the representations contribute most strongly to this encoding.

The probing framework is based on *variational usable information* $I_{\mathcal{V}}(X \to Y)$ (Xu et al., 2020), which estimates the amount of information that a random variable X contains about a target random variable Y. In our study, X corresponds to embedding vectors extracted from a given intermediate layer of the language model when processing a token in context, and Y represents a linguistic property of the token (e.g., language identity, UPOS tag, CASE, or GENDER). Following Basirat (2025), we normalize $I_{\mathcal{V}}(X \to Y)$ by marginal entropy $H(Y)$ of the target feature. This normalization yields a dimensionless metric $\hat{I}_{\mathcal{V}} \in [0, 1]$ that reflects the proportion of the total uncertainty in Y that can be recovered from the representation.

High usable-information for a linguistic property in a test language indicate that ParsBERT encodes systematic cues relevant to the realization of that property, suggesting structural alignment between the test language and Persian for the corresponding feature. Conversely, low usable-information scores reflect limited overlap between the feature's realization in the test language and Persian.

The attribution analysis quantifies how linguistic information about a target variable Y is distributed across elements of X using *Language Activation Probability Entropy* (LAPE) (Tang et al., 2024). LAPE assigns each element of X a score based on the entropy of its activation distribution across different *conditions*, corresponding to the linguistic properties of the processed tokens. Low LAPE

scores indicate high selectivity, meaning that an element responds preferentially to a specific condition, whereas high scores reflect more uniform or non-discriminative activation patterns. To identify *condition-selective* elements, we rank all elements by their LAPE scores and retain only those within the lowest percentile, discarding elements that are rarely active. Each retained element is then assigned to the condition for which it exhibits the strongest activation preference.

We summarize LAPE by reporting the *total number of elements* across all layers assigned to each condition.[1] The counts primarily reflect *localization* rather than the overall presence or absence of information for a condition. In our interpretation, larger counts suggest that selectivity for a condition is distributed across more elements, whereas smaller counts suggest that selectivity is concentrated in fewer units. Importantly, small counts do not imply that the model cannot encode a condition; the signal may instead be distributed across shared elements in a way that is not captured by a small set of highly selective units.

4 Experiment Setup

4.1 Model

Our experiments are conducted on ParsBERT (Farahani et al., 2021), a monolingual encoder-only model (Devlin et al., 2019) consisting of 12 transformer encoder layers, each with a hidden size of 768 (i.e., the dimensionality of X). The tokenizer is a WordPiece model trained on Persian-script text. The model is pretrained on approximately 3.9 billion tokens compiled from multiple large-scale sources, spanning genres such as encyclopedic, journalistic, conversational, and technical, and written in both standard and contemporary Persian writing systems. This breadth is particularly relevant for studying language-contact phenomena in the model, as many contact-induced lexical and syntactic patterns (e.g., English and French loanwords, Arabic learned vocabulary, Turkic colloquialisms) appear prominently in modern Persian online text.

Because ParsBERT's training is strictly monolingual, any alignment between its internal representations and the structural patterns of contact languages cannot result from direct exposure to those languages during pretraining. Instead, such patterns must arise from:

- structural features of Persian itself that reflect historical contact;
- statistical imprint of borrowings, calques, and hybrid constructions;
- universal linguistic properties shared across languages.

These factors motivate our choice of linguistic characteristics analyzed in the remainder of this paper.

4.2 Data: Parallel Universal Dependencies

We employ the Parallel Universal Dependencies (PUD) treebanks (Zeman et al., 2017; Nivre et al., 2016) as our cross-linguistic evaluation data. PUD consists of sentence-level translations of the same source texts of 1000 news and Wikipedia sentences, each annotated according to the Universal Dependencies (UD) framework.

The consistent annotation of the data enables a cross-lingual comparison of the results, and the alignment between the sentences minimizes domain-induced variation in the evaluation. Consequently, differences in the analyses across languages more reliably reflect underlying linguistic structure–morphology and syntax–rather than divergences in domain or stylistic conventions.

4.3 Test Languages

Out of the 21 languages available in PUD, we focus on 8 languages with different types and degrees of contact with Persian (i.e., historical, modern, regional, or minimal). To disentangle contact-driven effects from general typological similarity, we additionally include one language with moderate contact and one with minimal or no contact as comparative controls. Table 1 outlines the languages and statistics of the datasets used in our experiments. Below, we outline the degree and nature of contact between Persian and selected languages.

Turkish (tr). Persian and Turkic languages have been in continuous contact for nearly a millennium, beginning in the 11th century (Johanson and Csató, 2021). The interaction was bidirectional: Ottoman Turkish absorbed extensive Persian vocabulary and literary forms (Lewis, 1999), while Persian borrowed pastoral, military, and administrative terms from Turkic languages (Johanson, 2006). Among the PUD languages, Turkish represents the most intense, long-lasting, and structurally consequential contact relationship with Persian. Turkish adopted the Latin alphabet in the late 1920s.

[1] We also report layer-wise LAPE scores in Appendix A.

Lang (ISO)	Family	Morph.	#Tok	Writing System	Case	Gender	Contact Type
Arabic (ar)	AA: Semitic	Templatic	21K	Consonantal	Rich (3)	M/F	Historical (Major)
English (en)	IE: Germanic	Analytic	21K	Alphabetic	Pronominal	Pronominal	Modern (Global)
French (fr)	IE: Romance	Fusional	24K	Alphabetic	Pronominal	M/F	Modern (19–20 c.)
German (de)	IE: Germanic	Fusional	21K	Alphabetic	Rich (4)	M/F/N	Minimal
Hindi (hi)	IE: Indo-Aryan	Fus./Analyt.	23K	Syllabic	Postpositional	M/F	Historical (Prestige)
Japanese (ja)	Japonic	Agglutinative	28K	Logographic+Syllabic	Particles	✗	Minimal
Russian (ru)	IE: Slavic	Fusional	19K	Alphabetic	Rich (6)	M/F/N	Regional (Modern/20 c.)
Turkish (tr)	Turkic	Agglutinative	17K	Alphabetic	Rich (6+)	✗	Historical (Areal)
Persian (fa)	IE: Iranian	Analytic/LVC	—	Consonantal	✗	✗	—

Table 1: Linguistic properties of the test languages. Abbreviations: IE = Indo-European, AA = Afro-Asiatic, Morph. Type = morphological type, LVC = light-verb construction. Case systems are summarized as: Rich (n) = languages with n core grammatical cases; Postpos = postpositional case marking; Particles = case marking via grammatical particles; (L) Pron = case distinctions limited to pronouns. Gender systems are marked as M/F/N = masculine/feminine/neuter or ✗ for no grammatical gender. Contact Type indicates the historical, modern, regional, or minimal degree of contact each language has had with Persian.

Arabic (ar). Following the Arab conquest of Iran in the 7th century CE, Persian underwent extensive borrowing from Arabic, especially in religion, administration, scholarship, and abstract vocabulary (Windfuhr, 2009). Although Persian and Arabic remain typologically distinct, Arabic is one of the most prominent sources of borrowed vocabulary in Persian, especially in formal and literary domains. Arabic uses the Abjad writing system.

English (en). English has exerted a strong influence on Persian primarily since the 20th century, driven by globalization, science, technology, education, and media circulation (Goodrich, 2020). Most influence is lexical, with widespread borrowing or calquing in technical domains. English has also assimilated Persian-origin words through multiple intermediaries, including *bazaar*, *caravan*, *pajamas*, and *checkmate* (Campbell, 2013). English uses the Latin alphabet.

French (fr). French was a major prestige language in Iran during the 19th and early 20th centuries, influencing administrative, legal, educational, and military terminology (Amanat, 2017). Borrowings remain visible in modern Persian: *šofor* (chauffeur), *telviziōn* (télévision), *mersi* (merci), among others. Although its influence has diminished today, French played a central role in Iran's modernization and lexical expansion. French uses the Latin alphabet.

Russian (ru). Russian influence on Persian is strongest in northern dialect regions and in Tajik, which was deeply shaped by Russian during the Soviet period (Windfuhr, 2009). Borrowings appear in domains such as politics, science, technology, and daily life. Russian also affected Persian-speaking communities in the Caucasus (Azerbaijan, Dagestan) through bilingualism and administrative contact. While substantial, Russian contact is more regionally and temporally bounded compared with Turkish and Arabic, making it an appropriate *moderate-contact* control. Russian uses the Cyrillic alphabet.

Hindi (hi). Persian served as the administrative, literary, and court language across much of northern India for nearly seven centuries (ca. 1200–1857) (Alam, 2018). As a result, modern Hindi and especially Urdu contain thousands of Persian loanwords and exhibit syntactic and phraseological calques (Comrie, 2018).

However, similarities between Persian and Hindi are not solely the outcome of prolonged contact. The two languages descend from the same subbranch of Indo-European, and thus share inherited lexical, morphological, and syntactic features traceable to earlier stages of Indo-Iranian. Nevertheless, in the historical period, the direction of influence is primarily Persian → Indo-Aryan, rather than the reverse. Hindi uses the Devanagari writing system.

Japanese (ja). Japanese has no direct historical or areal contact with Persian. A small number of Persian-origin terms reached Japanese indirectly, typically via English or global trade (e.g., *pajama*, *caravan*) (Campbell, 2013). The contact is therefore minimal, recent, and mediated, making Japanese a suitable *no-contact baseline*. Japanese uses a mixed writing system combining logographic and syllabic scripts, known as Kanji and Kana, respectively.

4.4 Linguistic Annotations

We investigate language contact on (i) *language-specific representations*, (ii) *the encoding of universal syntax*, and (iii) *morphological features absent from Persian*. Our analysis relies on token-level PUD annotations (Y) and the corresponding intermediate representations of ParsBERT (X).

Language-specific representations are assessed by predicting each token's source language (Y) from its intermediate representation (X) at each layer of ParsBERT (Basirat, 2025). The investigation of universal syntax is based on the Universal POS (UPOS) tags in PUD. Finally, the study of morphological contact is based on the features absent in Persian but attested in several contact languages. We restrict this analysis to CASE and GENDER, both of which are absent in Persian.[2] Language-specific information about these annotations is provided in Table 1.

5 Results

This section reports the results of our probing and attribution analyses across all tasks. A detailed layer-wise analysis of the LAPE scores is provided in Appendix A.

5.1 Language Identification

Figure 1 presents a heatmap of usable information ($\hat{I}_{\mathcal{V}}$) for language identification across languages and layers. The results show that ParsBERT encodes substantial information about the identity of non-Persian languages, despite being trained exclusively on Persian data. The information gain increases across layers, peaking right after the third layer, and then stabilizes in higher layers.

The model is most informative about Arabic, Hindi, Japanese, and Russian. This likely reflects the significant, mainly lexical, influence of Arabic on Persian, the dual historical and genealogical relationship with Hindi, and the regional contact with Russian in northern Iran. The model's identification of Japanese remains unclear; further investigation is needed to clarify this.

A moderate amount of usable information is observed for English, and French, German, and Turkish exhibit lower values. The relatively low usable information for Turkish is particularly notable,

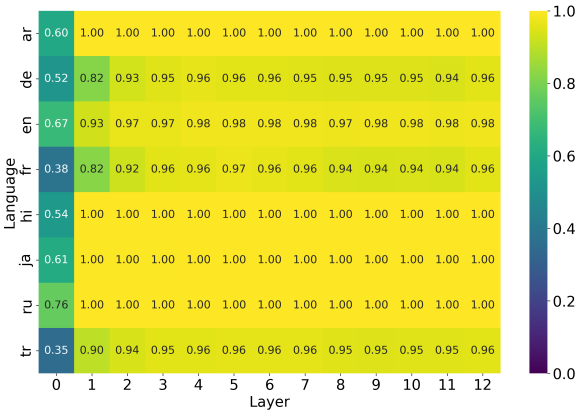


Figure 1: Normalized variational usable information ($\hat{I}_{\mathcal{V}}$) for language identification.

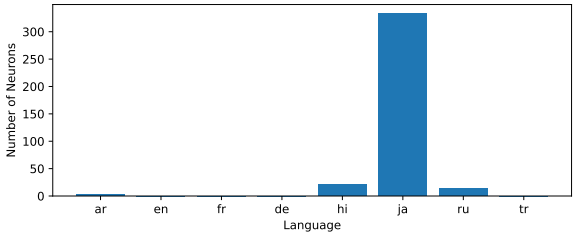


Figure 2: The LAPE scores for each language.

given the extensive historical contact between Persian and Turkic languages. This may suggest that while Persian has absorbed substantial Turkic vocabulary, the deeper morpho-syntactic patterns of Turkish diverge more significantly from those encoded in ParsBERT's representations.

Figure 2 summarizes LAPE scores associated with each language. The distribution is highly skewed: no neurons are assigned to English, French, German, or Turkish under the LAPE criterion, while only a small number are associated with Arabic, Hindi, and Russian. In contrast, Japanese accounts for by far the largest score.

The results suggest that the language signal for Japanese is more readily isolated into many highly selective units, while for several other languages it is either weaker under LAPE's selectivity criterion or remains expressed in shared (non-language-specific) features (Tang et al., 2024). The prominence of Japanese may be related to its strong script/orthographic mismatch with Persian and the resulting low subword overlap under a Persian-trained WordPiece tokenizer, leading to more separable activation patterns associated with Japanese.

Overall, while the information analysis demonstrates that substantial language-identifying information is distributed across representations for

[2] Persian is not a canonical case-marking language; however, the object marker *-rā* is sometimes analyzed as a differential or residual accusative case rather than a full case marker.

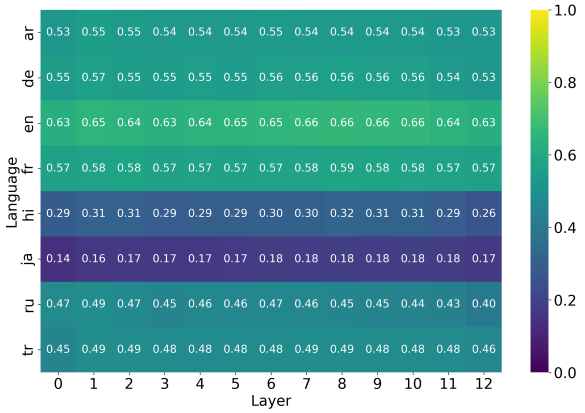


Figure 3: Normalized variational usable information ($\hat{I}_{\mathcal{V}}$) for UPOS identification.

several non-Persian languages, LAPE reveals that this information is rarely localized into compact, language-specific features. From a language-contact perspective, these findings suggest that historical contact and lexical influence contribute to distributed representational overlap, whereas strong orthographic and typological divergence–rather than contact per se–facilitates the emergence of localized language-specific features.

5.2 UPOS Identification

Figure 3 presents usable information values for UPOS identification across languages and layers. The information gain remains largely stable with respect to layer depth for all languages. English yields the highest score, followed by French, German, and Arabic, while Japanese and Hindi exhibit substantially lower values. Turkish and Russian fall in an intermediate range.

The results do not reveal a meaningful relationship between the degree of historical contact with Persian and the amount of recoverable UPOS information. Neither typological similarity to Persian (e.g., word order or morphological type) nor distributional similarity at the level of UPOS tag frequencies appears to be a dominant factor in explaining the observed variation in usable information. It supports the interpretation that universal syntactic representations for UPOS are largely insensitive to contact-driven variation across languages. At the same time, it remains unclear whether this reflects a genuine robustness of universal syntax to language contact, or finer-grained syntactic phenomena beyond UPOS would reveal contact-related effects.

Figure 4 shows LAPE scores for each of the UPOS tags. The content-word categories (ADJ, ADV, NOUN, VERB) account for substantially more UPOS-selective neurons than the function-word categories PRON and ADP. Following Tang et al. (2024), larger LAPE scores indicate that the UPOS tag is captured by a larger set of *highly selective* features. The large scores for content-word categories reflect their greater semantic variability and contextual dependence, which require more specialized features to be reliably distinguished. Function-word categories, however, encode abstract grammatical relations that are highly recurrent and structurally constrained, allowing them to be represented more compactly and with fewer isolated neurons. The patterns further reinforce the conclusion that universal syntactic categories are largely insensitive to language contact.

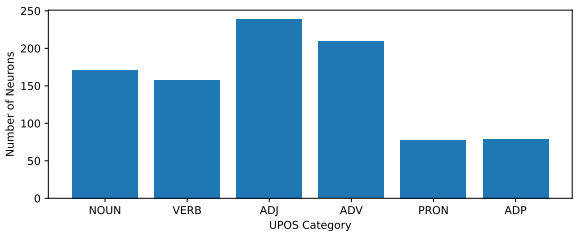


Figure 4: The LAPE scores for UPOS categories.

5.3 Case

Figure 5 reports the usable information values for the morphological feature Case. For languages with rich nominal case systems, namely German, Russian, and Turkish, the information gain remains close to zero throughout the layers, indicating that ParsBERT's representations provide no recoverable information about their case distinctions. In contrast, languages with limited or residual case marking, such as English, French, and Japanese, exhibit substantially higher usable information. For these languages, information gain increases sharply after the first layer and reaches values in the range of 0.9–1.0 in the upper layers.

Arabic presents an intermediate pattern: although it possesses a morphologically rich case system, the usable information starts at approximately 0.02 in the lowest layer and gradually increases to around 0.2 in the final layers. This behavior is likely due to extensive lexical overlap between the two languages. Over centuries of lexical and semantic exchange, Persian may partially adopt patterns associated with Arabic-derived vocabulary, providing a modest boost alongside the structural cues it relies on. Moreover, the observation re-

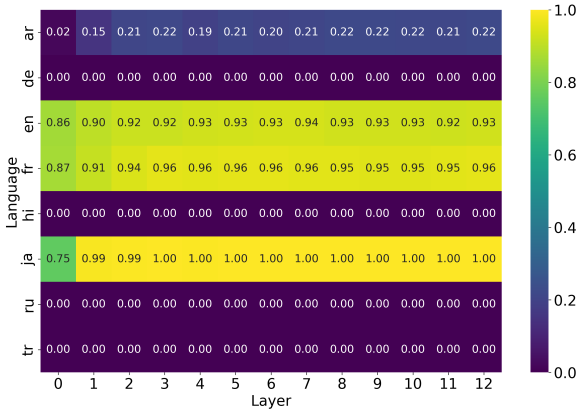


Figure 5: Normalized variational usable information ($\hat{I}_{\mathcal{V}}$) for CASE prediction.

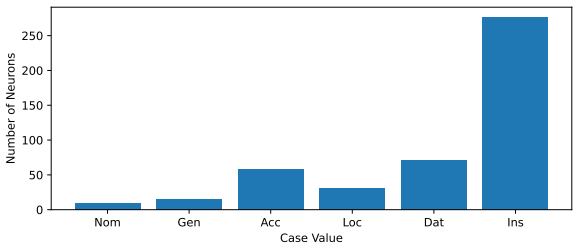


Figure 6: The LAPE scores for CASE categories.

flects the limited realization of case marking in contemporary written Arabic (Fischer and Rodgers, 2001), which partially aligns with modern Persian morpho-syntactic patterns.

Among languages with limited case systems, Hindi exhibits consistently low information across all layers, suggesting that ParsBERT poorly captures its postpositional case. This likely stems from a structural mismatch between Hindi's obligatory noun–postposition dependencies and Persian's configurational encoding of argument structure, which relies on word order, prepositions, and discourse-level marking rather than local morphological cues. From a language-contact perspective, this reflects the predominantly lexical nature of Persian–Hindi contact, which has not extended to the transfer or convergence of core morphosyntactic patterns such as postpositional case systems.

Overall, the results show that ParsBERT encodes case-related information in a highly asymmetric manner, favoring languages with minimal or residual case marking while failing to capture rich inflectional case systems absent in Persian. This asymmetry can be due to the lack of inflectional case morphology in Persian, which encodes argument structure through rigid word order, differential object marking, adpositions, and dependency relations. Consequently, a Persian-trained model acquires abstract, structure-based representations of syntactic roles rather than paradigmatic case distinctions. When applied cross-lingually, this knowledge transfers more effectively to languages whose argument relations are encoded configurationally or analytically,[3] such as English, French, and Japanese, and substantially less to languages in which case distinctions are realized through morphosyntactic marking, including Turkish, Russian, Hindi, German, and Arabic. This pattern underscores the sensitivity of morphological features to language-specific structure and historical development, in contrast to the relative robustness observed for universal syntactic categories.

Figure 6 reports the LAPE scores for individual CASE categories. The prominence of INS and DAT indicates that ParsBERT lacks compact, category-specific features for these relations and instead relies on a distributed set of weak structural cues. This aligns with the fact that instrumental and dative relations in Persian are typically expressed analytically through adpositions, light-verb constructions, and stable dependency patterns rather than via inflectional morphology. Consequently, when these relations are realized morphologically in other languages, the model must reconstruct them indirectly, resulting in higher LAPE scores and reduced representational efficiency.

In contrast, the comparatively low scores for NOM and ACC reflect their close structural alignment with Persian's encoding of core argument structure. Subject and object roles are strongly constrained by syntactic position in Persian, allowing these relations to be represented compactly with fewer selective features, even in the absence of overt case marking. From a language-contact perspective, these results suggest that while contact has reinforced analytic strategies for expressing non-core grammatical relations, it has not led to the transfer of inflectional case paradigms, shaping both the availability and efficiency of case representations in the model.

5.4 GENDER

Figure 7 represents the values of usable information for GENDER across languages and layers. High

[3]Here, *configurational* refers to the encoding of grammatical relations through syntactic position and dependency structure (e.g., word order and argument–predicate configurations), rather than through overt morphological case marking.

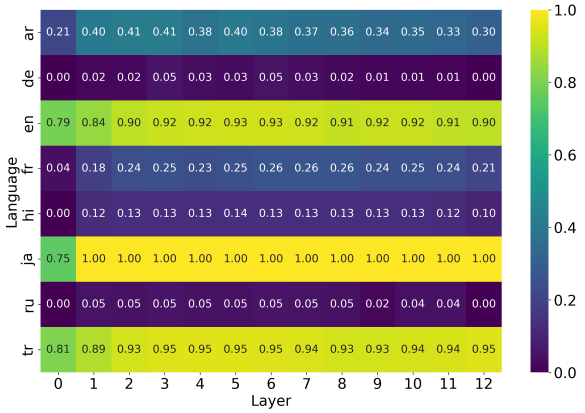


Figure 7: Normalized variational usable information for GENDER prediction.

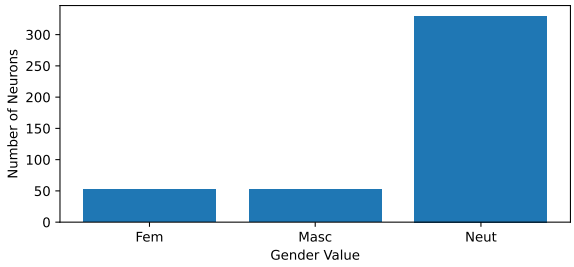


Figure 8: The LAPE scores for Gender categories.

usable-information for gender-neutral languages such as Japanese and Turkish, as well as for English, whose gender marking is largely restricted to pronominal forms, indicate that ParsBERT captures cues associated with the absence or marginal realization of grammatical gender, consistent with Persian's gender-neutral system. However, languages with fully grammaticalized gender systems exhibit lower and variable results. No recoverable information is available for three-gender languages, whereas two-gender languages, such as French, Hindi, and Arabic, exhibit moderate usable information, with Arabic resulting in the highest usable information among gendered languages, likely reflecting partial surface-level alignment and extensive lexical overlap with Persian.

Figure 8 reports LAPE scores for each gender category. Gender-related selectivity in ParsBERT is concentrated in NEUT-associated features. This observation is consistent with the usable-information results for three-gender systems and likely reflects the greater semantic diversity and lexical heterogeneity of neuter nouns. Because neuter nouns typically encompass a broad range of inanimate and abstract referents, a Persian-trained model distributes their representation across many features, each capturing a subset of the neuter class, whereas masculine and feminine nouns, which are often more semantically constrained, can be encoded with fewer selective features.

Overall, the better performance of two-gendered languages, compared to three-gendered ones, can be explained from two perspectives: the nature of the gender systems themselves and the type of language contact Persian has experienced. From a structural standpoint, two-gender systems tend to align more closely with general semantic and discourse cues such as animacy, humanness, and natural sex, which remain accessible even in a gender-neutral language like Persian and can therefore be exploited by ParsBERT. In contrast, three-gender systems rely on largely arbitrary noun-class assignments and agreement paradigms that must be learned morphologically and are not recoverable from syntactic configuration or semantics alone.

From a language-contact perspective, this asymmetry is further shaped by the predominantly lexical and semantic nature of Persian's contact with gendered languages, which facilitates surface-level alignment and the recoverability of semantically grounded gender distinctions typical of two-gender systems, without inducing the grammatical restructuring necessary to support fully paradigmatic gender systems.

6 Conclusion

We investigated whether contact-induced linguistic structure is reflected in the internal representations of a monolingual Persian-trained language model. The results reveal a clear asymmetry between syntactic and morphological features: universal syntactic categories (UPOS) show little sensitivity to historical contact, whereas morphological features such as CASE and GENDER are strongly shaped by language-specific structure. In particular, the model favors analytic and configurational encodings aligned with Persian, while failing to capture rich inflectional paradigms absent from the training language. The attribution analysis further indicates that contact-related alignment is primarily reflected through distributed cues rather than compact, category-specific features, pointing to surface-level or semantic alignment rather than grammatical transfer. Overall, these findings suggest that monolingual language models implicitly encode selective traces of language contact, constrained by the structural properties of the training language.

Limitations

A central limitation of our study is that the PUD collection does not include several languages that have played crucial roles in the historical and areal development of Persian. Notably absent are major Turkic varieties such as Azerbaijani and Turkmen, which have exerted long-term and deeply structural influence on Persian, often more directly than modern Turkish. Similarly, Urdu, the closest Indo-Aryan successor to the Persianate linguistic tradition and a primary locus of Persian lexical and syntactic transfer, is missing from PUD, preventing a fuller assessment of Persian's influence in South Asia. Important regional contact languages such as Armenian and Kurdish, both of which share extensive areal features with Persian, are also unavailable. The exclusion of these languages constrains the breadth of our analysis, as the strongest cases of sustained bilingualism, bidirectional borrowing, and structural convergence cannot be directly evaluated within the PUD framework. Future work incorporating UD treebanks or comparable resources for these languages would enable a more comprehensive and historically aligned investigation of Persian language contact.

References

Muzaffar Alam. 2018. *The Languages of Political Islam in India, c.1200–1800*. Permanent Black.

Abbas Amanat. 2017. *Iran: A Modern History*. Yale University Press.

Ali Basirat. 2025. Multilingual learning strategies in multilingual large language models. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 507–518, Suzhuo, China. Association for Computational Linguistics.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Lyle Campbell. 2013. *Historical Linguistics: An Introduction*. MIT Press, Cambridge, MA.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Bernard Comrie. 2018. *The World's Major Languages*. Routledge.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186.

Mehdi Farahani, Mostafa Gharachorloo, Mohammad Manthouri, and Mehrnoush Shamsfard. 2021. ParsBERT: Transformer-based model for persian language understanding. In *Proceedings of the 2nd Workshop on NLP for Internet Freedom (NLP4IF)*, pages 1–10.

Wolfdietrich Fischer and Jonathan Rodgers. 2001. *A Grammar of Classical Arabic*. Yale University Press, New Haven.

Negin Hosseini Goodrich. 2020. English in iran. *World Englishes*, 39(3):482–499.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Lars Johanson. 2006. Structural factors in turkic language contacts. In Yaron Matras and April McMahon, editors, *Language Contact and Areal Linguistics*, pages 24–43. Cambridge University Press, Cambridge.

Lars Johanson and Éva Á. Csató, editors. 2021. *The Turkic Languages*, 2 edition. Routledge, London.

Geoffrey Lewis. 1999. *The Turkish Language Reform: A Catastrophic Success*. Oxford University Press.

Yaron Matras. 2009. *Language Contact*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Yaron Matras and Jeanette Sakel, editors. 2007. *Types of loan: Matter and pattern*, pages 15–30. De Gruyter Mouton, Berlin, New York.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Andrea Schröter and Ali Basirat. 2025. Universal patterns of grammatical gender in multilingual large language models. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 34–46, Suzhuo, China. Association for Computational Linguistics.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Sarah Grey Thomason and Terrence Kaufman. 2023. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley.

Frans Van Coetsem. 2000. *A General and Unified Theory of the Transmission Process in Language Contact*. Winter, Heidelberg.

Hartger Veeman, Marc Allassonnière-Tang, Aleksandrs Berdicevskis, and Ali Basirat. 2020. Cross-lingual embeddings reveal universal and lineage-specific patterns in grammatical gender assignment. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 265–275, Online. Association for Computational Linguistics.

Gernot Windfuhr, editor. 2009. *The Iranian Languages*, 1 edition. Routledge, London.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

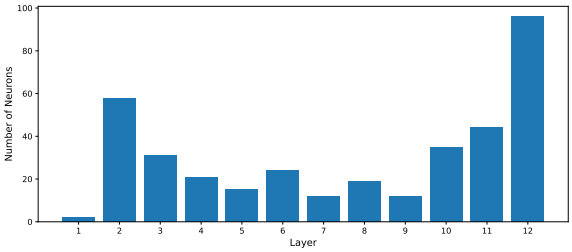


Figure 9: Layer-wise distribution of language-specific neurons identified by LAPE.

A Layer-wise Distribution of LAPE

In this section, we provide an overview of the LAPE scores across layers and task categories. Detailed analyses for each task are presented in the following subsections.

A.1 Language Identification

Figure 9 shows the total number of language-specific neurons identified at each layer. The results show that language-specific neurons are concentrated in a small number of layers, most notably Layer 2 and the top layers, rather than being evenly distributed throughout the network.

Figure 10 further breaks down the results by language. The dominance of Japanese-specific neurons across almost all layers, particularly in the upper layers, indicates that ParsBERT allocates a large number of highly specialized neurons to processing Japanese tokens. This likely reflects strong script-level and orthographic differences between Japanese and Persian, which require distinct representational pathways despite the model's monolingual training.

Hindi and Russian exhibit a more moderate but still noticeable concentration of language-specific neurons, primarily in higher layers. This aligns with the results from usable information, where both languages show relatively high recoverable language identity information. In contrast, English, French, German, and Turkish are associated with very few or no language-specific neurons. This

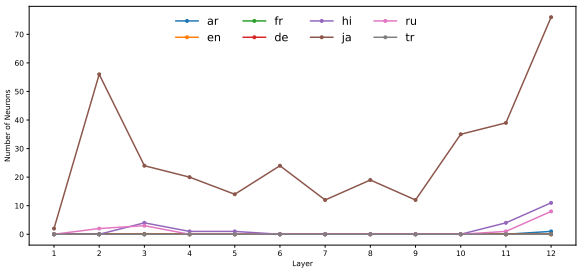


Figure 10: Layer-wise distribution of language-specific neurons identified by LAPE.

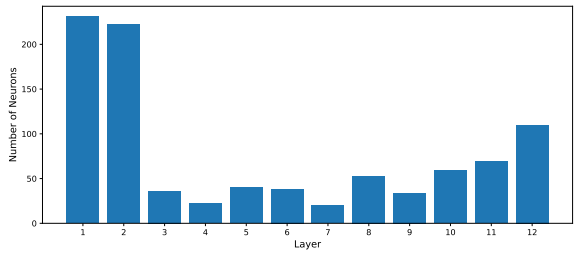


Figure 11: Layer-wise distribution of UPOS-specific neurons identified by LAPE.

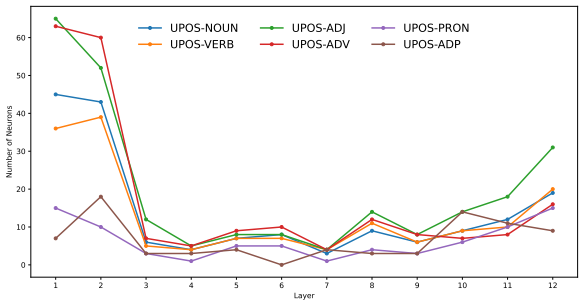


Figure 12: Layer-wise distribution of UPOS-specific neurons identified by LAPE.

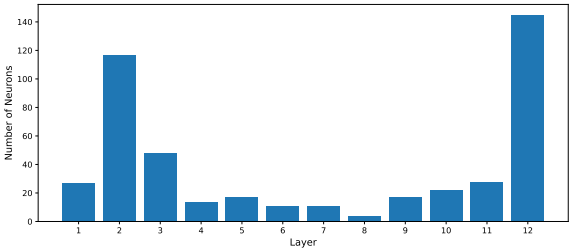


Figure 13: Layer-wise distribution of Case-specific neurons identified by LAPE.

suggests that for these languages, language identity is either weakly encoded or distributed across shared neurons rather than being localized in highly specific units.

A.2 UPOS Identification

Figure 11 shows the layer-wise LAPE scores for UPOS identification. The LAPE results indicate that UPOS selectivity is present both very early and again in the upper part of the network, with weaker neuron-level selectivity in the intermediate layers.

The per-category breakdown in Figure 12 shows that ADJ and ADV contribute the largest numbers of UPOS-selective neurons across layers, followed by NOUN and VERB. In contrast, PRON and ADP contribute substantially fewer selective neurons throughout the network. This indicates that the overall LAPE profile is driven primarily by content-word categories, whereas function-word categories yield comparatively fewer neurons that meet the LAPE selectivity criterion. From the perspective of language contact, this asymmetry may indicate that ParsBERT encodes function-word categories in a more abstract and transferable manner across languages, whereas content-word representations are more language-specific and therefore less directly aligned across Persian and the test languages.

A.3 Case

Figure 13 presents the layer-wise distribution of Case-specific neurons, and Figure 14 further decomposes this distribution by individual case values. The number of selected neurons is low in Layer 1, peaks sharply in Layer 2, drops substantially across the middle layers, and then increases again toward the final layer, where it reaches its maximum. This indicates that case selectivity is concentrated in a small subset of layers rather than being evenly distributed across the network depth.

The per-case breakdown in Figure 14 shows that this pattern is driven primarily by INS, which accounts for the largest share of case-selective neurons and exhibits pronounced peaks in Layers 2 and 12. Other case values contribute far fewer neurons overall: DAT shows a moderate peak in the early layers, while ACC and LOC increase mainly in the final layer; NOM and GEN remain sparse throughout. Overall, the LAPE analysis indicates that, when case-related selectivity is observed, it is localized to a limited number of layers and concentrated on a small subset of case values.

A.4 Gender

Figure 15 shows the layer-wise distribution of gender-selective neurons identified by LAPE. The results show that gender selectivity is not uniformly

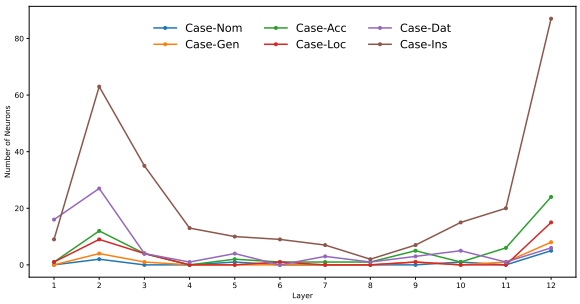


Figure 14: Layer-wise distribution of Case-specific neurons identified by LAPE.

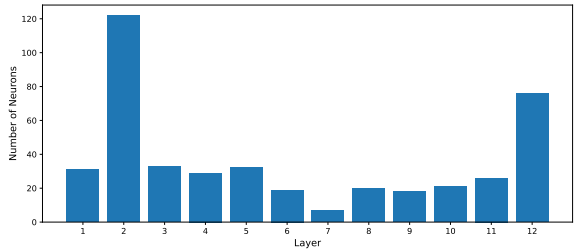


Figure 15: Layer-wise distribution of Gender-specific neurons identified by LAPE.

distributed across depth, but concentrates in a small number of layers, especially early (Layer 2) and late (Layer 12).

The per-category breakdown in Figure 16 indicates that the aggregate pattern is driven primarily by the NEUT category, which accounts for the majority of gender-selective neurons in nearly every layer and shows strong peaks in Layers 2 and 12. In contrast, FEM and MASC contribute comparatively few neurons and remain low across the network, with only small increases in the upper layers. Overall, the LAPE analysis suggests that neuron-level selectivity for gender is dominated by the NEUT label, while selectivity for FEM/MASC distinctions is comparatively sparse.

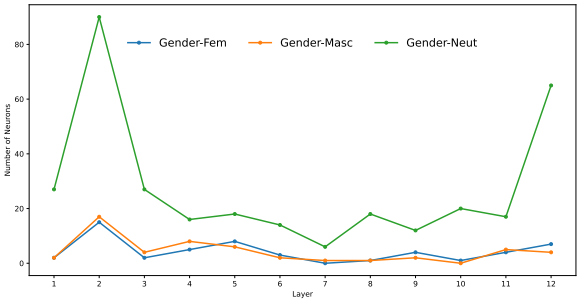


Figure 16: Layer-wise distribution of Gender-specific neurons identified by LAPE.

Online Polarization Detection in Persian (Farsi) Social Media

Saeedeh Davoudi
Information Retrieval Lab
Computer Science Department
Georgetown University
saeedeh@ir.cs.georgetown.edu

Nazli Goharian
Information Retrieval Lab
Computer Science Department
Georgetown University
nazli@ir.cs.georgetown.edu

Abstract

Polarization detection in low-resource and mid-resource languages remains a significant challenge for social understanding. This paper presents the first comprehensive benchmark to evaluate transformer-based models for detection of polarized language in Persian (also called Farsi) social media. The aim is to evaluate 1) how and if finetuning the pre-trained models have substantial impact; 2) how Persian specific monolingual models compare to multilingual for this task; 3) how and if transfer learning from models trained on other languages such as culturally-distant English, and culturally-close[er] Turkish, and Arabic can be of interest for this task; and 4) how competitive Large Language Models (LLMs) are in a zero-shot setting. Our evaluation of ten transformer-based models and two LLMs on a publicly available Farsi polarization dataset shows promising findings, highlighting both the strengths and limitations of each approach.

1 Introduction

Polarization is the fragmentation of society into antagonistic groups with fundamentally opposed values and identities (Stewart and Tingley, 2020). Polarization in online discourse has emerged as a challenge for social media platforms and content moderation systems (Shen and Rose, 2019). Polarization is often discussed under the umbrella of toxic or hateful language, but it captures a different phenomenon. Hate speech focuses on how a target is spoken about, while polarization focuses on how sharply groups are separated in public discourse. For example, a polarized sentence in Persian might say:

«کسی که از این حکومت دفاع می‌کند، هیچ‌وقت نمی‌تواند درد مردم معترض را بفهمد؛ دنیای ما کاملاً از هم جداست.»[1]

This sentence clearly separates us from them and presents the groups as unable to understand each other, but they do not directly insult each other. In contrast, a hate-oriented sentence might be:

«معترض‌ها آدم‌های کثیفی هستند و باید از جامعه جمع شوند.»[2]

Here, a group is directly attacked, which turns the statement into hateful or abusive content.

While significant progress has been made in detecting polarization in high-resource languages like English (Juvino Santos et al., 2025), low-resource and mid-resource languages such as Persian remain understudied despite having large, politically active online communities. This gap represents a critical limitation: millions of Farsi speaking users express polarized views on Twitter, Instagram, and other platforms regarding politics, religion, women's rights, and social issues. Yet, tools to systematically detect and analyze this polarization remain largely unavailable. The recent expansion of the POLAR dataset to include Persian (Naseem et al., 2025) presents an opportunity to evaluate polarization detection models for this language.

This paper presents the first comprehensive benchmark for the detection and categorization of Persian polarization in both binary and multilabel settings. We investigate four key research questions. First, to what extent does fine-tuning improve performance compared to pretrained models? This question examines whether domain-specific training substantially enhances model capabilities. Second, how do monolingual (specifically pretrained for Persian) models compare with multilingual models on Persian polarization detection and categorization? Third, does cross-lingual transfer from models trained on languages such as Arabic and Turkish, which are spoken in countries

[1] Translation: "Someone who defends this government can never understand the pain of the protesting people; our worlds are completely separate."

[2] Translation: "The protesters are filthy people and should be cleared away from society."

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 50–59

with some level of cultural similarity, outperform transfer from the culturally distant English language? Fourth, how well are LLMs for both binary polarization detection and multilabel polarization-type classification?

In the rest of this paper, we first review related works in Section 2. Section 3 and Section 4 present the datasets used in our study, the models employed, and the experimental setup and results. Finally, Section 5 outlines the limitations and directions for future work, and Section 6 concludes the paper.

2 Related Work

Polarization detection has emerged as an important research area in Natural Language Processing (NLP), particularly given its implications for understanding online discourse and social dynamics. Early work focused primarily on English-language contexts, examining ideological expression in political debates and social media (Kubin and Von Sikorski, 2021; Hohmann et al., 2023; Kreiss and McGregor, 2024). (Naseem et al., 2025) introduced POLAR, the largest multilingual polarization benchmark with more than 23,000 instances across 7 languages, including multi-level annotations for binary detection, type classification, and manifestation identification. The dataset addresses a critical gap in multilingual NLP by providing balanced coverage of diverse languages and social phenomena related to political, ethnic, religious, and gender-based polarization. The recent expansion of the POLAR framework includes a Persian dataset, offering an unprecedented opportunity to evaluate polarization detection models for low and mid-resource languages.

Persian NLP has seen growing attention, particularly in toxic language detection, hate speech detection, and offensive language identification. A foundational contribution to Persian hate speech research is the PHATE dataset (Delbari et al., 2024), which consists of over 7,000 manually-annotated Persian tweets with multi-label hate speech annotations. What distinguishes PHATE from previous Persian datasets is its inclusion of annotators' explanations that justify each hate speech label alongside information about targeted groups. This enriched annotation approach facilitates not only supervised classification but also research on model interpretability. The dataset encompasses a hierarchical multi-label structure with three subcategories: violence, vulgarity, and hate. (Kebriaei et al., 2023) contributed to Persian offensive language detection research by constructing a large-scale 38,000-tweet corpus using keyword-based data selection techniques combined with crowdsourced annotations and a curated Persian insulting lexicon. This dataset represents one of the largest openly available resources for Persian harmful language detection and demonstrates the feasibility of scaling Persian offensive language annotation. The study employed both classical machine learning approaches and modern transformer-based models to establish baselines on this dataset. Their work shows that Persian-specific models outperform multilingual alternatives, reinforcing the importance of language-specific pre-training for this task.

Cross-lingual transfer has proven effective for many NLP tasks. Recent work by (Bokaei et al., 2025) provides particularly relevant insights for cross-lingual transfer learning in Persian toxic language detection. In their comprehensive study on the PHATE dataset, they investigated the role of cultural context in transfer learning effectiveness. Critically, they found that transfer from languages originating from culturally similar countries (Arabic, Indonesian) yields significantly better results than transfer from culturally distant yet resource-rich languages like English.

However, the focus of prior work on toxic language and hate speech detection has left polarization detection, a distinct phenomenon with different manifestations and social implications, largely unexplored in Persian.

3 Experimental Plan

We structure our experimental plan around the following research questions on Persian polarization detection and polarization categorization:

- **RQ1:** Does fine-tuning on Persian polarization data improve performance compared to using pretrained models alone? Which fine-tuned model outperforms the others among monolingual and multilingual models?
- **RQ2:** How do Persian-specific monolingual models compare with multilingual models?
- **RQ3:** Does cross-lingual transfer from culturally related languages, such as Arabic and Turkish, outperform transfer from culturally distant languages such as English?

Table 1: Dataset statistics for polarization detection and category classification. For Persian, values are averages over 5 CV folds with 80% for training and 20% test (10% of training data is considered as validation dataset). Samples can belong to multiple categories.

		Polarization Detection			Category Classification				
Language	**Split**	**Total**	**Non-Polarized**	**Polarized**	**Political**	**Racial**	**Religious**	**Gender**	**Other**
Persian	Train	2,372	632	1,740	1,043	58	229	142	574
	Validation	264	70	194	116	6	25	16	64
	Test	659	171	488	289	16	63	39	160
English	Train	2,676	1,674	1,002	996	264	106	67	121
Turkish	Train	2,364	1,209	1,155	1,057	400	360	113	114
Arabic	Train	3,379	1,868	1,511	780	583	283	369	565

- **RQ4:** Are zero-shot LLMs performing better than transformer-based models in both binary polarization detection and multilabel polarization-type classification?

Polarization Detection We first examine whether a given text contains polarizing content. Each sample is assigned a binary label indicating the presence (1) or absence (0) of polarization. This setting reflects a common real-world scenario in which systems must distinguish polarized discourse from neutral content before any further analysis.

Polarization Categorization We then study a more fine-grained setting that focuses on identifying the types of polarization expressed in a text. Each sample is evaluated independently across five dimensions: Political, Racial/Ethnic, Religious, Gender/Sexual, and Other. The Other category captures content that does not fall into the predefined categories. Since multiple forms of polarization can co-occur within the same text, this is formulated as a multi-label classification problem. For example, gender-related polarization may overlap with religious arguments, which makes this setting more challenging than binary detection.

Dataset Overview We use POLAR dataset (Naseem et al., 2025) for both polarization detection and categorization tasks. Table 1 presents dataset statistics across Persian, English, Turkish, and Arabic languages. It is shown that the dataset is highly imbalanced among labels, and models need to tackle this challenge. Figure 1 presents the sentence length distribution across different languages. Word counts range from an average of 11 words in English to 22 words in Turkish, while character counts vary more, from 70 characters (English) to 172 characters (Turkish). This difference in character/word count reflects the distinct morphological properties of different languages. Persian exhibits a unique pattern where non-polarized content is longer compared to polarized content. In contrast, English, Turkish, and Arabic show the opposite trend, with polarized content being longer than non-polarized content.

Model Selection We evaluate ten transformer-based models, grouped into monolingual Persian models and multilingual models. This design allows us to directly compare language-specific pretraining with multilingual representations and to analyze their strengths and limitations for Persian polarization analysis. In addition, we include two zero-shot LLM baselines, Gemini-2.5 flash lite (Google Cloud) and GPT-5 nano (OpenAI) to assess how instruction-tuned LLMs perform without task-specific training. Table 2 shows all transformer-based models and their categories used in this study.

Monolingual Persian Models We include four models that are pretrained exclusively on Persian data. These models are expected to capture Persian-specific syntax, vocabulary, and discourse patterns more effectively than multilingual models.

ParsBERT is a BERT-style encoder pretrained on a large and diverse Persian corpus and has shown strong performance across multiple Persian NLP tasks (Farahani et al., 2021). ALBERT-fa is a lightweight alternative based on ALBERT, pretrained on billions of Persian tokens, and is included to study the trade-off between model size and performance (Farahani, 2022). RoBERTa-fa follows the RoBERTa training strategy adapted to Persian, enabling us to examine the impact of more aggressive pretraining compared to standard BERT

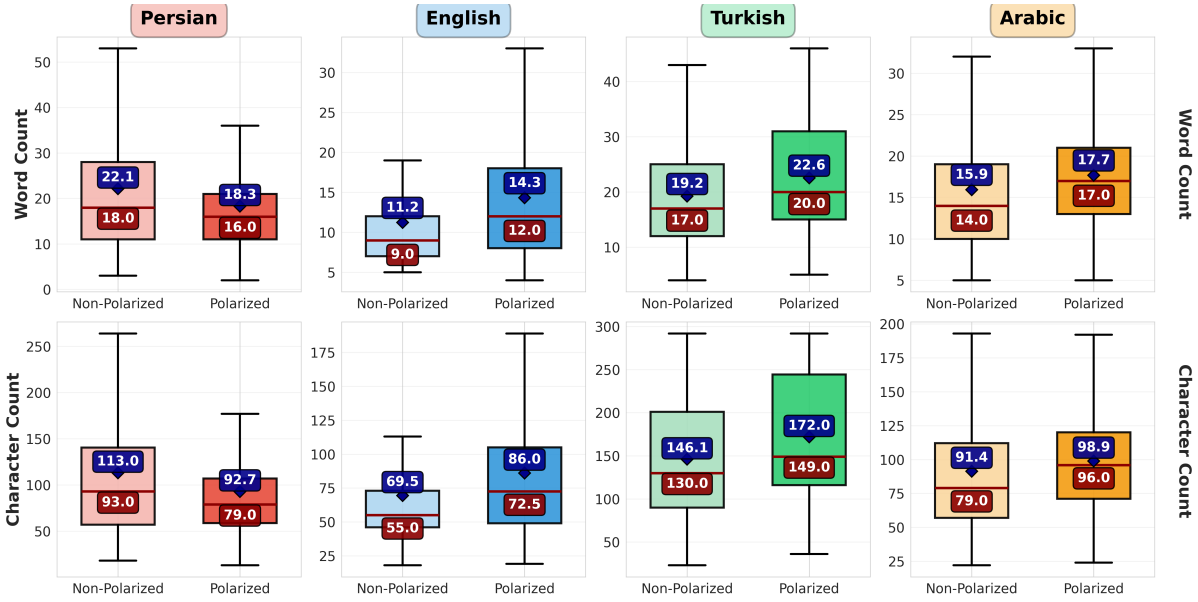


Figure 1: Comparison of sentence length distributions between non-polarized and polarized data across four languages. The top row shows word count distributions, while the bottom row displays character count distributions. Each language is represented with a distinct color scheme: Persian (red), English (blue), Turkish (green), and Arabic (orange), with lighter shades for non-polarized content and darker shades for polarized content. Numbers in small red boxes are median and numbers in small blue boxes are mean.

(HooshvareLab, 2021b). DistilBERT-fa applies knowledge distillation to produce a smaller and faster Persian model while retaining much of the representational power of BERT (HooshvareLab, 2021a).

These models have been widely used and shown strong performance in related classification and representation learning tasks, making them ideal candidates for polarization analysis.

Multilingual Models We also evaluate six multilingual models that differ in architecture, pretraining objectives, and cross-lingual design. These models enable us to assess how well multilingual representations transfer to Persian and how architectural choices affect polarization detection.

mBERT is the original multilingual BERT model trained on Wikipedia data from over 100 languages and serves as a strong baseline for multilingual transfer (Devlin et al., 2019). XLM-R extends mBERT by using larger and more diverse training data and is widely regarded as a strong multilingual encoder (Conneau et al., 2020). InfoXLM builds upon XLM-R by incorporating information-theoretic objectives and parallel data, making it particularly suitable for cross-lingual transfer scenarios (Chi et al., 2021). RemBERT revisits multilingual embedding design and rebalances pretraining data across languages, achieving strong performance on cross-lingual benchmarks (Chung et al., 2020). mDeBERTaV3 enhances cross-lingual generalization through disentangled attention and ELECTRA-style pretraining (Replaced Token Detection), which is particularly effective in low-resource settings (He et al., 2023). Finally, LaBSE is included as a sentence-level encoder trained with translation-ranking objectives, allowing us to study how language-agnostic sentence representations perform compared to token-level encoders (Feng et al., 2022).

Large Language Models Gemini-2.5 flash lite is a cost and latency optimized Gemini 2.5 model designed for high throughput with multimodal input support. GPT-5 nano is the fastest and most cost efficient GPT-5 variant, aimed at high volume workloads like classification and summarization, with text (and image) inputs and text outputs.

Overall, this selection is based on POLAR benchmark and covers a diverse range of model sizes, architectures, and pretraining strategies.

Training Configuration We performed 5-fold crosss-validation with 72-8-20 splits for training, validation, and test. For cross-lingual training, we used a 90-10 split of the source language data for training and validation, then evaluated the trained models on all 5 Persian test folds to assess cross-

Table 2: Overview of models evaluated in this study. The Cross-lingual column indicates if the model was fine-tuned on other languages (English, Turkish, Arabic, and Turkish+Arabic) in addition to Persian. In total, we evaluate 10 pretrained models, 10 finetuned models on Persian, and 24 cross-lingual models.

Model Family	Size	Pretrained Models	Fine-tuned Models on Persian	Cross-lingual
Multilingual Models				
mBERT	178M	mBERT	mBERT-fa	✓
XLM-RoBERTa	278M	XLM-R	XLM-R-fa	✓
RemBERT	575M	RemBERT	RemBERT-fa	✓
mDeBERTa	278M	mDeBERTa	mDeBERTa-fa	✓
InfoXLM	270M	InfoXLM	InfoXLM-fa	✓
LaBSE	471M	LaBSE	LaBSE-fa	✓
Monolingual (Persian) Models				
ParsBERT	118M	ParsBERT	ParsBERT-fa	–
ALBERT	12M	ALBERT-fa	ALBERT-fa-ft	–
DistilBERT	66M	DistilBERT-fa	DistilBERT-fa-ft	–
RoBERTa	125M	RoBERTa-fa	RoBERTa-fa-ft	–

lingual transfer capabilities. It is important to note that samples can belong to multiple polarization categories simultaneously, as reflected in the overlapping category counts shown in Table 1.

All models were fine-tuned with early stopping (patience=5) based on validation performance, with a maximum of 50 training epochs. We used the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 32, and a maximum sequence length of 128 tokens. We also applied class weights during training to address label imbalance. Binary cross-entropy loss was used to handle the multi-label nature of categorization task. [3]

For both binary polarization detection and multi-label categorization, we use the weighted F1 score as the primary evaluation metric, which provides a balanced measure of precision and recall. It accounts for class imbalance by weighting each category's F1 score by its number of true instances, making it more suitable for datasets with varying category frequencies.

4 Results and Analysis

[4] In this section, we present the results and address our research questions. Figure 2 compares pretrained vs finetuned models' performance on two related but distinct tasks: detecting whether content is polarized (binary classification), identifying which specific categories it belongs to (multi-label classification). The evaluation indicates that fine-tuning is essential for both tasks. While pretrained models struggle, fine-tuned models achieve strong results, with the best models reaching 82.7% and 77.9% F1 scores, respectively (RQ1). Also, different models excel at different tasks; category identification is inherently more difficult than polarization detection, with all models showing a consistent performance gap between the two tasks. RoBERTa-fa performs best at detecting polarization (82.7%), while LaBSE-fa leads in identifying specific categories (77.9%) (RQ1). Monolingual models outperform multilingual models in polarization detection. This suggests that for simpler settings such as binary classification, monolingual models are sufficient and well suited to the task (RQ2).

Figure 3 examines whether models trained on polarized content in other languages can detect polarization in Persian without ever seeing Persian training examples. We trained models on English, Arabic, Turkish, and combinations of Arabic and Turkish languages, then tested them on Persian data. The results show that cross-lingual transfer works remarkably well. For binary polarization detection, training on English data produces the best results, with XLM-R achieving 71.5% weighted F1 on Persian test data. This is impressive considering the model never saw any Persian examples dur-

[3] The source code is available at https://github.com/dsaeedeh/Polarization_Detection

[4] When evaluated on the official SemEval 2026 Task 9 blind test set, our fine-tuned system achieved highly competitive results. Specifically, our submission ranked 7th among 44 systems in the first subtask (XLM-RoBERTa model), and 14th among 27 in the second (ParsBERT model).

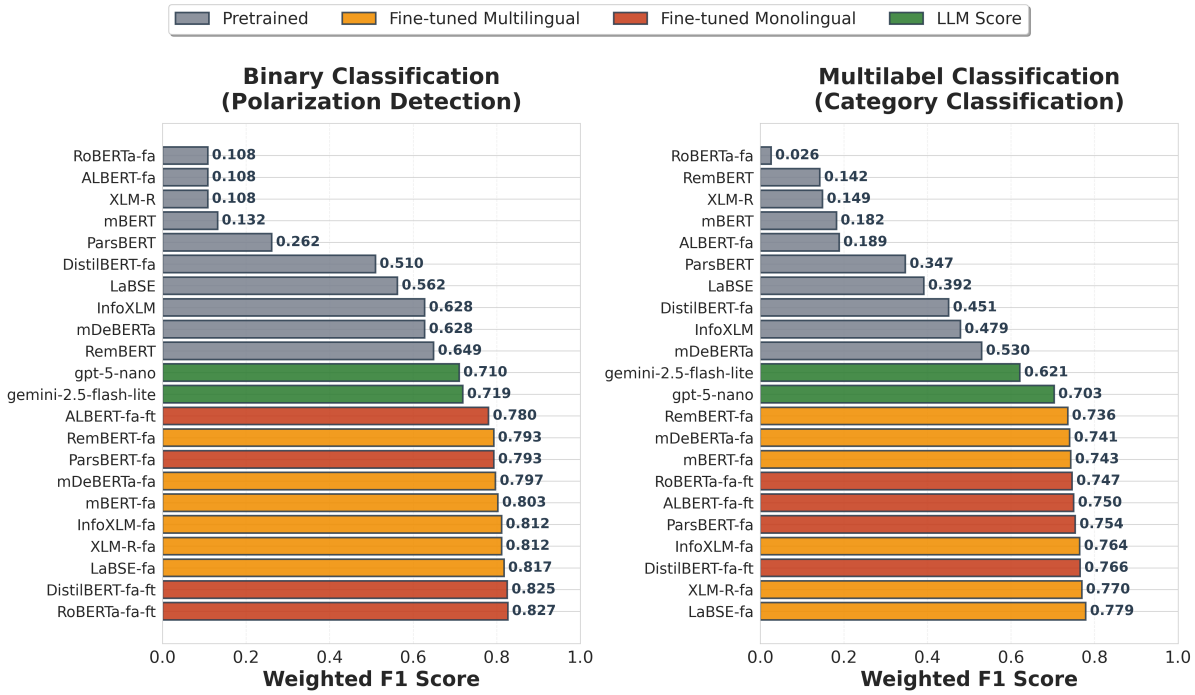


Figure 2: Performance comparison between transformer baselines and zero-shot LLMs for polarization detection (left chart) and categorization (right chart). The plots show that fine-tuning consistently improves performance over pretrained models, while zero-shot LLMs provide strong training-free baselines. Identifying specific categories remains more challenging than detecting whether content is polarized. All bars report 5-fold cross-validation averages, and LLM results are computed by aggregating predictions across folds.

ing training. For multilabel category classification, combining Turkish and Arabic training data yields the best transfer performance at 71.9% weighted F1 using RemBERT (RQ3).

Different models show varying cross-lingual capabilities. XLM-R and RemBERT consistently perform well across different source languages, while other models show more variability. Interestingly, the performance of the "Turkish+Arabic" transfer model likely benefits from cultural proximity. These languages share similar political discourse patterns, religious references, and social issues with Persian, making polarization strategies more transferable across these culturally connected regions (RQ3). These findings show that polarization detection systems can work in new languages without requiring labeled training data in those languages. The 71-72% weighted F1 scores achieved through cross-lingual transfer fall midway between pretrained (37-53%) and fully fine-tuned models (78-83%), offering a practical middle ground when labeled data is scarce.

Figure 4 presents the best performing models for each of the three approaches we evaluated. For binary polarization detection, the results show a clear advantage for fine-tuning. RoBERTa-fa achieves the highest performance at 82.7% when fine-tuned on Persian data. Among pretrained models, RemBERT performs best at 64.9%. Cross-lingual transfer using XLM-R trained on English data achieves 71.5% weighted F1, falling between pretrained and fine-tuned approaches. Turning to multi-label category classification, LaBSE proves to be the superior model, achieving a weighted F1 of 77.9% after fine-tuning. In the pretrained setting, mDeBERTa demonstrates the strongest capability, reaching 53.0%, though a significant gap remains between the fine-tuned and pretrained baselines. Finally, cross-lingual transfer from "Turkish+Arabic" languages using RemBERT yields 71.9%. This pattern can be further understood by examining the cross-lingual category confusion heatmap (see Figure 5). The heatmap reveals that models trained on Turkish and Arabic languages transfer more accurately across culturally salient categories such as Political and Religious, where discourse patterns and topical content are regionally shared with Persian. In contrast, English-trained models exhibit higher confusion, particularly misclassifying Religious, Gender/Sexual, and Political categories. This suggests that the English models lack exposure to the specific cultural and thematic nuances

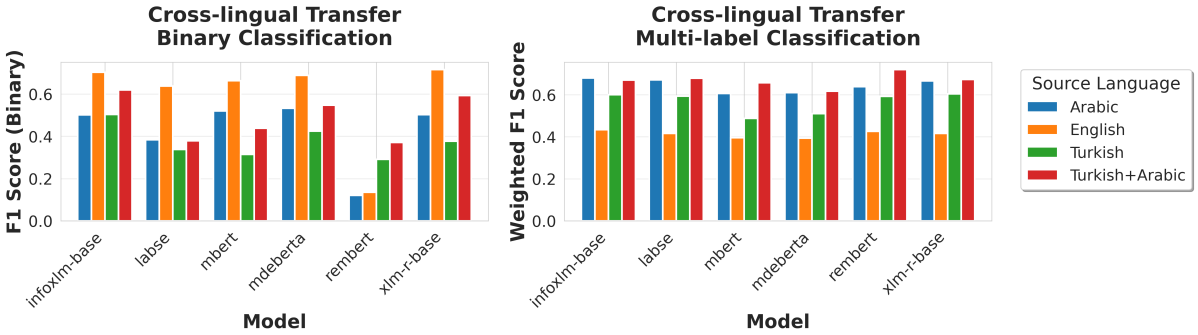


Figure 3: Cross-lingual transfer learning results showing how models trained on other languages perform when tested on Persian. Different colored bars represent training on English, Arabic, Turkish, or Turkish and Arabic data together. For binary polarization detection (left), English training data works best, with XLM-R reaching 71.5% F1. For multi-label categorization (right), Turkish and Arabic combined training achieves 71.9% F1 using RemBERT.

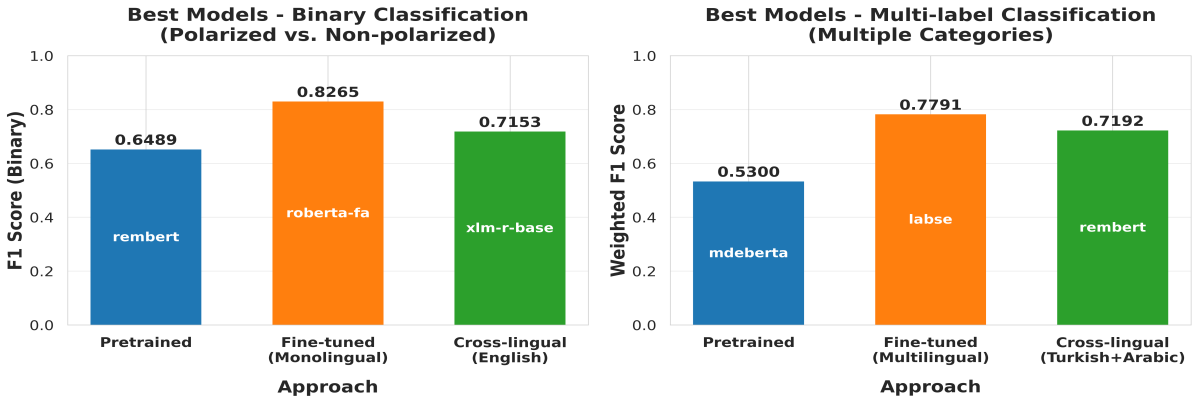


Figure 4: Best performing models for each approach. Each bar shows the top model from three approaches: pretrained, fine-tuned on Persian data, and cross-lingual transfer for binary polarization detection (left) and multi-label category classification (right).

present in Middle Eastern contexts, leading to more frequent misclassifications in categories that are less prominent or differently framed in Western discourse. These findings support the hypothesis that cultural proximity, reflected in shared regional issues and discourse, enhances cross-lingual transfer for fine-grained category classification.

In addition to these transformer-based baselines, we evaluated two zero-shot LLMs, gemini-2.5-flash-lite and gpt-5-nano (Figure 2. For binary polarization detection, gemini-2.5-flash-lite reaches 71.9% weighted F1 and gpt-5-nano reaches 71.0%, placing them close to the cross-lingual transfer result with English-trained XLM-R. This indicates that zero-shot LLMs can provide a strong training-free alternative for polarization detection, outperforming the best pretrained transformer baseline (RemBERT), but still falling notably short of the best fine-tuned model (RoBERTa-fa)(RQ4).

Turning to multi-label category classification, gpt-5-nano achieves 70.3% weighted F1, approaching the culturally proximate transfer result and substantially outperforming the pretrained baseline, while gemini-2.5-flash-lite reaches 62.1%, improving over pretrained transformers but remaining clearly below cross-lingual transfer and fine-tuning models (RQ4).

These results highlight four key findings: First, fine-tuning consistently yields the best performance for both tasks, improving over pretrained models by 27-47% relative. Second, cross-lingual transfer is particularly effective for multi-label classification, closing most of the gap between pretrained and fine-tuned approaches. Third, two zero-shot LLMs provide strong training-free baselines: they outperform the best pretrained transformers, though both remain below fine-tuning. Last, different models excel at different tasks and approaches: RoBERTa-fa dominates binary classification when fine-tuned, while LaBSE leads in multi-label classification, and RemBERT shows strong cross-lingual transfer capabilities.

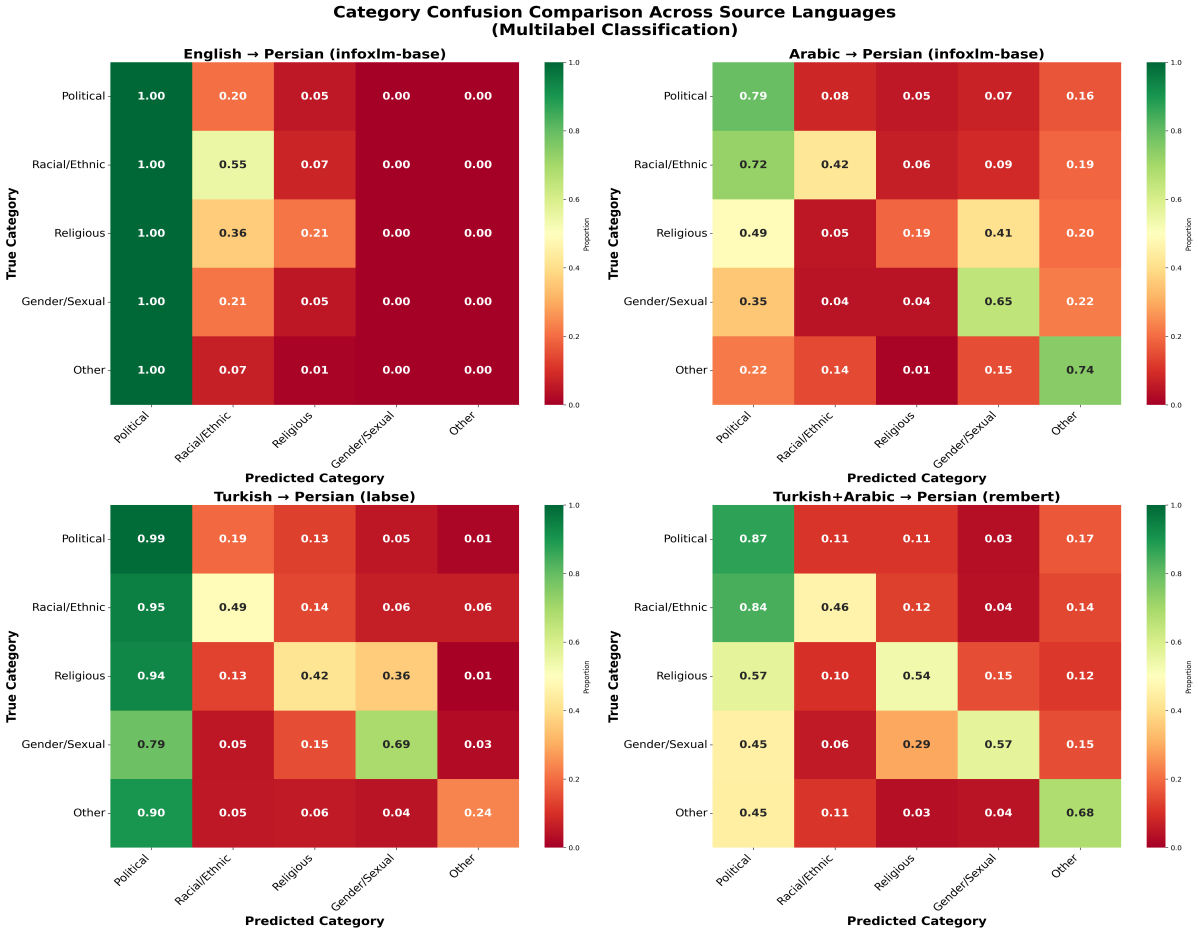


Figure 5: Cross-lingual category confusion matrix comparing best models trained on English, Turkish, Arabic, and Turkish+Arabic. The heatmap highlights that Turkish+Arabic models achieve lower confusion and higher accuracy in culturally salient categories (Political, Religious), while English-trained models more frequently misclassify these categories, especially as Other or Political, reflecting the impact of cultural proximity on fine-grained polarization classification.

5 Limitations

While our study provides comprehensive insights into polarization detection across multiple approaches, several limitations should be acknowledged. Our evaluation focuses on Persian as the primary target language, with cross-lingual experiments on English, Arabic, and Turkish. Expanding to additional languages, particularly those from different language families and cultural contexts, would strengthen claims about cross-lingual transferability. Despite strong gains from fine-tuning, there remains room for improvement, especially for multi-label categorization. Our zero-shot LLM evaluation is limited to two models and a specific prompt setup; performance may vary with prompting strategies, decoding choices, and model versions, and we do not study robustness to such factors. In future work, we plan to extend the LLM setting to few-shot and to develop explainability methods to better understand which linguistic and cultural cues models rely on when detecting polarization across languages.

6 Conclusion

This study benchmarks polarization detection in Persian social media through binary classification (polarized vs. non-polarized) and multi-label categorization (specific polarization types). To compare performance across different learning stages, we tested 10 transformer models under pretrained, fine-tuned, and cross-lingual transfer settings and 2 zero-shot LLMs (gemini-2.5-flash-lite and gpt-5-nano), totaling 46 model instances. Fine-tuning achieves the best results, reaching 82.7% F1 for binary detection and 77.9% weighted F1 for categorization. When labeled data is unavailable, cross-lingual transfer provides a practical alternative, achieving 71-72% F1 and performing particu-

larly well for multi-label categorization when transferring from culturally related languages (Turkish+Arabic). Zero-shot LLMs offer a complementary training-free baseline. They are competitive for binary detection and, for categorization. Overall, these findings show that effective polarization detection is possible in low and mid-resource languages, either through fine-tuning when data is available or through culturally informed cross-lingual transfer and strong zero-shot LLM baselines when data is scarce.

References

Zahra Bokaei, Walid Magdy, and Bonnie Webber. 2025. Culture matters in toxic language detection in Persian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9290–9304, Vienna, Austria. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *ArXiv*, abs/2010.12821.

Alexis Conneau and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Zahra Delbari, Nafise Sadat Moosavi, and Mohammad Taher Pilehvar. 2024. Spanning the spectrum of hatred detection: A persian multi-label hate speech dataset with annotator rationales. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17889–17897.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Mehrdad Farahani. 2022. Albert-persian: A lite bert model for persian. HuggingFace model card m3hrdadfi/albert-fa-base-v2.

Mehrdad Farahani, Mohammad Gharachorloo, and 1 others. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Google Cloud. Gemini 2.5 flash-lite. `https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash-lite`. Accessed: 2026-02-10.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Marilena Hohmann, Karel Devriendt, and Michele Coscia. 2023. Quantifying ideological polarization on a network using generalized euclidean distance. *Science Advances*, 9(9):eabq2044.

HooshvareLab. 2021a. Distilbert-fa: A distilled persian bert model. HuggingFace model card HooshvareLab/distilbert-fa-zwnj-base.

HooshvareLab. 2021b. Roberta-fa-zwnj-base: A roberta model for persian language understanding. HuggingFace model card HooshvareLab/roberta-fa-zwnj-base.

Lucas Raniére Juvino Santos, Leandro Balby Marinho, Claudio Elizio Calazans Campelo, Filippo Menczer, and Alessandro Flammini. 2025. Can large language models effectively mitigate polarization in social media text? In *Proceedings of the 17th ACM Web Science Conference 2025*, Websci '25, page 348–357, New York,

NY, USA. Association for Computing Machinery.

Emad Kebriaei, Ali Homayouni, Roghayeh Faraji, Armita Razavi, Azadeh Shakery, Heshaam Faili, and Yadollah Yaghoobzadeh. 2023. Persian offensive language detection. *Mach. Learn.*, 113(7):4359–4379.

Daniel Kreiss and Shannon C McGregor. 2024. A review and provocation: On polarization and platforms. *New Media & Society*, 26(1):556–579.

Emily Kubin and Christian Von Sikorski. 2021. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206.

Usman Naseem, Juan Ren, Saba Anwar, Sarah Kohail, Rudy Alexandro Garrido Veliz, Robert Geislinger, Aisha Jabr, Idris Abdulmumin, Laiba Qureshi, Aarushi Ajay Borkar, and 1 others. 2025. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *arXiv preprint arXiv:2505.20624*.

OpenAI. Gpt-5 nano model. `https://developers.openai.com/api/docs/models/gpt-5-nano`. Accessed: 2026-02-10.

Qinlan Shen and Carolyn Rose. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 58–69, Florence, Italy. Association for Computational Linguistics.

Brandon M Stewart and Dustin H Tingley. 2020. The nature and origins of mass opinion. *Journal of Politics*, 82(3):757–778.

ParsCORE: The Persian corpus of online registers

Alireza Razzaghi and Erik Henriksson and Veronika Laippala
TurkuNLP, University of Turku
{alireza.razzaghi, erik.henriksson, mavela}@utu.fi

Abstract

Despite recent advances in automatic web register (genre) labeling and its applications to web-scale datasets and LLM development, the effectiveness of these tools for digitally low-resource languages remains unclear. This study introduces ParsCORE, the first large-scale collection of Persian web registers (genres), and evaluates deep learning models for register classification and keyword analysis across major registers. Using 2,000 human-annotated documents, the models achieved a micro F1-score of 0.76. The findings provide a foundation for future research on the linguistic and cultural specificities of Persian registers.

1 Introduction

Register, sometimes referred to as genre, denotes text classes such as news articles, advertisements, and how-to pages (Biber, 1988; Biber and Conrad, 2019). Registers are defined based on the relationship between the typical linguistic features of a text and its situational context. Recently, register has also been reconceptualized as a cultural construct, highlighting the need to understand the specificities of the communicative context in the modeling of registers (Biber and Egbert, 2023).

Recent advances in web register and web genre identification have resulted in reliable register identification tools (Lepekhin and Sharoff, 2021; Kuzman and Ljubešić, 2023; Henriksson et al., 2024) that can be used to add register metadata to Web datasets such as HPLT (Burchell et al., 2025). This substantially improves the usability of such datasets. In line with FineWeb-Edu (Lozhkov et al., 2024), register information provides an effective mechanism for web data sampling, including the selection of documents for LLM training (Myntti et al., 2025). Additionally, register-labeled web datasets offer unprecedented possibilities for linguistics, with particular relevance for lower-resourced and minority languages.

Existing register and genre identification tools have demonstrated significant multilingual capacities (Henriksson et al., 2024; Repo et al., 2021; Lepekhin and Sharoff, 2021; Kuzman et al., 2023). However, these tools are largely based on European languages, such as English, Finnish, and Slovenian, with other languages represented only in small evaluation examples. The characteristics of online registers and their variation in non-European languages still remain largely unknown. In this paper, we target this lack by presenting resources for examining Web registers in Persian.

We present ParsCORE v0.1, the first large-scale manually annotated corpus of web registers in Persian. ParsCORE targets the entire unrestricted web, following the annotation standards established in the Multilingual CORE Corpora (Henriksson et al., 2024; Erten-Johansson et al., 2024; Skantsi and Laippala, 2023). These standards define a hierarchical register scheme of 9 main and 14 subregister classes. This provides a solid basis for both automatic web register identification and the linguistic analysis of language use online. In this paper, we present the corpus compilation process, its register distribution, and the linguistic characteristics of the registers using text dispersion keyness (Egbert and Biber, 2018). Finally, we report initial experiments applying the corpus for automatic register identification. The ParsCORE, codes, and results are available on GitHub.

2 Related Work

Web registers (or genres) have been studied in many corpora (see Kuzman and Ljubešić 2023 for a survey). Notably, the Corpus of Online Registers of English (CORE) (Biber and Egbert, 2018) was the first to target the full unrestricted web rather than predetermined categories. CORE's data-driven annotation scheme allows hybrid documents to be assigned multiple registers simultane-

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 60–73

ously (Biber et al., 2020, 2015). This design addresses challenges posed by the unrestricted web, where many documents combine features of several registers.

Despite the general underrepresentation of digitally low-resource languages, several web register corpora have been developed for such languages. Examples include GINCO, a web genre corpus for Slovenian (Kuzman et al., 2022), the Finnish Corpus of Online Registers (Skantsi and Laippala, 2023), the Turkish web register corpus (Erten, 2023), and Swedish and French web registers (Hellström et al., 2025). Multilingual CORE (Henriksson et al., 2024) extended this line of work by applying the CORE taxonomy across 16 languages, integrating several language-specific resources. Similarly, Sharoff (2018) proposed the Functional Text Dimensions (FTD) approach. This method models texts by their similarity to functional prototypes to enable analysis of hybrid documents and varying degrees of register specificity in five languages.

Keyword analysis—used here to examine Persian registers (Section 5.1)—is a corpus-linguistic method for identifying distinctive features of a text collection (Scott, 1997; Bondi, 2010; Culpeper and Demmen, 2015). We apply the dispersion-based method (Egbert and Biber, 2018), which identifies vocabulary distributed across documents rather than concentrated in a few outliers (Gries, 2021; Sönning, 2023). Previous cross-linguistic studies demonstrate that register variation can involve language-specific grammatical features, such as honorific marking in Korean (Biber, 1995). Register variation can also reflect culturally specific features, such as the strategic use of a perfective suffix in Turkish (Erten, 2023). Hellström et al. (2025) showed that cross-linguistic divergences often arise from grammatical realizations and register-specific traits. These findings highlight the importance of cross-linguistic register analysis. The present study contributes a Persian perspective.

Recent studies show strong performance on automatic register identification using manually annotated corpora. Henriksson et al. (2024) reported a micro F1 of 79% on the Multilingual CORE. This performance is obtained using multi-label classification with 25 registers and outperforms previous approaches such as Kuzman et al. (2022). Their results show that multilingual training benefits languages and register classes with limited data. Kuzman et al. (2022) reported a micro F1 of 78% on twelve genre classes aligned with the CORE schema using XLM-R (Conneau et al., 2020). In a cross-lingual experiment, they trained models on machine-translated English versions of GINCO and evaluated them on CORE texts, achieving a micro F1 of 63%. Kuzman et al. (2023) further examined different annotation schemes, including CORE, FTD (Sharoff, 2018), and GINCO. The best performance was achieved on the X GENRE dataset, which integrates all three corpora into nine classes. Using XLM-R Base with single-label classification, the model reached a micro F1 of 80%.

Building on this work, the present study focuses on Persian, a digitally low-resource language lacking extensive web register data and multi-label models. The Multilingual CORE (Henriksson et al., 2024) contains only 69 Persian instances. In contrast, our study presents a larger Persian web register corpus and implements multi-label classification, representing an initial step toward large-scale multilingual register analysis that incorporates Persian.

Narrative NA	359 | 569
News report, Sports report, Narrative blog, Other narrative	
Opinion OP	149 | 353
Review, Opinion blog, Advice, Denominational religious blog / sermon, Other opinion	
Informational Description IN	251 | 439
Description of a thing or a person, Research article, Encyclopedia article, FAQ, Legal terms and conditions, Other Informational description	
Interactive Discussion ID	43 | 190
How-to HI	35 | 146
Recipe, Other how-to	
Informational Persuasion IP	287 | 438
Description with intent to sell, News & opinion blog or editorial, Other informational persuasion	
Lyrical LY	28 | 199
Spoken SP	12 | 153
Interview, Other spoken	
Machine Translated MT	20 | 32

Table 1: Main registers and sub-registers with the number of documents before and after upsampling.

3 Data

We employed a randomized 2,000-document sample from the Persian portion of HPLT 3.0, a large-scale web-crawled monolingual dataset covering 198 languages and drawing on both general web content and internet archives. It contains approximately 3.3 petabytes of so-called "wide crawls" from the Internet Archive, covering the period from 2012 to 2020 and 57 full snapshots from Common Crawl spanning 2014 to 2025. Together, these sources contribute to a total volume of about 7.2 petabytes of raw web archive data. The Persian section of HPLT 3.0 contains 124.02M documents, and a MinHash-based global near-deduplication is implemented. Moreover, HPLT 3.0 provides automatic register annotations for Persian, assigning each document a confidence score between 0.0 and 1.0 for each register. We chose HPLT over raw Common Crawl data because it is cleaned and deduplicated, yielding less noisy and more reliable text for register analysis. The register annotations reported here apply only to the specific sample used in this study.

The sample is human-annotated based on the hierarchical CORE register scheme. The definition is available at the annotation guidelines, and the abbreviations are presented in Table 7 (Appendix). The annotation followed a two-step procedure, in which documents were first accepted or rejected and then assigned register labels. A total of 700 documents were rejected during annotation. The purpose of rejection was to restrict register annotation to full, coherent texts. Documents were rejected if they (1) consisted mainly of headline-style fragments, lists, download pages, or captions; (2) contained fewer than two complete sentences or lacked coherent text; (3) were dominated by boilerplate or 'junk' content; (4) were poorly extracted, not in the target language; or (5) consisted mostly of special characters or numbers.

As a document might have multiple registers, a multi-label tagging approach was taken with a holistic perspective. The proportion of a register across the whole document has been considered for the multi-label tagging. A hybrid document could be in multiple forms. One type is a single coherent document about a topic that has multiple registers. Another is a document consisting of short paragraphs about different topics or different aspects of a topic that are united.

In addition to register labels, a separate "Tags" column is defined for metadata. For accepted documents, this column can be empty, "OTHLI" (code-switching to other languages such as English or Arabic), or "FINGLISH" (Persian transliterated in the Latin alphabet). Rejected documents are tagged as "JNK" (junk), "MLPT" (more than three main registers), or "OTHL" (text in another language).

"MLPT" texts are longer sections where coherence may be lacking, and the register can shift, sometimes including more than three distinct registers. "JNK"texts are short, headline-like sentences that, while complete, cannot establish a register. Although texts with excessive main registers may still be suitable for next-token prediction tasks, they do not constitute coherent register instances and were therefore excluded from register annotation. Examples of the most frequent Tags are presented in Table 2. Figure 1 shows the proportion of tags within each main register, highlighting that code-switching varies across registers. Code-switching is less frequent in the Narrative register, but more common in How-to and Informational Persuasion.

In the first batch of 1,100 randomly extracted instances from HPLT3, we observed a highly skewed class distribution, with some main registers occurring fewer than 100 times.. To address this imbalance, we applied an upsampling strategy. URLs of rejected texts were excluded from further sampling, and a minimum of 145 instances was set for underrepresented register classes. To reach this threshold, we selected instances from HPLT3 with automatic labeling confidence between 0.2 and 0.4. These scores fall below the 0.4 confidence threshold established in Burchell et al. (2025) and were intended to target difficult cases for manual annotation. Table 1 presents the hierarchical register schema along with instance counts from random sampling and after upsampling; hybrid documents were counted multiple times, once for each register they contain.

4 Methods

4.1 Keyness

In this section, we focus on how keyness was calculated for the registers in our corpus. We use the term "token" rather than "word" because text was segmented by whitespace rather than processed using a linguistically informed tokenization method. Considering the nature of the Persian writing sys-

Translation	Text	TAG
*[…]*Crowdedness, reason for not burying Morteza Pashaei Ali Daei's resemblance to the Hollywood actor + Photo The Asian champion peddling on the street*[…]*	ازدحام جمعیت عامل عدم دفن مرتضی پاشایی شباهت علی دایی با بازیگر هالیود + عکس قهرمان آسیا در کنار خیابان دست فروشی می کند	JNK
Puffer fish contain a toxin called tetrodotoxin, which is 1,200 times more deadly to humans than cyanide*[…]* It's not fair that in this big city, you are the proverbial needle in a haystack. Three days ago, after making the necessary plans, Persepolis Club sent this program to the representatives of the Milan Stars team, and *[…]*	ماهی پف کننده حاوی سمی به نام تتراتوکسین است که برای انسانها 1200 بار کشنده تر از سم سیانور می باشد... اِنصاف نباشد که در این شهرِ دَرَندَشت ضرب المثل سوزن در کاه تو باشی. باشگاه پرسپولیس سه روز پیش پس از برنامه ریزي هاي لازم اين برنامه را براي نمايندگان تيم ستارگان ميلان ارسال كرد و...	MLPT
*[…]*The rotation speed of this hard drive is 7200 revolutions per minute (RPM) and uses a high-speed SATA III interface with a speed of 6 Gbps for data transfer. -Memory: 2 TB -Size: 3.5 inch -Connection: SATA 3.0	...سرعت چرخش این هارد درایو 7200 دور در دقیقه (RPM) بوده و از رابط پرسرعت $SATAIII$ با سرعت 6 گیگابیت بر ثانیه، برای انتقال اطلاعات بهره می برد -Memory: 2 TB -Size: 3.5 inch -Connection: SATA 3.0	OTHLI

Table 2: Tags Examples.

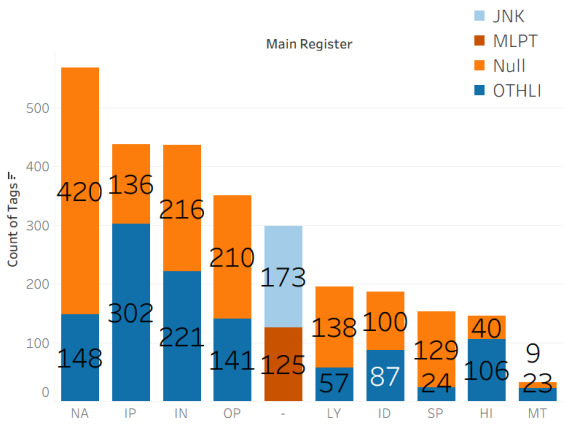


Figure 1: Distribution of Tags on registers

tem, where letters may be concatenated or written separately depending on orthographic conventions, certain elements (e.g., the present tense marker "می") may appear as independent tokens, and a single word may be split across multiple tokens. For each token, both document frequency and term frequency were calculated. Tokens were then ranked according to their log likelihood (G^2) scores to identify most distinctive vocabulary of each register.

4.2 Register labeling

Since web documents frequently exhibit characteristics of multiple registers, we formulate register identification as a multi-label classification task.

Automatic register labeling was performed using XLM-R Large (Conneau et al., 2020) and BGE-M3 (Chen et al., 2024), with maximum input sequence lengths of 512 and 1024 tokens, respectively. These models were selected for two main reasons. First, XLM-R Large has been shown to achieve state-of-the-art performance in multilingual web register identification, consistently outperforming monolingual and smaller multilingual

models across a wide range of languages and experimental settings (Henriksson et al., 2024). Second, XLM-R Large and BGE-M3 provide strong multilingual representations. In addition, BGE-M3 was included for its longer context window (up to 8,192 tokens), which is particularly important for register classification, as register cues may be distributed across extended document spans rather than concentrated locally.

We evaluated the models under four training and evaluation settings:

1. **Monolingual**: trained and evaluated on Persian.

2. **Multilingual-1**: trained on a mixture of Persian and Multilingual CORE, evaluated on Persian.

3. **Multilingual-2**: trained and evaluated on a mixture of Persian and Multilingual CORE.

4. **Zero-Shot**: trained on Multilingual CORE and evaluated on Persian.

The data was split for training and evaluation. In the Persian-only setting, 70% of the data was used for training, while the remaining 30% was reserved for evaluation purposes and model tuning. In the zero-shot setting, 80% of the data was allocated for training. For the Multilingual-1 setting (see Section 4.2), the full multi-CORE corpus was combined with 50% of the Persian data for training, with the remaining Persian data used for evaluation. In the Multilingual-2 setting, 90% of the multi-CORE corpus was combined with 50% of the Persian data for training, and the remaining data was held out for evaluation.

Model optimization was carried out using Bayesian hyperparameter optimization with the Optuna default optimizer, which is the Tree-structured Parzen Estimator (TPE) sampler. Continuous hyperparameters were sampled from predefined ranges, including the learning rate $[5 \times 10^{-6}, 3 \times 10^{-5}]$ on a logarithmic scale, weight decay $[0.0, 0.1]$, and warmup ratio $[0.0, 0.15]$. Discrete hyperparameters included batch size $4, 8$ and gradient accumulation steps $4, 8$. The maximum number of training epochs was set to 10.

Register	Translation	Keyword	Keyness
NA_ne	Report	گزارش	109.884
NA_ne	Added	افزود	103.847
NA_ne	He/She	وی	82.510
NA_ne	Deputy	معاون	81.432
NA_ne	Reporter	خبرنگار	79.825
NA_ne	Head of/Boss	رئیس	69.856
NA_ne	Specify/State	تصریح	66.140
NA_ne	Minister	وزیر	64.012
NA_ne	(X) Council	شورای	56.164
NA_ne	Ministry	وزارت	53.543
IP_ds	Product	محصول	76.842
IP_ds	Comments/Opinions	دیدگاهی	50.065
IP_ds	Products	محصولات	40.0852
IP_ds	Specifications	مشخصات	37.653
IP_ds	Capability/Capable	قابلیت	33.982
IP_ds	Product	کالا	29.970
IP_ds	Dimensions	ابعاد	29.370
IP_ds	Material	جنس	29.200
IP_ds	Frame/Body	بدنه	28.821
IP_ds	Warranty	ضمانت	28.375
IP_oe	Download	دانلد	138.783
IP_oe	(Down)Load	لوڈ	138.783
IP_oe	Download	داونلود	138.783
IP_oe	Download	ddl	138.783
IP_oe	Down(load)	ڈاؤن	138.783
IP_oe	Download	donload	138.783
IP_oe	usnet	usnet	138.783
IP_oe	Download	nhkgn	138.783
IP_oe	Download	danload	138.783
IP_oe	uznet	یوزنت	138.783
IN_dtp	Qajar	قاجار	20.471
IN_dtp	1080p	1080p	19.778
IN_dtp	Herbert	هربرت	19.778
IN_dtp	Jordan	اردن	17.191
IN_dtp	It has been	شدهاست	15.560
IN_dtp	Architecture	معماری	14.793
IN_dtp	Fame	شهرت	13.412
IN_dtp	2020	2020	13.363
IN_dtp	(X) Studio	استودیوی	13.363
IN_dtp	Small stick (Gillidanda)	پیل	13.199
OP_av	Benefits	فواید	39.407
OP_av	Hormonal	هورمونی	33.001
OP_av	Don't	نکنید	31.440
OP_av	Avoid	اجتناب	30.996
OP_av	Fat	چربی	30.0467
OP_av	Diabetes	دیابت	29.656
OP_av	Body	بدن	29.259
OP_av	Do/Give	دهید	27.824
OP_av	Inflammatory	التهابی	24.363
OP_av	You can	بتوانید	23.874

Table 3: 10 top keywords from the five highest-percentage registers

5 Evaluation

5.1 Keyness

Table 3 presents the top ten keywords for the most frequent registers in our data: News, Description with Intent to Sell, Other Informational Persuasion (which broadly persuades but lacks characteristics of any more specific register), Description of a Thing or Person, and Advice. These registers were selected for analysis based on their frequency; the

top 20 registers are presented in Table 4, showing that hybrids are less common. The full distribution of registers is available in the Appendix, Table 8. The results exhibit a long-tail distribution, in which primary registers account for most documents, while diverse hybrid combinations occur less frequently.

Register	Percentage
-	14.95%
NA_ne	10.46%
IP_ds	5.80%
IP_oe	4.15%
IN_dtp	3.70%
OP_av	3.05%
LY	3.00%
NA_on	2.90%
SP_os	2.30%
SP_it	1.65%
MT	1.55%
IN_ra	1.35%
NA_nb	1.35%
ID	1.30%
NA_sr	1.20%
LY+ID	1.10%
IN_dtp+OP_av	1.00%
IP_ds+IN_fi	0.90%
NA_on+LY	0.85
IN_oi	0.85%

Table 4: Distribution of registers.

In the "News" register, most keywords are nouns related to administration and governance, a pattern also observed for Turkish (Erten, 2023). In contrast to findings for Turkish and French (Hellström et al., 2025), our analysis did not identify past-tense verbs as salient keywords. This absence may be due to tokenization issues in Persian, where verb constructions can be split into multiple tokens by spaces or half-spaces. For example, the noun "گزارش" ('report') becomes a verb in the construction "گزارش کرد" ('reported'), which may prevent such forms from being captured as single verbal units in keyword analysis. Additionally, the pronoun "وی", equivalent to "he" or "she", is characteristic of narrative usage, where it is commonly employed in recounting events. The same characteristic of the Narrative register occurs in English, concerning recounting events (Biber and Egbert, 2018).

In the "Description with the intention to sell" class, a product or a service is described, with the purpose of selling and overt marketing. The keywords include nouns related to products or their features, such as "جنس" {'Material'}, "ابعاد" {'Dimensions' }, and "مشخصات"" {'specifications'}. Unlike English product descriptions, which tend to make extensive use of adjectives and evaluative language (Biber and Egbert, 2018), Persian tends toward noun repetition, despite the grammatical possibility of noun-adjective modification.

The "Other informational persuasion" register is characterized by frequent occurrences of the word "download", either as a single token or split according to orthographic conventions, and contains the highest proportion of non-Persian tokens. The token "nhkgn" represents the Persian word for "download" typed with a Persian keyboard while the input language remains English, a practice common on software distribution websites. Similarly, "usnet" and "uznet" are alternative transliterations of "یوزنت", a broadband service provider.

In the "Description of a thing or a person" register, proper names such as Qajar, Jordan, and Herbert appear among the keywords. Some items point to Persian-specific historical or cultural references. For example, "Qajar" refers to a formerly aristocratic Iranian dynasty that ruled Iran from 1789 to 1925, while "Jordan" reflects the long, complex historical relationship between Iran and Jordan, dating back to the Achaemenid Empire according to Wikipedia. The word "پیل" denotes the "Gilli" in the traditional game Gillidanda "پیل دسته", also known as "الک دولک".

Other tokens likely reflect sampling randomness rather than register-specific properties. For instance, "Herbert" appears in reference to multiple entities, including Herbert Le Porrier, Herbert Hoover, and the Herbert Berghof Studio. Similarly, tokens such as "1080p" and "2020" typically indicate video resolution or production year, reflecting broader media-description conventions rather than Persian-specific features.

In the "Advice" register, which is based on opinions that suggest actions to solve a particular problem, keywords are primarily health-related nouns, content-focused and topic-driven, rather than narrative. At the same time, the relatively high diversity of verbs reflects the directive nature of the advice register, where actions and recommendations are foregrounded, often through imperative forms (e.g., negative imperatives such as "don't"). Moreover, lexical borrowing with phonological adapta-

tion has happened to health-related nouns, for instance, like “Hormonal” and “Diabetes”, instead of writing them in English, and in contrast with the “Informational Persuasion” category.

It is essential to note that the symbol (X) in Table 3 marks the presence of the suffix ی, which allows a word to function as a head noun and take dependents such as adjectives or nouns. However, this only happens when a word ends with a vowel, such as “شورا” {“Council”}, while for other words like “وزارت” {“Ministry”} that end with a consonant, adding a {/e/} vowel at the end of the word will do the same. Moreover, in the Persian orthography of Iran, writing vowels {/æ/}, {/e/}, and {/o/} is omitted. Therefore, it cannot be generalized that using head nouns that take dependents is much frequent in a specific register by just looking at separate words.

Taken together, the analysis reveals both features that align with patterns reported for other languages and keywords that reflect language-specific characteristics of Persian. This highlights the need for language-specific approaches for modeling registers.

5.2 Register labeling

Model(max_lenght)	Setting	F1 Score(μ)
XLM-R (512)	Monolingual	0.72
XLM-R (512)	Multilingual-1	0.75
XLM-R (512)	Multilingual-2	0.75
XLM-R (512)	Zero-shot	0.76
bge-m3 (1024)	Monolingual	0.74
bge-m3 (1024)	Multilingual-1	0.75
bge-m3 (1024)	Multilingual-2	0.75
bge-m3 (1024)	Zero-shot	0.76

Table 5: Model performances.

We assess model performance using micro F1, computed by aggregating true positives, false positives, and false negatives across all labels before calculating precision and recall. In a multi-label setting with imbalanced class distributions, micro F1 is well-suited as a performance measure, as it weights labels proportionally to their frequency and reflects the model’s effectiveness across all individual label decisions. The results of the automatic register labeling are presented in Table 5. The multilingual and zero-shot settings achieved comparable performance, with micro F1 scores close to 0.75. The multilingual training improved model performance, with the monolingual Persian-only model achieving 0.72 micro F1.

In addition to the overall micro F1 score, the classification report in Table 6 shows that performance varies substantially across registers. Frequent and well-represented registers (e.g., NA, IP) achieve higher F1 scores, while sparse registers (e.g., ed, en, lt) show lower performance. Among the main registers, NA (F1 = 0.84), SP (F1 = 0.92), and IP (F1 = 0.74) show a relatively balanced precision–recall trade-off, indicating stable performance for these categories. In contrast, other main registers exhibit larger discrepancies between precision and recall, suggesting less consistent detection. General performance on subregisters is weaker and more variable. Subregister classes such as ds (F1 = 0.82) and ne (F1 = 0.85) achieve comparatively strong results with a balanced precision–recall trade-off. However, several low-resource subregisters show low F1 scores and large precision–recall imbalances. In several low-support classes, high precision but low recall suggests the model makes conservative predictions, reflecting the difficulty of learning reliable decision boundaries in a multi-label setting. Results for classes with very limited support are less reliable and further highlight the difficulty of fine-grained subregister classification under data-sparse conditions. Therefore, additional and cleaner data might be required to achieve better performance.

Previously, Henriksson et al. (2024) reported a best average micro F1 of 0.77 across all languages (0.79 on main labels) using XLM-R, XLMR-XL, and BGE-M3 (2048 tokens) in the multilingual setting. However, the multilingual CORE dataset is approximately ten times larger than the Persian corpus, and Persian may exhibit distinctive linguistic features. Moreover, even among the best models in Henriksson et al. (2024), performance varies substantially across languages, with micro F1 scores ranging from 0.72 to 0.81 (Table 3).

In addition, based on manual analysis, many instances in our Persian sample from HPLT3 contained noisy content, such as irrelevant page tags generated for search engine indexing, short clickable links to other pages, and brief advertising texts. This noise may have hindered model learning and introduced confusion, even when longer context windows were used.

class	precision	recall	f1-score	support
HI	0.59	0.83	0.69	35
ID	0.80	0.68	0.74	47
IN	0.68	0.62	0.65	110
IP	0.73	0.76	0.74	111
LY	0.87	0.94	0.90	50
MT	1.00	0.33	0.50	9
NA	0.84	0.85	0.84	139
OP	0.62	0.49	0.55	87
SP	0.92	0.92	0.92	38
av	0.65	0.54	0.59	37
ds	0.80	0.85	0.82	52
dtp	0.69	0.39	0.50	74
ed	0.50	0.27	0.35	11
en	0.00	0.00	0.00	3
fi	1.00	0.45	0.62	11
it	0.88	0.79	0.83	19
lt	0.50	0.33	0.40	3
nb	0.87	0.76	0.81	17
ne	0.86	0.85	0.85	78
ob	0.60	0.25	0.35	12
ra	0.83	0.62	0.71	8
re	0.71	1.00	0.83	5
rs	0.67	0.25	0.36	16
rv	1.00	0.62	0.77	8
sr	0.89	0.80	0.84	10

Table 6: Classification report

6 Conclusion

This study addresses a notable gap in register-based research by focusing on Persian, a digitally low-resource language that has so far been underrepresented in large-scale web register corpora and multilingual register identification efforts. By combining manual annotation with computational modeling, the work contributes both empirical data and methodological insights relevant to cross-linguistic register analysis. We presented ParsCORE, the first large-scale, human-annotated corpus of Persian web registers, and evaluated register identification models under three different settings as an initial step toward automatic identification. In addition, we conducted keyword analysis across major registers. The results show that model performance is comparable to that reported for digitally high-resource languages. However, additional data is required, and further investigation is needed to identify potential linguistic and cultural specificities of Persian registers. The keyness analysis provides insights into differences across registers and categories. The manually annotated dataset and the model optimization pipeline are publicly available on GitHub. Taken together, ParsCORE provides a foundation for future research on Persian web registers, including the development of more robust multi-label models, improved handling of morphologically complex constructions, and broader cross-lingual comparisons involving both high- and low-resource languages.

Limitations and further work

Regarding inter-annotator agreement, this study follows the taxonomy and annotation schema established in prior work to ensure methodological consistency. Rather than conducting full parallel annotation, only challenging or ambiguous instances were discussed with experts who had previously applied this framework to other languages. This approach was adopted due to the difficulty of identifying experts with specific experience in Persian register variation who are also available and willing to perform manual annotation. Consequently, formal inter-annotator agreement measures were not calculated. Another limitation is that only the beginnings of documents (512 or 1024 tokens) were used due to token constraints. Future research should explore the use of document endings, combined beginning–end segments, and alternative windowing approaches.

Future studies could apply clustering methods to the semantic embeddings of classified documents in order to elucidate register relationships and to qualitatively investigate both correctly classified and misclassified documents (Santini, 2005; Gries, 2021). Moreover, keyness analysis was limited to the top 10 keywords for the five most frequent registers. Future work could extend this analysis to include all registers and a larger set of keywords, providing a more comprehensive view of register-specific lexical patterns across the corpus.

Acknowledgments

Alireza Razzaghi received funding from the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No 101177564—HAIF. Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. Veronika Laippala and Erik Henriksson received funding from the Research Council of Finland through "FIN-CLARIAH research infrastructure" (project 358720, which has also received funding from the European Union – NextGenera-

tionEU instrument), "Mechanisms of register variation in massively multilingual web-scale corpora" (project 362459), and "Green NLP - controlling the carbon footprint in sustainable language technology" (project 353167). Furthermore, we also wish to acknowledge CSC – IT Center for Science Ltd. for providing computational resources.

References

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*. Cambridge University Press, Cambridge.

Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press.

Douglas Biber and Jesse Egbert. 2023. What is a register? accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies*, 5(1):1–22.

Douglas Biber, Jesse Egbert, and Mark Davies. 2015. Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora*, 10(1):11–45.

Douglas Biber, Jesse Egbert, and Daniel Keller. 2020. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16(3):581–616.

Marina Bondi. 2010. *Perspectives on keywords and keyness: An introduction*, pages 1–18. John Benjamins Publishing Company.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jonathan Culpeper and J. Demmen. 2015. *Keywords*, pages 90–105.

Jesse Egbert and Doug Biber. 2018. Incorporating text dispersion into keyword analyses. *Corpora*, 14(1):77–104. Publisher Copyright: © Edinburgh University Press.

Selcen Erten. 2023. Exploring register variation in turkish web corpus. *14–15 September 2023, University of Mannheim, Germany*, page 60.

Selcen Erten-Johansson, Valtteri Skantsi, Sampo Pyysalo, and Veronika Laippala. 2024. Linguistic variation beyond the Indo-European Web: Analyzing Turkish Web registers in TurCORE. *Register Studies*.

Stefan Gries. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9:1–33.

Saara Hellström, Valtteri Skantsi, Anna Salmela, and Veronika Laippala. 2025. From keywords to key embeddings – contrasting french and swedish web registers using multilingual deep learning. *Corpus Linguistics and Linguistic Theory*.

Erik Henriksson, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. 2024. Automatic register identification for the open web using multilingual deep learning. *arXiv preprint arXiv:2406.19892*.

Taja Kuzman and Nikola Ljubešić. 2023. Automatic genre identification: A survey. *Language Resources and Evaluation*.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO training dataset for web genre identification of documents out in the wild. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1584–1594, Marseille, France. European Language Resources Association.

Mikhail Lepekhin and Serge Sharoff. 2021. Experiments with adversarial attacks on text genres. *CoRR*, abs/2107.02246.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. 2025. Register always matters: Analysis of LLM pretraining data through the lens of language variation. In *Second Conference on Language Modeling*.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the english web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*.

Marina Santini. 2005. Genres in formation? An exploratory study of web pages using cluster analysis: Proceedings of the 8th annual colloquium for the UK special interest group for computational linguistics (CLUK05). In *Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK05)*, Manchester.

Mike Scott. 1997. Pc analysis of key words — and key key words. *System*, 25(2):233–245.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Valtteri Skantsi and Veronika Laippala. 2023. Analyzing the unrestricted web: The finnish corpus of online registers. *Nordic Journal of Linguistics*, page 1–31.

Lukas Sönning. 2023. Evaluation of keyness metrics: performance and reliability. *Corpus Linguistics and Linguistic Theory*, 20.

A Appendix

Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb
Other narrative	on
Opinion	OP
Review	rv
Opinion blog	ob
Advice	av
Denominational religious blog / sermon	rs
Other opinion	oo
Informational Description	IN
Description of a thing or a person	dtp
Research article	ra
Encyclopedia article	en
FAQ	fi
Legal terms and conditions	lt
Other Informational description	oi
Interactive Discussion	ID
How-to	HI
Recipe, Other how-to	
Informational Persuasion	IP
Description with intent to sell	ds
News & opinion blog or editorial	ed
Other informational persuasion	oe
Lyrical	LY
Spoken	SP
Interview	it
Other spoken	os
Machine Translated	MT
Tags	
Junk	JNK
Other language included	OTHLI
Other language	OTHL

Table 7: Registers abbreviation

Table 8: Distribution of registers

Register	Count	Percentage	Tag	Tag Count	Tag Percentage
-	173	8.65%	JNK	173	57.86%
-	125	6.25%	MLPT	125	41.81%
-	1	0.05%	OTHL	1	0.33%
NA_ne	209	10.46%	OTHLI	40	19.14%
IP_ds	116	5.80%	OTHLI	84	72.41%
IP_oe	83	4.15%	OTHLI	67	80.72%
IN_dtp	74	3.70%	OTHLI	43	58.11%
OP_av	61	3.05%	OTHLI	20	32.79%
LY	60	3.00%	OTHLI	18	30.00%
NA_on	58	2.90%	OTHLI	16	27.59%
SP_os	46	2.30%	OTHLI	3	6.52%
SP_it	33	1.65%	OTHLI	4	12.12%
MT	31	1.55%	OTHLI	22	70.97%
IN_ra	27	1.35%	OTHLI	13	48.15%
NA_nb	27	1.35%	OTHLI	1	3.70%
ID	26	1.30%	OTHLI	14	53.85%
NA_sr	24	1.20%	OTHLI	4	16.67%
LY+ID	22	1.10%	OTHLI	2	9.09%
IN_dtp+OP_av	20	1.00%	OTHLI	8	40.00%
IP_ds+IN_fi	18	0.90%	OTHLI	14	77.78%
NA_on+LY	17	0.85%	OTHLI	2	11.76%
IN_oi	17	0.85%	OTHLI	5	29.41%
OP_ob	16	0.80%	OTHLI	5	31.25%
IP_oe+LY	16	0.80%	OTHLI	15	93.75%
OP_oo	14	0.70%	OTHLI	2	14.29%
NA_ne+IN_dtp	14	0.70%	OTHLI	5	35.71%
IP_oe+HI_oh	14	0.70%	OTHLI	14	100.00%
IP_ds+ID	13	0.65%	OTHLI	9	69.23%
HI_oh	13	0.65%	OTHLI	12	92.31%
IN_dtp+HI_oh	13	0.65%	OTHLI	11	84.62%
IP_ed	13	0.65%	OTHLI	3	23.08%
IN_dtp+LY	13	0.65%	OTHLI	4	30.77%
IN_dtp+ID	12	0.60%	OTHLI	5	41.67%
HI_re	11	0.55%	OTHLI	1	9.09%
IP_oe+ID	11	0.55%	OTHLI	9	81.82%
IP_ds+HI_oh	10	0.50%	OTHLI	7	70.00%
NA_nb+LY	10	0.50%	OTHLI	2	20.00%
OP_rs	9	0.45%	OTHLI	3	33.33%
LY+IN_dtp	9	0.45%	OTHLI	1	11.11%
IP_ds+IN_dtp	9	0.45%	OTHLI	6	66.67%
OP_av+ID	9	0.45%	OTHLI	4	44.44%
HI_oh+ID	9	0.45%	OTHLI	8	88.89%
NA_ne+IN_oi	8	0.40%	OTHLI	3	37.50%
NA_ne+OP_rs	8	0.40%	OTHLI	2	25.00%
NA_on+OP_oo	8	0.40%	OTHLI	4	50.00%
OP_rv	8	0.40%	OTHLI	4	50.00%
IN_dtp+IP_ds	7	0.35%	OTHLI	6	85.71%
IN_dtp+OP_rs	7	0.35%	OTHLI	3	42.86%
IP_oe+IN_dtp	7	0.35%	OTHLI	3	42.86%

Register	Count	Percentage	Tag	Tag Count	Tag Percentage
SP_os+OP_rs	7	0.35%	OTHLI	2	28.57%
IN_en	7	0.35%	OTHLI	5	71.43%
NA_ne+IP_ed	6	0.30%	OTHLI	3	50.00%
NA_on+OP_av	6	0.30%	OTHLI	4	66.67%
NA_on+IN_dtp	6	0.30%	OTHLI	4	66.67%
IN_dtp+NA_on	6	0.30%	OTHLI	1	16.67%
OP_ob+ID	6	0.30%	OTHLI	2	33.33%
IN_dtp+IP_oe	6	0.30%	OTHLI	5	83.33%
NA_ne+OP_av	5	0.25%	OTHLI	2	40.00%
IN_dtp+OP_rv	5	0.25%	OTHLI	4	80.00%
NA_nb+NA_on	5	0.25%	OTHLI	1	20.00%
NA_ne+SP_it	5	0.25%	OTHLI	2	40.00%
NA_ne+OP_oo	5	0.25%	OTHLI	2	40.00%
IP_ed+SP_it	5	0.25%	OTHLI	1	20.00%
IN_dtp+OP_ob	4	0.20%	OTHLI	3	75.00%
NA_on+SP_it	4	0.20%	OTHLI	2	50.00%
LY+OP_rs	4	0.20%	OTHLI	2	50.00%
OP_av+IP_oe	4	0.20%	OTHLI	2	50.00%
NA_nb+OP_rs	4	0.20%	OTHLI	1	25.00%
OP_av+HI_oh	4	0.20%	OTHLI	2	50.00%
NA_on+OP_rs	4	0.20%	OTHLI	2	50.00%
HI_oh+IN_dtp	4	0.20%	OTHLI	3	75.00%
IP_oe+NA_sr	4	0.20%	OTHLI	4	100.00%
IN_oi+OP_av	3	0.15%	OTHLI	2	66.67%
NA_nb+OP_oo	3	0.15%	OTHLI	1	33.33%
LY+HI_oh	3	0.15%	OTHLI	3	100.00%
IP_oe+SP_it	3	0.15%	OTHLI	2	66.67%
IP_ed+ID	3	0.15%	OTHLI	1	33.33%
IN_oi+ID	3	0.15%	OTHLI	1	33.33%
IN_dtp+HI_oh+ID	3	0.15%	OTHLI	3	100.00%
OP_oo+ID	3	0.15%	OTHLI	1	33.33%
NA_nb+ID	3	0.15%	OTHLI	1	33.33%
HI_oh+IN_oi	3	0.15%	OTHLI	1	33.33%
IN_dtp+SP_os	3	0.15%	OTHLI	2	66.67%
IN_dtp+HI_oh+OP_av	3	0.15%	OTHLI	3	100.00%
OP_av+HI_re	3	0.15%	OTHLI	3	100.00%
IP_oe+NA_ne	3	0.15%	OTHLI	2	66.67%
IP_ds+NA_on	2	0.10%	OTHLI	1	50.00%
IN_dtp+IN_lt	2	0.10%	OTHLI	1	50.00%
ID+IN_dtp+OP_av	2	0.10%	OTHLI	1	50.00%
IN_oi+IP_oe	2	0.10%	OTHLI	1	50.00%
OP_rv+NA_on	2	0.10%	OTHLI	1	50.00%
NA_on+IP_oe	2	0.10%	OTHLI	2	100.00%
IP_ds+OP_av	2	0.10%	OTHLI	1	50.00%
OP_rs+LY	2	0.10%	OTHLI	1	50.00%
IP_ds+LY	2	0.10%	OTHLI	1	50.00%
IP_ds+OP_rv+ID	2	0.10%	OTHLI	1	50.00%
IN_fi	2	0.10%	OTHLI	1	50.00%
NA_on+ID	2	0.10%	OTHLI	1	50.00%
IP_oe+IN_oi+HI_oh	2	0.10%	OTHLI	1	50.00%
IP_ds+IN_oi	2	0.10%	OTHLI	1	50.00%

Register	Count	Percentage	Tag	Tag Count	Tag Percentage
IP_ds+HI_re	2	0.10%	OTHLI	2	100.00%
IP_ds+HI_oh+ID	2	0.10%	OTHLI	2	100.00%
OP_oo+HI_oh	2	0.10%	OTHLI	2	100.00%
IP_oe+HI_oh+ID	2	0.10%	OTHLI	2	100.00%
IN_fi+IP_ds	2	0.10%	OTHLI	2	100.00%
IP_ds+IN_en	2	0.10%	OTHLI	1	50.00%
IN_dtp+NA_ne	2	0.10%	OTHLI	1	50.00%
IP_ds+IP_oe	2	0.10%	OTHLI	1	50.00%
OP_av+IN_dtp	2	0.10%	OTHLI	1	50.00%
OP_ob+NA_ne	2	0.10%	OTHLI	1	50.00%
NA_sr+IP_oe	2	0.10%	OTHLI	1	50.00%
NA_ne+IP_oe	2	0.10%	OTHLI	2	100.00%
IN_ra+IP_ds	1	0.05%	OTHLI	1	100.00%
OP_av+OP_rs	1	0.05%	OTHLI	1	100.00%
SP_os+OP_rs+IN_dtp	1	0.05%	OTHLI	1	100.00%
IN_dtp+IN_fi+IP_ds	1	0.05%	OTHLI	1	100.00%
NA_ne+IN_lt	1	0.05%	OTHLI	1	100.00%
SP_it+OP_oo	1	0.05%	OTHLI	1	100.00%
NA_on+IN_dtp+OP_oo	1	0.05%	OTHLI	1	100.00%
NA_on+LY+NA_nb	1	0.05%	OTHLI	1	100.00%
OP_av+OP_oo	1	0.05%	OTHLI	1	100.00%
NA_nb+NA_on+OP_ob	1	0.05%	OTHLI	1	100.00%
OP_oo+NA_on	1	0.05%	OTHLI	1	100.00%
LY+HI_oh+IP_oe	1	0.05%	OTHLI	1	100.00%
OP_rv+LY+ID	1	0.05%	OTHLI	1	100.00%
HI_oh+IP_oe+LY	1	0.05%	OTHLI	1	100.00%
IN_dtp+IP_ds+LY	1	0.05%	OTHLI	1	100.00%
IN_en+ID	1	0.05%	OTHLI	1	100.00%
IN_oi+ID+HI_oh	1	0.05%	OTHLI	1	100.00%
IN_dtp+ID+OP_rv	1	0.05%	OTHLI	1	100.00%
NA_nb+HI_oh+ID	1	0.05%	OTHLI	1	100.00%
IP_oe+IN_dtp+ID	1	0.05%	OTHLI	1	100.00%
ID+OP_rs	1	0.05%	OTHLI	1	100.00%
NA_nb+OP_oo+ID	1	0.05%	OTHLI	1	100.00%
OP_rs+IN_oi+ID	1	0.05%	OTHLI	1	100.00%
ID+IN_oi+OP_oo	1	0.05%	OTHLI	1	100.00%
IP_ds+ID+OP_rv	1	0.05%	OTHLI	1	100.00%
HI_oh+OP_av	1	0.05%	OTHLI	1	100.00%
NA_ne+SP_os+OP_oo	1	0.05%	OTHLI	1	100.00%
OP_av+NA_ne	1	0.05%	OTHLI	1	100.00%
NA_nb+SP_it+IN_dtp	1	0.05%	OTHLI	1	100.00%
OP_rs+SP_os	1	0.05%	OTHLI	1	100.00%
SP_os+IP_ds	1	0.05%	OTHLI	1	100.00%
HI_oh+ID+IN_dtp	1	0.05%	OTHLI	1	100.00%
IN_oi+HI_oh	1	0.05%	OTHLI	1	100.00%
IN_fi+HI_oh+IP_oe+NA_ne	1	0.05%	OTHLI	1	100.00%
IN_fi+HI_oh	1	0.05%	OTHLI	1	100.00%
IP_oe+HI_oh+IN_oi	1	0.05%	OTHLI	1	100.00%
IN_fi+HI_oh+ID	1	0.05%	OTHLI	1	100.00%
IN_dtp+IP_oe+HI_oh	1	0.05%	OTHLI	1	100.00%
NA_on+HI_oh	1	0.05%	OTHLI	1	100.00%

Register	Count	Percentage	Tag	Tag Count	Tag Percentage
HI_re+IN_dtp	1	0.05%	OTHLI	1	100.00%
IP_ds+IN_dtp+ID	1	0.05%	OTHLI	1	100.00%
IN_dtp+OP_av+HI_re	1	0.05%	OTHLI	1	100.00%
HI_oh+IN_dtp+IN_fi	1	0.05%	OTHLI	1	100.00%
NA_nb+HI_re+IP_oe	1	0.05%	OTHLI	1	100.00%
IN_dtp+IN_fi+HI_oh+ID	1	0.05%	OTHLI	1	100.00%
NA_on+IN_oi+HI_oh+ID	1	0.05%	OTHLI	1	100.00%
HI_oh+IN_fi+ID	1	0.05%	OTHLI	1	100.00%
OP_ob+HI_oh	1	0.05%	OTHLI	1	100.00%
IN_dtp+HI_oh+IP_ds	1	0.05%	OTHLI	1	100.00%
OP_av+IN_dtp+LY	1	0.05%	OTHLI	1	100.00%
NA_ne+IN_dtp+ID	1	0.05%	OTHLI	1	100.00%
IN_dtp+IN_lt+NA_ne	1	0.05%	OTHLI	1	100.00%
NA_ne+IN_oi+OP_oo	1	0.05%	OTHLI	1	100.00%
NA_ne+IN_dtp+OP_oo	1	0.05%	OTHLI	1	100.00%
NA_ne+OP_ob	1	0.05%	OTHLI	1	100.00%
IN_ra+IP_oe	1	0.05%	OTHLI	1	100.00%
IP_ds+OP_rs	1	0.05%	OTHLI	1	100.00%
IN_ra+HI_oh	1	0.05%	OTHLI	1	100.00%
OP_rs+IN_dtp	1	0.05%	OTHLI	1	100.00%
OP_rs+NA_nb	1	0.05%	OTHLI	1	100.00%
OP_rv+ID	1	0.05%	OTHLI	1	100.00%
OP_av+NA_on	1	0.05%	OTHLI	1	100.00%
IP_oe+OP_rv	1	0.05%	OTHLI	1	100.00%
ID+HI_oh	1	0.05%	OTHLI	1	100.00%
NA_ne+OP_rs+HI_oh	1	0.05%	OTHLI	1	100.00%
OP_av+HI_oh+IP_oe	1	0.05%	OTHLI	1	100.00%
NA_nb+OP_av	1	0.05%	OTHLI	1	100.00%
NA_on+IP_ds+IN_dtp	1	0.05%	OTHLI	1	100.00%
IP_ds+MT	1	0.05%	OTHLI	1	100.00%
IP_ed+OP_rs	1	0.05%	OTHLI	1	100.00%
OP_rs+OP_ob	1	0.05%	OTHLI	1	100.00%
IP_ds+OP_rv	1	0.05%	OTHLI	1	100.00%

PMWP: A Benchmark for Math Word Problem Solving in Persian

Marzieh Abdolmaleki, Mehrnoush Shamsfard†, Veronique Hoste and Els Lefever
LT3 – Ghent University, Belgium
†NLP Lab – Shahid Beheshti University, Iran
{marzieh.maleki,veronique.hoste,els.lefever}@ugent.be
m-shams@sbu.ac.ir

Abstract

Mathematical reasoning captures fundamental aspects of human cognitive ability. Although recent advances in LLMs have led to substantial improvements in automated mathematical problem solving, most existing benchmarks remain focused on English. As a result, robust mathematical reasoning remains a challenging and insufficiently explored capability for underrepresented languages including Persian. To address this gap, we introduce PMWP, the first dataset of 15K elementary-level Persian math word problems that supports both supervised training and evaluation of reasoning models. By expanding mathematical reasoning resources beyond English, PMWP contributes to the development of multilingual AI systems with stronger reasoning capabilities. In this work, we conduct a systematic evaluation of the Persian math word problem solving capabilities of different state-of-the-art LLMs. Our results indicate that DeepSeek-V3 exhibits reduced language bias when problem texts are translated into English, while Gemini-2.5-Flash achieves the highest equation value accuracy (72.02%) in Persian. In addition, we investigate parameter-efficient adaptation for equation generation by applying LoRA-based fine-tuning to LLaMA-3-8B and Qwen-2.5-7B. Our results show that, following fine-tuning, these open-weight models achieve 91.65% and 92.53% exact equation match accuracy, respectively. Overall, our findings provide insights into the comparative strengths and limitations of proprietary and open-weight models for mathematical reasoning in Persian.

1 Introduction

Math Word Problem (MWP) solving can be defined as a Question Answering task in which a problem description includes quantitative information alongside a related question. In many cases, particularly in algebraic word problems, the question concerns an unknown variable that must be computed using the information in the problem statement. A solver system must therefore either compute the final numerical answer directly or generate a mathematical equation that leads to the correct numerical answer. As such, MWP solving requires a combination of natural language understanding, numerical reasoning, and symbolic manipulation, making it an especially desirable capability for large language models (LLMs). An example of an MWP is shown in Table 1.

English Text:
Nikoo invited 10 people to her birthday party. They each ate 8 pieces of pizza. How many pieces of pizza did they eat in total?
Equation: $x = 10 \times 8$ **Answer:** 80

Table 1: An English translated example of a math word problem from the PMWP dataset.

Recent models have demonstrated substantial improvements on mathematical reasoning benchmarks. GPT-4 (OpenAI, 2023) achieves over 92% accuracy on grade-school mathematics problems in GSM8K (Cobbe et al., 2021) when using chain-of-thought prompting (Wei et al., 2023), reflecting notable progress in both linguistic understanding and logical reasoning. Despite these advances, mathematical reasoning remains unevenly studied across languages. While numerous benchmark datasets have been developed to assess mathematical reasoning capabilities, most are limited to English (Patel et al., 2021; Miao et al., 2020; Koncel-Kedziorski et al., 2016). A small number of datasets have been introduced for other languages, including Chinese (Wang et al., 2017; Zhao et al., 2020; Qin et al., 2021), as well as Bangla (Prama et al., 2025), and Romanian (Cosma et al., 2025).

In contrast, despite the widespread use of LLMs by Persian speaking communities, particularly for educational purposes, mathematical reasoning in

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 74–82

Persian remains largely unexplored. The existing resources (Abaskohi et al., 2024) are designed solely for evaluation and provide no training data, limiting their usefulness for robust model development and systematic analysis.

To address this gap, we introduce the PMWP dataset, consisting of 15,588 elementary school-level MWPs annotated with their corresponding equations and numerical answers. PMWP is the first Persian benchmark with training, development, and test splits, enabling both model training and systematic evaluation. Two key characteristics of a high-quality MWP dataset are *scale*, measured by the number of problems, and *diversity*, reflected in vocabulary size. However, previous work has shown that models can often achieve high accuracy by exploiting shallow heuristics rather than genuine reasoning (Patel et al., 2021). To mitigate this issue, we apply data augmentation techniques that treat each quantitative component of a problem as a variable that can be recomputed under different conditions (Kumar et al., 2022). This design encourages models to rely less on memorized patterns and more on underlying reasoning processes. Moreover, it enables the same problem structure to be assessed from multiple perspectives, providing a more robust evaluation of the ability to reason mathematically.

Finally, we evaluated the performance of several widely used LLMs, including both commercial API-based models and open-weight models, on the PMWP dataset. This evaluation provides insights into the current capabilities and limitations of state-of-the-art LLMs for mathematical reasoning in Persian, and allows us to analyze language bias toward English in both answer prediction and equation generation settings. Our contributions are summarized as follows:

- We introduce PMWP, the first dataset for Persian MWP solving, designed for solving problems through equation generation and reasoning evaluation, with standardized training, validation, and test splits.
- We provide a systematic evaluation of popular open-weight and proprietary LLMs, including DeepSeek-V3, Gemini-2.5-Flash, and GPT-4o, on mathematical reasoning in Persian, and analyze language bias toward English for both answer prediction and equation generation.
- We study parameter-efficient adaptation of open-weight LLMs, including Llama-3-8B and Qwen-2.5-7B, using LoRA for Persian MWP solving through equation generation.
- The dataset and fine-tuned models are publicly available[1] to support future research on mathematical reasoning in Persian.

2 Related Work

Early work in this area focused on elementary-level arithmetic problems involving single-unknown equations, leading to foundational datasets such as ADDSUB (Hosseini et al., 2014), SINGLEEQ (Koncel-Kedziorski et al., 2015), and MULTIARITH (Roy and Roth, 2015). These datasets established early task formulations for mapping natural language descriptions to linear equations and laid the groundwork for subsequent MWP solvers in English.

Building on these foundations, larger and more diverse datasets were introduced to support systematic evaluation. In English, MAWPS (Koncel-Kedziorski et al., 2016) consolidated several early datasets into a unified benchmark. In Chinese, MATH23K (Wang et al., 2017) marked a significant milestone by providing over 23,000 elementary-level MWPs with equation annotations, and has since become a standard benchmark for equation generation and arithmetic reasoning. Subsequent datasets such as ASDIV (Miao et al., 2020) and SVAMP (Patel et al., 2021) emphasized problem diversity and robustness.

In parallel with dataset development, recent advances in LLMs have substantially improved performance on mathematical reasoning tasks. Techniques such as chain-of-thought prompting (Wei et al., 2023) and equation or program generation (Romera-Paredes et al., 2024; Imani et al., 2023) have enabled models such as GPT-4 (OpenAI et al., 2024) to achieve strong results on established English benchmarks. However, these evaluations remain largely English-centric, with Chinese datasets serving as the primary non-English alternative.

Prior work has explored benchmarking multilingual LLMs on Persian mathematical tasks, highlighting both the potential of large models and their limitations in handling Persian linguistic structure and numerical reasoning (Abaskohi et al., 2024). The existing Persian resources used in this benchmark consist of two small evaluation-focused

[1] https://github.com/marzieh-abdolmaleki/PMWP

datasets. One includes 50 multiple choice elementary level mathematics questions without contextual problem descriptions, while the other contains 179 questions drawn from 7th and 10th grade entrance examinations and selected translations from the MATH dataset (Hendrycks et al., 2021). Despite these efforts, Persian-focused studies largely rely on small-scale evaluation sets and do not provide publicly available datasets that support supervised training. As a result, dataset scarcity remains a primary bottleneck for systematic evaluation and model adaptation in Persian.

Translation-based approaches have been proposed as a low-cost alternative for expanding multilingual coverage, including automatically translated variants of GSM8K (Chen et al., 2024). However, prior work on Persian NLP consistently shows that naïve translation often fails to capture language-specific structures and cultural conventions that are especially important for educational tasks (Khashabi et al., 2021). These limitations motivate dataset construction strategies that combine translation with targeted linguistic validation and controlled augmentation. To the best of our knowledge, no large-scale, publicly available Persian MWP dataset currently exists that supports both training and systematic evaluation. Existing Persian resources are limited in scale and designed solely for evaluation. In contrast, the PMWP dataset combines machine translation from a large, well-established benchmark (MATH23K) (Wang et al., 2017) with structured data augmentation that systematically reassigns the unknown variable within each problem. All translations are revised by Persian native speakers and culturally adapted by modifying names, entities, and events. This design enables scalable dataset construction while preserving mathematical correctness and advancing research on Persian mathematical reasoning.

3 Persian MWP Dataset

In this section, we describe our methodology to construct the Persian MWP dataset. The goal is to collect Persian MWPs at the primary school level so that problems are solvable via linear equations with one unknown. Linear equations can only include the four arithmetic operators (addition, subtraction, multiplication, and division). In this regard, machine translation (Section 3.1) and data augmentation (Section 3.2) methods are used.

Prototype Problem
Nikoo invited 10 people to her birthday party. They each ate 8 pieces of pizza. How many pieces of pizza did they eat in total?
Equation: $X = 10 \times 8$ Answer: 80
Transformed Problem
Nikoo invited 10 people to her birthday party. They each ate 8 pieces of pizza. They ate 80 pieces of pizza in total.
Equation: $80 = 10 \times 8$
Candidate Problem 1
Nikoo invited X people to her birthday party. They each ate 8 pieces of pizza. They ate 80 pieces of pizza in total.
Equation: $X = 80 \div 8$ Answer: 10
Candidate Problem 2
Nikoo invited 10 people to her birthday party. They each ate X pieces of pizza. They ate 80 pieces of pizza in total.
Equation: $X = 80 \div 10$ Answer: 8
Candidate Problem 3
Nikoo invited 10 people to her birthday party. They each ate 8 pieces of pizza. They ate X pieces of pizza in total.
Equation: $X = 10 \times 8$ Answer: 80

Table 2: An example of the data augmentation method.

3.1 Machine Translation

We randomly sampled 8,000 Chinese mathematical word problems from the MATH23K dataset and translated them into Persian using the Google Translate API. We selected Chinese–Persian transfer because MATH23K is a widely used benchmark for this task, contains problem types closely aligned with our target domain, and provides substantially more data than English-based datasets such as MAWPS. The translated problems were manually reviewed and corrected when necessary by qualified expert linguists holding master's or Ph.D. degrees. All reviewers were native Persian speakers from Iran and culturally adapted the problems by modifying names, entities, and events. We also identified and corrected incorrect or inconsistent equations in the original Chinese dataset. Both the translated texts and their corresponding equations were revised during this process. After quality control, we retained 5,000 Persian math word problems for our experiments.

3.2 Data Augmentation

To extend the dataset, we produced new problems from the translated prototype problems taking the following steps:

- **Question sentence to informative sentence transformation:** Interrogative sentences beginning with terms like "what," "how much," "how many", etc., undergo handwritten rule-based transformations to incorporate an unknown symbol "x" as informative sentences. Then unknown symbols in both the text and the corresponding equation are replaced with the final answer.
- **New unknown selection:** after replacing the original unknown symbol in both the text and equation with the final answer, each existing quantity in the equation which also appears in the context is used as a new unknown candidate.
- **New equation construction for the new unknown:** the altered equation is changed so that the new unknown symbol occurs at the left side of the equation.

Because the augmentation method is driven by existing quantities in equations, it operates without errors. An example is provided in Table 2. While the data augmentation method may affect vocabulary diversity, generating different variations of the same underlying MWPs helps models avoid relying on shallow heuristics or memorizing common problem patterns. Using this approach, we produced 10,588 new Persian MWPs.

Text	
Correct	91%
Low Readability	5%
Need Correction	4%
Equation	
Correct	97%
Wrong	3%
Answer	
Correct	95%
Wrong	5%

Table 3: Validity results of PMWP.

To validate the quality of the dataset, expert linguists revised 600 randomly chosen problems with their corresponding final equations and answers. As shown in Table 3, texts, equations, and final answers in the Persian dataset yield an accuracy of 91%, 97%, and 95%, respectively. Text evaluation is divided into three categories: Correct, Low Readability, and Need Correction. A Correct text demonstrates a combination of natural language fluency and sound mathematical reasoning. On the other hand, a Low Readability text may have accurate mathematical content but is difficult to comprehend due to its sentence structure. Texts that cannot be supported by mathematical reasoning are considered for correction. In addition, a correct final answer must align with the question posed in the corresponding problem text. Similarly, a correct equation is one that accurately reflects the mathematical principles within the corresponding problem text and that yields the correct final answer.

# of Problems	15,588
# of Tokens	522,298
# of Types	4,671
# of Equation Templates	133

Table 4: Statistical information of PMWP.

	Text	**Equation**
Min. Length	6	5
Max. Length	73	16
Avg. Length	31.8	8.2

Table 5: Token-Based statistical analysis of problem text and equation components in PMWP. Equation tokens include both numbers and operators.

The novel Persian MWP collection is entitled PMWP. The statistical information of the dataset is shown in Table 4 and Table 5. The dataset is divided into train, test, and validation partitions, by the ratio of 80, 10, 10, respectively.

4 Experimental Setup

In this section, we outline the experimental setup for evaluating mathematical reasoning on the PMWP dataset. We conduct zero-shot evaluations to analyze the impact of input language and output format on model performance, followed by parameter-efficient fine-tuning experiments using LoRA to assess improvements in symbolic equation generation.

Model	AnsAcc (Persian)	AnsAcc (English)	EqValueAcc (Persian)	EqValueAcc (English)
DeepSeek-V3 (%)	**71.66**	71.31 (−0.35)	70.49	69.43 (−1.06)
Gemini-2.5-Flash (%)	69.66	**73.60** (+3.94)	72.02	71.43 (−0.59)
GPT-4o (%)	53.26	**71.31** (+18.05)	66.49	67.78 (+1.29)

Table 6: Zero-shot performance of different models on the PMWP test set. AnsAcc denotes numeric answer accuracy and EqValueAcc denotes value-based equation accuracy. Values in parentheses indicate the change after translation to English.

4.1 Zero-shot Evaluation

We evaluate the zero-shot mathematical reasoning capabilities of LLMs without any task-specific fine-tuning or in-context examples. All experiments are conducted on the PMWP test set, which consists of Persian MWPs requiring the computation of a target variable x.

To disentangle the effects of output representation and input language, we define two zero-shot evaluation settings based on the required output format: numeric answer prediction and symbolic equation generation. Within each setting, we consider both direct Persian input and translation-based English input. In the direct Persian setting, models generate final answers or equations directly from the original Persian problem. In the translation-based English setting, each Persian problem is first translated into English using the same model and then solved under an English instruction. While this setting does not isolate translation quality, it reflects realistic end-to-end usage of LLMs as multilingual problem solvers.

Zero-shot Numeric Reasoning. In the numerical setting, models are instructed to solve the problem and output the final numerical value of the target variable. The corresponding prompt template, translated into English, is shown below:

> You are given a math word problem involving the variable x, enclosed in < >.
>
> Instructions:
>
> - Explain the reasoning steps clearly and completely.
> - All explanations must appear before the final line.
> - The output must end with exactly one final line.
> - The final line must contain only the final numerical answer in the following format:
>
> x = answer
>
> - Do not add any text, symbols, whitespace, or punctuation after the final line.
>
> Problem: <problem text>
>
> Answer:

Zero-shot Symbolic Reasoning. In the symbolic setting, models are required to generate a symbolic equation that represents the solution, without numerically simplifying it. The prompt template, translated into English, used in this setting is shown below:

> You are given a math word problem involving the variable x, enclosed in < >.
>
> Instructions:
>
> - Explain the reasoning steps clearly and completely.
> - All explanations must appear before the final line.
> - The output must end with exactly one final line.
> - The final line must contain only the equation solution in the following format:
>
> x = equation
>
> - Do not numerically simplify the final equation (e.g., keep $x = 10/2$, not $x = 5$).
> - Do not add any text, symbols, whitespace, or punctuation after the final line.
>
> Problem: <problem text>
>
> Equation:

Models. We conduct zero-shot evaluations using three LLMs: DeepSeek-V3 (DeepSeek-AI et al., 2025), GPT-4o (OpenAI et al., 2024), and Gemini-2.5-Flash (Comanici et al., 2025). DeepSeek-V3 is an instruction-tuned model optimized for reasoning-intensive tasks. GPT-4o is a proprietary general-purpose model with strong multilingual and reasoning capabilities. Gemini-2.5-Flash is a latency-optimized model designed for efficient inference while maintaining competitive reasoning performance. This selection allows comparison across open and proprietary models. All models are queried via their respective APIs using a unified

prompting strategy and identical decoding configurations to ensure fair comparison. We use greedy decoding with temperature set to zero and do not provide chain-of-thought exemplars. A consistent system prompt defines the model as a mathematical problem solver.

Model predictions are compared against the gold annotations in the PMWP test set using complementary metrics. For numeric reasoning, we report answer accuracy (AnsAcc) for both Persian and translation-based English inputs. For symbolic reasoning, we report value-based equation accuracy (EqValueAcc), which measures whether the numerical solution obtained by solving the generated equation matches the gold answer. As shown in Table 6, translation-based inference generally improves numeric answer accuracy, most notably for GPT-4o, indicating a strong sensitivity to input language in zero-shot reasoning. GPT-4o is also the only model that shows an improvement in equation value accuracy after translation, whereas the performance of the other models decreases under this setting. This observation is consistent with prior findings that English translated prompts often yield better performance than Persian prompts in multilingual evaluations (Abaskohi et al., 2024). However, the improvement observed for DeepSeek-V3 after translation is minimal, suggesting that this model is less sensitive to input language than the other evaluated models. This difference may also be attributed to the models' internal Persian-to-English translation capabilities, as they are evaluated in an end-to-end setting rather than under controlled translation conditions. Nevertheless, the consistent performance gains in answer accuracy observed across models indicate a systematic bias toward English inputs. In contrast, symbolic equation generation appears to be less affected by input language choice, although it remains a more challenging task overall. For direct Persian input, DeepSeek-V3 achieves stronger performance in numeric answer accuracy, while Gemini-2.5-Flash performs better in equation value accuracy, highlighting differences in model behavior across output formats and reasoning settings.

4.2 LoRA Fine-tuning

To assess the impact of parameter-efficient adaptation on symbolic mathematical reasoning, we fine-tune autoregressive language models using Low-Rank Adaptation (LoRA) (Hu et al., 2021). In contrast to the zero-shot setting, fine-tuning is performed exclusively for symbolic equation generation, where the goal is to produce an explicit equation defining the variable x.

Models. We fine-tune two open-weight autoregressive models: Qwen2.5-7B (Yang et al., 2024) and LLaMA-3-8B (AI@Meta, 2024). Both models are instruction-capable causal language models with general reasoning abilities. We select these models to evaluate the effectiveness of LoRA-based adaptation on medium scale architectures under limited computational budgets.

We again use the PMWP dataset with predefined train, validation, and test splits. Each training instance consists of a Persian problem statement and its corresponding gold equation. The fine-tuning task is formulated as conditional generation of a symbolic equation. Given a Persian math word problem, the model is instructed to output only the equation that determines the value of x, without providing explanations or numerical simplification. All outputs are constrained to the format:

```
x = <equation>
```

This formulation directly targets structural reasoning and equation construction.

LoRA Configuration and Training Setup. We apply LoRA adapters to the query, key, value, and output projection matrices of the self-attention layers. For all experiments, we use a rank of $r = 16$, scaling factor $\alpha = 32$.

Models are fine-tuned using the AdamW optimizer for three epochs. The effective batch size is 32. Training is performed in half-precision (FP16) with gradient checkpointing enabled.

Metric	Qwen2.5-7B	LLaMA-3-8B
EqMatchAcc (%)	91.65	92.53
EqMismatch (%)	0.88	0.35
Errors (%)	7.47	7.11

Table 7: Performance comparison of Qwen2.5-7B and LLaMA-3-8B on the PMWP test set. EqMatchAcc denotes exact equation match accuracy. EqMismatch refers to cases where the generated equation differs textually from the gold equation but yields the correct numerical result. Errors indicates the percentage of incorrect or invalid generated equations.

Table 7 presents the performance of Qwen2.5-7B and LLaMA-3-8B on the PMWP test set. Both models achieve high exact equation match accuracy, exceeding 91%, indicating that parameter-

efficient fine-tuning enables strong symbolic reasoning performance on elementary-level Persian MWPs. LLaMA-3-8B slightly outperforms Qwen2.5-7B in exact equation matching, while both models exhibit comparable rates of errors.

5 Conclusion

We introduced PMWP, a Persian dataset for elementary-level math word problem solving with explicit equation annotations and standardized training, validation, and test splits. The dataset is constructed by translating problems from an established benchmark and fully validating all translated instances through expert human review, followed by structured data augmentation to increase scale while preserving mathematical correctness. Our zero-shot evaluation of open-source and proprietary LLMs reveals differing sensitivities to input language, with symbolic equation generation showing greater robustness to translation than direct answer prediction. In addition, our LoRA-based fine-tuning experiments demonstrate that open-weight models can achieve high equation generation accuracy on PMWP with a limited number of trainable parameters. Overall, this work provides a systematic assessment of current LLM capabilities for mathematical reasoning in Persian and establishes PMWP as a benchmark to support future research in multilingual and reasoning.

Limitations

PMWP focuses on elementary-level math word problems solvable with single-variable linear equations. This scoped design enables controlled evaluation of foundational mathematical reasoning in Persian, but does not cover more advanced problem types such as multi-variable reasoning, non-linear equations, geometry, or probability. As a result, the findings may not directly generalize to higher-level mathematical reasoning tasks; however, PMWP provides a solid foundation for future extensions in Persian, similar to earlier studies conducted in English.

Acknowledgments

This work was supported by the Special Research Fund of Ghent University under grant number BOF.BAF.2024.0248.01. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO, and the Flemish Government – department EWI.

References

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2189–2203, Torino, Italia. ELRA and ICCL.

AI@Meta. 2024. Llama 3 model card.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Adrian Cosma, Ana-Maria Bucur, and Emilian Radoi. 2025. RoMath: A mathematical reasoning benchmark in Romanian. In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pages 95–111, Suzhou, China. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533, Doha, Qatar. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. Association for Computational Linguistics.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabadi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, and 6 others. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.

Vivek Kumar, Rishabh Maheshwary, and Vikram Pudi. 2022. Practice makes a solver perfect: Data augmentation for math word problem solvers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4194–4206, Seattle, United States. Association for Computational Linguistics.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2023. Gpt-4 technical report. `https://arxiv.org/abs/2303.08774`. ArXiv:2303.08774.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Tabia Tanzin Prama, Christopher M. Danforth, and Peter Dodds. 2025. BanglaMATH : A Bangla benchmark dataset for testing LLM mathematical reasoning at grades 6, 7, and 8. In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pages 134–149, Suzhou, China. Association for Computational Linguistics.

Jinghui Qin, Xiaodan Liang, Yining Hong, Jianheng Tang, and Liang Lin. 2021. Neural-symbolic solver for math word problems with auxiliary tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5870–5881, Online. Association for Computational Linguistics.

Bernardino Romera-Paredes and 1 others. 2024. Mathematical reasoning with large language models. *Nature*, 625:468–475.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854, Copenhagen, Denmark. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wei Zhao, Mingyue Shang, Yang Liu, Liang Wang, and Jingming Liu. 2020. Ape210k: A large-scale and template-rich dataset of math word problems. *Preprint*, arXiv:2009.11506.

APARSIN: A Multi-Variety Sentiment and Translation Benchmark for Iranic Languages

Sadegh Jafari[1], Tara Azin[2], Farhad Roodi[3,*], Zahra Dehghani Tafti[4,*], Mehrdad Ghadrdan[5,*], Elham Vatankhahan Esfahani[6,*], Aylin Naebzadeh[16,*], Mohammadhadi Shahhosseini[7,*], Ghafoor Khan[12,*], Kazem Forghani[9,*], Danial Namazi[4,*], Seyed Mohammad Hossein Hashemi[8,*], Farhan Farsi[10,*], Mohammad Osoolian[9,*], Maede Mohammadi[11,*], Mohammad Erfan Zare[4,*], Muhammad Hasnain Khan[9,*], Muhammad Hussain[13,*], Nooreen Zaki[14,*], Joma Mohammadi[10,*], Shayan Bali[13,*], Mohammad Javad Ranjbar[4,*], Els Lefever[1,+], Veronique Hoste[1,+]

[1]Ghent University, Belgium, [2]Carleton University, Canada, [3]Purdue University, USA, [4]University of Tehran, Iran, [5]Islamic Azad University Science and Research, Iran, [6]Tarbiat Modares University, Iran, [7]University of Milan, Italy, [8]Shahid Beheshti University, Iran, [9]Iran University of Science and Technology, Iran, [10]Amirkabir university of technology, Iran, [11]Ferdowsi university of Mashhad, Iran, [12]Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan, [13]Yuan Ze University, Taiwan, [14]Polytechnical University, Xian, China [15]King's College London, England, [16]University of Hull, England

*** Equal contribution. + Supervisor. Correspondence:** sadegh.jafari@ugent.be

Abstract

The Iranic language family includes many underrepresented languages and dialects that remain largely unexplored in modern NLP research. We introduce APARSIN, a multi-variety benchmark covering 14 Iranic languages, dialects, and accents, designed for sentiment analysis and machine translation. The dataset includes both high- and low-resource varieties, several of which are endangered, capturing linguistic variation across them. We evaluate a set of instruction-tuned Large Language Models (LLMs) on these tasks and analyze their performance across the varieties. Our results highlight substantial performance gaps between standard Persian and other Iranic languages and dialects, demonstrating the need for more inclusive multilingual and dialectally diverse NLP benchmarks.

1 Introduction

The Iranian plateau and its surrounding regions are among the most linguistically diverse areas in the world, home to numerous Iranic languages across multiple branches of the Indo-Iranian family (Kent, 1953; Windfuhr, 2009). This linguistic richness has emerged through centuries of cultural exchange, migration, and sustained language contact across contemporary Iran, Afghanistan, Pakistan, and parts of Iraq (Witzel, 2003; Windfuhr, 2006). While widely spoken languages such as Iranian Persian, Dari, and Pashto serve as lingua francas and administrative languages across much of the region, dozens of smaller language varieties and dialects remain vital to the cultural identity of millions of speakers. Despite this linguistic diversity and extensive prior work on Standard Persian (Farsi et al., 2025a; Abaskohi et al., 2024), Iranic languages other than Standard Persian have been severely underrepresented in natural language processing (NLP) research (Kamaly, 2025), especially in interpretive tasks such as translation and sentiment analysis.

Sentiment analysis is a core area of NLP research with applications across e-commerce, social media monitoring, political analysis, and digital humanities (Liu, 2015). The widespread use of social media platforms has generated large amounts of opinionated text across languages worldwide, creating opportunities to study public sentiment and language use in authentic communicative contexts (Liu, 2022). However, despite substantial advances in sentiment analysis, low-resource and underrepresented languages continue to face significant barriers, mainly due to the scarcity of annotated datasets and the lack of standardized orthographies for wide varieties (Joshi et al., 2020).

For Iranic languages, these challenges are particularly pronounced. Wide varieties lack stan-

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 83–97

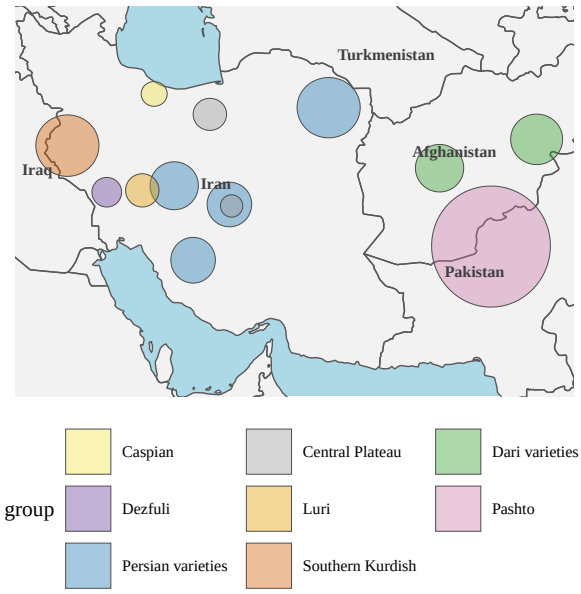


Figure 1: Approximate geographic distribution of Iranic varieties in APARSIN. The map conveys comparative geographic spread rather than exact boundaries, and speaker densities and colors indicate linguistic groups. Based on Glottolog reference points (Hammarström et al., 2025) and generated using Glottospace (Norder et al., 2022).

dardized writing systems (Farsi et al., 2025b) and have historically been transmitted through oral tradition. When written, they typically rely on adapted forms of the Perso-Arabic script with varying orthographic conventions, which often result in inconsistency in digital text. Today, social media serves as an important platform for written communication in these varieties and dialects, yet speakers frequently code-switch between standard Persian and local forms, adapt Persian orthography to represent dialectal features, or use non-standard spellings. These characteristics make sentiment analysis for Iranic varieties particularly challenging. To the best of our knowledge, no comprehensive benchmark currently exists to support research in this domain. You can find our GitHub page here.[1]

To address this gap, we present APARSIN [2], the first multi-variety sentiment and translation analysis benchmark for Iranic languages. The dataset comprises 1,400 annotated social media comments across 14 Iranic languages and dialects: Pashto, Hazaragi, Kabuli Dari, Standard Persian, Shirazi, Khorasani, Esfahani, Yazdi, Kalhori Kurdish, Luri Bakhtiari, Dezfuli, Tonekaboni, Semnani, and Zoroastrian Dari. The dataset represents both major branches of the Iranian language family (Western and Eastern Iranic) and includes both widely spoken languages with millions of speakers (Persian, Pashto, Kurdish) and endangered dialects and language varieties that have received no computational attention (Semnani, Tonekaboni, Zoroastrian Dari). Each sample is annotated for sentiment polarity (positive, negative, neutral) by native speakers and includes translations into Persian and English as well as transliterations for cross-variety and cross-lingual analysis. Figure 1 illustrates the geographic distribution of the languages included in our dataset.

Our contributions are as follows:

- We introduce APARSIN, the first sentiment analysis benchmark covering 14 Iranic languages and dialects, consisting of 1,400 annotated social media comments with corresponding translations.
- Our work presents comprehensive baseline experiments on language identification, machine translation, and sentiment classification using instruction-tuned LLMs for these underrepresented dialects and language varieties.
- The dataset, annotation guidelines, and evaluation code will be released publicly to support future research on sentiment analysis and other NLP tasks for low-resource languages.

2 Related Work

While sentiment analysis achieves high performance for major languages, low-resource and underrepresented languages face substantial challenges due to noisy social media text, non-standardized orthography, and limited training data (Joshi et al., 2020). Developing sentiment analysis resources for such languages remains critical for supporting their speaker communities and advancing cross-linguistic understanding of sentiment expression. For Iranic varieties beyond standard Persian, computational resources remain severely limited. The primary focus of existing work has been on standard Persian spoken in Iran, with several sentiment and emotion analysis datasets developed in recent years. Notable examples include ArmanEmo

[1] https://github.com/SilkRoadAparsin

[2] Pronounced *apārsin* اَپارسین. The name is adapted from Pahlavi sources, meaning "above Simorgh's flight range", where Simorgh is a legendary bird in Persian mythology (Bahar, 1995). The name is used descriptively to evoke the highland regions extending across the Hamoun Lake area (in Iran) and the Hindu Kush.

(Mirzaee et al., 2025) and PersianEmo (Hussiny et al., 2024) for emotion analysis, and SentiFars (Dehkharghani, 2019) for sentiment polarity. However, these resources focus exclusively on standard varieties and do not address the substantial linguistic diversity present within the Iranian language family. Beyond standard Persian, sentiment analysis work for other Iranic varieties is extremely sparse. For Pashto, Kamal et al. (2016) developed a lexicon-based system achieving 73.2% accuracy on social media text. For Kurdish, Badawi (2023) introduced the KMD emotional dataset for Sorani (Central Kurdish), but other Kurdish branches, such as Northern and Southern Kurdish, remain unaddressed. For Central Plateau languages (Semnani, Zoroastrian Dari), Luri varieties, and Caspian languages like Mazandarani, no sentiment resources exist whatsoever. Large-scale multilingual sentiment benchmarks provide important precedents for our work. AfriSenti (Muhammad et al., 2023) covers 14 African languages, and MD-ArSenTD (Baly et al., 2017) addresses multiple Arabic dialects, demonstrating both the feasibility and importance of creating sentiment resources for underrepresented language communities. These efforts highlight common challenges such as non-standardized orthographies, limited digital corpora, and the need for native speaker annotation. Computational approaches to sentiment analysis have also evolved alongside these dataset development efforts. Early work was largely based on rule-based and dictionary-based methods (Mohammad et al., 2013; Taboada et al., 2011; Turney, 2002). Subsequent research shifted toward classical machine learning approaches that relied on manually engineered features (Agarwal and Mittal, 2016; Le and Nguyen, 2015). Advances in deep learning (Yadav and Vishwakarma, 2020; Zhang et al., 2018) and the adoption of pretrained language models have since reshaped the field. Current state-of-the-art systems employ multilingual pretrained models such as XLM-R (Conneau et al., 2020) and mDeBERTaV3 (He et al., 2021), as well as instruction-tuned LLMs that demonstrate strong performance across diverse languages and domains (Zhang et al., 2024). In this work, we evaluate several instruction-tuned LLMs on sentiment classification, language identification, and machine translation tasks for Iranic varieties, providing the first comprehensive assessment of these models' capabilities for this language family.

3 Dataset Overview

APARSIN includes 14 Iranic languages and dialects that represent both major branches of the Iranian language family: Western and Eastern Iranic. The selected varieties cover a wide geographic range, from the mountainous regions of Afghanistan and Pakistan to Iran's verdant Caspian coast, with representation extending into Iraq. This selection includes both major languages with millions of speakers (e.g., Persian, Pashto, Kurdish) and smaller, often endangered varieties (e.g., Semnani, Tonekaboni, Zoroastrian Dari) that have received little to no attention in NLP research. Despite shared Iranic roots, these languages show substantial diversity, which makes them well-suited for studying sentiment analysis across related yet distinct varieties. Table 1 provides an overview of all languages included in the dataset, their classifications, and geographic distributions. Representative examples for each language variety are provided in Table 8 in Appendix.

3.1 Language Family

The Iranian language family, which forms the western branch of the Indo-Iranian languages within Indo-European, is represented in our dataset through three Western Iranic subgroups and one Eastern Iranic branch:

Central Plateau Iranic languages (Semnani and Zoroastrian Dari) belong to the Northwestern Iranic branch and are spoken in central Iran. These languages are associated with some of the longest-standing communities of Iranic speech in the region (see Appendix A). They preserve historically archaic features and have received little to no attention in computational research, in part due to their smaller speaker populations.

Southwestern Iranic constitutes the largest group in our dataset, including Persian and its regional varieties (Kabuli, Hazaragi, Shirazi, and Khorasani), Luri, Dezfuli, and Kalhori (a variety of Southern Kurdish within the Kurdic branch). Standard Persian (also known as Iranian Persian) serves as the standard literary language and lingua franca across much of the region.

Western Iranic (Caspian) is represented by Tonekaboni, a Caspian variety spoken in Mazandaran Province that is commonly treated as distinct from Mazandarani dialects.

Eastern Iranic is represented by Pashto, one of the major languages of Afghanistan and Pakistan,

Language / Variety	ISO	Iranic branch	Region(s)
Persian (Iranian variety)	pes	West Iranic - Southwestern	Iran
Shirazi	pes	West Iranic - Southwestern	Southern Iran (Fars)
Yazdi	pes	West Iranic - Southwestern	Central Iran (Yazd)
Esfahani	pes	West Iranic - Southwestern	Isfahan
Khorasani	pes	West Iranic - Southwestern	Northeastern Iran (Greater Khorasan)
Kabuli	prs	West Iranic - Southwestern	Afghanistan (Kabul and surrounding areas)
Hazaragi	haz	West Iranic - Southwestern	Afghanistan (central highlands)
Pashto	pus	East Iranic	Afghanistan, Pakistan
Kalhori	sdh	West Iranic - Kurdic	Western Iran (Kermanshah), Iraq (Khanaqin)
Luri Bakhtiari	bqi	West Iranic - Southwestern	Southwestern Iran (Zagros region)
Dezfuli	def	West Iranic - Southwestern	Southwestern Iran (Khuzestan)
Tonekaboni	mzn	West Iranic - Caspian	Northern Iran (Caspian coast, Mazandaran)
Semnani	smy	West Iranic - Central Plateau	Central Iran (Semnan)
Zoroastrian Dari	gbz	West Iranic - Central Plateau	Iran (Yazd, Kerman)

Table 1: Iranic languages and dialects included in our dataset. ISO 639-3 codes are reported at the language level (varieties without distinct ISO codes share the parent language code).

with substantial phonological and morphological differences from the Western Iranic varieties.

3.2 Geographic Scope

Geographically, the dataset covers Iran, Afghanistan, Iraq, and Pakistan. Many of these languages show significant dialectal variation across regions. Persian varieties, for instance, differ substantially between Iran (Shirazi, Khorasani, Esfahani, Standard Persian, Yazdi) and Afghanistan (Kabuli, Hazaragi), showing both geographic separation and distinct ethnic community identities. Kurdish varieties are also found across multiple countries, with Kalhori (Southern Kurdish) spoken primarily in Iran's Kermanshah province and extending into Iraq's Khanaqin region (Azin and Ahmadi, 2021).

3.3 Writing Systems

All languages in our dataset use the Perso-Arabic script, except Zoroastrian Dari, which is provided only using transliteration due to its exclusively oral transmission among the community from which our data were collected. It includes additional letters not found in Arabic to represent phonemes such as /p/, /tʃ/, /ʒ/, and /g/. While the script is shared across varieties, individual languages use distinct orthographic conventions and, in some cases, additional diacritics reflecting local phonological features. The widespread use of Perso-Arabic script shows shared historical and cultural influences, as well as Persian's longstanding role as a prestige and administrative language. Moreover, since many smaller Iranic languages and dialects have historically been transmitted through oral traditions, in contemporary social media contexts, speakers adapt standard Persian orthography to represent such varieties. This results in non-standard spellings, orthographic variation, and code switching between standard and dialectal forms.

4 Dataset Collection and Processing

Collecting authentic and representative data for Iranic languages poses significant challenges, particularly for low-resource and endangered varieties. Many of these languages have a limited digital presence, lack standardized orthography, and are primarily transmitted through oral traditions. To address these challenges and ensure ecological validity, we adopted a multi-step data collection strategy that combines existing resources, automated web crawling, and manual data collection with the help of native speakers.

4.1 Speaker Recruitment

Native speakers of different Iranic language varieties were recruited through social media platforms such as LinkedIn and other community-based online networks. All recruited speakers participated voluntarily. As a form of acknowledgment and compensation, contributors who provided substantial data are included as co-authors in this paper.

This recruitment strategy was chosen to ensure the collection of real-world, authentic language use as it naturally appears in social media and everyday communication. Such an approach is particularly important for low-resource and endangered Iranic languages, for which curated datasets are scarce or entirely unavailable, and where linguistic knowledge is often maintained within small speaker communities.

4.2 Dataset Collection

We employed three complementary data collection approaches depending on the availability and digital footprint of each language variety.

4.2.1 Published Datasets

For well-resourced and standardized Iranic languages, such as Standard Persian and Pashto, we rely on existing publicly available datasets. These resources provide relatively large volumes of high-quality textual data and serve as a solid foundation for languages with established writing systems and sufficient online presence. In this work, we make use of the English–Pashto Language Dataset (EPLD) (Khan et al., 2025), the Pashto–English Bilingual Sentiment Corpus[3], as well as two large-scale Persian sentiment analysis datasets, namely Digikala Sentiment [4] and SnappFood Sentiment (Farahani et al., 2020).

4.2.2 Web Crawling

For under-resourced languages, dialects, and accent varieties, we developed an automated web-crawling framework to gather naturally occurring language data from the internet. The framework targets sources such as social media platforms, personal and community blogs, as well as linguistic, cultural, and regional websites.

Automated Crawling Pipeline. The crawling process followed a multi-stage pipeline powered by LLMs. First, we prompted an LLM to generate language-specific search keywords, including alternative spellings, colloquial forms, and community-specific identifiers. These keywords were then used to query search engines and online platforms automatically. Next, for each retrieved document, the LLM was used to extract candidate sentences written in the target language variety, along with their corresponding translations when available. All experiments reported in this work used GPT-5 (OpenAI, 2025a) as the underlying LLM.

Human Annotation. To ensure data quality, we employed native speakers of each language variety to annotate the collected samples manually. For each sentence, annotators verified (i) whether the original utterance is valid and genuinely belongs to the target language variety, and (ii) whether the provided translation is correct. Both checks were annotated using binary True/False labels. Table 2 reports the aggregated results of this annotation process.

[3] https://www.kaggle.com/datasets/farhadkhan66/pashto-translated-corpus

[4] https://www.digikala.com/opendata/

Results and Discussion. Table 2 summarizes the crawling and annotation outcomes for different language varieties. The number of collected samples varies substantially across languages, reflecting differences in online presence and community activity. Several varieties, such as *Dezfuli*, *Khorasani*, and *Yazdi* exhibit very high original validity rates, indicating that the LLM-guided keyword generation and sentence extraction were effective in identifying genuine language samples.

In contrast, languages such as *Luri* and *Semnani* show lower validity and translation correctness rates, highlighting the challenges of noisy web data and limited standardized written forms. Notably, for *Kalhori*, while most original utterances were valid, translation correctness was considerably lower, suggesting difficulties in obtaining reliable translations for certain varieties even when the original text is available. Finally, for some endangered languages with extremely limited or nonexistent online presence (e.g., Zoroastrian varieties), web crawling yielded little to no usable data, underscoring the limitations of purely web-based collection methods for severely under-documented languages.

Language	Samples	Orig. Valid	Trans. Correct
Dezfuli	764	99.738%	85.602%
Hazaragi	383	80.940%	58.486%
Esfahani	294	83.333%	82.993%
Khorasani	1216	99.753%	99.342%
Luri	1040	31.538%	31.635%
Semnani	470	60.213%	58.511%
Shirazi	563	65.187%	71.048%
Kalhori	403	99.504%	25.310%
Tonekaboni	437	79.863%	64.073%
Yazdi	697	98.278%	96.987%

Table 2: Dataset statistics per language variety. **Orig. Valid** denotes the proportion of samples whose original utterance was verified by native speakers as valid and belonging to the target language, while **Trans. Correct** indicates the proportion of samples with a correct human-verified translation.

4.2.3 Manual Data Collection

For languages where automated methods were ineffective, we relied on manual data collection by volunteer native speakers. Contributors were asked to collect short, common, and naturally occurring text samples from social media platforms such as Facebook, Twitter/X, and messaging forums, or to provide original examples representative of everyday usage. Clear guidelines and collection protocols were provided to ensure consistency, ethical data

handling, and linguistic authenticity. This manual approach was essential for preserving endangered languages that are rarely written and poorly represented in digital spaces.

4.3 Topic Modeling and Sample Selection

We apply a topic modeling pipeline based on BERTopic (Grootendorst, 2022) to discover semantic patterns across multilingual Iranic language data. Sentence-level representations are obtained using the BGE-M3 (Chen et al., 2024) multilingual embedding model, while topic representations are generated using GPT-4o-Mini (OpenAI, 2025b) to improve interpretability. Dimensionality reduction is performed using UMAP (McInnes et al., 2018) with 15 nearest neighbors, a cosine distance metric, and a 10-dimensional projection for clustering, followed by a two-dimensional projection for visualization. Topics are identified using HDBSCAN (Campello et al., 2013) with a minimum cluster size of 10, enabling density-based discovery without predefining cluster shapes. The model is configured to extract 10 distinct topics, corresponding to the 10 languages and dialects in the dataset. For balanced sample selection, one representative sample is selected for each topic–language pair, resulting in a total of 100 selected samples (for more details see Appendix B).

4.4 Translation and Transliteration Approaches

In order to create a comprehensive and high-quality dataset for low-resource Iranic varieties, we employed a combination of translation and transliteration strategies to capture authentic, real-world language use from social media.

4.4.1 Translation Approach

To address the limited digital presence of low-resource and endangered Iranic varieties, part of the dataset was created through translation. For varieties lacking sufficient naturally occurring written data, sentences were translated from high-resource languages such as Persian and English. Translations were performed by native speakers and professional translators, depending on availability and linguistic needs. The translation direction was chosen on a case-by-case basis to ensure semantic accuracy and cultural relevance. Quality was ensured through cross-checking by additional native speakers, targeted reviews by linguistically trained annotators, and consistency checks across sentiment annotations.

4.4.2 Transliteration Approach

All sentences were also transliterated into a unified Latin-based format using a standardized scheme adapted from the Iranian Studies transliteration guidelines[5]. Annotators followed the guidelines to maintain consistency across languages.

This transliteration supports cross-linguistic analysis, improves accessibility for researchers unfamiliar with Perso-Arabic scripts, facilitates integration with NLP models, and enables future reuse of the dataset in multilingual settings.

4.5 Sentiment Annotation Process

After selecting 100 samples and translating them into each target language, we asked three native volunteer annotators per language to label the sentiment of each sentence as *Negative*, *Neutral*, or *Positive*. To assess annotation reliability, we computed inter-annotator agreement using Krippendorff's α (KA) and the average pairwise Cohen's κ (CK).

Table 3 reports the agreement scores across the Iranic language varieties. We observe substantial agreement for Tonekaboni ($\alpha = 0.913$), indicating highly consistent sentiment interpretation among annotators. Moderate agreement is achieved for languages such as Kalhori, Esfahani, and Semnani, while lower agreement is observed for Pashto, Shirazi, and Dezfuli. These variations likely reflect differences in dialectal ambiguity, sentiment expression, and the limited availability of standardized sentiment cues in low-resource language varieties. Overall, the results highlight both the feasibility of sentiment annotation and the inherent challenges of achieving high agreement across diverse Iranic languages.

5 Experiments

We conduct our experiments using eight instruction-tuned LLMs spanning four model families: OpenAI: GPT-4o and GPT-4o-mini (Achiam et al., 2023); Google: Gemma-3-12B-IT and Gemma-3-27B-IT (Team et al., 2025); Meta: Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct (Grattafiori et al., 2024); and Qwen: Qwen3-14B and Qwen3-32B (Yang et al., 2025).

[5] https://github.com/SilkRoadAparsin/Translation

Language	KA	CK
Tonekaboni	0.913	0.913
Zoroastrian Dari	0.423	0.444
Kalhori	0.579	0.581
Pashto	0.197	0.215
Esfahani	0.541	0.544
Shirazi	0.384	0.391
Semnani	0.510	0.510
Dezfuli	0.330	0.332

Table 3: Inter-Annotator Agreement (IAA) scores for sentiment annotation across Iranic language varieties. **KA** denotes Krippendorff's α, measuring overall annotation reliability across multiple annotators, while **CK** denotes the average pairwise Cohen's κ, reflecting annotator consistency at the pair level. Higher values indicate stronger agreement.

5.1 Language Detection

We evaluate LLMs on identifying the *language* (L), *dialect* (D), and *accent* (A) of samples in APARSIN, a challenging setting due to the close relatedness of Iranic varieties, shared Perso-Arabic script, and non-standardized orthography. As shown in Table 4, all models perform poorly at the language level (macro-F1 ≤ 0.18), frequently confusing closely related dialects such as Persian, Kurdish, and Luri. Dialect identification shows slightly higher but still limited performance (macro-F1 up to 0.22), with better results for more lexically distinctive varieties such as Semnani and Dezfuli, while Persian regional dialects are often conflated. Accent identification achieves the highest scores (macro-F1 up to 0.48), particularly for varieties with salient non-standard lexical or phonological cues reflected in writing (e.g., Hazaragi, Southern Kurdish), whereas *General* accents remain difficult to detect.

5.2 Machine Translation

Each model was prompted to translate sentences from the Iranic languages into English, and the resulting machine-generated translations were compared against the gold English references available in our dataset. Translation quality was measured using BERTScore by computing the semantic similarity between the embeddings of the reference English translations and the model-generated English outputs, as well as BLEU for surface-level n-gram overlap. As shown in Tables 5 and 6, GPT-4o and LLaMA-70B consistently achieve the highest translation quality across Iranic languages, while Qwen models perform poorly. Languages such as Khorasani and Pashto yield better scores under both BERTScore and BLEU.

5.3 Sentiment Classification

As shown in Table 7, larger models generally achieve higher macro F1 scores across languages, dialects, and accents, with GPT-4o and Gemma-27B obtaining the best average performance, while smaller and Qwen models show noticeably lower results.

6 Conclusion

In this work, we presented APARSIN, a multi-variety benchmark for sentiment analysis and machine translation across 14 Iranic languages and dialects. By explicitly modeling variation at the language, dialect, and accent levels, our dataset exposes significant disparities in model performance, particularly for low-resource and endangered varieties. Experimental results with instruction-tuned LLMs show that strong performance on standard Persian does not generalize to other Iranic varieties. We hope that APARSIN will encourage future research on inclusive evaluation, data collection, and modeling approaches for underrepresented language communities.

Limitations

This work has several limitations. First, due to recent internet blackouts in Iran, communication with a subset of annotators was temporarily disrupted, resulting in the loss of a small portion of annotation results. These missing annotations will be recovered and incorporated into a future release of the dataset once connectivity with Iran is fully restored. Second, although APARSIN covers a broad range of Iranic languages and dialects, the dataset remains relatively small in scale, with 1,400 annotated samples. This limits the extent to which conclusions can be generalized, particularly for training data-intensive models. Finally, our experiments focus on sentiment analysis and machine translation using instruction-tuned LLMs. While these models provide a strong baseline, they may not reflect the performance of task-specific or fine-tuned models, nor do they fully capture other important NLP challenges for Iranic languages, such as morphological analysis or syntactic variation. Future work should address these limitations by expanding the dataset and incorporating additional tasks.

Lang	Dialect	Accent	Models																							
			GPT-4o			GPT-4o-mini			Qwen-32B			Qwen-14B			Gemma-27B			Gemma-12B			LLaMA-70B			LLaMA-8B		
			L	D	A	L	D	A	L	D	A	L	D	A	L	D	A	L	D	A	L	D	A	L	D	A
Caspian	Mazandarani	Tonekaboni	0.00	0.29	0.09	0.00	0.15	0.30	0.00	0.21	0.23	0.00	0.01	0.00	0.00	0.01	0.28	0.00	0.04	0.20	0.00	0.16	0.32	0.00	0.44	1.00
Central Iran	Zoroastrian Dari	Yazdi	0.00	0.00	1.00	0.00	0.00	0.46	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.33	0.00	0.00	1.00	0.00	0.00	0.49
Kurdish	Southern	Southern	0.32	0.03	1.00	0.24	0.03	1.00	0.14	0.00	0.35	0.01	0.00	0.01	0.49	0.21	1.00	0.17	0.14	1.00	0.24	0.04	1.00	0.23	0.00	1.00
Luri	Southern	General	0.11	0.30	0.00	0.00	0.49	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.33	0.00	0.00	1.00	0.00	0.05	0.25	0.00	0.00	1.00	0.00
Pashto	Central	General	0.33	0.03	0.00	0.23	0.00	0.00	0.15	0.19	0.00	0.02	0.01	0.00	0.50	0.02	0.00	0.19	0.02	0.00	0.24	0.06	0.44	0.16	0.28	0.48
Persian	Dari	General	0.16	0.10	0.00	0.32	0.07	0.00	0.32	0.05	0.19	0.08	0.00	0.00	0.24	0.24	0.19	0.24	0.13	0.23	0.16	0.03	0.00	0.24	0.00	0.03
Persian	Hazaragi	Hazaragi	0.06	0.09	1.00	0.14	0.06	0.49	0.15	0.00	0.37	0.03	0.00	0.07	0.22	0.00	1.00	0.13	0.03	1.00	0.10	0.05	1.00	0.31	0.00	1.00
Persian	Iranian	General	0.50	0.32	0.50	0.25	0.24	0.00	0.36	0.15	0.13	0.07	0.03	0.03	0.33	0.29	0.19	0.50	0.26	0.30	0.33	0.33	0.33	0.33	0.50	0.19
Persian	Iranian	Isfahani	0.14	0.10	0.00	0.15	0.09	0.12	0.13	0.08	0.00	0.03	0.03	0.00	0.32	0.04	0.00	0.19	0.02	0.01	0.22	0.14	0.02	0.32	0.24	0.00
Persian	Iranian	Shirazi	0.19	0.14	0.09	0.24	0.11	0.00	0.18	0.17	0.03	0.04	0.00	0.00	0.24	0.07	0.07	0.24	0.06	0.06	0.22	0.15	0.01	0.24	0.24	0.01
Persian	Iranian	Yazdi	0.16	0.17	0.00	0.22	0.11	0.00	0.17	0.17	0.00	0.01	0.01	0.00	0.23	0.10	0.00	0.23	0.05	0.00	0.18	0.20	0.00	0.22	0.31	0.00
Semnani	Semnani	Semnani	0.00	0.48	1.00	0.00	0.49	1.00	0.00	0.35	0.38	0.00	0.07	0.04	0.00	1.00	1.00	0.00	0.49	1.00	0.00	0.23	1.00	0.00	0.50	1.00
Unclassified	Dezfuli	Dezfuli	0.00	0.46	1.00	0.01	0.47	0.50	0.00	0.21	0.33	0.00	0.00	0.01	0.00	0.43	1.00	0.00	0.41	1.00	0.00	0.50	1.00	0.00	0.18	1.00
	Macro-F1		0.14	0.22	0.42	0.18	0.20	0.48	0.13	0.14	0.31	0.02	0.02	0.02	0.15	0.18	0.45	0.13	0.17	0.47	0.15	0.19	0.45	0.16	0.22	0.47

Table 4: Evaluation of LLMs on the APARSIN dataset for identifying *language* (L), *dialect* (D), and *accent* (A) across Iranic varieties.

Language	GPT-4o	GPT-4o-mini	Qwen-32B	Qwen-14B	Gemma-27B	Gemma-12B	LLaMA-70B	LLaMA-8B
Shirazi	0.4590	0.4489	-0.2614	-0.2790	0.0906	0.4514	0.3887	0.1631
Yazdi	0.4660	0.4147	-0.2208	-0.2483	0.0554	0.4572	0.3463	0.0894
Esfahani	0.4339	0.4192	-0.2661	-0.2890	0.0372	0.4381	0.3360	0.0784
Khorasani	0.6259	0.6364	-0.2284	-0.2363	0.4200	0.6870	0.6536	0.4685
Kabuli	0.3014	0.3208	-0.2694	-0.2805	0.1508	0.3451	0.3254	0.1877
Hazaragi	0.3947	0.3482	-0.2412	-0.2747	0.0496	0.3732	0.3414	0.0258
Pashto	0.5747	0.5540	-0.2327	-0.2937	0.4033	0.5652	0.5212	0.3158
Kalhori	0.3168	0.2493	-0.3115	-0.3203	0.1073	0.2672	0.2332	0.0367
Luri Bakhtiari	0.3252	0.2720	-0.2469	-0.2687	0.0597	0.3011	0.2567	0.1014
Dezfuli	0.3695	0.3221	-0.2747	-0.2735	-0.0260	0.3520	0.2723	0.0280
Tonekaboni	0.4452	0.3758	-0.2855	-0.2924	-0.0095	0.4281	0.3602	0.0109
Semnani	0.3298	0.2688	-0.3080	-0.3146	-0.0823	0.2764	0.2495	-0.0264
Zoroastrian Dari	0.2306	0.1870	-0.2145	-0.2276	-0.1515	0.2276	0.1718	-0.0036

Table 5: BERT Score results for translation from Iranic languages into English across different LLMs.

Language	GPT-4o	GPT-4o-mini	Qwen-32B	Qwen-14B	Gemma-27B	Gemma-12B	LLaMA-70B	LLaMA-8B
Shirazi	0.2078	0.1899	0.0016	0.0029	0.0258	0.1091	0.0950	0.0211
Yazdi	0.2073	0.1556	0.0028	0.0029	0.0187	0.1005	0.0826	0.0216
Esfahani	0.1482	0.1602	0.0016	0.0019	0.0177	0.0942	0.0714	0.0219
Khorasani	0.3642	0.4195	0.0101	0.0077	0.1368	0.3425	0.2909	0.1326
Kabuli	0.1213	0.1647	0.0028	0.0030	0.0603	0.1170	0.1130	0.0492
Hazaragi	0.1049	0.0783	0.0012	0.0004	0.0147	0.0585	0.0553	0.0061
Pashto	0.2671	0.2457	0.0030	0.0029	0.1173	0.1782	0.1959	0.0613
Kalhori	0.0599	0.0294	0.0008	0.0012	0.0201	0.0313	0.0346	0.0074
Luri Bakhtiari	0.0755	0.0461	0.0010	0.0007	0.0128	0.0505	0.0555	0.0093
Dezfuli	0.1105	0.0734	0.0017	0.0016	0.0153	0.0854	0.0468	0.0132
Tonekaboni	0.1437	0.1029	0.0021	0.0016	0.0130	0.0791	0.0789	0.0088
Semnani	0.0545	0.0608	0.0006	0.0004	0.0058	0.0260	0.0205	0.0022
Zoroastrian Dari	0.0431	0.0152	0.0005	0.0003	0.0024	0.0123	0.0176	0.0025

Table 6: BLEU scores for translation from Iranic languages into English.

Language	GPT-4o	GPT-4o-mini	Qwen-32B	Qwen-14B	Gemma-27B	Gemma-12B	LLaMA-70B	LLaMA-8B
Tonekaboni	0.6363	0.6286	0.2132	0.3033	0.6551	0.6412	0.5044	0.3104
Zoroastrian Dari	0.0550	0.0550	0.2258	0.2782	0.0550	0.0550	0.1925	0.1082
Kalhori	0.4557	0.5305	0.2938	0.2342	0.6188	0.3678	0.4666	0.2507
Pashto	0.6849	0.7105	0.1764	0.1705	0.6576	0.6618	0.4242	0.3489
Esfahani	0.6410	0.6413	0.3160	0.3754	0.5847	0.5676	0.5489	0.3992
Shirazi	0.6693	0.6500	0.3252	0.2796	0.6552	0.6730	0.5381	0.3237
Semnani	0.5199	0.4847	0.2582	0.2772	0.4909	0.4696	0.3692	0.3544
Dezfuli	0.5999	0.5490	0.2828	0.2670	0.5815	0.5820	0.3752	0.3955
AVERAGE	**0.5327**	**0.5312**	**0.2614**	**0.2732**	**0.5374**	**0.5023**	**0.4274**	**0.3114**

Table 7: Sentiment performance across Iranic dialects and related languages. All scores are macro F1.

References

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. Benchmarking large language models for persian: A preliminary study focusing on chatgpt. *Preprint*, arXiv:2404.02403.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Basant Agarwal and Namita Mittal. 2016. Machine learning approach for sentiment analysis. In *Prominent Feature Extraction for Sentiment Analysis*, pages 21–45. Springer.

Zahra Azin and Sina Ahmadi. 2021. Creating an electronic lexicon for the under-resourced southern varieties of kurdish language. In *Proceedings of Seventh Biennial Conference on Electronic Lexicography (eLex 2021)*.

Soran Badawi. 2023. Kmd: A new kurdish multilabel emotional dataset for the kurdish sorani dialect. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing*, pages 308–315, Online. Association for Computational Linguistics.

Mehrdad Bahar. 1995. *Bundahišn*. Toos Publications, Tehran, Iran.

Ramy Baly, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Khaled Bashir Shaban, and Wassim El-Hajj. 2017. Comparative evaluation of sentiment analysis methods across arabic dialects. *Procedia Computer Science*, 117:266–273.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Rahim Dehkharghani. 2019. Sentifars: A persian polarity lexicon for sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2):1–12.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding. *arXiv*, abs/2005.12515.

Farhan Farsi, Farnaz Aghababaloo, Shahriar Shariati Motlagh, Parsa Ghofrani, MohammadAli Sadraei-Javaheri, Shayan Bali, Amir Hossein Shabani, Farbod Bijary, Ghazal Zamaninejad, AmirMohammad Salehoof, and Saeedeh Momtazi. 2025a. MELAC: Massive evaluation of large language models with alignment of culture in Persian language. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1933–1950, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Farhan Farsi, Parnian Fazel, Farzaneh Goshtasb, Nadia Hajipour, Sadra Sabouri, Ehsaneddin Asgari, and Hossein Sameti. 2025b. PahGen: Generating Ancient Pahlavi text via grammar-guided zero-shot translation. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 171–182, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Mehrdad Ghadrdan. 2007. *A Linguistic and Grammatical Study of the Zoroastrian Dialect of Sharifabad, Ardakan, Yazd*. Rakhshid Publishing, Shiraz, Iran.

Saloumeh Gholami. 2018. Remnants of zoroastrian dari in the colophons and sālmargs of iranian avestan manuscripts. *Iranian Studies*, 51(2):195–211.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. Glottolog 5.2. http://glottolog.org. Accessed 2025-12-31.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *Proceedings of the International Conference on Learning Representations*.

Mohammad Ali Hussiny, Mohammad Arif Payenda, and Lilja Øvrelid. 2024. Persianemo: Enhancing farsidari emotion analysis with a hybrid transformer and recurrent neural network model. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group*

on Under-resourced Languages @ LREC-COLING 2024, pages 257–263, Torino, Italy. ELRA and ICCL.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Uzair Kamal, Imran Siddiqi, Hammad Afzal, and Arif ur Rahman. 2016. Pashto sentiment analysis using lexical features. In *Proceedings of the Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, Tebessa, Algeria.

Arta Modarres Kamaly. 2025. Towards inclusive nlp: Evaluating llms on low-resource indo-iranian languages. In *5th Muslims in ML Workshop co-located with NeurIPS 2025*.

Roland G Kent. 1953. *Old Persian: Grammar. Texts. Lexicon*, volume 33. American Oriental Society.

Rabia Khan, Huzaifa Saleem Khan, and Shireen Ijaz. 2025. English-pashto language dataset (epld). *Mendeley Data*, (V1).

Binh Le and Huy Nguyen. 2015. Twitter sentiment analysis using machine learning techniques. In *Advanced Computational Methods for Knowledge Engineering*, volume 358 of *Advances in Intelligent Systems and Computing*, pages 279–288. Springer.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Cambridge, UK.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Leland McInnes, John Healy, and James Melville. 2018. Umap: uniform manifold approximation and projection for dimension reduction. arxiv. *arXiv preprint arXiv:1802.03426*, 10.

H. Mirzaee, J. Peymanfard, H. Habibzadeh Moshtaghin, and 1 others. 2025. Armanemo: A persian dataset for text-based emotion detection. *Language Resources and Evaluation*, 59:2565–2587.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiao-Dan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, and 8 others. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Sietze Norder, Laura Becker, Hedvig Skirgård, Leonardo Arias, Alena Witzlack-Makarevich, and Rik van Gijn. 2022. glottospace: R package for the geospatial analysis of linguistic and cultural data. *Journal of Open Source Software*, 7(77):4303.

OpenAI. 2025a. Gpt-5. OpenAI model documentation. https://platform.openai.com/docs/models/gpt-5.

OpenAI. 2025b. Introducing gpt-4o-mini. https://platform.openai.com/docs/models/gpt-4o-mini. Accessed: 2025-06-04.

Ebrahim Rezapour and Mona Pishgou. 2013. A typological study of causative constructions in the semnani language. *Journal of Linguistic and Rhetorical Studies*, (23):175–214.

Arash Salar. 2019. The case system in the semnani language within a categorial framework. M.a. thesis, Faculty of Persian Literature and Foreign Languages, University of Tehran, Tehran, Iran.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Gernot Windfuhr. 2009. *The Iranian Languages*, volume 2009. Routledge London.

Gernot L. Windfuhr. 2006. Iran vii. non-iranian languages (6) in islamic iran. In *Encyclopaedia Iranica*, volume 13, pages 393–396. Encyclopaedia Iranica Foundation.

Michael Witzel. 2003. *Linguistic evidence for cultural exchange in prehistoric western Central Asia*. 129. Department of East Asian Languages and Civilizations, University of ….

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, 53(6):4335–4385.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

A Endangered languages in the Dataset

Zoroastrian Dari (also known as Behdini or Gavruni) is an endangered Iranic language spoken by Zoroastrian communities primarily in Yazd and surrounding areas of Iran, with smaller populations in Kerman and Tehran. The language exhibits substantial dialectal variation and is used almost exclusively in spoken form, with no standardized written tradition (Gholami, 2018). Within the Yazd Province, Zoroastrian Dari includes multiple dialects specific to individual villages and Zoroastrian quarters. These varieties differ in phonetics, phonology, morphology, syntax, and vocabulary. In this study, we focus specifically on the Zoroastrian dialect of Sharif Abad (Ardakan, Yazd), for which the monograph by Ghadrdan (2007) provides the earliest and most comprehensive linguistic description available to date. Owing to the absence of publicly available corpora and the exclusively oral nature of the dialect, this study relies on examples and elicited materials documented in that work as the primary source for dataset construction.

Sentiment annotation for this dialect was challenging due to its limited use among younger speakers. Since many sentences were translated into Sharif Abad's Zoroastrian Dari and written in our transliteration scheme, annotators were unfamiliar with reading them. We addressed this by reading examples aloud during voice-based annotation sessions. Utterances were delivered with neutral prosody to minimize potential annotation bias, and repetitions were used when necessary to support accurate understanding. These methodological adaptations indicate the challenges of working with an endangered, predominantly oral language and should be considered when interpreting the sentiment annotations.

Semnani is another endangered Iranic language in our dataset spoken in and around the city of Semnan in north central Iran. From a genetic and historical perspective, Semnani is rooted in Parthian (also known as Arsacid Pahlavi) and belongs to the Northwestern branch of Iranic languages (Salar, 2019). The language holds a distinctive position within the Iranian linguistic landscape due to its high degree of structural preservation and authenticity, while at the same time facing the risk of extinction (Rezapour and Pishgou, 2013). Semnani comprises several closely related but distinct varieties spoken in Semnan and neighboring areas, including Sorkhe'i, Lasgerdi, Sangsari, and

Aftarabadi.

The language lacks publicly available corpora and is transmitted primarily through oral use, though written forms use the Perso-Arabic script. However, ongoing community-based revitalization efforts have supported continued engagement with the language, including emerging literacy among younger speakers. Accordingly, sentiment annotation was conducted by three literate community members (two aged 26 and one aged 43), demonstrating the viability of written annotation despite the language's endangered status.

B Topic Modeling and Sample Selection

To identify semantically coherent patterns across Iranic languages and dialects, we employ a topic modeling framework that integrates semantic representation learning, non-linear dimensionality reduction, density-based clustering, and large language model–based topic labeling.

First, all textual samples are encoded into a shared semantic space using a multilingual sentence embedding model. This representation captures cross-lingual and intra-dialectal semantic similarity, enabling meaningful comparison between closely related language varieties as well as more distant ones. The use of a transformer-based encoder ensures robustness to lexical variation and supports low-resource dialects.

Given the high dimensionality of the resulting embeddings, a non-linear manifold learning technique is applied to project the representations into a lower-dimensional space. The dimensionality reduction is guided by neighborhood-based hyperparameters that preserve local semantic structure while maintaining global separability between emerging topics. A higher-dimensional projection is used to support stable clustering, while a two-dimensional projection is reserved exclusively for visualization and qualitative analysis.

Topic discovery is performed using a density-based clustering algorithm. This approach does not assume a predefined number of clusters and is well-suited for data with irregular cluster shapes. A minimum cluster size constraint is imposed to ensure that identified topics correspond to semantically meaningful groupings rather than noise. Samples that do not strongly belong to any dense region are treated as outliers, preventing forced topic assignments.

For topic interpretability, each discovered cluster is labeled using a large language model. Instead of relying solely on keyword frequency, the model generates concise, human-readable topic descriptors based on representative samples within each cluster. This strategy significantly improves interpretability, particularly in multilingual and dialectally diverse settings.

The final output is a two-dimensional semantic map in which samples are grouped and labeled according to their inferred topics. Representative points are selected using medoid-based labeling to reduce visual clutter and enhance readability. This visualization supports informed sample selection and qualitative inspection of linguistic variation across the dataset.

The dataset used in this study covers ten Iranic languages and dialects, including multiple varieties of Persian as well as Pashto, Kurdish, Mazandarani, Semnani, Dezfuli, and Luri. This diversity enables the model to capture both cross-language semantic structure and fine-grained dialectal distinctions.

C Additional Prompt Templates

All classification tasks are framed as **closed-set** problems with deterministic decoding (temperature $= 0$).

C.1 Language, Dialect, and Accent Detection

Language-related identification is evaluated using a three-stage prompting pipeline.

System Message

You are a classifier. Given a text and an allowed label list, choose exactly one label from the list. Return a JSON object.

Stage 1: Language Identification

Task: Identify the language of the given text.
Output format (JSON): {"language": "<LABEL>"}
Constraint: The value for "language" must be exactly one item from the allowed list.
Allowed languages (choose ONE): [...]
Text: <TEXT>

Stage 2: Dialect Identification

Task: Identify the dialect of the text (language is given).
Output format (JSON): {"dialect": "<LABEL>"}
Constraint: The value for "dialect" must be exactly one item from the allowed list.
Language: <LANGUAGE>
Allowed dialects (choose ONE): [...]
Text: <TEXT>

Stage 3: Accent Identification

Task: Identify the accent of the text (language and dialect are given).
Output format (JSON): {"accent": "<LABEL>"}
Constraint: The value for "accent" must be exactly one item from the allowed list.
Language: <LANGUAGE>
Dialect: <DIALECT>
Allowed accents (choose ONE): [...]
Text: <TEXT>

C.2 Translation

System Message

You are a professional translator with expertise in Iranic languages. Just return the translation.

Text Translation

Task: Translate the given text into the target language, considering the provided language, dialect, and accent.
Output format: Plain text (translation only, no explanations).
Source Language: <LANGUAGE>
Dialect: <DIALECT>
Accent: <ACCENT>
Target Language: <TARGET_LANGUAGE>
Text: <TEXT>

C.3 Sentiment Classification

Sentiment classification is conducted using metadata-aware prompting.

System Message

You are a classifier. Given a text and an allowed label list, choose exactly one label from the list. Return a JSON object.

Sentiment Classification

Task: Identify the sentiment of the text (language, dialect, and accent are given).
Output format (JSON): {"sentiment": "<LABEL>"}
Constraint: The value for "sentiment" must be exactly one item from the allowed list.
Language: <LANGUAGE>
Dialect: <DIALECT>
Accent: <ACCENT>
Allowed sentiments (choose ONE): [...]
Text: <TEXT>

Language/Dialect	Example	Transliteration	English	Standard Persian
Semnani	خوشتُنَه بَخُو، مَرتُمُنَه تُون کَه.	khoshtona bakhow, martomona town ka	Eat for yourself, dress for others.	برای خودت بخور، برای مردم بپوش.
Dezfuli	سی چ؟	si che	For what?	برای چه؟
Esfahani	حجمی ساندویچش کلا نصف شده‌س.	hajmi sāndevichesh kolan nesf shodees	The size of his sandwich has completely halved!	حجم ساندویچش کلا نصف شده!
Yazdi	موخوم یققم پاره کنم.	mokhom yagham pāra konam	I want to tear my collar (from despair/anger)!	می‌خواهم یقه‌ام را پاره کنم!
Zoroastrian Dari	not available	tow kovari barem za	From which direction has the sun risen?	آفتاب از کدام سمت درآمده است؟
Tonekaboni	اَما هم آنیم.	amā ham ānim	We are coming too.	ما هم می‌آیم.
Kalhori	ای گیان ئەڕا شارگان.	ey gyān arrā shāragamān	Oh, our dear city.	ای جان برای شهرمان.
Luri Bakhtiari	هر چه دیش عاقل بی خوش کلوعه.	har che deish āqel bi khosh kalu'e	Unlike her mother, who was sensible, she seems to be a fool.	برعکس مادرش که عاقل بود، خودش انگار دیوانه است.
Shirazi	خو پَ چرو نمیوی؟	kho pa chero namioy	Then why don't you come?	آخه پس چرا نمی‌آیی؟
Khorasani	یره چقد خیت رِفت.	yare cheqad khit reft	Bro, that was such a waste!	داداش، چقدر ضایع شد.
Kabuli Dari	خدا خر ره دیده که برش شاخ نداده.	Khodā khar ra dida ke barash shākh nadāda	God knew the donkey well, so He didn't give it horns.	خدا خر را دیده که به او شاخ نداده.
Pashto	هوا ناڅاپه سړه شوه۔	hawā nātsāpeh sra shoeh	The weather suddenly became cold.	هوا ناگهان سرد شد.
Hazaragi	کلو جالب خاد بود.	kalo jāleb khād bud	It would be very interesting.	خیلی جالب خواهد بود.

Table 8: Examples across varieties: original sentence (Perso-Arabic script), transliteration, English translation, and Standard Persian.

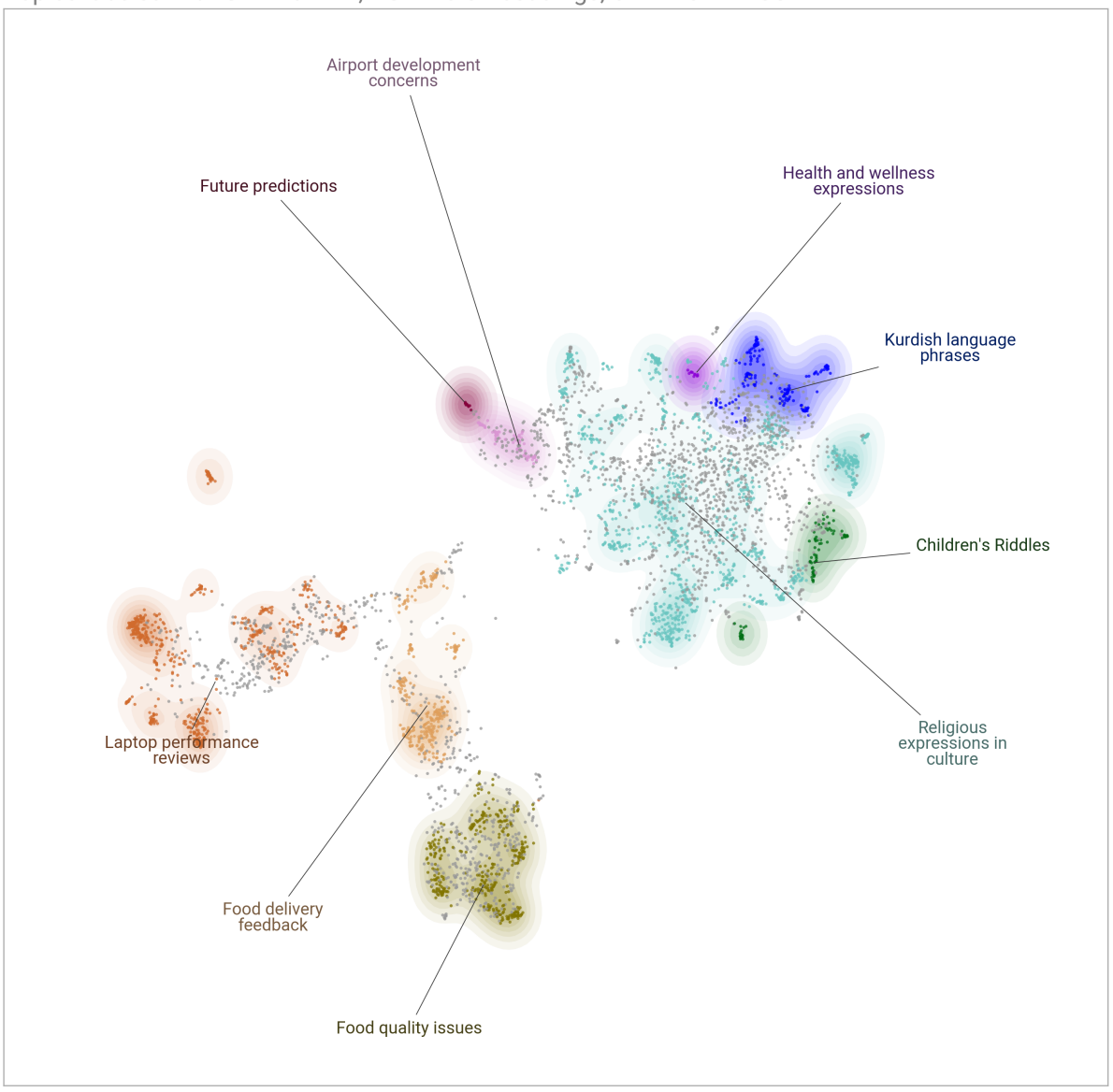


Figure 2: Topic modeling and visualization of Iranic languages and dialects. Semantic sentence embeddings are projected into a low-dimensional space using non-linear manifold learning and clustered via density-based methods. Automatically generated topic labels produced by a large language model enable interpretable analysis of cross-lingual and dialectal variation.

One Language, Three of Its Voices: Evaluating Multilingual LLMs Across Persian, Dari, and Tajiki on Translation and Understanding Tasks

Noor Mairukh Khan Arnob[1], **Abu Bakar Siddique Mahi** [2]
[1]Research & Innovation Lab, University of Asia Pacific
[2]BUBT Research Graduate School, Bangladesh University of Business and Technology
Correspondence: arnob@uap-bd.edu

Abstract

The Iranian linguistic family is pluricentric, encompassing Iranian Persian, Dari (Afghanistan), and Tajiki (Tajikistan). While Multilingual Large Language Models (MLLMs) claim broad coverage, their robustness across these regional variants and script differences (Perso-Arabic vs. Cyrillic) remains under-explored, particularly in the open-weight landscape. We evaluate five open-weight models from the Qwen, Bloomz, and Gemma families across four downstream tasks: Sentiment Analysis, Machine Translation (MT), NLI, and QA. Utilizing a dataset of over 240,000 processed samples, we observe severe performance disparities. While the fine-tuned `gemma-3-4b-persian` achieves promising results on Iranian Persian (77.3% accuracy in Sentiment), almost all tested models appear to suffer catastrophic degradation on Tajiki script (dropping to <1.0 BLEU). These findings highlight a critical "script barrier" in current open-weight MLLM development for Central Asian languages. Code and data available here.

1 Introduction

The Persian language (*Farsi*) serves as a lingua franca for millions across the Silk Road region, officially recognized as Persian in Iran, Dari in Afghanistan, and Tajiki in Tajikistan. While these varieties share a high degree of mutual intelligibility in their spoken forms, they diverge significantly in the written domain. Iranian Persian and Dari utilize the Perso-Arabic script with subtle lexical and morphological differences, whereas Tajiki adopted the Cyrillic script during the Soviet era (Windfuhr, 2009).

This pluricentric nature presents a unique challenge for Natural Language Processing (NLP). Current Multilingual Large Language Models (MLLMs) often treat Persian as a monolith, predominantly trained on data scraped from the Iranian webspace. This creates a representational bias (Blasi et al., 2022) that may marginalize Dari and Tajiki speakers.

In this work, we present a focused evaluation of five open-weight Large Language Models, ranging from 1.5B to 4B parameters. By focusing on models suitable for edge deployment rather than proprietary giants, we aim to understand the accessibility of Persian NLP. We focus on the following Research Questions (RQs):

- **RQ1 (Task Performance):** How do general-purpose multilingual models compare to language-specific fine-tunes on standard Persian tasks?
- **RQ2 (Script Robustness):** Does the change from Perso-Arabic to Cyrillic script (Tajiki) cause catastrophic generalization failure in models that claim Persian support?
- **RQ3 (Dialectal Variance):** How do models perform on Dari, which shares the script but differs in vocabulary and grammar?

We evaluate these questions across four distinct tasks: Sentiment Analysis, Natural Language Inference (NLI), Question Answering (QA), and Machine Translation (MT).

2 Related Work

2.1 Persian NLP Resources

The landscape of Persian NLP has expanded significantly in recent years, though it remains skewed towards the Iranian standard. The SentiPers corpus (Hosseini et al., 2018) established a baseline for polarity detection. However, modern evaluations require larger, diverse sources. An openly available dataset (Khayati, 2021) for persian sentiment analysis was collected from the popular iranian eccormerce website, Digikala. Another valuable resource provides comments from the food delivery service, Snappfood for sentiment analysis. The introduction of FarsTail (Amirkhani et al., 2023) for NLI and PQuAD (Darvishi et al., 2023) for

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 98–104

reading comprehension provided the first standardized benchmarks for reasoning in Persian. These datasets, however, are exclusively in Iranian Persian, limiting their utility for cross-variant assessment.

2.2 Multilingual LLM Evaluation

Evaluations of MLLMs such as BLOOM (Workshop et al., 2022) and Qwen (Bai et al., 2023) often report aggregate metrics on massive benchmarks (e.g., MMLU). While these models include Persian in their training data, detailed breakdown by dialect is rarely provided. Recent works have highlighted the "curse of multilinguality," (Chang et al., 2024) where increasing the number of languages can dilute performance on low-resource variants. Fine-tuned approaches, such as the *Gemma-Persian* model (Shojaei, 2025) adapted from Google's Gemma family, attempt to mitigate this by focusing on specific language families. Our work contextualizes these models within the specific linguistic constraints of the Iranian family.

3 Experimental Setup

3.1 Models

We selected five open-weight models based on their accessibility and claimed multilingual support. We chose moderate sized LLMs for future deployability in edge-devices. All models were loaded using 4-bit quantization (NF4) on an NVIDIA RTX 4090 GPU to simulate resource-constrained environments, with the exception of Gemma-3, which utilized bfloat16 precision based on architecture requirements.

1. **Qwen2.5-Instruct (1.5B & 3B):** Alibaba's diverse multilingual models (Team, 2024), known for strong instruction following.
2. **Bloomz (1.7B & 3B):** The instruction-tuned variants (Muennighoff et al., 2023) of the BigScience BLOOM model, trained on xP3.
3. **Gemma-3-4b-persian:** A specialized fine-tune of Google's Gemma-3, specifically optimized for Persian instruction following (Shojaei, 2025).

3.2 Data Processing and Prompts

We utilized a unified data loading pipeline to process raw datasets into a standardized JSONL format.

Preprocessing: Text normalization was applied to handle common Persian orthographic variations (e.g., unifying Arabic/Persian *Ye* (ی) and *Kaf* (ک), normalizing zero-width non-joiners). For Sentiment Analysis, we mapped diverse label spaces (e.g., 5-star ratings, -2 to +2 scales) to a unified {Negative, Neutral, Positive} schema.

Prompt Engineering: We utilized English-language zero-shot prompts for all tasks to maintain consistency across the multilingual models. However, for the Machine Translation task, we explicitly conditioned the models on the specific variant. The prompts followed the template: *"Translate the following [Source Language] text to English"*, where [Source Language] was dynamically populated as "Dari", "Tajik", or "pes" (Standard Persian) based on the FLORES-200 subset. This ensures that performance degradation on variants (e.g., Tajiki) is due to model capability gaps rather than ambiguous instructions. English prompts were chosen for cross-model consistency since all models are instruction-tuned in English, but Persian-language prompts could yield different results. (See Appendix A for full templates).

3.3 Tasks and Datasets

We construct an evaluation benchmark spanning over 240,000 samples across four tasks, from which 1,000 test instances per task are sampled for model evaluation. Details of the datasets, their sources, and the specific variants covered are detailed in Table 1. For deterministic evaluation considering time and computational constraints, we sampled a stratified test set of $N = 1000$ per task (Seed=42).

Task	Dataset	Domain	Total Size	Variant
Sentiment	SentiPers (Hosseini et al., 2018)	Digital Reviews	15.6k	Iranian
	Digikala (Khayati, 2021)	E-commerce	98.4k	
	SnappFood (Farahani et al., 2021)	Food Delivery	70.0k	
MT	FLORES-200 (Team et al., 2022)	Wiki/News	3.0k	Pes/Prs/Tgk
	Tatoeba (Tiedemann, 2020)	General	13.7k	Iranian
NLI	FarsTail (Amirkhani et al., 2023)	General	10.3k	Iranian
QA	PQuAD (Darvishi et al., 2023)	Wikipedia	60.3k	Iranian

Table 1: Summary of datasets used in this study. "Pes", "Prs", and "Tgk" refer to Iranian Persian, Dari, and Tajiki respectively. Total Size refers to available samples before test split sampling.

3.4 Evaluation Metrics

For classification tasks, we use Accuracy and Macro-F1. We incorporated a strict parsing logic: model outputs that did not strictly match the label space were categorized as "unknown," penalizing the accuracy score. For QA, we report Exact Match (EM) and F1. For MT, we report BLEU and chrF scores using `sacrebleu`. chrF is particularly im-

portant for Persian due to its agglutinative morphology, where token-based BLEU may penalize valid variations.

4 Results & Analysis

4.1 Overall Performance (RQ1)

Table 2 summarizes the performance across all tasks. Addressing **RQ1**, the fine-tuned **Gemma-3-4b-persian** demonstrates superior performance compared to general-purpose baselines, achieving the highest scores in Sentiment (77.3%), QA (41.0 EM), and MT (23.3 BLEU). This suggests that for languages with high morphological richness like Persian, general multilingual pre-training is perhaps less effective than targeted fine-tuning.

4.2 Sentiment and NLI Analysis

In Sentiment Analysis (Figure 1), the Qwen and Gemma models performed consistently well. Notably, a source breakdown revealed that models performed best on SnappFood (short, informal reviews) compared to SentiPers (formal). For example, `Qwen2.5-3B` achieved 78.7% on SnappFood but only 61.8% on SentiPers, suggesting modern LLMs are better aligned with web-style informal text.

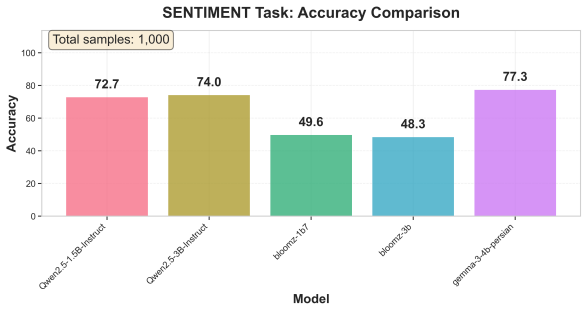


Figure 1: Sentiment Accuracy by Model. Gemma and Qwen significantly outperform the Bloomz baselines.

To better understand the failure modes of these models, we analyzed the specific error types in the sentiment task, as shown in Figure 2. The stacked bar chart reveals a distinct pattern: while the `gemma-3-4b-persian` model maintains a relatively low error count overall, the baseline models (particularly Bloomz) exhibit a massive disparity. The high "parsing failure" rate in older models is absent in Qwen and Gemma, yet the latter still struggle with distinguishing between *Neutral* and *Negative* sentiment. This typically occurs in Persian reviews where polite formalities (Taarof) often mask negative feedback, confusing models that lack deep cultural fine-tuning.

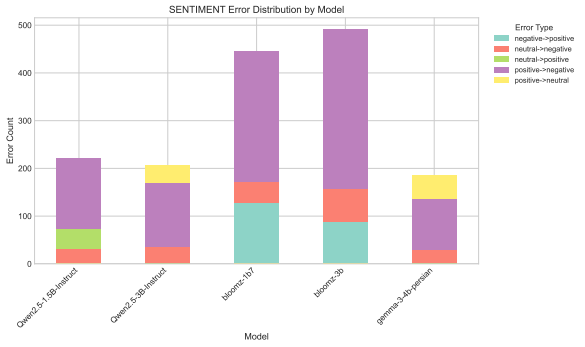


Figure 2: Distribution of top error types in Sentiment Analysis. While newer models avoid formatting errors, conflation of Neutral and Negative classes remains a persistent linguistic challenge.

The Bloomz family struggled significantly with NLI, yielding accuracy scores near the random baseline (33%). Analysis of the logs reveals a high parsing failure rate for Bloomz (up to 1.8%), indicating that the models often failed to generate the specific labels requested, instead hallucinating continuations of the premise.

4.3 The Script Barrier: Tajiki Robustness (RQ2)

Addressing **RQ2**, our findings indicate a severe cross-script degradation of models on the Tajiki variant. As illustrated in Figure 3, there is a stark disparity between performance on Perso-Arabic scripts (Persian, Dari) and Cyrillic (Tajiki).

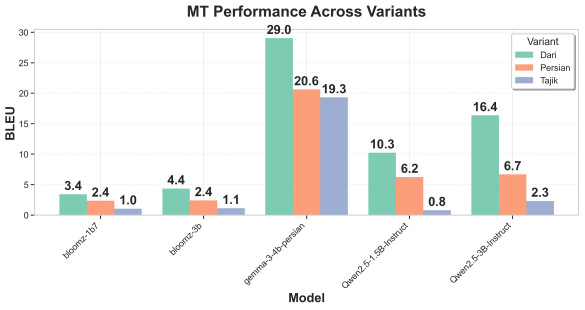


Figure 3: Performance across Persian variants (MT Task). Note the near-zero performance on Tajiki for Qwen and Bloomz.

For the `Qwen2.5-1.5B` model, performance drops from 6.2 BLEU on Persian to 0.8 BLEU on Tajiki. This implies that the model probably treats Tajiki as an entirely foreign language, despite it being linguistically identical to Persian in syntax and morphology. However, it is to be noted that we did

Model	Sentiment		NLI		MT (Avg)		QA	
	Acc	F1	Acc	F1	BLEU	chrF	EM	F1
`Bloomz-1b7`	49.6	35.3	33.9	20.3	2.4	21.2	10.1	23.6
`Bloomz-3b`	48.3	33.8	35.4	24.8	2.7	22.6	17.5	32.5
`Qwen2.5-1.5B`	72.7	57.9	56.2	51.0	4.6	32.0	10.8	35.2
`Qwen2.5-3B`	74.0	62.9	**63.9**	**59.3**	7.6	37.4	6.1	37.4
`gemma-3-4b-persian`	**77.3**	**67.8**	49.4	44.2	**23.3**	**50.8**	**41.0**	**68.4**

Table 2: Main results on Persian evaluation tasks. `gemma-3-4b-persian` dominates in generation tasks (MT, QA) and Sentiment, while `Qwen2.5-3B` shows strong reasoning capabilities in NLI.

not conduct a controlled transliteration experiment to disentangle whether the observed performance gap is primarily due to script mismatch (Cyrillic tokenization) or genuine dialectal divergence. The `gemma-3-4b-persian` model is the outlier, maintaining a score of 19.3 BLEU on Tajiki. This suggests that the base Gemma model's pre-training may have included Cyrillic data (perhaps Russian or Central Asian text) that allows for cross-script transfer, a capability preserved during fine-tuning.

4.4 Question Answering Capabilities

The QA task (PQuAD) proved most difficult. The Qwen models, despite good instruction following, struggled with the extractive nature of the task. They often paraphrased the answer rather than extracting the span, leading to low Exact Match scores.

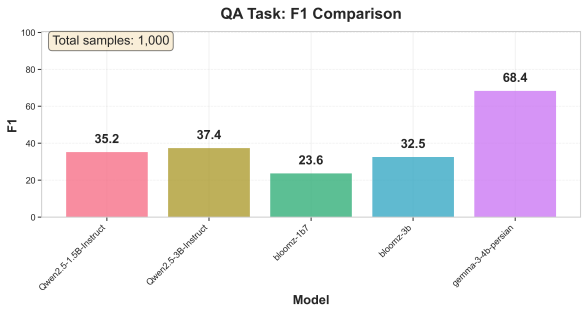


Figure 4: F1 Scores on PQuAD Question Answering. The fine-tuned Gemma model shows superior span extraction capabilities compared to general multilingual baselines.

As shown in Figure 4, textttgemma-3-4b-persian achieved a dominating F1 score of 68.4, nearly double that of the nearest competitor. This indicates that the fine-tuning process for Gemma likely included Persian reading comprehension tasks, aligning it better with the cultural context and structural requirements of PQuAD.

4.5 The Dari Divergence: Lexical vs. Structural (RQ3)

Addressing **RQ3**, qualitative analysis reveals that Dari divergence is primarily lexical rather than structural, stemming from English and Pashto loanwords (e.g., نوتهوپ (*Pohantoon*) for University) versus Iranian Persian's French influence. As detailed in Table 3, weaker models like **Bloomz** struggled with these dialectal markers, frequently hallucinating (e.g., misinterpreting *Pohantoon* as unrelated entities). In contrast, **Qwen2.5** and `gemma-3-4b-persian` demonstrated robust zero-shot generalization. Additionally, while the fine-tuned Gemma model respected Dari orthography (e.g., *mekonand*), base models often over-corrected text to the Tehrani standard during generation, highlighting a persistent training bias despite high aggregate BLEU scores.

Concept	Iranian Persian	Dari (Target)	Model Handling
University	Dāneshgāh (هاگشناد)	Pohantoon (پوهنتون)	**Gemma:** ✓ | **Bloomz:** ×
Company	Sherkat (شرکت)	Kompani (کمپنی)	**All Models:** ✓
Technology	Fanāvari (فناوری)	Teknālozhi (تکنالوژی)	**Qwen:** ✓ | **Bloomz:** ×
Policy	Siāsat (سیاست)	Pālisi (پالیسی)	**Gemma:** ✓ | **Qwen:** ✓

Table 3: Qualitative analysis of lexical divergence in Dari. Stronger models (Gemma/Qwen) successfully bridge the lexical gap, while weaker baselines (Bloomz) suffer from hallucinations when encountering non-Iranian specific vocabulary (e.g., Pashto loanwords like *Pohantoon*).

5 Reproducibility

To encourage future work and enable other researchers to replicate our results, we release the code for this manuscript in this link.

6 Limitations and Future Work

This study has several limitations that should be considered when interpreting our findings. All experiments used English zero-shot prompts for consistency; Persian-language prompts may yield

different results. Our model selection was limited to small open-weight models (1.5B–4B), and the absence of larger or proprietary baselines limits its generalizability. The `gemma-3-4b-persian` model used bfloat16 precision while others used NF4 quantization, which may partially explain its performance advantage. Our Tajiki results conflate script mismatch with dialectal divergence; a controlled transliteration experiment would be needed to disentangle these factors, and our findings should be read as indicating insufficient Cyrillic coverage rather than a definitive failure to understand Tajiki itself. Finally, results are from a single 1,000-sample split without confidence intervals, and bootstrap resampling would strengthen statistical reliability.

7 Acknowledgment

We would like to extend our gratitude to the Research and Innovation Lab, Department of Computer Science and Engineering, University of Asia Pacific for graciously providing ample computational resources for this research work.

8 Conclusion

Our investigation reveals a stark dichotomy in the landscape of Persian NLP within the open-weight model ecosystem: while fine-tuning has successfully adapted LLMs to the Iranian standard, a formidable "script barrier" appears to isolate Tajiki speakers, and dialectal nuances in Dari remain under-utilized. Since this study focuses on smaller open-weight LLMs, further work is required to determine if these limitations persist in larger, proprietary models. The superior performance of `gemma-3-4b-persian` underscores that generic multilingual pre-training is likely insufficient for complex tasks like QA; rather, achieving true linguistic inclusivity across the Iranian continuum requires curated, multi-script instruction tuning that treats these variants not as separate low-resource languages, but as a unified linguistic heritage.

References

Hossein Amirkhani, Mohammad AzariJafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, Zohreh Pourjafari, and Azadeh Amirak. 2023. Farstail: A persian natural language inference dataset. *Soft Computing*, pages 1–13.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. 2023. Pquad: A persian question answering dataset. *Computer Speech & Language*, 80:101486.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. 2018. Sentipers: A sentiment analysis corpus for persian. *arXiv preprint arXiv:1801.07737*.

Armin Khayati. 2021. Digikala-comments-sentiment-analysis. `https://github.com/Arminkhayati/Digikala-comments-sentiment-analysis`. GitHub repository, accessed 14 January 2026.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Mohammad Shojaei. 2025. Gemma 3-4B Persian (v0) Model. `https://huggingface.co/mshojaei77/gemma-3-4b-persian-v0`. Hugging Face model repository, accessed 15 January 2026.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jörg Tiedemann. 2020. The tatoeba translation challenge—realistic data sets for low-resource and multilingual mt. *arXiv preprint arXiv:2010.06354*.

Gernot Windfuhr. 2009. *The Iranian Languages*, volume 2009. Routledge London.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

A Prompt Templates

We utilized the following zero-shot prompts for our experiments. For the Machine Translation task, the prompt was dynamically adjusted based on the specific variant (Persian/Dari/Tajik) in the FLORES-200 dataset to ensure models were not penalized for script confusion. It is to be noted that English prompts bias toward instruction-tuned models.

Task	Template
Sentiment	Classify the sentiment of the following Persian text as one of: negative, neutral, positive. Text: {text} Respond with only the sentiment label, nothing else. Sentiment:
MT	Translate the following {source_lang} text to {target_lang}. {source_lang} text: {text} {target_lang} translation:
NLI	Given the premise and hypothesis below, determine the relationship between them. Choose one of: entailment, neutral, contradiction. Premise: {premise} Hypothesis: {hypothesis} Respond with only the label, nothing else. Relationship:
QA	Answer the question based on the given context. Extract the answer directly from the context. Context: {context} Question: {question} Answer:

Table 4: Zero-shot prompt templates used for evaluation.

B Reproducibility Details

Experiments were conducted using 4-bit quantization (NF4) for Qwen and Bloomz models to simulate consumer-grade hardware constraints. We utilized python 3.10.19 for running our scripts. The Gemma-3 model was loaded in bfloat16 precision as it was an architectural requirement for loading Gemma. We utilized the `bitsandbytes` python package for quantization and the `transformers` python pacakge for inference. Deterministic generation was enforced by setting the temperature to 0.0 and `do_sample=False` for all tasks. We acknowledge that a single 1,000-sample split without confidence intervals is a limitation, especially for noisy MT metrics. Bootstrap confidence intervals should be adopted in future works.

C Detailed Experimental Results

C.1 Sentiment Analysis by Domain

Table 5 presents the performance breakdown across the three source datasets used in the aggregated Sentiment Analysis task.

- **SnappFood:** Short, informal food delivery reviews.
- **Digikala:** Product reviews, semi-formal/colloquial.
- **SentiPers:** Formal/Academic corpus.

The results indicate that general-purpose multilingual models (Qwen, Bloomz) struggle significantly with the formal register of SentiPers (e.g., Qwen2.5-1.5B drops from 76.6% on SnappFood to 48.0% on SentiPers). In contrast, the fine-tuned **gemma-3-4b-persian** maintains high robustness (74.5% on SentiPers), suggesting successful alignment across different registers of Iranian Persian. A visual depiction of the LLMs' performance across different datasets is shown in Figure 5.

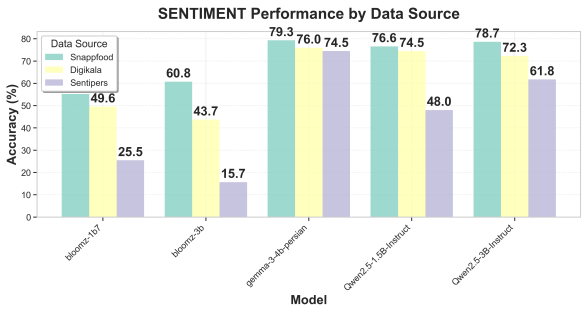


Figure 5: Performance of Multilingual LLMs across datasets of different nature, indicating more recent LLMs are better adept to handle web data.

C.2 Machine Translation by Variant and Source

Table 6 presents a granular breakdown of BLEU scores across language variants and data sources.

Model	**SnappFood** (Informal)		**Digikala** (Semi-Formal)		**SentiPers** (Formal)	
	Acc	F1	Acc	F1	Acc	F1
`Bloomz-1b7`	55.3	37.1	49.6	34.6	25.5	19.8
`Bloomz-3b`	60.8	39.7	43.7	30.7	15.7	12.9
`Qwen2.5-1.5B`	76.6	51.4	74.5	56.8	48.0	41.0
`Qwen2.5-3B`	78.7	52.6	72.3	55.6	61.8	55.7
`gemma-3-4b-persian`	**79.3**	**53.4**	**76.0**	**62.1**	**74.5**	**70.0**

Table 5: Detailed performance breakdown on Sentiment Analysis datasets. The *SentiPers* dataset proved most challenging for base models, while the fine-tuned Gemma model demonstrated cross-domain robustness.

Model	**Variant Performance (BLEU)**			**Source Breakdown (BLEU)**	
	Persian	**Dari**	**Tajiki**	**FLORES-200**	**Tatoeba**
`Bloomz-1b7`	2.4	3.4	1.0	2.5	0.2
`Bloomz-3b`	2.4	4.4	1.1	2.9	0.1
`Qwen2.5-1.5B`	6.2	10.3	0.8	5.0	0.1
`Qwen2.5-3B`	6.7	16.4	2.3	9.1	0.1
`gemma-3-4b-persian`	**20.6**	**29.0**	**19.3**	**25.0**	**0.2**

Table 6: Detailed Machine Translation results. The left section highlights the "Script Barrier" where performance collapses on Tajiki for non-fine-tuned models. The right section shows the model performance averaged across variants on specific datasets.

Three critical patterns emerge from this data:

The Script Barrier (Tajiki): The most salient result is the catastrophic failure on Tajiki. Although Tajiki and Persian are highly mutually intelligible and share core syntax and morphology, the Cyrillic script poses a major barrier for base multilingual models. Both `Qwen2.5-1.5B` and Bloomz collapse to near-zero performance (< 1.1 BLEU), effectively treating Tajiki as unseen. In contrast, the fine-tuned gemma-3-4b-persian remains usable at 19.3 BLEU, indicating that targeted instruction tuning can bridge the Perso-Arabic–Cyrillic script gap where general multilingual pre-training fails.

The Dari Inversion: Contrary to expectations, models often perform better on Dari than on Standard Persian (e.g., Gemma-3: 29.0 vs. 20.6 BLEU). Qualitative analysis suggests this is not due to better dialectal modeling, but to test-set artifacts: the Dari subset of FLORES-200 contains simpler, less metaphorical sentences that models translate more literally and accurately.

Domain Brittleness (Tatoeba vs. FLORES): We observe a variation between data sources. While models achieve respectable performance on FLORES-200 (Wiki/News domain), they fail almost completely on the Tatoeba dataset (averaging < 0.2 BLEU) (Figure 6). Tatoeba consists largely of short, context-free, and idiomatic sentences. The models' inability to handle these samples indicates a high sensitivity to domain and sentence length; they struggle to ground short, ambiguous text without the extensive context provided in Wikipedia-style paragraphs.

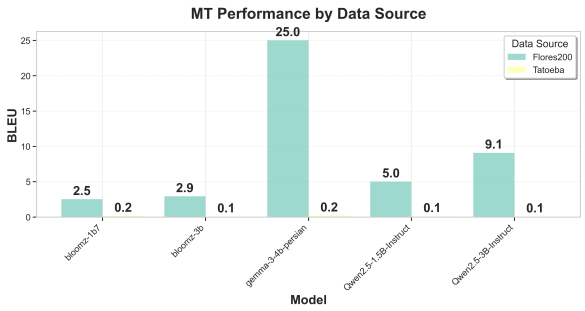


Figure 6: Comparison of Machine Translation performance (BLEU) across data sources. While models achieve reasonable scores on the standard FLORES-200 benchmark, they exhibit near-total failure on the Tatoeba dataset, highlighting a lack of robustness to diverse sentence structures and community-sourced data.

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, Tehran, Iran
[2] Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
{mohammadjranjbar, hfaili, shakery}@ut.ac.ir

Abstract

Punctuation restoration is essential for improving the readability and downstream utility of automatic speech recognition (ASR) outputs, yet remains underexplored for Persian despite its importance. We introduce **PersianPunc**, a large-scale, high-quality dataset of 17 million samples for Persian punctuation restoration, constructed through systematic aggregation and filtering of existing textual resources. We formulate punctuation restoration as a token-level sequence labeling task and fine-tune ParsBERT to achieve strong performance. Through comparative evaluation, we demonstrate that while large language models can perform punctuation restoration, they suffer from critical limitations: over-correction tendencies that introduce undesired edits beyond punctuation insertion (particularly problematic for speech-to-text pipelines) and substantially higher computational requirements. Our lightweight BERT-based approach achieves a macro-averaged F1 score of 91.33% on our test set while maintaining efficiency suitable for real-time applications. We make our dataset and model publicly available to facilitate future research in Persian NLP and provide a scalable framework applicable to other morphologically rich, low-resource languages.[1]

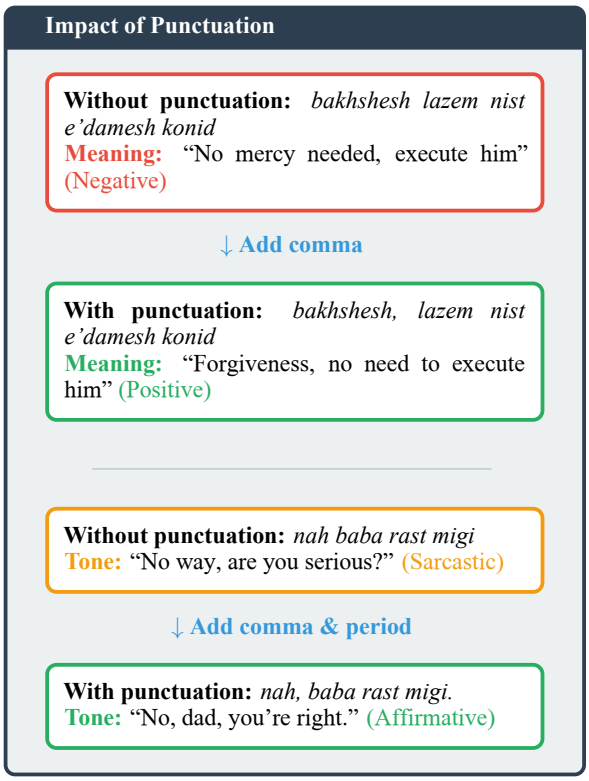


Figure 1: Persian punctuation restoration dramatically affects semantic interpretation. Minimal punctuation changes transform sentence meaning from negative to positive sentiment.

1 Introduction

Punctuation restoration represents an essential task in natural language processing, particularly for languages with limited computational resources. The absence of punctuation in raw text—whether from automatic speech recognition, informal digital communication, or historical documents—severely impacts the performance of downstream NLP tasks including machine translation, text summarization, and sentiment analysis.

Despite the growing maturity of Persian NLP, punctuation restoration has received limited attention compared to other languages. The critical importance of punctuation in Persian is evidenced by dramatic semantic changes that occur with minimal punctuation modifications, as shown in Figure 1. Existing Persian studies have been constrained by small-scale datasets, domain-specific applications, or lack of publicly available models, highlighting the critical need for comprehensive approaches that can handle the full complexity of Persian text across diverse domains and writing styles.

[1]Our resources are publicly available: full dataset (17M samples), training subset, and fine-tuned model.

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 105–113
March 29, 2026

This work addresses these challenges through a comprehensive approach to Persian punctuation restoration using fine-tuned BERT models and large-scale dataset curation. We present a dataset curation methodology that systematically aggregates multiple Persian text sources, resulting in a high-quality corpus for training robust punctuation restoration models. Our main contributions are:

- We present **PersianPunc**, a large-scale Persian punctuation restoration dataset containing 17 million filtered and deduplicated samples spanning diverse domains, sourced from six complementary corpora covering both formal and informal Persian text.
- We provide a systematic dataset curation framework including detailed preprocessing, quality filtering, and train/validation/test splits, with comprehensive analysis of punctuation distribution patterns in Persian.
- We achieve strong performance on Persian punctuation restoration with a fine-tuned ParsBERT model, demonstrating competitive results compared to large language models while requiring significantly lower computational resources and avoiding over-correction issues.

2 Related Work

Our research is situated at the intersection of punctuation restoration, Transformer-based NLP, and Persian text processing. This section reviews the evolution of methodologies for this task, starting from general approaches and progressively narrowing the focus to the specific challenges and prior work in the Persian language.

2.1 Punctuation Restoration as a Sequence Modeling Task

Historically, punctuation restoration was tackled with statistical methods, including n-gram language models (Beeferman et al., 1998; Gravano et al., 2009) and models incorporating prosodic features from speech to predict boundaries (Christensen et al., 2001; Kim and Woodland, 2003). These early systems laid the groundwork but were often limited by the scope of their handcrafted features and statistical models.

The advent of deep learning marked a significant shift. Recurrent Neural Networks (RNNs), particularly models using Bidirectional Long Short-Term Memory (BiLSTM) units, became the standard, framing the problem as a sequence labeling task where each token is classified with a punctuation mark (or none) (Xu et al., 2016). This paradigm was often enhanced with Convolutional Neural Networks (CNNs) to capture local character-level features (Tündik and Szaszák, 2018; Zelasko et al., 2018), leading to substantial performance gains over classical methods. Our work follows this successful sequence labeling formulation, leveraging a more powerful neural architecture.

2.2 Transformer-Based Approaches for Punctuation Restoration

The introduction of the Transformer architecture (Vaswani et al., 2017) and pre-trained language models like BERT (Devlin et al., 2019) revolutionized NLP. For punctuation restoration, fine-tuning BERT-based models quickly became the state-of-the-art approach in high-resource languages, demonstrating superior performance in capturing long-range dependencies crucial for understanding sentence structure and punctuation placement (Courtland et al., 2020; Yi et al., 2020; Nagy et al., 2021). These models are typically lightweight and efficient, making them suitable for real-time applications like ASR post-processing.

More recently, Large Language Models (LLMs) have demonstrated impressive capabilities in a zero-shot or few-shot capacity for various text generation and correction tasks (Brown et al., 2020). However, their application to focused tasks like punctuation restoration comes with potential drawbacks, including high computational inference costs and a tendency for over-correction, where they may alter the source text beyond simply adding punctuation. A key part of our contribution is to rigorously evaluate these trade-offs against a fine-tuned, specialized model.

2.3 State of Persian Text Processing and Punctuation

Early work on Persian punctuation restoration was pioneering but limited in scale. Hosseini and Sameti (2017) introduced the first known corpus for this task, achieving an F1-score of 69.0% with a Conditional Random Field (CRF) model. While foundational, this work highlighted the need for larger and more diverse datasets and more powerful models.

More recently, Farokhshad et al. (2021) proposed ViraPart, a multi-task text refinement framework for Persian that handles punctuation restoration, Zero-Width Non-Joiner (ZWNJ) recognition, and Ezafe construction. Using ParsBERT (Farahani et al., 2020), they achieved a strong F1-score of 92.13% for punctuation on the Bijankhan corpus (Bijankhan, 2004). Their work demonstrated the effectiveness of Transformer-based models for Persian text refinement. However, ViraPart focuses on multiple text refinement tasks simultaneously, and was evaluated on a smaller, single-domain corpus.

Despite these advances, critical gaps remain. First, there is a lack of a large-scale, publicly available dataset specifically curated for punctuation restoration across diverse domains. Most existing efforts rely on smaller or general-purpose corpora like Bijankhan. Second, the capabilities and limitations of modern LLMs for Persian punctuation restoration have not been systematically studied, particularly regarding over-correction behavior.

3 Methodology

3.1 Dataset Construction

3.1.1 Data Sources and Collection Strategy

We construct a comprehensive Persian punctuation restoration dataset by systematically aggregating high-quality corpora spanning diverse domains and registers. Our multi-source approach addresses the linguistic diversity challenges of Persian NLP through careful curation. We selected source datasets through manual inspection, verifying that at least 100 random samples from each contained proper punctuation usage.

Our dataset combines sources across two primary categories:

Formal Academic Text: Bijankhan-Peykare Corpus (Bijankhan et al., 2011), Persian Medical QA (Kalahroodi et al., 2025), and Persian Wikipedia (MaralGPT, 2023) provide standardized punctuation patterns in formal contexts, covering literary, medical, and encyclopedic domains.

Contemporary Informal Text: Persian Telegram Channels (Shojaei, 2023), Farsi Stories (Pasban, 2023), and Blog Dataset V2 (Lab, 2023) capture modern conversational patterns and varied punctuation usage, representing social media, narrative fiction, and personal blogging styles.

3.1.2 Preprocessing and Quality Control

Normalization Pipeline All texts undergo systematic preprocessing to ensure consistency:

1. **Punctuation standardization:** English punctuation marks (comma, semicolon, question mark) are converted to their Persian equivalents (Persian comma ،, Persian semicolon ؛, Persian question mark ؟).

2. **Character filtering:** Non-Persian characters are removed while preserving common Persian script variants and Arabic letters used in Persian text.

3. **Whitespace normalization:** Multiple spaces are collapsed, and leading/trailing whitespace is removed.

Sentence Segmentation and Filtering We first segment each source corpus into sentence-level units using end-of-sentence punctuation marks (period, exclamation mark, question mark). Each candidate sentence then undergoes multi-stage filtering:

- **Structural requirements:** Minimum length of 10 characters; at least two target punctuation marks from the set {., ،, , !, ;, :}; proper sentence termination with period, exclamation mark, or question mark.

- **Content filtering:** Removal of sentences containing URLs, email addresses, social media handles (@ mentions), emojis, excessive special symbols (more than 20% non-alphabetic characters), or substantial mixed-language content (more than 30% non-Persian text).

- **Linguistic quality:** Pattern-based detection and removal of repetitive punctuation (e.g., "....", "!!!!"), enumerative sequences (numbered lists, bullet points), and fragmented text (sentences with more than 50% single-character tokens).

Rationale for Filtering Criteria The requirement for at least two punctuation marks ensures that samples present meaningful punctuation restoration challenges beyond simple sentence termination. While this filtering criterion does exclude simple sentences (which are underrepresented in formal Persian writing), it ensures the

dataset focuses on the core challenge of internal sentence punctuation, which is critical for ASR applications where sentence boundaries are often detected separately. We acknowledge this as a dataset characteristic rather than a limitation, as it creates a focused benchmark for comma, colon, and question mark insertion—the most challenging and impactful aspects of Persian punctuation restoration. Future work could address simple sentence coverage through stratified sampling or separate evaluation sets.

3.1.3 Deduplication and Dataset Splitting

To ensure dataset quality and prevent data leakage, we perform exact deduplication across all source corpora. Due to the large dataset size (initial pool of over 20 million samples), we implement an efficient SHA-256 hash-based deduplication strategy with whitespace normalization. Each sentence is normalized (lowercased, whitespace-collapsed) before hashing to detect duplicates that differ only in formatting.

After deduplication, our final dataset contains 17,102,014 unique samples. For model training and evaluation, we randomly sample a 1M subset stratified by source corpus. This subset is split into training (989,000 samples), validation (10,000 samples), and test (1,000 samples) sets. The sampling strategy maintains the source distribution proportions to ensure representativeness across domains.

3.2 Dataset Statistics and Punctuation Analysis

We conducted a comprehensive punctuation analysis on the complete dataset of 17,102,014 samples to understand the characteristics of Persian punctuation usage in our corpus.

3.2.1 Punctuation Distribution

Table 1 presents the distribution of punctuation marks across the entire dataset. All samples contain at least one punctuation mark (by design), with an average of 2.51 punctuation marks per sentence.

The distribution reflects typical Persian text characteristics, where commas are heavily used for clause separation and complex sentence structures.

3.3 Punctuation Restoration Model and Training Setup

We formulate punctuation restoration as a token-level sequence labeling problem. Given an input sequence of tokens without punctuation, the model predicts a punctuation label for each token position. We define five classes: `EMPTY` (no punctuation), `COMMA` (،), `QUESTION` (؟), `PERIOD` (.), and `COLON` (:). Note that we focus on the four most common and semantically important punctuation marks, excluding exclamation marks and semicolons which are less frequent and often interchangeable with periods and commas in Persian.

Table 1: Distribution of punctuation marks in the complete dataset (17M samples).

Mark	Total Count	% of Total
Persian comma (،)	21,291,632	50.13%
Period (.)	15,076,946	35.50%
Colon (:)	4,228,554	9.96%
Exclamation (!)	1,209,227	2.85%
Persian question (؟)	665,841	1.57%
Total	**42,472,200**	**100.00%**

Model Architecture Our model architecture consists of a pre-trained ParsBERT encoder (Farahani et al., 2020) followed by a linear classification layer with dropout regularization. ParsBERT is a monolingual Persian BERT model pre-trained on a large Persian corpus, making it well-suited for Persian NLP tasks.

Given an input sentence with punctuation removed, we:

1. Tokenize using ParsBERT's WordPiece tokenizer
2. Pass tokens through ParsBERT to obtain contextualized embeddings
3. Apply dropout (p=0.1) for regularization
4. Project embeddings to 5-dimensional class logits via a linear layer
5. Assign the predicted punctuation class to each token position

For subword tokens generated by WordPiece tokenization, we assign punctuation labels only to the first subword of each word, ignoring continuation subwords during both training and evaluation. This aligns the token-level predictions with word-level punctuation placement.

Training Configuration We train on the 1M sample subset described in Section 3.1.3. The model is optimized using AdamW with a learning rate of 2×10^{-5}, weight decay of 0.01, and trained for 3 epochs. We use a batch size of 85 with gradient accumulation over 8 steps (effective batch size

of 680). The loss function is cross-entropy computed over all token positions.

Evaluation Metrics We employ standard sequence labeling metrics:

- **Per-class metrics:** Precision, recall, and F1-score for each punctuation class (`COMMA`, `PERIOD`, `QUESTION`, `COLON`)
- **Macro-averaged F1:** Arithmetic mean of per-class F1-scores, giving equal weight to each punctuation type regardless of frequency
- **Micro-averaged F1:** F1-score computed from the sum of per-class true positives, false positives, and false negatives, effectively weighting classes by their frequency
- **Full Sentence Match (FSM) Rate:** Percentage of test sentences where the predicted punctuation sequence exactly matches the gold standard. This metric is particularly important for evaluating LLMs, as it captures whether the model made any edits beyond punctuation insertion (over-correction).

Throughout this paper, unless otherwise specified, "F1-score" refers to the macro-averaged F1-score, which provides an overall measure of punctuation restoration accuracy giving equal weight to each punctuation type regardless of frequency.

4 Results and Analysis

4.1 Overall Performance

Our fine-tuned ParsBERT model achieves a macro-averaged F1-score of 91.33% and a micro-averaged F1-score of 97.28%. The macro-averaged score is lower due to the class imbalance (periods and commas dominate the dataset), but performance remains strong across all punctuation types as shown in Table 6 in the appendix.

4.2 Comparison with Prior Work

It is important to note that the CRF and ViraPart results shown in Table 2 are evaluated on different datasets (the Hosseini et al. corpus and Bijankhan corpus, respectively), and therefore cannot be directly compared to our results. We include these numbers for reference to situate our work within the Persian punctuation restoration literature, but we make no claims of superiority over these methods without evaluation on the same test set.

The substantial improvement over the CRF baseline (69.00% vs 91.33%) likely reflects both the advancement in modeling approaches (Transformer-based vs. CRF) and potential differences in dataset difficulty. The ViraPart score (92.13%) is more competitive, though direct comparison remains inappropriate due to the different evaluation sets.

4.3 Comparison with Large Language Models

We evaluated two variants of GPT-4o on our test set using the zero-shot prompt shown in Appendix 2. The prompt explicitly instructs the model to only add punctuation without modifying the source text.

Our ParsBERT model achieves 91.33% macro F1, outperforming both GPT-4o (85.96%) and GPT-4o-mini (79.54%) on the same test set. More importantly, the FSM Rate reveals a critical limitation of LLM-based approaches: GPT-4o achieves only 50.10% exact matches, while our model achieves 61.80%.

Analysis of the mismatches reveals that GPT-4o exhibits over-correction in approximately 5% of samples, making undesired edits such as:

- Removing words deemed unnecessary
- Replacing informal words with formal equivalents
- Correcting perceived spelling or grammatical errors

Notably, we observed no cases of word additions, only deletions and substitutions. This over-correction behavior is particularly problematic for ASR post-processing pipelines, where the source text (transcribed speech) should be preserved verbatim with only punctuation added.

Additionally, GPT-4o requires substantially higher computational resources for inference compared to our lightweight ParsBERT model, making it less suitable for real-time applications or deployment in resource-constrained environments.

4.4 Analysis of Model Performance

Table 6 (Appendix) provides detailed per-class performance metrics. The model performs exceptionally well on periods (F1: 98.71%), which is expected given their high frequency and relatively consistent usage patterns. Performance on other punctuation types remains strong: colons (90.45%), question marks (88.89%), and commas (80.03%).

Model	Test Set	Macro F1 (%)	FSM (%)
CRF (Hosseini and Sameti, 2017)	Hosseini et al. corpus	69.00	—
ViraPart (Farokhshad et al., 2021)	Bijankhan corpus	92.13	—
GPT-4o-mini	**Our test set**	79.54	38.01
GPT-4o (OpenAI, 2023)	**Our test set**	85.96	50.10
Our Model (ParsBERT)	**Our test set**	**91.33**	**61.80**

Table 2: Comparison of punctuation restoration performance across models. Note that CRF and ViraPart results are on different test sets and are not directly comparable. GPT-4o and our model are evaluated on the same test set from PersianPunc.

The lower performance on commas reflects their more nuanced usage in Persian, where comma placement can be somewhat flexible and context-dependent, leading to greater ambiguity in the gold standard annotations themselves.

5 Conclusion and Future Work

This work presents PersianPunc, a large-scale dataset of 17 million samples for Persian punctuation restoration, constructed through systematic aggregation and quality filtering of diverse Persian text sources. We demonstrate that a fine-tuned ParsBERT model achieves strong performance (91.33% macro F1) while avoiding the over-correction issues and computational overhead of large language models.

Our primary contribution is the dataset itself, which addresses a critical gap in Persian NLP resources. The curation methodology, including detailed preprocessing pipelines, quality filtering criteria, and comprehensive punctuation analysis, provides a framework applicable to other low-resource languages.

Future work should explore several directions. The development of domain-specific models for Persian literature, news, and social media text could address the variation in punctuation usage across different domains. Additionally, incorporating prosodic information from Persian speech could improve punctuation restoration for speech-to-text applications. Furthermore, extending the model to jointly handle punctuation restoration and Zero-Width Non-Joiner (ZWNJ) insertion would address the broader text normalization challenges specific to Persian writing systems.

6 Limitations

This work has several limitations that should be acknowledged. First, the dataset creation process relies on existing Persian texts, which may contain punctuation errors or inconsistencies that could propagate to the trained model. Second, the model's performance is optimized for contemporary Persian writing styles and may not generalize well to historical or highly specialized Persian texts. Third, our evaluation is limited to 1,000 test sentences due to resource constraints, including expensive API costs for commercial LLM evaluation and lack of GPU access for extensive experimentation. More extensive evaluation with larger test sets, multiple training runs, and statistical significance testing would strengthen our findings but was not feasible given these constraints.

References

D Beeferman, A Berger, and J Lafferty. 1998. Cyberpunc: a lightweight punctuation annotation system for speech. In *ICASSP*, pages 689–692.

M Bijankhan. 2004. The role of the corpus in writing a grammar: An introduction to a software. *Iranian Journal of Linguistics*, 19(2):48–67.

Mahmood Bijankhan, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. 2011. Lessons from building a persian written corpus: Peykare. *Language Resources and Evaluation*, 45(2):143–164.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901.

H Christensen, Y Gotoh, and S Renals. 2001. Punctuation annotation using statistical prosody mod-

els. In *Proc Isca Workshop on Prosody in Speech Recognition and Understanding*.

M Courtland, A Faulkner, and G McElvain. 2020. Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 272–279.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and M Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding. *arXiv preprint arXiv:2005.12515*.

Narges Farokhshad, Milad Molazadeh, Saman Jamalabbasi, Hamed Babaei Giglou, and Saeed Bibak. 2021. Virapart: A text refinement framework for automatic speech recognition and natural language processing tasks in persian. In *arXiv preprint arXiv:2110.09086v3*.

A Gravano, M Jansche, and M Bacchiani. 2009. Restoring punctuation and capitalization in transcribed speech. In *ICASSP*, pages 4741–4744.

Mohammadsaleh Hosseini and Hossein Sameti. 2017. Creating a corpus for automatic punctuation prediction in persian texts. In *Proceedings of the 2017 Iranian Conference on Electrical Engineering (ICEE)*, Tehran, Iran.

Mohammad Javad Ranjbar Kalahroodi, Amirhossein Sheikholselami, Sepehr Karimi, Sepideh Ranjbar Kalahroodi, Heshaam Faili, and Azadeh Shakery. 2025. Persianmedqa: Evaluating large language models on a persian-english bilingual medical question answering benchmark. *Preprint*, arXiv:2506.00250.

J Kim and P Woodland. 2003. A combined punctuation generation and speech recognition system and its performance enhancement using prosody. *Speech Communication*, 41:563–577.

RohanAI Lab. 2023. Persian blog dataset version 2 for text analysis. `https://huggingface.co/datasets/RohanAiLab/persian_blog_V2`.

MaralGPT. 2023. Persian wikipedia dataset for large language model training. `https://huggingface.co/datasets/MaralGPT/persian-wikipedia`.

Attila Nagy, Bence Bial, and Judit Ács. 2021. Automatic punctuation restoration with bert models. *arXiv preprint arXiv:2101.07343*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sina Pasban. 2023. Farsi tiny stories: A persian dataset for language model training. `https://huggingface.co/datasets/sinap/FarsiTinyStories`.

Mohammad Shojaei. 2023. Persian telegram channels dataset for nlp research. `https://huggingface.co/datasets/mshojaei77/PersianTelegramChannels`.

Márk Á Tündik and György Szaszák. 2018. Joint word- and character-level embedding cnn-rnn models for punctuation restoration. In *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 135–140.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Kaituo Xu, Lei Xie, and Kaisheng Yao. 2016. Investigating lstm for punctuation prediction. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5.

Jingsi Yi, Jianhua Tao, Zheng Tian, Ye Bai, and Cunli Fan. 2020. Focal loss for punctuation prediction. In *Proc. Interspeech*, pages 721–725.

Piotr Zelasko, Piotr Szymanski, Jan Mizgajski, Adrian Szymczak, Yishay Carmiel, and Najim Dehak. 2018. Punctuation prediction model for conversational speech. In *Proc. Interspeech*, pages 3603–3607.

A Punctuation Analysis Details

In this appendix, we provide comprehensive analysis of punctuation patterns and model performance.

A.1 Punctuation Co-occurrences

Analysis of punctuation co-occurrence reveals common patterns in Persian writing. Table 3 shows the most frequent punctuation pairs appearing together in the same sentence.

Table 3: Most frequent punctuation co-occurrences in the dataset.

Punctuation Pair	% of Sentences
Period + Persian comma	79.43%
Period + Colon	16.91%
Colon + Persian comma	10.94%
Exclamation + Persian comma	4.34%

The combination of period and Persian comma appears in nearly 80% of sentences, indicating that most sentences contain multiple clauses separated by commas before the final period.

A.2 Sentence-Level Punctuation Coverage

Table 4 shows the percentage of sentences containing each punctuation mark (counted once per sentence regardless of frequency).

Table 4: Percentage of sentences containing each punctuation mark.

Punctuation	Coverage
Period (.)	15,076,946 (88.94%)
Persian comma (،)	14,585,086 (86.04%)
Colon (:)	4,036,797 (23.81%)
Persian question (؟)	665,841 (3.93%)

A.3 Distribution of Punctuation Counts per Sentence

Table 5 presents the distribution of the number of punctuation marks per sentence. The majority of sentences (68.37%) contain exactly 2 punctuation marks, which is a direct consequence of our filtering criterion requiring at least 2 marks per sentence combined with the natural distribution in source texts.

Table 5: Distribution of punctuation counts per sentence.

# Punctuations	# Sentences	Percentage
2	11,589,324	68.37%
3	3,420,146	20.18%
4	1,034,579	6.10%
5	362,609	2.14%
6+	511,518	3.02%
Total	**17,102,014**	**100.00%**

A.4 Punctuation-Specific Performance

Table 6 presents a detailed analysis of per-class performance. The macro-averaged F1-score of 91.33% demonstrates strong overall performance across all punctuation classes.

Table 6: Per-class performance metrics for punctuation restoration on the test set (1,000 sentences).

Punctuation	Precision	Recall	F1-Score
Persian Comma (،)	0.8408	0.7635	0.8003
Period (.)	0.9855	0.9886	0.9871
Question (؟)	0.8750	0.9032	0.8889
Colon (:)	0.9137	0.8955	0.9045
Macro Average	**0.9038**	**0.8877**	**0.9202**
Micro Average	**0.9729**	**0.9727**	**0.9728**

B LLM Evaluation Prompt

We used the prompt shown in Figure 2 to evaluate GPT-4o and GPT-4o-mini. The temperature was set to 0, and maximum tokens were set to 2048 to accommodate longer outputs. Prompts were issued in English, as we found LLMs demonstrate better instruction-following in English compared to Persian.

Evaluation Prompt for LLMs

Role: You are a punctuation restoration system for Persian text.
Task: Add appropriate punctuation marks to the given Persian text.
Rules:

- Do NOT fix, correct, or modify ANY words in the text.
- Do NOT change the order of words.
- Do NOT add or remove any words.
- ONLY add punctuation marks where appropriate.
- Use these punctuation marks: . (period), ، (Persian comma), ؟ (Persian question mark), : (colon)
- Return the result as a JSON object with a single key "text" containing the punctuated text.

Input text (without punctuation): *{text}*
Output format:

```
{"text": "your punctuated text here"}
```

Important: Keep ALL words EXACTLY as they are in the input. Do NOT fix spelling, grammar, or anything else. ONLY add punctuation.

Figure 2: Prompt used for zero-shot evaluation of GPT-4o and GPT-4o-mini on Persian punctuation restoration. The system is explicitly instructed to only add punctuation marks without altering the original text in any way.

Shughni Machine Translation Enhanced by Donor Languages

Dmitry Novokshanov[1], Innokentiy S. Humonen[2, 3, 4], Ilya Makarov[2,3, 4],
[1]HSE University, [2]AXXX, [3]iMak Lab [4]Trusted AI Research Center, RAS
Correspondence author: danovokshanov@gmail.com

Abstract

This paper presents the first machine translation system for Shughni, an extremely low-resource Eastern Iranian language spoken in Tajikistan and Afghanistan. We fine-tune NLLB-200 models and explore auxiliary language selection through typological similarity and "super-donor" experiments. Our final Shughni–Russian model achieves a chrF++ score of 36.3 (45.7 on BivalTyp data), establishing the first computational translation resource for this language. Beyond reporting system performance, this work demonstrates a practical path toward supporting languages with virtually no prior MT resources.

Our demo system with Shughni-Russian-English translation (Russian serves as a pivot language for the Shughni-English pair) is available on HuggingFace (https://huggingface.co/spaces/Novokshanov/Shughni-Translator).

1 Introduction

Machine translation (MT) has advanced rapidly with the emergence of neural MT (NMT) (Bahdanau et al., 2014), particularly transformer-based architectures (Vaswani et al., 2017). Yet these breakthroughs mostly benefit high-resource language pairs, while low-resource languages remain severely underserved (Koehn and Knowles, 2017).

To mitigate data scarcity, low-resource MT typically leverages transfer learning (Zoph et al., 2016), multilingual models (Johnson et al., 2017), and data augmentation such as back-translation (Sennrich et al., 2015) or pivoting (Cheng et al., 2016; Cheng, 2019; Kim et al., 2019). Multilingual systems show that parameter sharing across related languages can substantially improve translation quality, raising the question of how to select the most beneficial auxiliary languages. A major step in multilingual NMT has been the No Language Left Behind (NLLB) project, which released a single model for 200 languages, including Iranian languages such as Pashto, Tajik, Dari, and Persian (Team et al., 2022), and provided a strong foundation for fine-tuning.

Building on these insights, we target the Shughni–Russian translation pair. Based on our examination of the impact of auxiliary languages — chosen by typological similarity or "super-donor" status — and of isolated factors such as token overlap, we present the first Shughni–Russian translation model, obtaining scores of 39.3 METEOR, 14.6 BLEU, and 36.3 chrF++.

Shughni, spoken by approximately 90,000 people in Tajikistan and Afghanistan (Kalandarov, 2018; Edel'man and Jusufbekov, 2000; Wendtland, 2009), is an Eastern Iranian language with extremely limited written resources. Its lack of standardized orthography and the small number of available texts make it a quintessential low-resource case.

Current large language models perform poorly: in tests with recent proprietary systems[1] via the LLM Arena platform[2], only trivial sentences (e.g., *Karamsho read an interesting book*) are translated from Shughni to Russian with minor errors, while translation into Shughni fails entirely. Even the latest ChatGPT 5.2 translated a simple sentence into Tajik instead of Shughni, as shown in Figure 1. We provide a comparison of our system's metrics with those of different open-source LLMs on our test set in the Results section.

2 Data

2.1 Training and validation

Our parallel Shughni–Russian data comes from two sources: a Shughni corpus (Makarov et al., 2022) and a Shughni–Russian dictionary (Makarov et al., 2022). The language corpus contains oral and writ-

[1]Model names: *claude-sonnet-4-20250514*, *o4-mini-2025-04-16*, *YandexGPT 5 Pro*, *GigaChat 2 Max*.

[2]LLM Arena v0.4.2, https://llmarena.ru/.

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 114–120

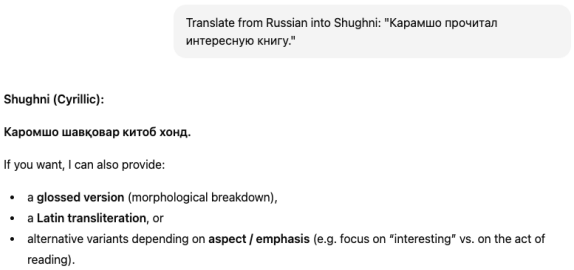


Figure 1: Example of translation into Shughni by ChatGPT 5.2. The correct translation is given in Figure 2.

ten texts of different genres, including fairy tales, stories, prose fiction, poetry, and riddles. Some texts are only annotated via a morphological parser and do not include translations. For our purposes, we take oral texts and fairy tales translated into Russian by native speakers, as well as Gospel fragments aligned with the Russian version (4,256 sentences in total). The Shughni–Russian dictionary contains ordinary entries (21,871) and example sentences (43,300). For dictionary entries, only the first translation equivalent was retained. After preprocessing, we obtained 69,227 training pairs.

Preprocessing included removal of annotations, special symbols, and stress marks; lowercasing and whitespace normalization; and orthographic unification. Since Shughni lacks a standardized script, we normalized all data into one of the accepted Cyrillic scripts using the converter by (Makarov et al., 2022) in order to maximize token overlap with Russian and Tajik and to ensure consistency across resources. In our demo, we also added support for the Latin script for both input and output.

Auxiliary language data were added from OPUS corpora (Tiedemann, 2012), prioritizing similar domains and excluding sources used in NLLB pretraining. For each auxiliary language, we sampled datasets of equal size to the Shughni corpus (69k parallel sentences).

2.2 Test sets

In the present study, two test sets are used. The first was obtained from the BivalTyp project (Say, 2020), a typological database of bivalent predicates and their encoding frames, which contains 123 Shughni-English sentence pairs (Chistiakova and Ryzhova, 2023). The sentences are quite short but varied. All sentences from this set include two arguments and a predicate connecting them; at the same time, the predicate meanings are selected variously in order to illustrate different argument encoding strategies. Thus, we also check how well the argument structure of the source sentence is preserved in translation. The sentences were translated into Russian manually from English. Figure 2 presents an example of a parallel sentence from this set.

Russian	**Shughni (Cyrillic)**
Карамшо прочитал интересную книгу.	Карамшойи аҷойӣб китоб ҳе̂йд.

Figure 2: Example parallel sentence from the BivalTyp test set. English translation: *Karamsho read an interesting book.*

For the second test set, we manually selected 110 sentences from NLLB-MD (the NLLB Multi-Domain Dataset) (Team et al., 2022) with diverse syntactic structures. After minimal adaptation to Pamiri realities, we asked several native speakers from Khorog, Gorno-Badakhshan Autonomous Region of the Republic of Tajikistan, to translate them into Shughni. The Shughni portions of both test sets were transliterated into the same uniform Cyrillic orthography used in the training dataset.

3 Methodology and Experiments

We framed our experiments around fine-tuning multilingual NMT models with auxiliary languages. The design was as follows: combine Shughni–Russian data with equal-sized corpora from an auxiliary language and train a shared model in both directions.

All models were trained in the same virtual environment and with the same training parameters that were empirically found to be optimal: AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-3, weight decay of 1e-3, and a total batch size of 1024 examples. Full training configurations are available upon request.

3.1 Baseline

As a starting point, we compared the performance of three models traditionally used in NMT: NLLB-200 (600M), mT0-small, and ByT5-small, and Gemma 3 (4B) as a solid multilingual baseline LLM. Even though these models are multilingual, we believe they have not seen any Shughni data. For fair comparison, missing Shughni tokens were added to the NLLB-200 and mT0 tokenizers to ensure all of our limited data receive desired representations. ByT5 is a token-free model; it operates directly on raw texts and does not require

tokenization. On Shughni→Russian translation, NLLB-200-distilled clearly outperformed the alternatives (see Table 1), and we selected the larger 1.3B distilled version as our baseline due to its greater stability.

Model	Meteor	chrF++
nllb-200	**0.384**	**33.6**
mt0-small	0.117	12.5
byt5-small	0.246	26.6
Gemma 3 (4B)	0.278	25.2

Table 1: Base model evaluation metrics on the Shughni → Russian validation set. Best in bold.

3.2 Auxiliary languages

Protasov et al. (Protasov et al., 2024) investigated whether morphological similarity enhances cross-lingual transfer in multilingual masked language modeling. They adopted features from The World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) as predictors of positive knowledge transfer in mT5 pretraining. Authors identified the most important typological features and overall best donor languages.

To explore transfer learning in machine translation, we experimented with several auxiliary languages. Two were chosen for typological proximity (Pashto, Somali) and two for typological distance (Vietnamese, Zulu). Table A.1 contains WALS (Dryer and Haspelmath, 2013) features of Shughni and donor languages. This illustrates that the most important features for transfer learning in language modeling (discovered in (Protasov et al., 2024)) are similar between Shughni, Pashto, and Somali. For Shughni, Vietnamese, and Zulu, on the other hand, these features are mostly opposite. We also considered two "super donors" reported to benefit many languages (Afrikaans, Slovenian), as discovered by (Protasov et al., 2024).

3.3 Final model

For our final system, we use the auxiliary language that performed best in both translation directions and enhance the training data with backtranslation (Sennrich et al., 2015) of a Shughni novel (8,138 unique sentences) and additional Shughni–Tajik Gospel fragments (195 sentences). This configuration yielded the strongest overall results, establishing a usable Shughni–Russian MT system.

3.4 Evaluation

To assess the impact of our transfer learning and backtranslation, we report two widely used automatic metrics implemented in the *evaluate*[3] library: **METEOR** and **chrF++**. We report METEOR scores multiplied by 100 for consistency. We compare metrics on our two test sets combined.

Additionally, to verify that automatic gains reflect genuine quality, we conducted a human evaluation using the **XSTS** protocol (AI et al., 2022) on 50 sentences randomly chosen from the combined test set. Native speakers from Khorog (3 university professors and 3 students with high Russian language proficiency) rated the final system against the baseline in both directions.

4 Results

4.1 Main experiment

We evaluated models both separately and jointly on the BivalTyp and NLLB-MD test sets. Table 2 summarizes key results: the NLLB-200 baseline, systems enhanced by auxiliary languages one at a time (with Pashto yielding the best mean performance), and our final system in two model sizes. The latter is enhanced by Pashto as an auxiliary language, as well as backtranslation and a small amount of Shughni–Tajik data. Pashto brought moderate gains over the baseline in both directions, while Slovenian was useful only when translating into Russian. The final model reached a chrF++ score of 45.7 for translation into Russian on the BivalTyp test set (36.3 on the combined test set), establishing a substantial advance over all alternatives.

The 600-million-parameter model excels at translating into Shughni, where the 1.3B model is better into Russian. We suggest that given fixed training dataset of a very limited size, 600M model has a better parameter-to-data ratio for the decoder to learn a new language. See (Caillaut et al., 2024) for more on translation-focused decoder scaling laws. Encoder, on the other hand, is less dependent on size scaling (Ghorbani et al., 2021). Table 3 compares our systems with the most recent open-source LLMs and shows their superiority. Moreover, NLLB models are much more compact than even the lightest LLMs, which allows them to be run on personal devices. This is especially important in hard-to-reach areas.

[3] `evaluate` v0.4.2, https://github.com/huggingface/evaluate.

	Meteor		chrF++	
Source	**sh**	**ru**	**sh**	**ru**
Target	**ru**	**sh**	**ru**	**sh**
Baseline	35.6	23.9	33.2	23.7
Afrikaans (aux)	37.5	24.0	34.4	22.6
Vietnamese (aux)	36.3	24.4	34.1	23.6
Zulu (aux)	36.1	25.2	33.6	24.3
Somali (aux)	35.6	24.2	33.6	23.5
Pashto (aux)	*38.3*	*24.1*	*34.6*	*23.9*
Slovanian (aux)	38.9	23.0	34.9	22.8
Final 1.3B	**39.3**	22.7	**36.3**	23.4
Final 600M	35.6	**25.8**	34.0	**26.2**

Table 2: Evaluation on the combined test set. Best results in bold. Best performance among auxiliary languages is underlined. Second-best performance among auxiliary languages is italicized.

	Meteor		chrF++	
Source	**sh**	**ru**	**sh**	**ru**
Target	**ru**	**sh**	**ru**	**sh**
Baseline	35.6	23.9	33.2	23.7
Qwen 3 (235B)	15.8	10.4	17.5	14.4
Gemma 3 (27B)	16.1	10.3	18.1	13.3
GPT-OSS (120B)	10.1	6.5	3.9	3.1
Final 1.3B	**39.3**	22.7	**36.3**	23.4
Final 600M	35.6	**25.8**	34.0	**26.2**

Table 3: Comparison of open-source LLMs and our systems. Evaluation on combined test set.

Human evaluation was conducted comparing the final system (1.3B) and the NLLB-200 baseline (also 1.3B). The final system scored higher than the baseline in both directions, confirming the practical usability of our Shughni–Russian MT system, as shown in Table 4. For inter-annotator agreement, we report weighted Cohen's kappa (Cohen, 1968). Following common interpretive conventions (Landis and Koch, 1977), our results indicate substantial agreement under quadratic weighting (κ = 0.65). The most common errors fall into two categories: lexical errors and grammatical time expression errors.

5 Demo description

We additionaly present a demo translation app via HuggingFace Spaces. The model is optimized for sentence-level translation and the Cyrillic Shughni script. Therefore, the demo application automatically performs segmentation and orthography conversion. Moreover, for convenience and wider use, our demo includes translation into English. Translation in this direction is implemented through Russian as a pivot language due to the almost total absence of Shughni–English data and reaches a chrF++ score of 42.2 on the BivalTyp data.

	Mean XSTS	
Source	**sh**	**ru**
Target	**ru**	**sh**
Combined test set		
Baseline	2.85	3.15
Final (1.3B)	**3.21**	**3.61**

Table 4: Human evaluation on Shughni → Russian and Russian → Shughni translation (Mean XSTS scores).

6 Conclusion

We present the first neural MT system for the Shughni language, an endangered Eastern Iranian language with virtually no prior machine translation support. Our experiments show that auxiliary languages can provide meaningful gains, though not always in line with standalone factors: Pashto proved most helpful not only because it is typologically similar but also as a close relative with potentially shared lexical units, despite differing writing systems. By combining Pashto auxiliary data with backtranslation, we achieve the best overall results and provide a translation model in two sizes. Human evaluation confirms that our improvements translate into practical usability.

7 Future work

Future research will examine auxiliary language choice more systematically, testing a broader set of candidates and features. Moreover, we will expand the number of language models, including the latest translation-oriented LLMs. We also plan to utilize Shughni–Russian dictionary entries more effectively and to benefit from annotations in the corpus (e.g., morphemes, glosses, part of speech, and general meaning). We believe these two sources can be a valuable source of syntactic data and can be integrated into the translation system itself. Furthermore, while Shughni is the local lingua franca, we aim to extend MT development to other Pamiri languages as part of ongoing efforts in their doc-

umentation and digital support. In the meantime, we encourage researchers and native speakers of Pamiri languages to collaborate on the expansion of the corpus and the machine translation project.

Acknowledgments

The authors thank everyone who participated in digitizing the Shughni–Russian dictionary and collecting the Shughni corpus. We would like to express our special appreciation to the native speakers in Khorog who translated the test dataset and participated in the human evaluation.

References

Daniel Licht META AI, Cynthia Gao META AI, Janice Lam META AI, Francisco Guzmán META AI, Mona Diab, and Philipp Koehn. 2022. Consistent human evaluation of machine translation across language pairs. *Volume 1: MT Research Track*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Gaëtan Caillaut, Raheel Qader, Mariam Nakhlé, Jingshu Liu, and Jean-Gabriel Barthélemy. 2024. Scaling laws of decoder-only models on the multilingual machine translation task. *Preprint*, arXiv:2409.15051.

Yong Cheng. 2019. Joint training for pivot-based neural machine translation. In *Joint training for neural machine translation*, pages 41–54. Springer.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016. Neural machine translation with pivot languages. *arXiv preprint arXiv:1611.04928*.

Daria Chistiakova and Daria Ryzhova. 2023. Bivalent patterns in Shughni. *BivalTyp: Typological database of bivalent verbs and their encoding frames*, (5):(Available online at https://bivaltyp.info, Accessed on 26 May 2025.).

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Edel'man and Jusufbekov. 2000. Shugnanskij jazyk [Shughni Language]. In *Jazyki mira: Iranskie jazyki. III. Vostochnoiranskie jazyki*, page 225.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. 2021. Scaling laws for neural machine translation. *Preprint*, arXiv:2109.07740.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, and 1 others. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Tokhir Kalandarov. 2018. Pamirskie narody, ih jazyki i perepis': etnicheskij diskurs [The Peoples of Pamir, Their Languages and the Census: The Ethnic Discourse]. *Etnograficheskoe obozrenie*, (5):162–178.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-english languages. *arXiv preprint arXiv:1909.09524*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Yury Makarov, Maksim Melenchenko, and Dmitry Novokshanov. 2022. Digital resources for the Shughni language. In *Proceedings of the workshop on resources and technologies for Indigenous, endangered and lesser-resourced languages in Eurasia within the 13th language resources and evaluation conference*, pages 61–64.

Vitaly Protasov, Elisei Stakovskii, Ekaterina Voloshina, Tatiana Shavrina, and Alexander Panchenko. 2024. Super donors and super recipients: Studying cross-lingual transfer between high-resource and low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 94–108.

Sergey Sergeevič Say. 2020. *BivalTyp: Typological database of bivalent verbs and their encoding frames. (Available online at https://bivaltyp.info, Accessed on 1 January 2026.)*. Institute for Linguistic Studies Russian Academy of Sciences.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and

20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Antje Wendtland. 2009. The position of the pamir languages within east iranian. *Orientalia Suecana*, 58:172–188.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

A WALS Features

Table A.1 contains selected WALS (Dryer and Haspelmath, 2013) features of Shughni and donor languages. The selection was based on (Protasov et al., 2024) and aimed not to show typological distance or similarity, but to illustrate our choice of donor languages in the context of machine translation and transfer learning only. This illustrates that the most important to us features between Shughni, Pashto, and Somali are nearly the same. This is the evidence of typological similarity between Shughni and Somali in one narrow fragment of the system, not in typological profiles of the entire languages.

WALS feature	Shughni	Pashto	Somali	Vietnamese	Zulu
Prefixing vs. suffixing in inflectional morphology	Strongly suffixing	Strongly suffixing	Strongly suffixing	Little affixation	Strong prefixing
Order of Object and Verb & Order of Adposition and NP	Other	OV & postpositions	Other	VO & prepositions	VO & prepositions
Order of Object and Verb	OV	OV	OV	VO	VO
Preverbal negative morphemes	[Neg–V]	[Neg–V]	Neg V	Neg V	[Neg–V]
Order of demonstrative and noun	Demonstrative–Noun	Demonstrative–Noun	Demonstrative suffix	Noun–Demonstrative	Mixed
Order of numeral and noun	Numeral–Noun	Numeral–Noun	Numeral–Noun	Numeral–Noun	—
Coding of nominal plurality	Plural suffix	Plural suffix	Plural suffix	Plural word	Plural prefix

Table A.1: Comparison of selected WALS features across five languages.

Segmentation Strategy Matters: Benchmarking Whisper on Persian YouTube Content

Reihaneh Iranmanesh[1]*, **Rojin Ziaei**[1]*, **Joe Garman**[1]
[1]Georgetown University
{ri164, nz204, jhg28}@georgetown.edu

Abstract

Automatic Speech Recognition (ASR) transcription accuracy remains highly sensitive to audio segmentation strategies, yet most benchmarks assume oracle timestamps unavailable in deployment. We systematically evaluate how audio segmentation affects Whisper's performance on 10 hours of Persian YouTube content, comparing transcript-aligned (oracle) versus silence-based (realistic) approaches across contrasting acoustic conditions. Results reveal striking content-type dependency: podcast content benefits from timestamp segmentation (33% lower mean WER), while entertainment content favors silence-based segmentation (8% lower mean WER). This finding demonstrates that optimal segmentation must be content-aware, with silence detection better capturing natural boundaries in acoustically heterogeneous media while avoiding mid-utterance splits. We publicly release our evaluation framework, 10 hours of audio with gold transcripts, and segmentation results here: https://github.com/ri164-bolleit/persian-youtube-whisper-benchmark

1 Introduction

Automatic Speech Recognition (ASR) systems have achieved remarkable progress through neural architectures, large-scale training data, and robust pretraining objectives [Prabhavalkar et al., 2023, Malik et al., 2021, Kheddar et al., 2024, Radford et al., 2023]. Despite these advances, transcription accuracy in real-world settings remains highly sensitive to upstream design choices, particularly how continuous audio streams are segmented prior to inference [Kuhn et al., 2024, Arriaga et al., 2024]. Segmentation affects not only computational efficiency but also linguistic context, temporal alignment, and the stability of downstream evaluation metrics.

Many existing benchmarks rely on oracle transcript timestamps to define segment boundaries, an assumption that rarely holds outside controlled evaluation settings [Faria et al., 2022, Sklyar et al., 2022, von Neumann et al., 2023]. This creates a gap between benchmark performance and real-world deployment accuracy, where segmentation must be inferred directly from acoustic signals.

1.1 Research Questions

We address three key questions:
(RQ1) How does segmentation strategy–oracle timestamps versus acoustic silence detection–impact ASR accuracy?
(RQ2) Does the optimal strategy vary by content type (structured podcasts versus noisy entertainment)?
(RQ3) How does model scale affect robustness to segmentation choices?

1.2 Contributions

Using OpenAI Whisper at three scales (small: 0.2B, medium: 0.8B, large: 1.5B parameters), we evaluate two segmentation pipelines on a 10-hour Persian YouTube dataset spanning contrasting acoustic conditions. Our contributions include: (1) first systematic evaluation of segmentation strategies for Persian ASR, (2) evidence that optimal approaches are fundamentally content-type dependent, and (3) a publicly released benchmark addressing the critical gap in Persian real-world ASR evaluation.

2 Related Work

2.1 ASR Benchmarking and Evaluation

The development of robust ASR benchmarks has emerged as a critical research area, with increasing recognition of their limitations in capturing real-world speech variability [Aksënova et al., 2021, Szymański et al., 2020]. Traditional benchmarks such as LibriSpeech [Panayotov et al., 2015] pri-

* Equal contribution.

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 121–130

marily focus on clean, well-structured read speech, rendering them inadequate for assessing ASR performance in spontaneous or conversational settings. Studies by [Del Rio et al., 2021] and [Cao et al., 2023] underscore the need for benchmarks that evaluate ASR systems under challenging conditions, including spontaneous speech, noisy environments, and multi-speaker interactions.

To address domain mismatch and improve cross-domain generalization, the End-to-End Speech Benchmark (ESB) [Gandhi et al., 2022] was introduced to evaluate ASR models across diverse domains without prior knowledge of data distributions. However, while ESB accounts for varied acoustic environments, it overlooks critical demographic factors such as speaker age, gender, and regional accents that significantly impact ASR performance [Aksënova et al., 2021].

2.2 Segmentation Strategies in ASR

While considerable attention has focused on model architecture and training data, the impact of audio segmentation strategies on ASR performance remains understudied. Most benchmarks assume oracle timestamps that align with transcript boundaries, an assumption that rarely holds in deployment [Faria et al., 2022, Sklyar et al., 2022, von Neumann et al., 2023]. Previous research [Kuhn et al., 2024, Arriaga et al., 2024] notes that standard benchmarks often fail to account for WER variability across segmentation approaches, creating a gap between benchmark performance and real-world accuracy.

This gap is particularly problematic as segmentation affects not only computational efficiency but also linguistic context preservation, temporal alignment quality, and the stability of evaluation metrics. In real-world deployments, segmentation must be inferred directly from acoustic signals using methods such as voice activity detection or silence-based splitting, yet most published results rely on oracle segmentation that provides an optimistic upper bound on achievable performance.

2.3 Persian ASR Resources and Challenges

For low-resource languages like Persian, comprehensive benchmarks remain scarce despite the language being spoken by over 100 million people. Existing Persian ASR datasets provide partial foundations but have significant limitations. The DeepMine dataset [Zeinali et al., 2019] offers a robust foundation with over 1,850 speakers and approximately 480 hours of audio, but its six-hour test set focuses primarily on formal speech, limiting applicability to informal or spontaneous contexts. Similarly, the Persian subset of Mozilla Common Voice [Ardila et al., 2019] suffers from demographic imbalances and data quality concerns due to its crowdsourcing approach. The FLEURS dataset [Conneau et al., 2023], while useful for few-shot learning, focuses on short, controlled utterances insufficient for evaluating conversational speech. [Sedghiyeh et al., 2025] introduce PSRB, a comprehensive benchmark for evaluating Persian ASR systems across diverse linguistic conditions including regional accents, speaker demographics, and acoustic environments. However, it does not explore segmentation strategies for handling long-form audio or optimal chunking approaches for Persian speech recognition.

Persian poses unique ASR challenges due to its linguistic diversity-spanning regional accents like Baluchi, Kurdish, and Dari-and features like variable word boundaries and the use of Zero Width Non-Joiner (ZWNJ) characters [Ghayoomi and Momtazi, 2009, Bijankhan et al., 2011]. These characteristics, combined with limited training data, exacerbate performance disparities in Persian ASR systems. Most existing Persian ASR research focuses on controlled conditions, with YouTube content representing a particularly challenging and underexplored domain combining spontaneous speech, variable acoustic quality, background music, and diverse speaking styles.

Our work addresses critical gaps in the literature by providing the first systematic evaluation of segmentation strategies for Persian ASR on real-world YouTube content, using two different segmentation approaches and various model scales.

3 Dataset

We curated a 10-hour audio–text dataset sourced from YouTube, designed to capture contrasting acoustic and conversational conditions. The dataset consists of two subsets, each totaling approximately five hours.

3.1 Roud Podcast Subset

Roud Podcast[1]: A conversational podcast featuring a single speaker in a structured format centered on book discussions. This content contains no background music, no speaker overlap, and min-

[1] https://www.youtube.com/@MojtabaShakoori

imal variability in acoustic conditions, providing a relatively clean and controlled speech environment.

Corpus Statistics: As shown in Table 1, this data includes six episodes totaling 5.1 hours containing 2,747 transcript-aligned segments. Segment lengths show low variance (mean: 16.9 words, SD: 5.5, range: 1–36), reflecting structured discourse with natural prosodic boundaries.

YouTube timestamps are generated by the content creator, naturally aligning with clause endings and prosodic boundaries. This production-time alignment creates favorable conditions for timestamp-based segmentation.

3.2 Kouman Entertainment Subset

Kouman Entertainment[2]: An entertainment-focused YouTube channel characterized by multiple hosts, frequent speaker changes, background music, and unpredictable audio content. This subset represents a substantially noisier and more heterogeneous speech setting.

Corpus Statistics: As shown in Table 2, this data includes ten episodes totaling 4.9 hours and containing 6,450 transcript segments. Segment lengths exhibit high variance (mean: 4.4 words, SD: 6.1, median: 1 word), reflecting fragmented timestamp-based splitting that produces single-word or phrase-level segments.

Episode	Segments	Words	Duration (min)
Episode 1	325	6,133	∼41
Episode 2	333	3,483	∼23
Episode 3	490	8,481	∼57
Episode 4	533	9,238	∼62
Episode 6	475	8,056	∼54
Episode 8	591	10,911	∼73
Total	**2,747**	**46,302**	**∼310**

Table 1: Roud Podcast corpus statistics. Type-Token Ratio: 0.142. Segment length: median=16, IQR=14–20 words. Timestamps created during production align with natural discourse boundaries.

Unlike Roud where timestamps align with discourse structure, Kouman timestamps are editorial markers added post-production for viewer navigation. These frequently split continuous speech during music-overlaid dialogue or multi-speaker exchanges, creating artificial mid-utterance boundaries. The mean 5.5:1 timestamp-to-silence ratio quantifies this fragmentation: timestamp segmentation produces more segments than acoustic boundaries warrant.

Episode	Segments	Words	Avg W/Seg	TS:Sil.
amazon	799	2,057	2.6	6.5:1
eshgh	655	3,624	5.5	6.9:1
hoosh	338	1,411	4.2	1.8:1
madrese	751	2,331	3.1	6.7:1
mia	693	1,391	2.0	7.2:1
nooshabe	901	4,827	5.4	6.6:1
norooz	718	3,191	4.4	7.7:1
safar	325	2,857	8.8	2.9:1
soal	827	2,514	3.0	6.9:1
youtuber	443	4,279	9.7	2.0:1
Total/Avg	**6,450**	**28,482**	**4.4**	**5.5:1**

Table 2: Kouman Entertainment corpus statistics. Ten episodes, approximately 4.9 hours. Type-Token Ratio: 0.276. TS:Silence Ratio indicates timestamp segmentation produces 5.5× more segments than silence-based segmentation (mean), revealing severe over-fragmentation. Segment length: min=0, max=64, median=1 word.

3.3 Data Preparation

For both subsets, the original YouTube videos were converted into WAV audio files (16 kHz, mono). Corresponding time-stamped transcripts were retrieved directly from YouTube and stored as CSV files aligned with each audio file. All transcripts were manually created by channel managers and content creators as YouTube lacks auto-transcription for Persian. All spoken content is in standard Persian.

4 Methodology

We evaluate automatic speech recognition performance using two distinct experimental pipelines that differ in how audio is segmented and aligned with reference human-generated transcripts. Our pipeline employs three OpenAI Whisper models [Radford et al., 2023] (small, medium, and large) and is evaluated using Word Error Rate (WER) and Character Error Rate (CER).

4.1 Transcript-Aligned Timestamp Segmentation

In the first setup, audio segmentation is directly derived from YouTube's time-stamped reference transcripts. Each transcript CSV consists of alternating timestamp and text lines, where timestamps indicate the start time of each spoken segment. Segment end times are inferred from the subsequent

[2] https://www.youtube.com/c/Kouman

timestamp, with the final segment extending to the end of the audio file.

Using these timestamps, the corresponding WAV audio files are segmented exactly to match the transcript boundaries. Each segment is then transcribed independently using Whisper. Since the audio segments and reference transcripts are perfectly aligned in time by construction, evaluation is performed via a one-to-one comparison between each reference segment and its corresponding Whisper transcription. This pipeline provides an optimistic estimate of transcription performance under ideal alignment conditions.

4.2 Silence-Based Segmentation

The second setup removes reliance on transcript-based segmentation and instead segments audio using acoustic silence detection. Audio is split into non-silent regions using energy-based thresholds, followed by post-processing to enforce segment duration constraints between 5 and 30 seconds. Short segments are merged, long segments are split, and brief silences are optionally retained at segment boundaries to preserve natural speech context.

Because these segments do not necessarily align with the reference transcript timestamps, evaluation requires a fuzzy time-alignment procedure. For each predicted audio segment, a reference text is constructed by concatenating all transcript entries whose timestamps overlap with the segment's time window. This pipeline more closely reflects real-world transcription scenarios.

5 Results

Here, we report results for the Roud podcast, a single-speaker, low-noise conversational dataset, and Kouman, a noisy, multi-speaker entertainment channel, using two segmentation strategies.

5.1 Segment Length and Granularity

The two segmentation strategies produce substantially different segment distributions for both datasets. Transcript-based segmentation yields a larger number of shorter segments, tightly aligned to sentence- or phrase-level transcript boundaries. In contrast, silence-based segmentation produces fewer but longer segments by merging contiguous speech between silence intervals and enforcing minimum and maximum duration constraints (5–30 seconds).

Timestamp-based segmentation produces 2.5–8× more segments per episode (319–901 segments) compared to silence-based segmentation (93–216 segments). This fragmentation creates shorter, more numerous boundaries that may interrupt mid-utterance in content with music, sound effects, and overlapping speech.

Across all evaluated episodes for both Roud and Kouman, silence-based segmentation consistently results in a lower total number of segments per episode, indicating longer average segment durations. This difference in granularity has direct implications for both transcription quality and evaluation stability, as longer segments preserve more linguistic context while increasing acoustic variability.

5.2 Roud Podcast: Clean Speech Results

Table 3 presents comprehensive results for the Roud podcast across six episodes and three model sizes. Across all episodes, performance improves monotonically with model size. Whisper large consistently achieves the lowest WER and CER, followed by medium and small.

We assess the statistical significance of performance differences using both parametric (independent two-sample t-tests) and non-parametric (Mann–Whitney U) tests on per-segment WER and CER distributions. Across evaluated episodes and models, differences between silence-based and timestamp-based segmentation are statistically significant at $\alpha = 0.05$ for both WER and CER.

Effect size analysis using Cohen's d indicates small-to-moderate effects, with stronger effects observed for WER than CER. In all cases, silence-based segmentation exhibits statistically significant degradation relative to transcript-based timestamp segmentation.

Paired Episode-Level Analysis:

Figures 1a and 1b show paired comparisons for the Whisper *large* model, while Figures 1c and 1d report the same analysis for the *medium* model.

For the large model, the mean paired improvement from silence-based to timestamp-based segmentation is 0.161 in WER (33% relative improvement) and 0.227 in CER (67% relative improvement), with paired t-tests yielding $p < 0.001$ and Wilcoxon signed-rank tests yielding $p = 0.031$. Similar trends are observed for the medium model, with mean paired improvements of 0.084 in WER and 0.167 in CER. The performance gap between segmentation strategies widens as model size increases, suggesting that larger models are more capable of exploiting high-quality segmentation

Episode	Model	Timestamp Seg.			Silence Seg.		
		WER	CER	Segs	WER	CER	Segs
1	large	**0.349**	0.114	325	0.505	0.339	139
1	medium	0.499	0.201	325	0.582	0.364	139
1	small	0.653	0.259	325	0.700	0.410	139
2	large	**0.337**	0.110	333	0.409	0.225	94
2	medium	0.480	0.187	333	0.494	0.258	94
2	small	0.656	0.282	333	0.623	0.310	94
3	large	**0.358**	0.122	490	0.527	0.373	266
3	medium	0.555	0.232	490	0.629	0.412	266
3	small	0.703	0.305	490	0.741	0.464	266
4	large	**0.331**	0.124	533	0.504	0.371	278
4	medium	0.500	0.227	533	0.585	0.400	278
4	small	0.653	0.285	533	0.697	0.440	278
6	large	**0.268**	0.098	475	0.484	0.372	263
6	medium	0.417	0.178	475	0.553	0.394	263
6	small	0.586	0.238	475	0.667	0.435	263
8	large	**0.307**	0.098	591	0.490	0.349	289
8	medium	0.476	0.184	591	0.590	0.383	289
8	small	0.640	0.254	591	0.700	0.427	289
Average (large)		**0.325**	0.111	458	0.487	0.338	222
Average (medium)		0.488	0.202	458	0.572	0.369	222
Average (small)		0.649	0.271	458	0.688	0.431	222

Table 3: Detailed results for Roud podcast across all episodes and models. Timestamp segmentation consistently outperforms silence-based segmentation across all three models. Whisper large achieves the lowest WER (and CER) across all episodes using timestamp segmentation.

while being penalized more by noisy segmentation.

5.3 Kouman Entertainment: Noisy Speech Results

Table 4 presents comprehensive results for the Kouman entertainment channel across ten episodes. Strikingly, the performance pattern reverses from the podcast results: silence-based segmentation now *outperforms* timestamp-based segmentation in 7 out of 10 episodes for both model sizes. (Results for Whisper Small are excluded as the model achieved WER values above 0.90.)

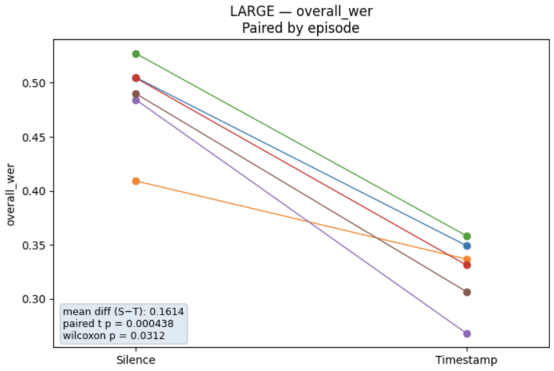
For the large model, WER differences are statistically significant (paired t-test $p = 0.020$, Wilcoxon $p = 0.027$), with silence-based segmentation achieving a mean improvement of 0.0604 (7.9% relative improvement) over timestamp-based segmentation. CER differences are not statistically significant (paired t-test $p = 0.476$, Wilcoxon $p = 1.000$), indicating that character-level accuracy is relatively stable across segmentation methods for larger models.

For the medium model, WER differences are highly significant (paired t-test $p = 0.007$, Wilcoxon $p = 0.027$), with silence-based segmentation showing a mean improvement of 0.0488 (5.7% relative improvement). CER differences approach significance (paired t-test $p = 0.114$, Wilcoxon $p = 0.037$), with a mean improvement of 0.0439.

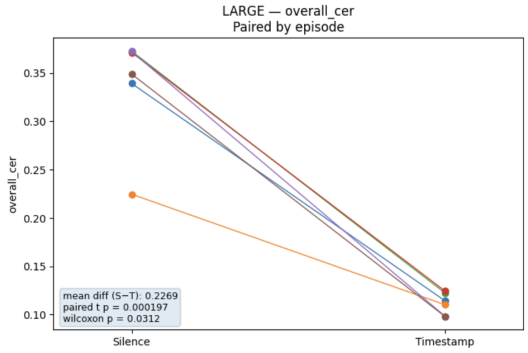
Paired Episode-Level Analysis:

Figures 2a and 2b show paired comparisons for the Whisper *large* model, while Figures 3a and 3b report the same analysis for the *medium* model. Each line corresponds to a single episode, connecting silence-based results to timestamp-based results.

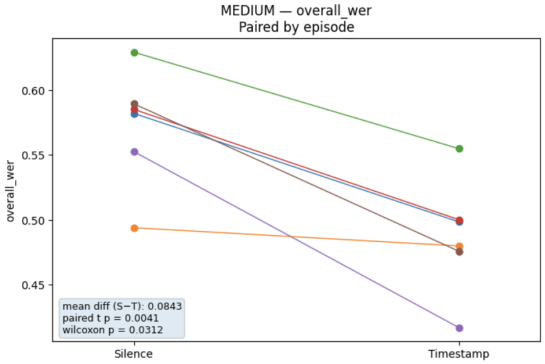
Across most (seven out of ten) episodes and both model sizes, silence-based segmentation yields lower or comparable WER and CER compared to timestamp-based segmentation. This trend contrasts with expectations that transcript-aligned segmentation would uniformly outperform silence-based approaches. Paired statistical tests confirm that for WER, silence-based segmentation significantly outperforms timestamp-based segmentation. For the *large* model, the mean paired difference (Silence minus Timestamp) is -0.0604 in WER and -0.0191 in CER, with WER differences reaching statistical significance at $\alpha = 0.05$. Similar trends are observed for the *medium* model, with mean paired differences of -0.0488 in WER and



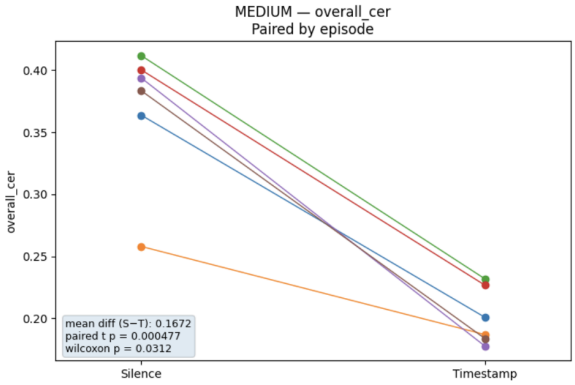
(a) Paired episode-level comparison of overall WER for the Whisper *large* model. Timestamp-based segmentation consistently yields lower WER across episodes.



(b) Paired episode-level comparison of overall CER for the Whisper *large* model. Timestamp-based segmentation achieves lower CER for every episode.



(c) Paired episode-level comparison of overall WER for the Whisper *medium* model. Improvements from timestamp-based segmentation remain consistent across episodes, but are smaller than for the *large* model, indicating increased sensitivity to segmentation at lower model capacity.



(d) Paired episode-level comparison of overall CER for the Whisper *medium* model. Timestamp-based segmentation consistently reduces CER across episodes, with statistically significant paired differences.

Figure 1: Episode-level paired comparison of WER (left) and CER (right) for the Whisper *large* (top) and *medium* (bottom) models. Each line corresponds to a single episode, connecting silence-based segmentation to YouTube timestamp–based segmentation. Reported statistics include the mean paired difference (Silence – Timestamp), paired t-test p-value, and Wilcoxon signed-rank p-value.

-0.0439 in CER, both showing statistically significant improvements for silence-based segmentation in WER.

These results suggest that for the Kouman dataset, silence-based segmentation produces segments that are better suited to the acoustic and linguistic characteristics of the content, possibly by avoiding mid-utterance boundaries introduced by timestamp-based segmentation.

5.4 Error Distribution Analysis

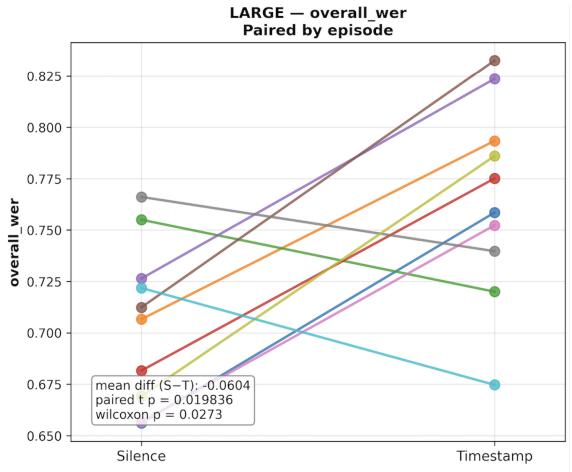
While silence-based segmentation generally outperforms timestamp-based segmentation for entertainment content, three episodes (Hoosh, Safar, YouTuber) exhibit degradation in average WER. The "YouTuber" episode shows the most degradation: for the large model, WER increased by 0.0471 (6.98%) and CER by 0.1129 (25.97%).

However, error distribution analysis reveals a critical engineering tradeoff. For the YouTuber episode, silence-based segmentation produces dramatically tighter distributions with lower variance compared to timestamp-based segmentation.

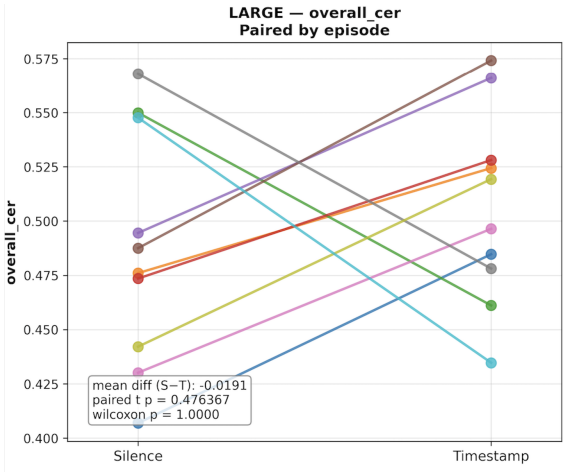
Variance-Bias Tradeoff: This pattern represents a fundamental tradeoff-silence-based segmentation trades marginally higher average error (6–26% degradation in these three episodes) for significantly improved consistency and predictability. The absence of extreme outliers suggests that even when silence detection produces suboptimal boundaries for content with rapid speech or heavy background audio, the resulting errors remain bounded.

Episode	Model	Timestamp Seg.			Silence Seg.		
		WER	CER	Segs	WER	CER	Segs
amazon	large	0.759	0.485	794	**0.656**	0.407	123
amazon	medium	0.848	0.563	794	0.769	0.466	123
eshgh	large	0.793	0.524	654	**0.707**	0.476	95
eshgh	medium	0.872	0.607	654	0.817	0.538	95
hoosh	large	**0.720**	0.461	337	0.755	0.550	183
hoosh	medium	0.819	0.527	337	0.833	0.594	183
madrese	large	0.775	0.528	748	**0.682**	0.474	111
madrese	medium	0.862	0.617	748	0.781	0.523	111
mia	large	0.824	0.566	692	**0.726**	0.495	96
mia	medium	0.895	0.637	692	0.817	0.537	96
nooshabe	large	0.833	0.574	901	**0.712**	0.488	137
nooshabe	medium	0.883	0.637	901	0.809	0.532	137
norooz	large	0.752	0.496	714	**0.657**	0.430	93
norooz	medium	0.854	0.588	714	0.772	0.489	93
safar	large	**0.740**	0.478	319	0.766	0.568	109
safar	medium	0.837	0.560	319	0.851	0.625	109
soal	large	0.786	0.519	821	**0.669**	0.442	119
soal	medium	0.861	0.594	821	0.777	0.509	119
youtuber	large	**0.675**	0.435	435	0.722	0.548	216
youtuber	medium	0.798	0.509	435	0.816	0.588	216
Average (large)		0.766	0.507	642	**0.705**	0.488	128
Average (medium)		0.853	0.584	642	0.804	0.540	128

Table 4: Detailed results for Kouman entertainment channel. Bold indicates lowest WER performance. Silence-based segmentation outperforms timestamp in 7/10 episodes for both models.



(a) Paired episode-level comparison of overall WER for the Whisper *large* model.



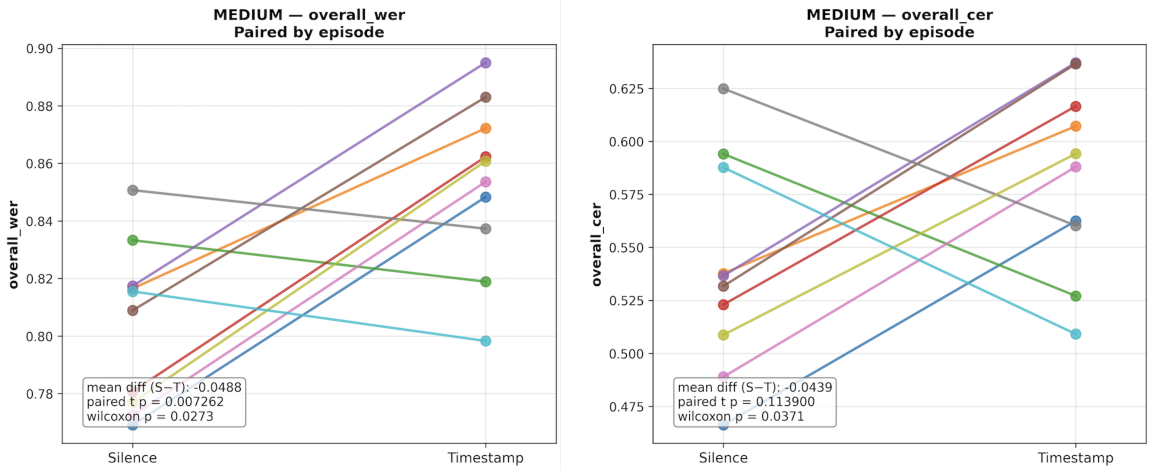
(b) Paired episode-level comparison of overall CER for the Whisper *large* model.

Figure 2: Kouman's Episode-level paired comparison of WER (left) and CER (right) for the Whisper *large* model. Each line corresponds to a single episode, connecting silence-based segmentation to YouTube timestamp–based segmentation. Reported statistics include the mean paired difference (Silence – Timestamp), paired t-test p-value, and Wilcoxon signed-rank p-value.

6 Discussion

6.1 Content-Type Dependency

Our results reveal that optimal segmentation strategy is fundamentally content-type dependent. For podcast content (Roud), timestamp-based segmentation significantly outperforms silence-based approaches across all models, with mean improvements of 33% in WER and 67% in CER.



(a) Paired episode-level comparison of overall WER for the Whisper *medium* model.

(b) Paired episode-level comparison of overall CER for the Whisper *medium* model.

Figure 3: Kouman's Episode-level paired comparison of WER (left) and CER (right) for the Whisper *medium* model. Each line corresponds to a single episode, connecting silence-based segmentation to YouTube timestamp–based segmentation. Reported statistics include the mean paired difference (Silence – Timestamp), paired t-test p-value, and Wilcoxon signed-rank p-value.

Conversely, for entertainment content (Kouman), silence-based segmentation achieves lower WER in 70% of episodes, with mean improvements of 7.9% for large models and 5.7% for medium models.

These results reflect fundamental differences in how timestamps relate to acoustic structure. Podcast timestamps, created during production, coincide naturally with speaker pauses and phrase boundaries–they capture the same acoustic events that silence detection would identify. Entertainment timestamps, added post-production, often split continuous audio arbitrarily, particularly during music, sound effects, or overlapping dialogue.

Silence detection inherently respects acoustic structure, avoiding mid-word or mid-phrase splits even when timestamps don't align with natural boundaries. This explains why longer, acoustically-grounded segments outperform shorter, arbitrarily-aligned ones in heterogeneous audio–they preserve utterance integrity and reduce boundary-induced errors.

6.2 Model Scaling Effects

Performance improves consistently as model size increases, regardless of content type or segmentation strategy. Importantly, larger models do not eliminate the content-type effect: podcasts continue to benefit from timestamp segmentation while entertainment content favors silence-based segmentation, even at the largest model size (1.5B parameters). This persistence confirms that the segmentation preferences reflect systematic differences rather than limitations that could be overcome with more model capacity.

For Roud, the absolute segmentation gap increases from small to large models, and the relative gap also widens indicating that while larger models achieve better absolute accuracy, they remain sensitive to segmentation quality. For Kouman, gaps remain proportionally consistent, suggesting that the acoustic boundary alignment effect scales with model capacity.

Larger models demonstrate superior generalization across both segmentation approaches, with lower variance in error rates across episodes. This robustness is particularly evident in Kouman's challenging acoustic conditions, where the large model shows more stable performance across diverse episodes.

While silence-based segmentation consistently underperformed for podcast content, its superiority for entertainment content appears counter-intuitive. The key difference lies in timestamp semantics: podcast timestamps are production-aligned, coinciding naturally with speaker turns and phrase boundaries, while entertainment timestamps are editorial markers added post-production for navigation (scenes, acts, highlights). These editorial timestamps frequently split continuous dialogue or complex audio scenes at arbitrary points that violate acoustic structure. Silence detection outperforms in entertainment precisely because it re-

spects natural speech boundaries, avoiding the mid-utterance splits that editorial timestamps introduce.

7 Conclusion

This work presents the first systematic evaluation of audio segmentation strategies for Persian YouTube content, revealing that optimal approaches are fundamentally content-type dependent. Our key findings: (1) podcast content favors timestamp segmentation with 33% lower WER, (2) entertainment content paradoxically benefits from silence-based segmentation with 8% lower WER, (3) this effect persists across model scales, confirming it reflects acoustic characteristics rather than capacity limitations, and (4) silence segmentation offers better worst-case guarantees through reduced error variance.

Our publicly released dataset addresses a critical gap in Persian ASR research, providing a systematic benchmark for real-world informal content. This enables researchers to build better tools and democratizes professional-quality transcription for Persian creators.

Future work should explore: (1) content-type classifiers coupled with adaptive segmentation systems, (2) hybrid timestamp-acoustic methods that validate boundaries acoustically, (3) alternative metrics beyond WER/CER, (4) parameter optimization per content type, and (5) extension to other languages and domains.

Limitations

While this study provides the first systematic evaluation of segmentation strategies for Persian ASR, several limitations warrant consideration. First, our analysis is restricted to two content types from Persian YouTube, conversational podcasts and entertainment videos, which may not generalize to other domains such as formal lectures, broadcast news, parliamentary proceedings, or telephone conversations where acoustic characteristics and speaking styles differ substantially. Second, we evaluate only OpenAI's Whisper models at three scales; comparisons with other state-of-the-art ASR systems (e.g., wav2vec 2.0, Conformer-based models, or Persian-specific architectures) would strengthen conclusions about whether content-type dependency is a general phenomenon or specific to Whisper's architecture and training. Third, our silence-based segmentation employs fixed energy thresholds and duration constraints (5–30 seconds) that were not systematically optimized; content-specific parameter tuning might yield different relative performance. Fourth, our evaluation relies solely on WER and CER metrics, which may not fully capture perceptual quality, semantic preservation, and complexity of Persian vocabulary. Finally, our dataset represents standard Persian from digital media creators; our findings may not extend to dialectal variations, code-switched content, or Persian spoken in different geographic regions or sociolinguistic contexts. Future work should address these limitations through multi-domain evaluation, cross-system comparison, and user-centered metrics.

References

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Carlos Arriaga, Alejandro Pozo, Javier Conde, and Alvaro Alonso. 2024. Evaluation of real-time transcriptions using end-to-end asr models. *arXiv preprint arXiv:2409.05674*.

Mahmood Bijankhan, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. 2011. Lessons from building a persian written corpus: Peykare. *Language Resources and Evaluation*, 45:143–164.

Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E. Margaret Perkoff, Michael Regan, Katharina Kann, James H. Martin, Martha Palmer, and Sidney D'Mello. 2023. A comparative analysis of automatic speech recognition errors in small group classroom discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 250–262.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Żelasko, and Miguel Jetté. 2021. Earnings-21: A practical benchmark for asr in the wild. *arXiv preprint arXiv:2104.11348*.

Arlo Faria, Adam Janin, Sidhi Adkoli, and Korbinian Riedhammer. 2022. Toward Zero Oracle Word Error Rate on the Switchboard Benchmark. In *Interspeech 2022*, pages 3973–3977.

Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush. 2022. Esb: A benchmark for multi-domain end-to-end speech recognition. *arXiv preprint arXiv:2210.13352*.

Masood Ghayoomi and Saeedeh Momtazi. 2009. Challenges in developing persian corpora from online resources. In *2009 International Conference on Asian Language Processing*, pages 108–113. IEEE.

Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *Information fusion*, 109.

Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. 2024. Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, 16(4):1–23.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nima Sedghiyeh, Sara Sadeghi, Reza Khodadadi, Farzin Kashani, Omid Aghdaei, Somayeh Rahimi, and Mohammad Sadegh Safari. 2025. Psrb: A comprehensive benchmark for evaluating persian asr systems.

Ilya Sklyar, Anna Piunova, and Christian Osendorfer. 2022. Separator-Transducer-Segmenter: Streaming Recognition and Segmentation of Multi-party Speech. In *Interspeech 2022*, pages 4451–4455.

Piotr Szymański, Piotr Żelasko, Mikołaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. Wer we are and wer we think we are. *arXiv preprint arXiv:2010.03432*.

Thilo von Neumann, Keisuke Kinoshita, Christoph Boeddeker, Marc Delcroix, and Reinhold Haeb-Umbach. 2023. Segment-less continuous speech separation of meetings: Training and evaluation criteria. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:576–589.

Hossein Zeinali, Lukáš Burget, and Jan Henrik Černockỳ. 2019. A multi purpose and large scale speech corpus in persian and english for speaker and speech recognition: the deepmine database. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 397–402. IEEE.

Multi-modal Neural Machine Translation for Low-Resource Classical Persian Poetry: A Culture-Aware Evaluation

Soheila Ansari, Mounir Boukadoum, Fatiha Sadat
Université du Québec à Montréal
{ansari.soheila, boukadoum.mounir, sadat.fatiha}@uqam.ca

Abstract

Persian poetry, particularly Rumi's *Masnavi-ye-Ma'navi*, is known for its complex form, mystical narrative style, rich cultural information, and linguistic nuances, and is considered a low-resource domain. Translating Persian poetry is a challenging task for neural machine translation (NMT) systems. To address this challenge, we present a novel multi-modal NMT system for Rumi's *Masnavi* in four stages. First, we built a new multi-modal parallel Persian-English corpus of 26,571 aligned verses from all six books of *Masnavi*, and all paired with aligned audio recitations. Second, a strong text-only baseline is developed by applying domain-adaptive fine-tuning to mBART-50, pre-trained on a large monolingual Persian poetry corpus, followed by training on the parallel *Masnavi* corpus (train set). Third, we extend this model to a multi-modal scenario by adding aligned audio representations using a cross-attention fusion mechanism. Fourth, we conduct a culture-aware evaluation. We propose a culture-specific item (CSI) evaluation approach by developing a CSI classification system and a Persian-English CSI dictionary alongside the standard MT metrics. Our findings demonstrate that integrating audio recitations increased the BLEU score from 9.85 to 17.95, and raised CSI-recall from 61.60% to 82.04%, suggesting greater consistency in producing culturally meaningful terms.

1 Introduction

Rumi's *Masnavi-ye-Ma'navi* has a central place in Persian cultural and literary heritage, representing a profound synthesis of theological and metaphorical discourse often referred to as the "Quran in Persian" (Sedaghat, 2020). This poetry is known for its mystical language, whose complexity often causes both human and machine translators to unintentionally impose ideological or structural biases that diminish the text's intentional indeterminacy (Sedaghat, 2020). Although neural machine translation (NMT) has made significant advancements, its application to classical poetry remains a demanding and largely unexplored area (Chakrabarty et al., 2021), (Ghassemiazghandi, 2023). Standard NMT models, typically trained on massive datasets like news and articles, frequently struggle with the low-resource nature of classical Persian and the dense figurative patterns found in the works of masters like Rumi (Ghassemiazghandi, 2023). As a result, they frequently fail to preserve the source text's distinctive "literary touch" and worldview (Ghassemiazghandi, 2023), (Chakrabarty et al., 2021). Research indicates that domain-adaptive fine-tuning on poetic corpora significantly outperform general-domain models, yet a critical gap remains: the lack of comprehensive, annotated datasets that capture the complex rhythmic and cultural rules of Persian language (Ghassemiazghandi, 2023), (Chakrabarty et al., 2021). A defining characteristic of Persian poetry is that it is intended to be heard rather than read silently. Recitation, through its unique rhythm, tone, and pauses, helps to resolve and clarify the ambiguities of the written Persian and reveal the poet's intent. In this context, multi-modal approaches that integrate audio features with text offer a game-changing ability for translation models, providing access to the aforementioned details that written words alone cannot convey (Shahrestani and Haghir Chehreghani, 2025). Furthermore, traditional evaluation metrics like BLEU often fail to track whether culture-specific items (CSIs), such as Sufi concepts, religious figures, or Quranic phrases are accurately preserved or erased during the translation process.

To address these limitations, this paper introduces, to the best of our knowledge, the first multi-modal NMT (MMNMT) model specifically designed for the *Masnavi*, supported by a recitation-aligned multi-modal Persian–English corpus. This dataset is made up of 26,571 aligned verse pairs

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 131–139

across all six books of *Masnavi*, accompanied by corresponding Persian audio recitations. A robust Persian poetry fine-tuned text-only baseline (PPFT) is established, and afterwards extended to a multi-modal model. We aim to better preserve the depth and artistic intent of Rumi's poetry after translation. Therefore, we propose a CSI-aware evaluation framework using a customized CSI lexicon to diagnose and enhance the cultural authenticity in machine translation outputs. Through this work, we demonstrate that multi-modal integration not only clarifies interpretive ambiguity but also substantially improves the accuracy of culturally grounded lexical realizations.

This paper is structured as follows: we present a review of the related work to our topic in Section 2. Section 3 is dedicated to describing our produced multi-modal dataset. In Section 4, we introduce our CSI taxonomy, annotation process, and the construction of a Persian–English CSI lexicon. Our proposed methodology is elaborated in Section 5. Section 6 presents results and CSI-aware evaluation and analysis. Finally, Section 7 discusses limitations and outlines directions for future work followed by concluding remarks.

2 Related Work

Recent studies outside the realm of Indo-European languages investigated on how using literary and poetic texts can serve as an important role to preserve and revitalize endangered Indigenous languages. For instance, Cadotte et al. (2024) demonstrated that when traditional bilingual data is scarce, literature and poetry can help to develop machine translation (MT) models, specifically for languages like Innu-Aimun. This study supports the fact that employing culturally rich texts is a practical approach for training specialized systems where generic data is insufficient.

Chakrabarty et al. (2021) showed improvements in neural poetry translation through multilingual fine-tuning on poetic corpora, emphasizing the preservation of poetic style and meaning. However, they evaluated their work by using standard metrics like BLEU and COMET. Such metrics struggle to preserve cultural aspects, which is the challenge our work aims to address by introducing a CSI lexicon for the evaluation phase. This enables a fair evaluation that takes into account the preservation of cultural weight throughout the translation process.

The evaluation of cultural fidelity in MT has recently moved toward tracking CSIs. Yao et al. (2024) introduced the culturally-aware machine translation (CAMT) corpus and the CSI-Match metric (a metric that determines how closely a model's output matches reference translations for CSIs, using a fuzzy matching approach) to assess how models handle cultural entities in modern prose, such as Wikipedia articles. Comparably, Thakur (2025) proposed culturally-grounded chain-of-thought (CG-CoT) for interpretative analysis of Yoruba proverbs using retrieval-augmented generation. Although these works highlight the inadequacy of standard metrics like BLEU for cultural tasks, they focus on modern domains or reasoning sequences (Yao et al., 2024), (Thakur, 2025). In another study, Sadr et al. (2025) found that AI models often fail to understand Persian social rules like Taarof because they are biased toward Western-style directness. This is known as a "cross-cultural pragmatics problem" (Stadler, 2012), in which the literal words used in a conversation are different from what the person actually means. Because these rituals feature a form of "polite verbal wrestling" (Rafiee, 1991), text-only models that are fixated on literal meanings cannot capture the true cultural intent. This emphasizes the importance of specialized tools to evaluate how well the cultural weight of Persian language is preserved during translation (Sadr et al., 2025). Our work aims to move beyond generic models by introducing a specialized Persian poetry fine-tuned architecture (PPFT) that addresses the linguistic challenges of classical Persian verse. As it will be elaborated in Section 5.1, we construct an 18-category CSI lexicon specifically for the *Masnavi*, providing a targeted "Recall" metric that measures the preservation of classical cultural weight.

Most current machine translation research typically relies on either image-text multi-modal approaches or text-only methods. This means that the integration of information extracted from speech has received comparatively less attention. Current multi-modal machine translation research is heavily biased toward image-text combinations for short captions. Parida et al. (2024) and Rajpoot et al. (2024) utilized visual context from images to resolve polysemy in languages like Hindi and Bengali. As noted in the survey by Lupascu et al. (2025), text-image pairs represent 63% of the literature, while audio-text integration for low-resource languages remains significantly unexplored.

While Shahrestani and Haghir Chehreghani (2025) utilized deep learning for prosody recognition in Persian poetry, they mostly focused on classifying metrical patterns rather than translation. Xiang et al. (2025) empirically quantified the "modality gap" between speech and text inputs in large language models, identifying differences in semantic comprehension performance. Their research concentrates on conversational and synthetic speech in English. We aimed to addresses this gap by using self-supervised Persian speech encoders (Wav2Vec2) (Babu et al., 2021) and a cross-attention fusion mechanism (Vaswani et al., 2017) to capture essential audio features for classical poetry *Masnavi* that written text alone cannot provide.

Recent advancements in large language models (LLMs) have enabled new approaches to poetic interpretation. Gao et al. (2024) utilized ChatGPT to translate Chinese classical poetry, demonstrating that prompting models to interpret symbols or provide explanations first can enhance translation quality and readability. Although our work also employs LLMs (i.e., GPT-4) within a combined annotation framework, we distinguish our contribution by introducing a structured CSI-Recall metric as a diagnostic tool. This enables measurable evaluation of cultural transfer that goes beyond the subjective "Overall Impression" used in previous human evaluations (Wang et al., 2024).

Compared to the state of the art, our contribution is distinguished by two key aspects: the integration of recitation audio to clarify ambiguities, and the development of a CSI lexicon that enables evaluation of cultural fidelity, a dimension overlooked by conventional metrics such as BLEU and standard domain adaptation techniques.

3 Multi-modal Corpus of *Masnavi*

We construct and introduce a multi-modal Persian-English corpus derived from *Masnavi-ye-Ma'navi*. Its linguistic structure is characterized by "anomalous speech," where conventional grammatical patterns are intentionally altered to enhance aesthetic and emotional expression (Khanmohammadi et al., 2023). In addition, Persian poetry functions within a hybrid semiotic system that intertwines linguistic meaning with rhythmic, musical, and transcendental dimensions, further increasing abstraction and interpretive ambiguity (Sedaghat, 2020). These properties make *Masnavi* a challenging yet representative testbed for culturally grounded machine translation, particularly in low-resource settings. As discussed, this area remains underrepresented in existing MT benchmarks for Iranian languages.

book_	persian_text	english_translation	audio_filename	language
4	همسال از علم خیزد هم جواب همچنانک خار و گل از خاک و آب	Both question and answer arise from knowledge just as the thorn and the rose from earth and water	4-16261.mp3	fa
2	گفت آن خر کاو به شب لاحول خورد جز بدین شیوه نداند راه کرد	He replied The ass that ate Lá hawl during the night cannot get along except in this manner	2-4433.mp3	fa
2	آن تقاضای دو چشم دل شناس کاو همیجوید ضیای بیقیاس	Recognise that that is the craving of the eyes of your heart which is seeking the immeasurable Light	2-4264.mp3	fa
3	هر رسول شاه باید جنس او آب و گل کو خالق افلاک کو	Every kings messenger must be of his kind where are water and clay in comparison with the Creator of the heavens	3-10957.mp3	fa
1	ای برادر چون ببینی قصر او چون که در چشم دلت رسته ست مو	O brother how wilt thou behold his palace when hair has grown in the eye of thy heart	1-1471.mp3	fa
2	مهر مومش حاکی انگشتری است باز آن نقش نگین حاکی کیست	The seal impressed on his wax is telling of the sealring of whom again does the device tell graven on the stone of the ring	2-5527.mp3	fa
4	چون جوالی بس گرانی میبری زان نباید کم که در وی بنگری	When you are carrying a very heavy sack you must not fail to look into it	4-14770.mp3	fa
6	گفت صد خدمت کنم پانصد سجود بندهای دارم تن اسپید و جهود	He replied I will perform a hundred services and five hundred prostrations I have a handsome slave but a Jew	6-22601.mp3	fa
2	رزق جویی را ز بالا خو گرم تو ز بالا بر گشودستی درم	I am accustomed to seeking daily bread from above Thou hast opened to me the door from above	2-8094.mp3	fa

Figure 1: Partial representation of the produced multi-modal dataset.

3.1 Text Sources and Alignment

Persian text and English translations were collected from publicly available *Masnavi* repositories that provide full access to the Persian poems and their translations (Javadi et al., 2015), (Lewis, 2015), (Nicholson and Sufism.org, 2013). Different Persian literature repositories (e.g., Ganjoor [1]) are used for cross-checking verse boundaries and book structure. Each example is indexed and aligned at the verse unit. This indexing is also used to identify the corresponding audio file. The final prepared corpus consists of aligned Persian source verses, their English translations, and corresponding Persian audio recitations. A partial representation of the dataset structure is shown in Figure 1.

3.2 Audio Modality

Audio recitations were obtained from *Masnavi* audio repositories designed for verse-level listening (Javadi et al., 2015). We map each audio clip to its verse identifier and resample audio to a consistent sampling rate (16 kHz).

3.3 Corpus Statistics

As noted in Table 1, the corpus contains 26,571 aligned Persian-English verse pairs and 26,571 corresponding Persian audio files, utilizing all six books. We split the whole corpus into three sets of train (21,255), validation (2,658), and test (2,658). Each set includes: `book_id`, `persian_text`, `english_translation`, `audio_filename`, and `language`. Because literary translations may vary in literalness and poetic style, we treat the English side as a *reference translation* rather than a unique legitimate target and complement standard MT metrics with culturally targeted evaluation that is further discussed in Section 5 and Section 6.

[1] https://ganjoor.net/moulavi/masnavi

	Text	Audio
Size	26,571 lines/files	
Train	21,255 lines/files	
Validation	2,658 lines/files	
Test	2,658 lines/files	

Table 1: Statistics of multi-modal *Masnavi* corpus dataset and its division to train, validation, and test sets.

4 Cultural Aspect in Persian Poetry

The Persian language poses intrinsic challenges for NMT that extend well beyond lexical sparsity (Sedaghat, 2020). Such challenges would raise up in the translation of Persian classical poetry, particularly in low-resource settings. Prior work has noted the phenomenon of *second-degree deformation*, where translators (humans) or machines apply filters (e.g., secularization or Islamization biases) that reduce the poet's intent into fixed interpretations (Sedaghat, 2020). In other words, literal translation strategies are especially inadequate for this complex task. Moreover, generic NMT systems can generate false content or degrade metaphorical expressions upon processing figurative language (Chakrabarty et al., 2021).

Recent research suggests that addressing poetic translation requires architectural and training adaptations rather than mere scaling. Applying approaches augmented with domain-adaptive objectives can better respect poetic limitations than generic NMT systems (Khanmohammadi et al., 2023). However, we believe that tracking cultural-aware aspect of these translations by employing CSI-evaluation, can compensate for the inadequacy and deficiency of these methodologies.

To achieve this, we introduce the CSI lexicon for formalizing and evaluating cultural specificity in Persian–English poetic translation (detailed in Section 5).

5 Proposed Methodology

5.1 CSI Lexicon

To address the aforementioned limitations, this work models cultural fidelity in Persian–English poetry translation through CSIs. CSIs represent explicit mentions of culturally grounded entities and concepts, such as Sufi concepts, Quranic references, doctrine, and symbolic animals that hold meaning, semantic, and inter-textual weight beyond their literal expression. In contrast, standard MT evaluation metrics are poorly suited for this matter.

- Step 1: We developed a detailed label for the CSIs by integrating concepts from Persian literary studies and cultural linguistics. The taxonomy includes 18 culture-specific categories, listed as: person, place, Sufi concept, mystical phrase, supernatural being, number/color, medical/scientific concept, Quranic reference, religious concept, virtue, divine attribute, foreign Arabic/Turkic term, animal symbol, natural element, object symbol, sound/music, main doctrine of love, and other. Each label is accompanied by definition and some examples. These labels were manually refined and cleaned through iterative review by using cultural references from classical sources.

- Step 2: We annotated *Masnavi* verses with these CSI labels. In order to do that, we adopted a hybrid approach combining manual supervision as well as an LLM assistance (i.e, GPT-4). A unique prompting pipeline was used for the LLM to extend the annotations. Prompts provided the verse, the full taxonomy, and detailed instructions. The model predicted relevant CSI spans and assigned appropriate labels. All LLM-generated annotations were reviewed and corrected by the author to ensure high quality. The resulting dataset consists of 1,000 annotated verses that were randomly selected from the training set using a fixed seed for reproducibility.

- Step 3: The cleaned, filtered annotated file from the previous step is converted into token-level BIO tagging format (each token is marked as Beginning (B), Inside (I), or Outside (O) of a labeled span).

- Step 4: We trained a CSI tagger using our PPFT model and the BIO-tagged dataset.

- Step 5: We ran the trained tagger on the *Masnavi* corpus. Therefore, our whole corpus is tagged. As a result, we have fully tagged *Masnavi* corpus with CSI labels and spans.

The final phase is to build the Persian-English CSI lexicon, where each Persian span is mapped to its aligned English translation. For each pair, we collected all aligned English phrases across the corpus (train set) and aggregated them into a candidate

set with noting the frequency and alignment confidence statistics. To improve precision and reduce noise, we performed several cleaning steps, such as canonicalization of English forms, frequency and rank filtering, and label consistency filtering.

After all these steps, we produced three lexicon versions denoted as strict, soft and broad. The first version, strict core lexicon, contains only high confidence spans that are verified by native human that serves as a gold-standard reference for evaluation. The second version, soft core lexicon, represents all tagged spans, that is a more relaxed version compared to the strict one, but still enforcing strong alignment confidence. The third version, broad lexicon, intends to maximize coverage. It includes all aligned CSI candidates above a minimal confidence threshold. It is noisier but still valuable for culture-aware evaluation.

We concluded that having three versions instead of a single lexicon was a better approach. A single version would risk being either too restrictive (i.e., missing valid translations) or non-restrictive by adding unwanted noise. This three-version approach enables us to demonstrate robustness across evaluation settings, and analyze the translation systems' response to progressively broader cultural criteria. It is worth mentioning that each lexicon is derived from the same alignment and tagging procedure, differing only in filtering thresholds over alignment confidence and frequency.

Finally, the lexicon offers substantial coverage of frequent and semantically significant CSIs. This structural approach allows the lexicon to serve as the following two objectives: either as a culture-aware evaluation framework, or as an analytical resource for examining cross-cultural transfer phenomena in automated translation systems. In this work, we employed the generated CSI lexicon to support the culture-aware evaluation for our machine translation models.

5.2 Multi-modal NMT system for Persian Poetry

While our work focuses on a neural framework, the choice between statistical machine translation (SMT) and NMT in low-resource settings continues to be a subject of discussion. Cadotte et al. (2024) found that for extremely limited datasets of fewer than 4,000 sentences, SMT outperformed NMT. However, because our *Masnavi* corpus consists of over 26,000 verse pairs, we can overcome such limitations and utilize a domain-adaptive neural baseline (denoted as PPFT) that is further enhanced by employing the audio modality.

The overview of our proposed multi-modal NMT framework is illustrated in Figure 2. After the preprocessing phase as discussed earlier, we implemented and evaluated a strong text-only unimodal baseline (denoted as the second phase in Figure 2). It is worth mentioning that almost all of general NMT systems are trained on the general domain of Persian language like news, and Wikipedia. Therefore, effective poetic translation requires changes in model architecture and training strategy. One way to have a more strong and meaningful translation is to have a strong baseline. However, culturally-based translation challenges, especially in low-resource settings, continue to persist even when using a stronger text-only baseline. In this regard, we pre-trained and fine-tuned the mBART50 (Tang et al., 2020) by using Persian poetry. We used a large monolingual Persian poetry corpus containing 1M lines, that is publicly available [2]. Then, we fine-tuned it on our parallel *Masnavi* Persian-English corpus (train set), denoted as Persian poetry fine-tuned text-only model (PPFT). We further used this model as the base for the rest of the multi-modal models.

As it is shown in Figure 2, the third phase regards developing the multi-modal NMT. This model is defined by the interaction between two high-dimensional latent spaces: the audio space (Wav2Vec 2.0) and the semantic poetry space (PPFT). In this system, two distinct encoders are employed. The audio encoder (Wav2Vec 2.0 XLS-R) (Babu et al., 2021) extracts phonetic representations, and the text encoder (PPFT), processes the Persian verses. We implemented a multi-modal fusion layer using a multi-head cross-attention mechanism (Vaswani et al., 2017). In result, the model can identify phonetic details in the Masnavi recitation to the related words or meaning in the verse. The resulting multi-model model is called MM-NMT model. Finally, fused multi-modal representation is converted into a target English sequence by using the mBART50 decoder (Tang et al., 2020).

6 Evaluation Framework

Tables 2 and 3 respectively report the results of the NMT/MMNMT models and the CSI-Evaluation on the main parallel Persian-English corpus. The CSI

[2]Persian_poems_corpus: https://github.com/amnghd/Persian_poems_corpus/tree/master

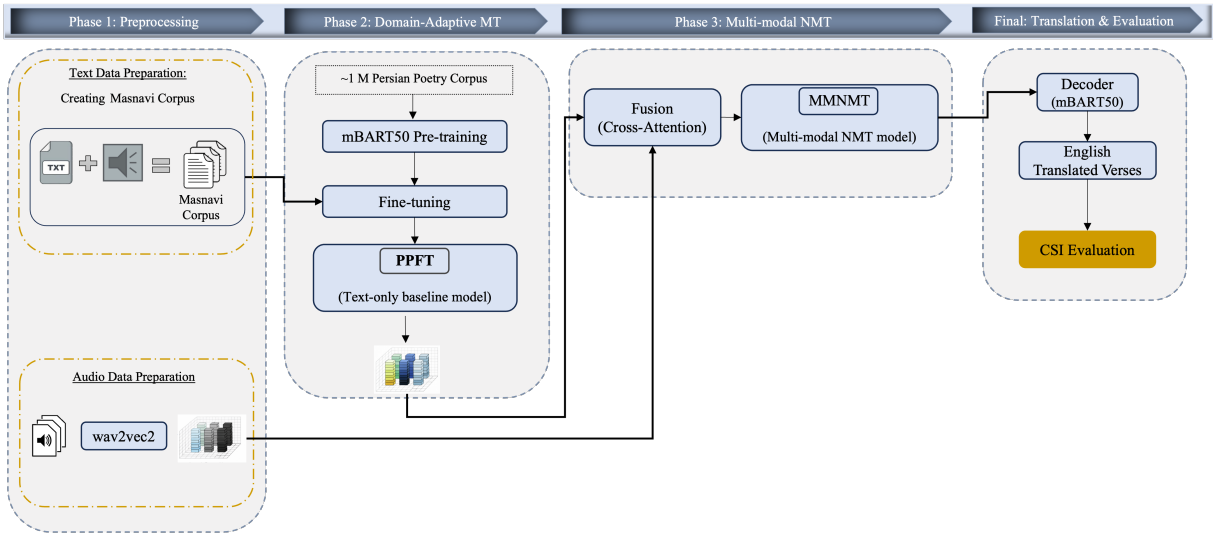


Figure 2: Overview of our proposed multi-modal translation system developed for Persian poetry translation to English.

Modality	Model	BLEU	ChrF++	BERTScore	COMET
Txt	PPFT	9.85	32.77	0.876	0.579
Txt + Audio	PPFT + Wav2Vec2	17.95	42.95	0.894	0.635

Table 2: Evaluation results for text-only baseline (PPFT) and multi-modal NMT (PPFT + Wav2Vec2) on Persian–English translation of the *Masnavi* corpus.

evaluation is done on the both uni-modal and multi-modal scenarios. Across all automatic metrics, the MMNMT demonstrates substantial and consistent improvements compare to the text-only model as elaborated in Table 2. Although the PPFT model performance metrics in Table 2 are reasonable for a low-resource poetic MT baseline, they are clearly limited. BLEU is conservative (paraphrase sensitivity), yet COMET and ChrF++ confirm the model is still missing many correct realizations.

By integration of the audio recitations and building up the proposed MMNMT system, the BLEU score and the the ChrF++ metric increased significantly. Such improvements indicate an enhanced management of morphological variations and character-level accuracy. Such character-level precision holds particular importance for Persian-to-English translation tasks. The semantic metrics like BERTscore and COMET are also presented, showing an increase from 0.876 to 0.894, and from 0.579 to 0.635, respectively. These two metrics are less sensitive to exact lexical overlap and more aligned with meaning and semantic preservation. This clearly indicates that our proposed MMNMT system produces translations that are semantically closer to the references. Furthermore, the consistent gains across all presented metrics, indicate that incorporating audio information enhanced both the fluency and the semantic adequacy of translations.

It is noteworthy that although the results have improved, the overall scores are still moderated compared to high-source prose translation tasks. This is because most automatic metrics can not fully measure the extent in which cultural and poetic meanings are preserved. For example, the conservative nature of metrics like BLEU often fails to capture the semantic and artistic depth of poetic translations. As noted by Cadotte et al. (2024), quantitative scores may not reflect the "actual quality" of poetic MT, as translations can be correct and culturally grounded even when they differ significantly from the reference string. Persian poetry includes culture-specific terms. Translating these correctly may not raise metric scores for word overlap or meaning similarity, but it is essential for understanding and keeping the cultural value of the text. This reinforces the necessity of defining a CSI-Recall metric, as investigated in the next section. Such metric provides a diagnostic of cultural fidelity that is not captured by standard evaluation metrics.

Model	Lexicon Scope	Coverage(%)	CSI-Recall (%)
PPFT	strict_core_top	24.16	61.60
PPFT + Wav2Vec2	strict_core_top	24.16	82.04
PPFT	soft_core_top	24.16	66.94
PPFT + Wav2Vec2	soft_core_top	24.16	89.04
PPFT	broad_rank	69.66	51.77
PPFT + Wav2Vec2	broad_rank	69.66	75.73

Table 3: CSI-aware evaluation of text-only baseline (PPFT) and multi-modal NMT (PPFT + Wav2Vec2) on the *Masnavi* corpus across three lexicon versions: strict_core (high-confidence gold-standard spans), soft_core (relaxer version with all tagged spans and strong alignment), and broad (maximum coverage with minimal confidence threshold).

6.1 CSI-Aware Evaluation

Table 3 presents the results of a culture-aware evaluation using our CSI taxonomy. This evaluation verifies whether culturally important Persian terms are correctly translated. Both the coverage and CSI-Recall metrics are reported. Coverage shows how many CSI words our dictionary knows. This coverage has nothing to do with our models, it only depends on the lexicon. Therefore, based on only what the lexicon can judge, the coverage of 24.16% seems low. This means that only 24% of CSI spans are lexicon-covered. Recall on the other hand means that out of the CSI words we can check, how many did the model translate correctly.

In all versions of our lexicons, the proposed MM-NMT (e.g., PPFT + Wav2Vec2 model) achieved higher recall than the uni-model PPFT. For instance, as it is shown in Table 3, in the strict core setting, CSI-Recall goes up from 61.60% to 82.04%. Similar improvements appear in the soft core and broad lexicon settings, where recall rises from 66.94% to 89.04%, and from 51.77% to 75.73%, respectively. These results show that the multi-modal model maintains considerably more cultural information when exact cultural rendering is required.

These gains illustrate that audio information helps guide the model to a more culturally appropriate translation. As discussed earlier, the act of recitation plays a vital role in how meaning is conveyed through rhythm, tone, pauses, and emphasis, elements that don't appear in writing, specifically for poetry that are hard to read and understand like *Masnavi* of Rumi. These sound-based features reveal much about the poem's structure, mood, and the poet's intentions. The written Persian language, particularly in classical poetry, is naturally ambiguous.

When audio features are integrated with text, translation models gain access to expressive and structural details that written words alone cannot capture. As supported by the results provided in both Table 2, and Table 3, the resultant multi-modal model enables translations that better preserve the depth, and intent of Persian poetry, especially in rich, culturally layered works from the classical tradition.

It is important to note two limitations of this evaluation. First, the core coverage is only 24%, which means the metric currently judges limited slice of CSI spans. Second, the CSI metric is currently recall-only over lexicon-covered spans (no precision), so it does not penalize hallucinated CSI words or incorrect usage. In order to validate that these gains are not caused by insignificant effects, we also demonstrated COMET/BERTScore improvements presented in Table 2. Based on the results, switching from text-only baseline to a multi-modal model that used the audio recitation has achieved a significant gain with a large and consistent improvement across all metrics.

7 Future Work

This work focused solely on using CSI lexicon as an evaluation step. In future work, we would like to investigate the impact of using our refined lexicon as an analytical resource for examining cross-cultural transfer phenomena in our proposed models. In addition, further refinement of the created CSI lexicon, particularly through expanded coverage, would strengthen its reliability. Finally, because automatic metrics such as BLEU, ChrF++, and BERTScore do not fully capture poetic fidelity, we will incorporate structured human evaluation protocols.

Conclusion

This study explores both the computational and cultural challenges of translating Rumi's *Masnavi-ye-Ma'navi*, a Persian mystical poetry known for its intentional ambiguity and rich symbolism. Since NMT systems show inadequacy in the low-resource poetic settings, we developed the first multi-modal translation framework specifically for classical Persian poetry. Our main contributions are: (1) creating a new parallel corpus with 26,571 pairs of Persian-English verse from all six books of the *Masnavi*, each accompanied by aligned audio recitations; (2) establishing a domain-adapted baseline through pre-training on Persian poetry corpus followed by fine-tuning on our *Masnavi* dataset (train set); (3) improving the model by integrating acoustic features via cross-attention fusion mechanisms; and (4) creating a CSI taxonomy and lexicon for the evaluation of cultural preservation in translation outputs.

The experiments show that adding audio greatly improves the results across multiple dimensions. Standard metrics show notable gains, BLEU increased from 9.85 to 17.95, while ChrF++ goes up from 32.77 to 42.95. More significantly, our CSI-aware evaluation demonstrates that multi-modal models better preserve cultural fidelity, with recall rising from 61.60% to 82.04% under strict lexicon constraints. These findings confirm that audio features and information help the model understand meaning more accurately and maintain cultural authenticity. This work establishes a foundation for culturally-aware machine translation of classical poetry. It demonstrates that multi-modal approaches can preserve the literary richness and mystical depth that are central to Persian poetry. Future research focus on expanding lexicon coverage, investigating audio's contribution through systematic ablation studies, and using the lexicon as an extra training resource to strengthen cultural understanding in translation models.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, and 1 others. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Antoine Cadotte, Nathalie André, and Fatiha Sadat. 2024. Machine translation through cultural texts: Can verses and prose help low-resource indigenous models? In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 121–127, Bangkok, Thailand. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruiyao Gao, Yumeng Lin, Nan Zhao, and Zhenguang G Cai. 2024. Machine translation of Chinese classical poetry: a comparison among ChatGPT, google translate, and deepl translator. *Humanities and Social Sciences Communications*, 11(1):1–10.

Mozhgan Ghassemiazghandi. 2023. Machine translation of selected ghazals of Hafiz from Persian into English. *AWEJ for Translation & Literary Studies*, 7(1).

Mojdeh Javadi, Mani Zarinebaf-Shahr, Alan Lewis, and Maryam Lewis. 2015. Masnavi.net: A full-text website of the masnavi in persian, english, and turkish. http://masnavi.net/. Accessed: 2026-01-11.

Reza Khanmohammadi, Mitra Sadat Mirshafiee, Yazdan Rezaee Jouryabi, and Seyed Abolghasem Mirroshandel. 2023. Prose2poem: The blessing of transformers in translating prose to persian poetry. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Alan Lewis. 2015. The masnavi. https://www.dar-al-masnavi.org/masnavi.html. Accessed: 2026-01-11.

Marian Lupascu, Ana-Cristina Rogoz, Mihai Sorin Stupariu, and Radu Tudor Ionescu. 2025. Large multimodal models for low-resource languages: a survey. *arXiv preprint arXiv:2502.05568*.

Reynold A. Nicholson and Sufism.org. 2013. Rumi: Masnavi, book i (version 1.0). https://sufism.org/wp-content/uploads/2013/12/Rumi-Book-I-Version-1.0.pdf. Accessed: 2026-01-11.

Shantipriya Parida, Ondřej Bojar, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, and Ibrahim Said Ahmad. 2024. Findings of WMT2024 English-to-low resource multimodal translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 677–683, Miami, Florida, USA. Association for Computational Linguistics.

Abdorreza Rafiee. 1991. *Variables of communicative incompetence in the performance of Iranian learners of English and English learners of Persian.* Ph.D. thesis, School of Oriental and African Studies (University of London).

Pawan Rajpoot, Nagaraj Bhat, and Ashish Shrivastava. 2024. Multimodal machine translation for low-resource Indic languages: A chain-of-thought approach using large language models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 833–838, Miami, Florida, USA. Association for Computational Linguistics.

Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoomian, Laleh Seyyed-Kalantari, and Ali Emami. 2025. We politely insist: Your LLM must learn the Persian art of Taarof. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1819–1838.

Amir Artaban Sedaghat. 2020. Translating Rumi through the prism of ideology. *Iran Namag*, 5(2):4–34.

Mohammadreza Shahrestani and Mostafa Haghir Chehreghani. 2025. Prosody recognition in persian poetry. *Speech Communication*, 170:103222.

Stefanie Stadler. 2012. Cross-cultural pragmatics. *The encyclopedia of applied linguistics*, pages 1–8.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Madhavendra Thakur. 2025. Culturally-grounded chain-of-thought (CG-CoT): Enhancing LLM performance on culturally-specific tasks in low-resource languages. *arXiv preprint arXiv:2506.01190*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. 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. Curran Associates Inc.

Shanshan Wang, Derek Wong, Jingming Yao, and Lidia Chao. 2024. What is the best way for ChatGPT to translate poetry? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043, Bangkok, Thailand. Association for Computational Linguistics.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. 2025. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5187–5202.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Author Index