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*