

Computationally Invisible: Charting the NLP Landscape of Iranian Languages

Hojjat Mokhtarabadi^{1*} Amir Hossein Kargaran^{2*} Ali Shakeri¹ Saba Shafiei¹

Nafiseh Nikeghbal³ Mohammad Hossein Manshaei⁴

¹Isfahan University of Technology ²LMU Munich ³TU Munich

⁴Hunter College and Graduate Center, City University of New York

h.mokhtarabadi@ec.iut.ac.ir,

amir@cis.lmu.de, mohammad.manshaei@hunter.cuny.edu

Abstract

The Iranian language family spans 101 ISO 639-3 coded varieties and over 200 million speakers across a vast arc from the Caucasus through the Middle East and Central Asia to South Asia and China. Yet this richness is barely reflected in modern NLP: resources, models, and benchmarks are overwhelmingly concentrated on Iranian Persian, leaving Pashto, Balochi, Gilaki, Zazaki, and the Pamir languages computationally invisible. This survey provides the first comprehensive landscape analysis of Iranian NLP, covering 495 papers published between 2020 and early 2026. We examine language coverage, NLP tasks, datasets, modeling approaches, and evaluation practices, and identify persistent structural gaps: script heterogeneity, dialectal fragmentation, data scarcity, and an evaluation monoculture that has saturated Persian while leaving near-zero baselines for every other variety. Together with this survey, we release a companion website providing interactive language maps, a curated resource browser, and a live publication index at <https://sigiranic.github.io>.

1 Introduction

The Iranian¹ (or Iranian²) language family is a branch of the Indo-Iranian subfamily of the Indo-European language family. Spoken across one of the most geographically expansive and politically diverse regions of the world, it stretches from the Caucasus in the west to Xinjiang in China in the east, and from the Russian North Caucasus in the north to the Persian Gulf and Oman in the south. Countries with significant Iranian-speaking populations include Iran, Afghanistan, Tajikistan, Pakistan, Iraq, Turkey, Azerbaijan, Russia (North

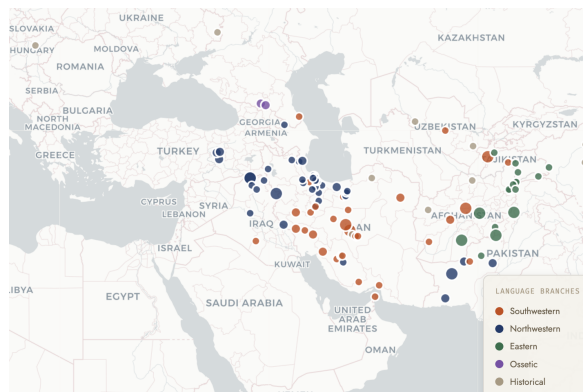


Figure 1: Geographic distribution of all 101 ISO 639-3 coded Iranian varieties, plotted at Glottolog centroid coordinates. Dot size reflects speaker population. The family spans from the Caucasus and Hungary in the west to Xinjiang in the east, and from the Russian North Caucasus to Oman in the south.

Ossetia–Alania), Georgia (South Ossetia), Uzbekistan, China, and Oman, among others. The family comprises 101 ISO 639-3 coded varieties (Eberhard et al., 2026), ranging from Persian, a language with over 100 million speakers and a rich written tradition spanning millennia, to Yagnobi, an endangered Sogdian descendant spoken by roughly 13,000 people in the Yagnob Valley of Tajikistan (Figure 1 shows the geographic distribution of all varieties).

NLP has undergone a transformative decade. Large-scale pretrained language models, from BERT (Devlin et al., 2019) to GPT-5 (Singh et al., 2026) and Llama 3 (Grattafiori et al., 2024), have redefined what is possible for high-resource languages. Yet these advances have largely bypassed the Iranian family. Most frontier models do not officially support any Iranian language beyond Persian, and even Persian support is often limited to the Iranian standard variety, failing to account for Afghan Dari or Tajik. Kurdish, the family’s second-largest language by speaker count, remains absent

*Equal Contribution.

¹ISO 639-2 / 5: ira; Glottolog: iran1269

²We use the terms Iranian and Iranian interchangeably throughout this paper.

from most major benchmarks. For Balochi, Zazaki, Talysh, and the Pamir languages, the research literature is nearly empty.

Despite this exclusion, Iranic NLP is active and growing. Iranian Persian has accumulated a substantial body of work in language modeling (Farahani et al., 2021a; Hosseinbeigi et al., 2025c), question answering (Khashabi et al., 2021; Darvishi et al., 2023), speech processing (Zeinali et al., 2019; Kalahroodi et al., 2025a), and LLM evaluation (Ghahroodi et al., 2024; Shamsfard et al., 2025). Other Iranic languages have also seen growing NLP research, with machine translation being one of the most active areas (Ahmadi et al., 2025b; Caswell et al., 2025).

This survey has three goals. First, we **systematically catalogue** Iranic NLP research across 495 papers, organizing coverage by language, task, and resource type. Second, we **audit the digital infrastructure** of the family, from web corpora and Wikipedia presence to morphological tools and parallel data, revealing the foundational constraints within which all NLP work must operate. Third, we **identify structural gaps and open challenges**, documenting which languages and tasks are critically underserved and what concrete steps the community can take. Taken together, these three goals make this survey an accessible entry point for newcomers to Iranic NLP, as well as a reference for researchers seeking to understand where the field stands and where it needs to go. Our analysis covers all 101 ISO 639-3 coded varieties across the four branches of the family.

2 Background: The Iranian Language Family

2.1 Taxonomy and Scope

ISO 639-3 assigns codes at two scope levels: macrolanguage and individual language. For the Iranic family, only four codes carry macrolanguage status: fas (Persian), kur (Kurdish), pus (Pashto), and bal (Balochi). The remaining 97 codes are all assigned as individual languages, regardless of whether linguists treat them as distinct languages, regional dialects, or historical varieties with no living speakers.

This formal taxonomy, however, diverges in significant ways from both popular usage and scholarly debate. The macrolanguage fas (Persian/Farsi) officially subsumes pes (Iranian Persian, the variety of Iran) and prs (Dari, the variety

of Afghanistan). It does not cover tgk (Tajik), as the latter is written in Cyrillic and has official status in Tajikistan, though the three varieties are largely mutually intelligible. Whether Persian, Dari, and Tajik are three registers of one language or three distinct languages is politically charged: in Iran, Afghanistan, and Tajikistan respectively, national identity and language policy intersect with the question. NLP systems that treat fas as a monolithic label conflate varieties whose orthography, lexicon, and speech patterns differ substantially. Similarly, kur covers Kurmanji (kmr), Sorani (ckb), and Southern Kurdish (sdh), which have limited mutual intelligibility and use different writing conventions, including Latin and Arabic-based scripts. The label “Kurdish” used in NLP papers without a qualifier is therefore systematically ambiguous.

The distinction between “language” and “dialect” is itself a social and political construction, not a purely linguistic one. Laki (lki) is classified by some as a Kurdish dialect and by others as a distinct Northwestern Iranian language; Glottolog places it in a Laki-Kurdish group. Bakhtiari (bqi) is ISO-coded as a separate variety of the Luri continuum (lrc, luz), but speakers of Standard Persian sometimes treat all Luri speech as a dialect of Persian. These ambiguities are not merely academic: they determine what data gets labeled, what models get built, and what communities receive NLP tools.

The 101 coded varieties span a vast range of vitality. At one extreme, Persian has nearly 60 million documents in FineWeb2 (Penedo et al., 2025) and over a million Wikipedia articles. At the other extreme, varieties like Natanzic (ntz), Nayini (nyq), and Sivandi (siy) have a few thousand speakers, no NLP resources, and face extinction within a generation.

2.2 Branches and Geographic Spread

The family is conventionally divided into four branches.

Southwestern Iranian. is the largest by speaker count, dominated by the Persian macrolanguage (fas/pes/prs/tgk) with approximately 130 million speakers across Iran, Afghanistan, Tajikistan, and diaspora communities. Also included are Luri, Bakhtiari, and Southern Luri (~5M speakers in western Iran); Laki (~680K in Kermanshah and Ilam); a cluster of highly endangered Central Plateau varieties in Isfahan and Hormozgan provinces; and Jewish Persian varieties Judeo-Persian (jpr) and Bukharic (bhh).

Language	ISO 639-3	Branch	Speakers (L1 / L2)	Script(s)	NLP
<i>Southwestern</i>					
Persian	fas / pes	SW	91M / 35M	Perso-Arabic, Cyrillic	H
Dari	fas / prs	SW	15M / 18M	Perso-Arabic	M
Tajik	tgk	SW	8M	Cyrillic	L
Luri / Bakhtiari	lrc / bqj / luz	SW	5M	Perso-Arabic	VL
Hazaragi	haz	SW	3M	Perso-Arabic	VL
Laki	lki	SW	0.7M	Perso-Arabic	VL
<i>Northwestern</i>					
Kurdish	kur / kmr / ckb / sdh	NW	26M	P-Arab/Lat/Cyr.	L
Balochi	bal / bcc / bgn / bgp	NW	9M	Perso-Arabic	VL
Mazanderani	mzn	NW	1.4M	Perso-Arabic	VL
Gilaki	glk	NW	3M	Perso-Arabic	VL
S. Kurdish	sdh	NW	3M	Perso-Arabic	VL
Zazaki	zza / diq / kiu	NW	2–4M	Latin	VL
Talysh	tly	NW	0.8M	P-Arab/Cyr/Lat.	VL
Gorani	hac	NW	0.4M	Perso-Arabic	VL
<i>Eastern</i>					
Pashto	pus / pbt / pbu / pst	E	51M / 5M	Perso-Arabic	L
Shughni	sgn	E	0.1M	Cyrillic, Latin	VL
Wakhi	wbl	E	0.06M	P-Arab/Cyr/Lat.	VL
Yaghnobi	yai	E	0.01M	Cyrillic	VL
<i>Ossetic</i>					
Ossetic	oss / osd	Oss.	0.6M	Cyrillic	VL

Table 1: Overview of major modern Iranian languages with ISO 639-3 codes, Wikipedia speaker counts (L1 / L2), and NLP resource level. NLP level: **H**igh, **M**edium, **L**ow, **VL** Very Low.

Northwestern Iranian. encompasses Kurdish (~26–30M speakers across Turkey, Iran, Iraq, and Syria), Balochi (~8.8M across Pakistan, Iran, and Afghanistan), the Caspian languages Mazanderani (~4M) and Gilaki (~3M), the Zaza-Gorani group including Zazaki (~2–4M) and Gorani (~400K), Talysh (~800K in Azerbaijan and Iran), and numerous small endangered Tatic varieties in Zanjan, Semnan, and East Azerbaijan provinces.

Eastern Iranian. is represented primarily by Pashto (pus/pbt/pbu/pst, ~51–60M speakers, co-official in Afghanistan) and the Pamir languages spoken across Tajikistan, Afghanistan, Pakistan, and Xinjiang, including Shughni (sgn, ~95K), Wakhi (wbl, ~60K), and Yaghnobi (yai, ~13K). Most Pamir languages are UNESCO-classified as endangered.

Ossetic. is the sole living descendant of the Scythian-Sarmatian branch, spoken in North Ossetia (Russia) and South Ossetia (Georgia), with two varieties: Iron (oss) and Digor (osd).

Historical varieties. including Old Persian (peo), Avestan (ave), Middle Persian (pal), Sogdian (sog), and Bactrian (xbc) are attested in inscriptions and manuscripts but fall outside the scope of applied NLP.

Table 1 summarizes the major modern Iranian languages covered in this survey, selected for their digital presence and NLP engagement. The complete 101-variety taxonomy appears in Appendix A.

2.3 Related Surveys and Community Context

Several recent surveys provide models for our work. For African languages, [Alabi et al. \(2025\)](#) show how community-led shared tasks catalyze progress across resource-scarce families. [Aji et al. \(2022\)](#) examine Indonesia’s 700+ languages, highlighting tensions between national language policy and minority variety support. [Veitsman and Hartmann \(2025\)](#) survey Central Asian Turkic NLP, a region geographically and sociolinguistically adjacent to the Iranian family. [Poria and Huang \(2025\)](#) cover South Asian low-resource languages, many sharing script conventions and contact phenomena with Northwestern Iranian varieties. [Lovenia et al. \(2024\)](#) aggregate community resources across Southeast Asian languages into a unified benchmark suite. For low-resource NLP methodology, [Hedderich et al. \(2021\)](#) survey transfer learning, data augmentation, and distant supervision approaches applicable across the Iranian family, while [Ranathunga et al. \(2023\)](#) provide a comprehensive treatment of low-resource neural machine translation directly relevant to the family’s mid- and lower-tier languages. [Arora et al. \(2022\)](#) argue that data scatteredness rather than scarcity is the primary obstacle in South Asian NLP, a diagnosis that applies equally here. Across all these surveys, common themes emerge: the dominance of one or two high-resource pivot languages, script challenges, evaluation gaps, and the critical role of community organization. The Iranian family has lacked such a systematic treatment until now.

The present survey is accompanied by a community website dedicated to NLP research and resources for the Iranian language family. The SilkRoadNLP workshop at EACL 2026 ([Merchant and Megerdooian, 2026](#)) was the first ACL-affiliated venue dedicated to the family, reflecting a growing momentum in the community.

3 Survey Methodology

Literature Collection and Annotation. We conducted a systematic search across Google Scholar, the ACL Anthology, and Semantic Scholar, querying all Iranian language names and ISO 639-3 codes combined with NLP task terms—including natural language understanding, entailment, sentiment analysis, question answering, summarization, language modeling, instruction tuning, benchmark, dataset, corpus, and toolkit—following the methodology of [Alabi et al. \(2025\)](#). Papers were required

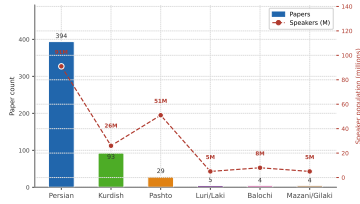


Figure 2: Paper count vs. speaker population (millions) for major Iranian languages. Papers covering several languages are counted once per language, so totals exceed 495. The mismatch is most striking for Pashto (51M speakers, 29 papers) and Balochi (8.8M speakers, 4 papers).

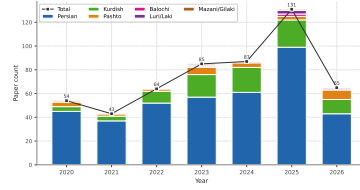


Figure 3: Iranian NLP paper counts by year (2020–2026). Growth is driven almost entirely by Persian; other languages show modest increases concentrated in Kurdish and Pashto. The 2026 bar is partial, reflecting a May 2026 search cutoff.

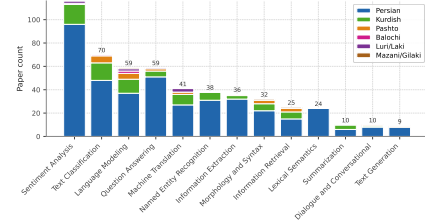


Figure 4: Paper counts by NLP task category. Sentiment analysis (116 papers) and text classification (70 papers) dominate; most attention goes to Persian. Papers spanning several tasks are counted once per task, so totals exceed 495.

to reference Iranian languages in the document body; the set was further augmented through targeted searches and citation tracing, yielding 5,878 candidate publications (2020–2026). The final literature search was conducted in May 2026. Papers were then processed through a two-stage pipeline using Qwen3.5-9B (Yang et al., 2025)³: the first stage filtered for relevance; the second annotated each paper for (1) language(s) covered, (2) NLP task category, (3) resource type (dataset, model, toolkit, benchmark, corpus, or method), (4) public availability of released resources, (5) publication year, and (6) venue. After filtering, deduplication, and accounting for multilingual studies, 495 papers remained. All annotations underwent manual verification by two authors; in cases of disagreement, a third author served as arbiter. Prompts and code are publicly available at <https://github.com/sigiranic/sigiranic.github.io>. Throughout the paper, claims that a resource does not exist should be read as “no publicly documented resource was found within the scope of this search” rather than as proof of non-existence.

Overview of the Corpus. Figure 2 shows the severe imbalance between NLP paper counts and speaker populations. Persian accounts for 394 papers (80% of the 495 papers in the corpus) despite representing roughly 60% of the family’s speakers. Kurdish, with 93 papers, receives reasonable attention relative to its speaker base. Pashto (29 papers), Luri/Laki (5), Balochi (4), and Mazanderani/Gilaki (4) are critically underserved given their tens of millions of speakers. For the remaining varieties, we could find no dedicated paper at all.

Figure 3 shows year-by-year growth: total output increased from 54 papers in 2020 to 131 in 2025, with Kurdish and Pashto growing modestly, while other varieties remain near zero throughout the period. Figure 4 shows the task distribution: sentiment analysis and text classification dominate, followed by language modeling, question answering, and machine translation. Figure 5 shows the flow from languages to research areas to resource types, highlighting that the Persian NLP ecosystem is broad and mature while all other varieties account for only a small portion of Iranian NLP.

4 Digital Infrastructure and Linguistic Resources

Appendix Table 8 reports the digital footprint of major Iranian languages across FineWeb2 corpus size, Wikipedia and Wiktionary coverage, UniMorph lemmata and forms, LID accuracy, speaker counts, and parallel data, along with the availability of grammar, WordNet, spell-checking, and ASR tools, with Arabic and Turkish as reference points.

Wikimedia Footprint. An important dimension of a language’s NLP readiness is its Wikimedia footprint. Among Iranian languages, ten have active Wikipedia subdomains (as of May 2026): Persian (fa, ~1M articles), Tajik (tg, ~118K), Kurmanji Kurdish (ku, ~92K), Sorani Kurdish (ckb, ~82K), Mazanderani (mzn, ~65K), Gilaki (glk, ~48K), Zazaki (diq, ~43K), Ossetian (os, ~22K), Pashto (ps, ~21K), and Talysh (tly, ~10K). A further eleven varieties have Wikimedia Incubator entries: Western Balochi (bgn), Bakhtiari (bqi), Southern Luri (luz), Laki (lki), Kirmanjki (kiu), Gorani (hac), Shughni (sgh), Wakhi (wbl), Yagh-

³<https://hf.co/Qwen/Qwen3.5-9B>

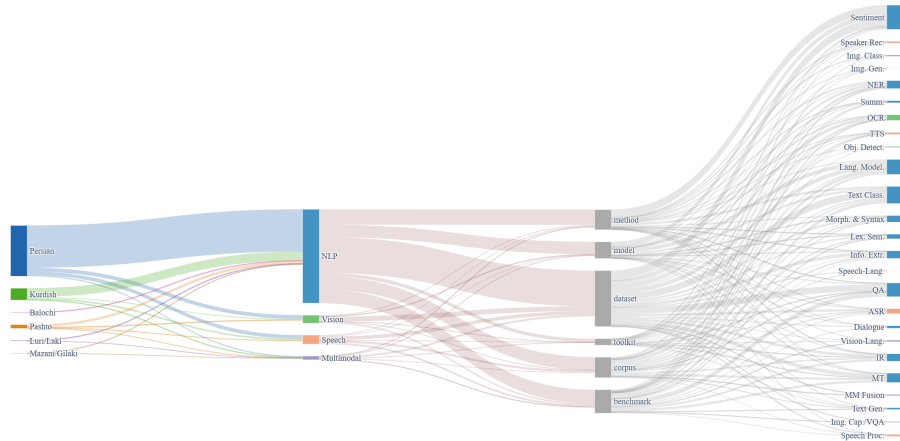


Figure 5: Flow from language to research domain, resource type, and NLP task. Persian dominates across all dimensions; other languages connect narrowly, primarily through corpora and datasets.

nobi (yai), Hazaragi (haz), and Larestani (lrl), though several of these contain only a single page and show no recent activity. The remaining 76 of the 97 individual varieties have no Wikipedia presence of any kind.

Two patterns from Table 8 are especially revealing. First, Gilaki and Mazanderani exhibit a *hollow digital presence*: each has tens of thousands of Wikipedia articles but fewer than 340 Wiktionary entries, indicating that quantitatively large corpora do not translate to the lexical infrastructure needed for lemmatization or semantic parsing. Second, the failure of the Northern Luri Wikipedia is instructive: an earlier attempt was abandoned after community review found that articles did not conform to authentic native grammar, likely because editors were writing in Persian with light lexical substitution. Multilingual corpora including GlotCC (Kargaran et al., 2024a), EMMA-500 (Ji et al., 2025), and MADLAD-400 (Kudugunta et al., 2023) cover between 8 and 16 Iranic variety-script pairs, but their quality for very low-resource varieties depends critically on language identification accuracy.

Morphological and Syntactic Resources. Iranic languages are morphologically complex: Persian has the Ezafe construction and rich verbal inflections; Kurdish has split ergativity; Pashto has polypersonal verbal agreement; Pamir languages have up to five agreement slots per verb. UniMorph coverage reveals a striking asymmetry: Northern Kurdish leads the family with 15,083 paradigms and 216,370 inflected forms, while Persian has only 273 paradigms despite its raw data advantage. For Balochi, Talysh, Mazanderani, Gilaki, Gorani, Luri,

and Laki, UniMorph coverage is zero. Persian morphological processing is supported by Hazm (Roshan Research, 2026), DadmaTools (Etezadi et al., 2022), and Parsivar (Mohtaj et al., 2018), among others; no equivalent tools exist for any other Iranic variety. Syntactic annotation is confined to a handful of treebanks: the Persian Universal Dependency treebank (Rasooli et al., 2020) and, for Kurdish, the UD Kurmanji (Gökırmak and Tyers, 2017) and Central Kurdish–Mukri (Asadpour, 2025) treebanks.

Language Identification. Language identification (LID) is the foundational prerequisite for corpus curation. For Iranic languages it is exceptionally difficult: Persian, Dari, and Tajik share large mutual vocabulary; Luri/Persian and Talysh/Azerbaijani are highly confusable. PALI (Ahmadi et al., 2023a) provides the first LID benchmark specifically for Perso-Arabic script languages including six Iranic varieties. GlotLID-V3 (Kargaran et al., 2023) is currently the broadest system, covering 16 Iranic variety-script pairs with near-perfect F1 for Persian and Kurdish but elevated false-positive rates for precarious pairs. MaskLID (Kargaran et al., 2024b) extends coverage to code-switched texts.

Parallel Corpora and Translation Resources. Machine translation has been the dominant NLP task for non-Persian Iranic languages. Table 9 presents sentence-pair availability across varieties; three tiers are visible. Persian dominates with over 600 million segment pairs with English (from web crawls, subtitles, and NLLB-200 (NLLB Team et al., 2022)). Kurdish holds a distant second with approximately 511K English pairs from the Kur-

dish Parallel Corpus (Ahmadi et al., 2022), supplemented by SMOL (Caswell et al., 2025) and PARME (Ahmadi et al., 2025b). Below this tier, every remaining Iranic variety has fewer than 40K pairs in total; Balochi has only 110 English pairs. PARME (Ahmadi et al., 2025b) is the most significant recent expansion, covering six Iranic languages. Gorani has 22K pairs from a literary translation corpus (Ahmadi et al., 2025a). The most critical gap is the scarcity of direct parallel data *between* regional varieties: Table 9 shows at most a few thousand pairs for any non-Persian pair, and Zazaki has none with any of the smaller varieties. Cross-lingual transfer paths therefore run almost entirely through Persian or English pivots. To our knowledge, the Kurmanji-Sorani corpus (Amini et al., 2021) remains the only substantial intra-Iranic MT resource.

5 NLP Tasks: Progress, Gaps, and Open Issues

Pre-training and Language Modeling. For Persian, the dominant paradigm of pretraining large encoders and fine-tuning downstream is well established. ParsBERT (Farahani et al., 2021a) remains the foundational monolingual encoder; the Matina corpus (Hosseini et al., 2025c) (73B tokens) and its accompanying LLM represent the current state of the art in Persian generative modeling. Several domain-adapted Persian encoders and instruction-tuned LLMs exist (see Appendix Table 12). FarsInstruct (Mokhtarabadi et al., 2025) covers 21 instruction datasets; HarfoSokhan (Sarvestani et al., 2026) provides 6M colloquial-to-formal Persian pairs, addressing the critical register gap. For Kurdish, KuBERT (muhealddin Awla et al., 2025) and KurdBERT (language-ml-lab, 2024) provide monolingual encoders; for Pashto, Ps-BERT (Haq et al., 2023). For all other Iranic languages, we found no monolingual pretrained transformer; the standard approach is multilingual encoders (XLM-R (Conneau et al., 2020)) or broader low-resource models such as Glot500 (Imani et al., 2023) and EMMA-500 (Ji et al., 2025). Frontier closed models perform reasonably on Standard Persian (Hosseini et al., 2025b; Ghahroodi et al., 2024) but we would expect substantial degradation for Dari and Tajik and near-chance performance for most other varieties, a hypothesis that current benchmarks cannot test.

Benchmarks and Evaluation. ParsiNLU

(Khashabi et al., 2021) established the first comprehensive Persian NLU benchmark (14,500+ instances, six tasks), with a critical methodological choice to author tasks natively. The Khayyam Challenge (Ghahroodi et al., 2024) evaluates subject-matter knowledge across 50+ academic domains. A growing set of benchmarks targets cultural alignment and pragmatic competence: ELAB (Pourbahman et al., 2025), MELAC (Farsi et al., 2025a), PerCul (Moosavi Monazzah et al., 2025), and FarsEval-PKBETS (Shamsfard et al., 2025); see Appendix for the full list. Belebele (Bandarkar et al., 2024) covers Persian, Tajik, Sorani, and Southern Pashto for reading comprehension. The APARSIN benchmark (Jafari et al., 2026) is a significant step toward multi-variety evaluation, spanning Persian dialects and Kalhori Kurdish, and is the first published benchmark for any Iranic language beyond the five most resourced varieties.

What is saturating vs. what is open. Standard Persian sentiment classification, Wikipedia-based reading comprehension, and NER are approaching saturation. The open frontiers for Persian are cultural and pragmatic competence, multi-hop reasoning, code-switching, and dialectal robustness. For all other Iranic languages, the open frontier is simply to build the first benchmark.

Question Answering. Persian QA has grown into a mature ecosystem, progressing from Wikipedia-based reading comprehension (PQuAD, Darvishi et al., 2023, 80K questions) to multi-hop reasoning (PersianMHQA, Taji et al., 2025, 7K questions) and domain-specific formats (Persian-MedQA, Kalahroodi et al., 2025b, 20K medical questions from residency exams). A catalogue of 25+ Persian QA datasets appears in Appendix Table 11. The central challenge is domain bias toward formal written Persian. Sorani Kurdish has a single news-domain dataset (Saeed, 2024). To our knowledge, no natively authored QA dataset exists for any other Iranic variety (Belebele (Bandarkar et al., 2024) covers four Iranic languages, but as translated items).

Sentiment Analysis and Affect. Sentiment analysis accounts for 116 papers (Figure 4), making it the most studied Iranic NLP task. Key resources include MirasOpinion (Ashrafi Asli et al., 2020) (938K reviews), SentiPers (Hosseini et al., 2021), ArmanEmo (Mirzaee et al., 2025), and EmoPars (Sabri et al., 2021); see Appendix Table 11 for the full list. The TaarofBench study (Sadr et al., 2025) demonstrates that LLMs score 40–48% be-

low native Persian speakers on taarof scenarios. APARSIN (Jafari et al., 2026) introduces the first multi-variety sentiment evaluation and should be treated as a template for expansion.

Speech Processing. Persian speech processing has a solid foundation: DeepMine (Zeinali et al., 2019) (480+ hours, ASR), ParsVoice (Kalahroodi et al., 2025a) (3,526 hours, TTS), ShEMO (Mohamad Nezami et al., 2019) (speech emotion), and PSRB (Sedghiyeh et al., 2025) (ASR benchmark). CommonVoice (Ardila et al., 2020) includes Persian and three other Iranian varieties. For Kurdish, JIRA (Veisi et al., 2022) and CORDI (Ahmadi et al., 2024) cover Sorani ASR and TTS. Massively multilingual speech encoders (MMS (Pratap et al., 2024), XEUS (Chen et al., 2024)) provide cross-lingual baselines, but typically evaluate only Persian. The DOLMA ASR initiative (Hameed et al., 2025) collected over 40 hours of speech for six further varieties—Gilaki, Laki, Hawrami, Mazanderani, Southern Kurdish, and Zazaki—though coverage is thin (Zazaki has 0.2 hours) and fine-tuned Whisper still reaches only 37.9–89.5% WER. For Balochi, Luri, Talysh, Shughni, Wakhi, and Yagh-nobi, we found no speech resources.

Vision-Language and OCR. Persian vision-language NLP is rapidly growing, covering VQA (Mobasher et al.; Farsi et al., 2025b), 3D visual reasoning (Vadaei et al., 2025), multimodal benchmarks (Ghahroodi et al., 2026), and image captioning (Thapliyal et al., 2022). Persian OCR spans printed text, handwriting, and scene text detection (see Appendix Table 11). For all other Iranian scripts, OCR resources are essentially absent. Glo-tOCR Bench (Kargaran et al., 2026) offers a starting point for measurement, evaluating OCR across 158 Unicode scripts; its findings are sobering, as most models perform well on fewer than ten scripts, and unfamiliar scripts induce hallucinated characters from visually similar systems—a failure mode that directly threatens Perso-Arabic varieties.

Traditional NLP Tasks. Persian NER is well studied (PEYMA (Shahshahani et al., 2019), ArmanPersoNERCorpus (Poostchi et al., 2018), PersoNER (Poostchi et al., 2016)). Text classification, summarization (Farahani et al., 2021b), and spell checking (Oji et al., 2021b) are also well covered. Script normalization for Kurmanji is addressed by Ahmadi and Anastasopoulos (2023); first steps toward pragmatic NLP include argumentation mining (Yeghaneh Abkenar and Stede, 2024) and Taarof-Bench (Sadr et al., 2025). For all other Iranian lan-

guages, none of these task categories have any annotated data.

6 Modeling Approaches and Evaluation

Transfer Learning and Cross-lingual Adaptation. Cross-lingual transfer is the primary strategy for low-resource Iranian varieties. Three types of proximity facilitate it: *lexical* (Persian, Dari, and Tajik share most vocabulary, enabling strong zero-shot transfer given correct script normalization); *script* (all Perso-Arabic-script varieties share tokenization conventions, though Unicode normalization diverges in ways that silently degrade transfer); and *grammatical* (varieties within the same branch share morphological patterns, as demonstrated for Kurmanji-to-Sorani MT (Sheykh Esmaili and Salavati, 2013)). Ahmadi et al. (2024) show that careful script normalization substantially improves transfer from larger Kurdish varieties to smaller ones. These efforts remain early; systematic intra-family transfer studies are lacking.

LLM Evaluation and Cultural Alignment. Benchmarking LLMs for Persian has emerged as its own subfield. The consistent finding across Khayyam (Ghahroodi et al., 2024), FarsEval-PKBETS (Shamsfard et al., 2025), and ELAB (Pourbahman et al., 2025) is that frontier models perform well on factual recall but show systematic gaps in cultural knowledge and pragmatic judgment. PerCul (Moosavi Monazzah et al., 2025) and MELAC (Farsi et al., 2025a) probe cultural alignment; Sakhaeirad et al. (2026) report a large gap between factual retrieval and cultural reasoning in Persian. MEXA (Kargaran et al., 2025) offers a language-agnostic cross-lingual alignment framework applicable to other Iranian varieties once evaluation parallel data becomes available.

7 Challenges, Discussion, and Future Directions

Resource concentration. Iranian Persian accounts for ~80% of surveyed papers, while Dari, Tajik, Kurdish, Pashto (51M speakers), Balochi (9M), Mazanderani (4M), and Gilaki (3M) remain severely underserved relative to their speaker populations. Broad multilingual efforts—NLLB-200 (NLLB Team et al., 2022), Glot500 (Imani et al., 2023), FLORES-200 (NLLB Team et al., 2022), MASSIVE (FitzGerald et al., 2023), FineWeb2 (Penedo et al., 2025), GlotCC (Kargaran et al., 2024a)—cover a handful of Iranian languages as

a byproduct of their broader mandate but are structurally insufficient: tokenizers are calibrated for high-resource languages, dialectal varieties are collapsed into standard forms, there is not enough data for these languages on the web, and a multilingual model covering Persian does not cover Dari unless Dari data was explicitly included. Iranic-specific efforts—PARME (Ahmadi et al., 2025b), SilkRoadNLP (Merchant and Megerdooian, 2026)—are beginning to fill this gap. Persian serves as the natural high-resource pivot; Kurdish occupies a second tier with sufficient resources for systematic model development; Pashto stands out as an outlier—51 million speakers but fewer resources than Kurdish—likely reflecting limited research infrastructure in Afghanistan and a smaller research community. The lesson from African NLP (Alabi et al., 2025) is that community-led shared tasks and centralized data hubs have an outsized catalytic effect.

Priority: Raw text. Publish Iranic-language text on the open web in identifiable form. Many varieties in Table 8 have no FineWeb2 presence at all, and web crawls are the entry point to every downstream corpus. Correct language tagging matters as much as volume: untagged or mistagged text is invisible to LID-filtered pipelines.

Priority: Translation benchmarks. Contribute Iranic varieties to FLORES+/OLDI.^a FLORES+ has grown from roughly 200 to 230 varieties under community stewardship, yet its Iranic coverage is unchanged from FLORES-200: the same six (pes, prs, tgtk, ckb, kmr, pbt). At roughly 2K sentences per variety and an open pull-request process, the bottleneck is contributor mobilization, not infrastructure.

Priority: Evaluation beyond translation. Natively authored NLU and QA benchmarks are absent for most Iranic varieties. At least one natively authored benchmark per variety is needed, and dedicated shared tasks are the most plausible route to producing them.

^ahttps://huggingface.co/datasets/openlanguagedata/flores_plus

Script heterogeneity and tokenization. Iranic languages use around a dozen scripts, most attested only historically; among living varieties, Perso-Arabic, Latin, and Cyrillic dominate, often in parallel for the same language. Within the Perso-Arabic writing scripts, Unicode normalization is incon-

sistent across Persian, Pashto, and Dari (differing *ya* and *kaf* code points), causing silent failures in cross-variety pipelines. Kurmanji appears in Latin, Arabic-based Sorani, and Cyrillic scripts depending on country context; Tajik uses Cyrillic exclusively. Zero-width non-joiners (ZWNJ) in Persian compound words are a persistent tokenization bug, and right-to-left rendering interacts poorly with mixed-script text. Standard subword tokenizers trained on Persian over-segment related varieties.

Priority: Develop a unified Iranic text normalization library; build multi-script representation models sharing parameters across writing systems for the same language.

Dialectal fragmentation. Persian, Kurdish, Balochi, and Luri are dialect continua. NLP systems trained on formal written Persian generalize poorly to colloquial register (Abdi Khojasteh et al., 2020), Dari, or regional varieties. Most benchmarks evaluate only standard forms, producing evaluation illusions. The language-dialect boundary is itself contested—Laki is variously treated as Kurdish, Luri, or an independent language—and these classifications determine what data gets labelled and what communities receive tools.

Priority: Extend benchmarks to colloquial and regional varieties; the APARSIN benchmark (Jafari et al., 2026) and HarfoSokhan (Sarvestani et al., 2026) are starting points.

Cultural misalignment and evaluation saturation. A disproportionate share of Persian benchmarks are translated from English: ParSQuAD from SQuAD. Cultural misalignment is documented (Moosavi Monazzah et al., 2025), and pragmatic phenomena such as *taarof* are absent from all existing frameworks (Sadr et al., 2025). At the same time, standard Persian sentiment classification, Wikipedia-based reading comprehension, and NER are approaching saturation. New evaluation paradigms are needed: culturally grounded commonsense reasoning, multi-hop reasoning over Persian knowledge graphs, pragmatic inference, and code-switching. For all other Iranic languages, the open frontier is simply to build the first benchmark.

Priority: Design benchmarks with native speaker communities as co-creators: QA from Persian, Kurdish, and Pashto literature; commonsense tasks grounded in regional cultural knowledge; pragmatic inference covering *taarof*, indirect refusal, and honorific register.

Endangered varieties and documentation.

The Pamir languages and Central Plateau Persian dialects face extinction within decades. Computational tools—automatic transcription, morphological analysis, glossing assistance—can support linguistic documentation even in very low-resource settings. Collaboration between NLP researchers, field linguists, and speaker communities is essential.

Priority: Apply massively multilingual speech models (MMS (Pratap et al., 2024), XEUS (Chen et al., 2024)) to Iranian languages; deploy community voice recording platforms with clear attribution and data governance.

Structural inequity. A substantial share of Iranian NLP research originates in Iran, where compute access, international API access, and conference participation are constrained by sanctions. A nationwide internet disruption prevented several Iranian research groups from submitting to SilkRoadNLP 2026 (Merchant and Megerdooian, 2026). Researchers on Kurdish, Pashto, Dari, and Pamir languages similarly lack institutional support and annotation budgets.

Priority: Advocate for compute access programs and open infrastructure analogous to the Lacuna Fund^a and the AI4D African Language Program (Siminyu et al., 2021) for African languages; address publication barriers for researchers in sanctioned or under-resourced contexts.

^a<https://lacunafund.org/>

8 Conclusion

This survey provides the first systematic landscape analysis of Iranian NLP. Our central finding is severe imbalance: despite 101 coded varieties and over 200 million speakers, 80% of documented research targets Iranian Persian. For most other varieties, the basic infrastructure—parallel corpora, morphological tools, evaluation benchmarks, speech data—is largely absent. At the same time, Persian NLP has matured to a point where the most impactful next steps are not more classification models but new evaluation frameworks probing cultural knowledge, pragmatic competence, and dialectal robustness.

Recent dedicated resources and the companion website released alongside this survey signal that the community is organizing. The tools exist; what is needed is the collective will to extend

NLP beyond Persian to the full range of Iranian languages, from Kurdish and Pashto to Zazaki, Gorani, Shughni, and Yaghnobi.

Limitations

Our survey focuses on papers from 2020 onward published in indexed NLP, speech, and AI venues, with a final literature search in May 2026, so 2026 is represented only partially. It likely underestimates the grey literature published in Persian-language journals and conference proceedings. English-only query terms may have missed research published primarily in Persian, Kurdish, Pashto, or Tajik. We did not conduct systematic quality assessment of cited datasets, and speaker count figures are approximations from multiple sources with different methodologies. The rapidly evolving LLM evaluation landscape means some model comparisons will be outdated quickly. Coverage of endangered and very-low-resource varieties is necessarily incomplete due to the near-absence of published research on many of them.

Ethical Considerations

This survey is observational and does not involve human subjects, new data collection, or model training. It synthesizes publicly available scholarly work and metadata about Iranian languages and their speakers. We nonetheless note several ethical dimensions.

Speaker counts and language status. Speaker estimates throughout this survey are drawn from heterogeneous web sources, including Wikipedia, Ethnologue, and the URL-NLP metadata dataset (van Esch et al., 2022), and should be read as approximations rather than authoritative figures. Reporting any single number for a contested variety can implicitly take a position in disputes over language identity and political recognition that we do not intend to adjudicate. Likewise, the language–dialect distinctions reflected in ISO 639-3 codes carry sociopolitical weight, and our use of these codes is descriptive rather than normative.

Endangered varieties. Several Iranian varieties discussed here are critically endangered. Computational documentation efforts on such languages must be conducted in collaboration with speaker communities, with explicit attention to consent, attribution, and data governance, and should prioritize community-defined goals rather than externally driven data extraction.

Structural inequities. Many researchers working on Iranic languages face constraints on compute, API access, and conference participation due to sanctions, internet restrictions, and limited institutional infrastructure. We acknowledge that the body of research surveyed here is partly shaped by these inequities and that any reading of "publication counts as effort" must take them into account.

Cited resources. Where datasets cited in this survey contain personal data, dialectal speech, or culturally sensitive content, responsibility for ethical use rests with the original creators' stated licenses and intended scopes; we have not independently audited every dataset.

Use of LLMs. We used large language models as writing aids for grammar correction and paraphrasing during the preparation of this manuscript. All technical claims, citations, and survey content were authored, verified, and edited by the human authors; LLMs were not used to generate analytical findings or to select papers for inclusion.

References

- Negin Abadani, Jamshid Mozafari, Afsaneh Fatemi, Mohammad Ali Nematbakhsh, and Arefeh Kazemi. 2021. [Parsquad: Machine translated squad dataset for persian question answering](#). In *2021 7th International Conference on Web Research (ICWR)*, pages 163–168. IEEE.
- 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.
- Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. [Persianllama: Towards building first persian large language model](#). *Preprint*, arXiv:2312.15713.
- Hadi Abdi Khojasteh, Ebrahim Ansari, and Mahdi Bohlouli. 2020. [LSCP: Enhanced large scale colloquial Persian language understanding](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6323–6327, Marseille, France. European Language Resources Association.
- Roshna Abdulrahman, Hossein Hassani, and Sina Ahmadi. 2019. [Developing a fine-grained corpus for a less-resourced language: the case of Kurdish](#). In *Proceedings of the 2019 Workshop on Widening NLP*, pages 106–109, Florence, Italy. Association for Computational Linguistics.
- David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. [SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.
- Majid Adibian, Hossein Zeinali, and Soroush Barmaki. 2025. [Deepmine-multi-tts: a persian speech corpus for multi-speaker text-to-speech](#). *Language Resources and Evaluation*, pages 1–20.
- Sina Ahmadi. 2019. [A rule-based kurdish text transliteration system](#). *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):1–8.
- Sina Ahmadi. 2020. [Building a corpus for the Zaza-gorani language family](#). In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 70–78, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Sina Ahmadi, Milind Agarwal, and Antonios Anastasopoulos. 2023a. [PALI: A language identification benchmark for Perso-Arabic scripts](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 78–90, Dubrovnik, Croatia. Association for Computational Linguistics.
- Sina Ahmadi and Antonios Anastasopoulos. 2023. [Script normalization for unconventional writing of under-resourced languages in bilingual communities](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14466–14487, Toronto, Canada. Association for Computational Linguistics.
- Sina Ahmadi, Zahra Azin, Sara Belelli, and Antonios Anastasopoulos. 2023b. [Approaches to corpus creation for low-resource language technology: the case of Southern Kurdish and Laki](#). In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 52–63, Dubrovnik, Croatia. Association for Computational Linguistics.
- Sina Ahmadi, Razhan Hameed, and Rico Sennrich. 2025a. [Literary translations and synthetic data for machine translation of low-resourced Middle Eastern languages](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 110–118, Vienna, Austria (in-person and online). Association for Computational Linguistics.
- Sina Ahmadi, Hossein Hassani, and Kamaladdin Abedi. 2020. [A corpus of the Sorani Kurdish folkloric lyrics](#).

- In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 330–335, Marseille, France. European Language Resources association.
- Sina Ahmadi, Hossein Hassani, and Daban Q. Jaff. 2022. [Leveraging multilingual news websites for building a kurdish parallel corpus](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).
- Sina Ahmadi, Daban Jaff, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2024. [Language and speech technology for Central Kurdish varieties](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10034–10045, Torino, Italia. ELRA and ICCL.
- Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahir Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, and 14 others. 2025b. [PARME: Parallel corpora for low-resourced Middle Eastern languages](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30032–30053, Vienna, Austria. Association for Computational Linguistics.
- Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasajo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. [One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.
- Masoud Akbari, Amir Hossein Karimi, Tayyeb Saeedi, Zeinab Saeidi, Kiana Ghezelbash, Fatemeh Shamsezat, Mohammad Akbari, and Ali Mohades. 2023. [A persian benchmark for joint intent detection and slot filling](#). *Preprint*, arXiv:2303.00408.
- Amir Mohammad Akhlaghi, Amirhossein Shabani, Mostafa Abdolmaleki, and Saeed Reza Kheradpisheh. 2025. [Persian-phi: Efficient cross-lingual adaptation of compact llms via curriculum learning](#). *Preprint*, arXiv:2512.07454.
- Jesujoba Oluwadara Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. [Charting the landscape of African NLP: Mapping progress and shaping the road ahead](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27807–27841, Suzhou, China. Association for Computational Linguistics.
- Alireza Alaei, P Nagabhushan, and Umapada Pal. 2011. A new dataset of persian handwritten documents and its segmentation. In *2011 7th Iranian conference on machine vision and image processing*, pages 1–5. IEEE.
- Zhila Amini, Mohammad Mohammadamini, Hawre Hosseini, Mehran Mansouri, and Daban Jaff. 2021. [Central kurdish machine translation: First large scale parallel corpus and experiments](#). *Preprint*, arXiv:2106.09325.
- Mostafa Amiri. 2024. Semi-alpaca instruction tuning dataset (persian). https://github.com/mostafaamiri/Persian_instruct_dataset.
- Ahmad Amirivojdan. 2025. [Shekar: A Python Toolkit for Persian Natural Language Processing](#). *Journal of Open Source Software*, 10(114):9128.
- Hossein Amirkhani, Mohammad AzariJafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, Zohreh Pourjafari, and Azadeh Amirak. 2026. FarsTail: A persian natural language inference dataset. *Soft Computing*, 30(2):891–903.
- Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitry Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. [TICO-19: the translation initiative for COvid-19](#). In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. [Common voice: A massively-multilingual speech corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2022. [Computational historical linguistics and language diversity in South Asia](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1396–1409, Dublin, Ireland. Association for Computational Linguistics.
- Fatemeh Zahra Arshia, Mohammad Ali Keyvanrad, Saeedeh Sadat Sadidpour, and Sayed Mohammad Reza Mohammadi. 2022. [Peqa: A massive persian question-answering and chatbot dataset](#). In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 392–397. IEEE.
- Fatemeh Asadi-zeydabadi, Ali Afkari-Fahandari, Amin Faraji, Elham Shabaninia, and Hossein Nezamabadi-pour. 2023. [Idpl-pfd2: A new large-scale dataset for printed farsi optical character recognition](#). *Preprint*, arXiv:2312.01177.

- Hiwa Asadpour. 2025. Central Kurdish-Mukri universal dependencies treebank. https://universaldependencies.org/treebanks/ckb_mukri/. Universal Dependencies v2.
- Seyed Arad Ashrafi Asli, Behnam Sabeti, Zahra Majdabadi, Preni Golazizian, Reza Fahmi, and Omid Momenzadeh. 2020. [Optimizing annotation effort using active learning strategies: A sentiment analysis case study in Persian](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2855–2861, Marseille, France. European Language Resources Association.
- Taha Shangipour Ataei, Kamyar Darvishi, Soroush Javdan, Behrouz Minaei-Bidgoli, and Sauleh Eetemadi. 2019. [Pars-absa: an aspect-based sentiment analysis dataset for persian](#). *Preprint*, arXiv:1908.01815.
- Mohammad Yasin Ayoubi, Sajjad & Davoodeh. 2021. Persianqa: a dataset for persian question answering. <https://github.com/SajjjadAyobi/PersianQA>.
- Hamed Babaei Giglou, Niloufar Beyranvand, Reza Moradi, Amir Mohammad Salehoof, and Saeed Bibak. 2022. [ParsSimpleQA: The Persian simple question answering dataset and system over knowledge graph](#). In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 59–68, Taipei, Taiwan. Association for Computational Linguistics.
- Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. [Xls-r: Self-supervised cross-lingual speech representation learning at scale](#). *Preprint*, arXiv:2111.09296.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabisa. 2024. [The belebele benchmark: a parallel reading comprehension dataset in 122 language variants](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.
- Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. [CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.
- BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Lucioni, François Yvon, and 1 others. 2023. [Bloom: A 176b-parameter open-access multilingual language model](#). *Preprint*, arXiv:2211.05100.
- Mahmood Bijankhan, J. Sheikhzadegan, and Mahmood R. Roohani. 1994. [Farsdat- the speech database of farsi spoken language](#).
- Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. [mhubert-147: A compact multilingual hubert model](#). *Preprint*, arXiv:2406.06371.
- Yasaman Boreshban, Hamed Yousefinasab, and Seyed Abolghasem and Mirroshandel. 2018. [Providing a religious corpus of question answering system in persian](#). *Signal and Data Processing*, 15(1).
- Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Djibrila Diane, Solo Farabado Cissé, Koulako Moussa Doumbouya, Edoardo Ferrante, Alessandro Guasoni, Christopher Homan, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, and 5 others. 2025. [SMOL: Professionally translated parallel data for 115 under-represented languages](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123, Suzhou, China. Association for Computational Linguistics.
- Tyler A. Chang, Catherine Arnett, Abdelrahman Eldeskey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanahun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, Aditi Gupta, Aftab Iyigun, Afonso Simplicio, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akshay Ramesh, Aleksei Dorkin, and 319 others. 2025. [Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures](#). *Preprint*, arXiv:2510.24081.
- Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2026. [Goldfish: Monolingual language models for 350 languages](#). *Preprint*, arXiv:2408.10441.
- William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. [Towards robust speech representation learning for thousands of languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10224, Miami, Florida, USA. Association for Computational Linguistics.
- Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. [Rethinking embedding coupling in pre-trained language models](#). In *International Conference on Learning Representations*.
- Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, Daan Van Esch, Vera Axelrod, Simran Khanuja, Jonathan H. Clark, Orhan Firat, Michael Auli, Sebastian Ruder, Jason

- Riesa, and Melvin Johnson. 2022. [Xtreme-s: Evaluating cross-lingual speech representations](#). *Preprint*, arXiv:2203.10752.
- 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.
- 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.
- Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. 2023. [PQuAD: A persian question answering dataset](#). *Computer Speech & Language*, 80:101486.
- Jon Dehdari and Deryle Lonsdale. 2008. [A link grammar parser for Persian](#). In Simin Karimi, Vida Samian, and Don Stilo, editors, *Aspects of Iranian Linguistics*, volume 1. Cambridge Scholars Press.
- Mohammad Dehghani and Heshaam Faili. 2024. [Persian typographical error type detection using deep neural networks on algorithmically-generated misspellings](#). *Preprint*, arXiv:2305.11731.
- 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, Minneapolis, Minnesota. Association for Computational Linguistics.
- David M. Eberhard, Gary F. Simons, and Alison J. Robinson. 2026. [Ethnologue: Languages of the World](#), 29 edition. SIL Global, Dallas, Texas.
- Nariman Esmailyfard. 2017. [ParsiAnalyzer: A Persian analysis plugin for Elasticsearch with tokenization, normalization, stop-word removal, and stemming](#). <https://github.com/NarimanN2/ParsiAnalyzer>.
- Romina Etezadi, Mohammad Karrabi, Najmeh Zare, Mohamad Bagher Sajadi, and Mohammad Taher Pilehvar. 2022. [DadmaTools: Natural language processing toolkit for Persian language](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 124–130, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.
- Romina Etezadi and Mehrnoush Shamsfard. 2020. [PeCoQ: A dataset for persian complex question answering over knowledge graph](#). In *2020 11th International Conference on Information and Knowledge Technology (IKT)*, pages 102–106.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. [Beyond english-centric multilingual machine translation](#). *Journal of Machine Learning Research*, 22(107):1–48.
- Mehrdad Farahani. 2020. [Summarization using bert2bert model on wikisummary dataset](#). <https://github.com/m3hrdadfi/wiki-summary>.
- Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021a. [ParsBERT: Transformer-based model for persian language understanding](#). *Neural Processing Letters*, 53(6):3831–3847.
- Mehrdad Farahani, Mohammad Gharachorloo, and Mohammad Manthouri. 2021b. [Leveraging parsbert and pretrained mt5 for persian abstractive text summarization](#). In *2021 26th International computer conference, computer society of Iran (CSICC)*, pages 1–6. IEEE.
- Farhan Farsi, Farnaz Aghababaloo, Shahriar Shariati Motlagh, Parsa Ghoofrani, MohammadAli Sadraei-Javaheri, Shayan Bali, Amirhossein 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](#). *Preprint*, arXiv:2508.00673.
- Farhan Farsi, Sadra Sabouri, Kian Kashfipour, Soroush Gooran, Hossein Sameti, and Ehsaneddin Asgari. 2024. [Syntran-fa: Generating comprehensive answers for farsi qa pairs via syntactic transformation](#).
- Farhan Farsi, Shahriar Shariati Motlagh, Shayan Bali, Sadra Sabouri, and Saeedeh Momtazi. 2025b. [Persian in a court: Benchmarking VLMs in Persian multi-modal tasks](#). In *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pages 52–56, Abu Dhabi, UAE. Association for Computational Linguistics.
- Christian Federmann, Tom Kocmi, and Ying Xin. 2022. [NTREX-128 – news test references for MT evaluation of 128 languages](#). In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.
- Weston Feely, Mehdi Manshadi, Robert E Frederking, and Lori S Levin. 2014. [The cmu metal farsi nlp approach](#). In *LREC*, pages 4052–4055.

- Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natara-jan. 2023. [MASSIVE: A 1M-example multilin-gual natural language understanding dataset with 51 typologically-diverse languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.
- Javad ForutanRad, Maryam HourAli, and Moham-madAli KeyvanRad. 2024. [Farsi question and an-swer dataset \(farsiquad\)](#). *Signal and Data Process-ing*, 20(4):107–120.
- Arash Ghafouri, Mohammad Amin Abbasi, and Hassan Naderi. 2023. [Ariabert: A pre-trained persian bert model for natural language understanding](#). *Research Square*.
- Arash Ghafouri, Mohammad Aghajani Asl, Hasan Naderi, and Mahdi Firouzmandi. 2025. [Islamicpcqa: A dataset for persian multi-hop complex question answering in islamic text resources](#). *IEEE Transac-tions on Audio, Speech and Language Processing*, 33:3801–3812.
- 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. 2026. [MEENA \(PersianMMU\): Multimodal-multilingual educational exams for n-level assessment](#). In *Find-ings of the Association for Computational Linguis-tics: EACL 2026*, pages 6457–6491, Rabat, Morocco. Association for Computational Linguistics.
- 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?](#) In *First Conference on Language Modeling*.
- Zahra Ghasemi and Mohammad Ali Keyvanrad. 2021. [Farsick: A persian semantic textual similarity and natural language inference dataset](#). In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 194–199. IEEE.
- Mehrdad Ghassabi, Pedram Rostami, Hamidreza Baradaran Kashani, Amirhossein Poursina, Zahra Kazemi, and Milad Tavakoli. 2025. [Leveraging online data to enhance medical knowledge in a small persian language model](#). *Preprint*, arXiv:2505.16000.
- Memduh Gökırmak and Francis M. Tyers. 2017. [A dependency treebank for Kurmanji Kurdish](#). In *Pro-ceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 64–72, Pisa, Italy. Linköping University Electronic Press.
- Preni Golazizian, Behnam Sabeti, Seyed Arad Ashrafi Asli, Zahra Majdabadi, Omid Momenzadeh, and Reza Fahmi. 2020. [Irony detection in Persian language: A transfer learning approach using emoji prediction](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2839–2845, Marseille, France. European Language Re-sources Association.
- 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](#). *Preprint*, arXiv:2407.21783.
- Puntis Jifroodian Haghighi, Nicola Nobile, Chun Lei He, and Ching Y. Suen. 2009. A new large-scale multi-purpose handwritten farsi database. In *Image Analysis and Recognition*, pages 278–286, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Muhammadreza Haghiri and Mahi Mohrechi. 2023. [Maral-7b-alpha-1](#).
- Razhan Hameed, Sina Ahmadi, Hanah Hadi, and Rico Sennrich. 2025. Automatic speech recognition for low-resourced middle eastern languages. In *Inter-speech*.
- Saied Alimoradi HamidReza Attar, Milad Lotfi. 2022. [Khoshnevis, a python library for persian text preprocess-ing](#).
- Ijazul Haq, Weidong Qiu, Jie Guo, and Peng Tang. 2023. [Pashto offensive language detection: a benchmark dataset and monolingual pashto bert](#). *PeerJ Com-puter Science*, 9:e1617.
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Is-lam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. [XL-sum: Large-scale multilingual abstractive summariza-tion for 44 languages](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computa-tional Linguistics.
- Mohammad Hasannejad and Pedram Porbaha. 2023. [A dataset of persian medical questions and answers](#).
- Mohammad Hasannejad and Pedram Porbaha. 2024. [A specialized dataset of persian pharmaceutical ques-tions and expert responses](#).
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embed-ding sharing](#). In *The Eleventh International Confer-ence on Learning Representations*.
- Michael A. Hedderich, Lukas Lange, Heike Adel, Jan-nik Strötgen, and Dietrich Klakow. 2021. [A survey](#)

- on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.
- Hamed Hematian Hemati, Atousa Toghyani, Atena Souri, Sayed Hesam Alavian, Hossein Sameti, and Hamid Beigy. 2023. [Pcoqa: Persian conversational question answering dataset](#). *Preprint*, arXiv:2312.04362.
- Sara Bourbour Hosseinbeigi, Sina Asghari, Mohammad Ali Seif Kashani, Mohammad Hossein Shalchian, and Mohammad Amin Abbasi. 2025a. [Advancing retrieval-augmented generation for persian: Development of language models, comprehensive benchmarks, and best practices for optimization](#). *Preprint*, arXiv:2501.04858.
- Sara Bourbour Hosseinbeigi, Behnam Rohani, Mostafa Masoudi, Mehrnoush Shamsfard, Zahra Saaberi, Mostafa Karimi Manesh, and Mohammad Amin Abbasi. 2025b. [Advancing Persian LLM evaluation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2711–2727, Albuquerque, New Mexico. Association for Computational Linguistics.
- Sara Bourbour Hosseinbeigi, Fatemeh Taherinezhad, Hesham Faili, Hamed Baghbani, Fatemeh Nadi, and Mostafa Amiri. 2025c. [Matina: A large-scale 73B token Persian text corpus](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9143–9157, Albuquerque, New Mexico. Association for Computational Linguistics.
- Pedram Hosseini, Poorya Hosseini, and David Broniatowski. 2020. [Content analysis of Persian/Farsi tweets during COVID-19 pandemic in Iran using NLP](#). In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.
- Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. 2021. [Sentipers: A sentiment analysis corpus for persian](#). *Preprint*, arXiv:1801.07737.
- Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. [Glot500: Scaling multilingual corpora and language models to 500 languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.
- Sadegh Jafari, Tara Azin, Farhad Roodi, Zahra Dehghani Tafti, Mehrdad Ghadrddan, Elham Vatankhahan Esfahani, Aylin Naebzadeh, Mohammadhadi Shahhosseini, Ghafoor Khan, Kazem Forghani, Dania Namazi, Seyed Mohammad Hossein Hashemi, Farhan Farsi, Mohammad Osoolian, Maede Mohammadi, Mohammad Erfan Zare, Muhammad Hasnain Khan, Muhammad Hussain, Nooreen Zaki, and 5 others. 2026. [APARSIN: A multi-variety sentiment and translation benchmark for iranic languages](#). In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*, pages 83–97, Rabat, Morocco. Association for Computational Linguistics.
- Pourya Jafarzadeh, Padideh Choobdar, and Vahid Mohammadi Safarzadeh. 2024. [Khayyam offline persian handwriting dataset](#). *Preprint*, arXiv:2406.01025.
- Naghme Jamali, Yadollah Yaghoobzadeh, and Hesham Faili. 2022. [PerCQA: Persian community question answering dataset](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6083–6092, Marseille, France. European Language Resources Association.
- Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. 2022. [Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data](#). *Preprint*, arXiv:2212.10440.
- Shaoxiong Ji, Zihao Li, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayán O’Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2025. [Emma-500: Enhancing massively multilingual adaptation of large language models](#). *Preprint*, arXiv:2409.17892.
- Safora Jolfaei, Azadeh Mohebi, and Mahla Hemmat. 2025. [PersianSciQA: A new dataset for bridging the language gap in scientific question answering](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2025)*, pages 31–40.
- Mohammad Javad Ranjbar Kalahroodi, Hesham Faili, and Azadeh Shakery. 2025a. [Parsvoice: A large-scale multi-speaker persian speech corpus for text-to-speech synthesis](#). *Preprint*, arXiv:2510.10774.
- Mohammad Javad Ranjbar Kalahroodi, Amirhossein Sheikholeslami, Sepehr Karimi, Sepideh Ranjbar Kalahroodi, Hesham Faili, and Azadeh Shakery. 2025b. [Persianmedqa: Evaluating large language models on a persian-english bilingual medical question answering benchmark](#). *Preprint*, arXiv:2506.00250.
- Zina Kamal and Hossein Hassani. 2020. [Towards Kurdish text to sign translation](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 117–122, Marseille, France. European Language Resources Association (ELRA).

- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. [GlotLID: Language identification for low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.
- Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schuetze. 2025. [MEXA: Multilingual evaluation of English-centric LLMs via cross-lingual alignment](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 27001–27023, Vienna, Austria. Association for Computational Linguistics.
- Amir Hossein Kargaran, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2026. [Glotocr bench: Ocr models still struggle beyond a handful of unicode scripts](#). *Preprint*, arXiv:2604.12978.
- Amir Hossein Kargaran, François Yvon, and Hinrich Schuetze. 2024a. [GlotCC: An open broad-coverage commoncrawl corpus and pipeline for minority languages](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Amir Hossein Kargaran, François Yvon, and Hinrich Schuetze. 2024b. [MaskLID: Code-switching language identification through iterative masking](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 459–469, Bangkok, Thailand. Association for Computational Linguistics.
- Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024c. [GlotScript: A resource and tool for low resource writing system identification](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7774–7784, Torino, Italia. ELRA and ICCL.
- Arefeh Kazemi, Jamshid Mozafari, and Mohammad Ali Nematbakhsh. 2022. Persianquad: the native question answering dataset for the persian language. *IEEE Access*, 10:26045–26057.
- Arefeh Kazemi, Zahra Zojaji, Mahdi Malverdi, Jamshid Mozafari, Fatemeh Ebrahimi, Negin Abadani, Mohammad Reza Varasteh, and Mohammad Ali Nematbakhsh. 2023. [Farsnewsqa: A deep learning-based question answering system for the persian news articles](#). *Inf. Retr.*, 26(1–2).
- Reza Khanmohammadi, Keivan Mirhoseini, Saeed Sayad, Pouya Kazemian, and Seyedabolghasem Mirroshandel. 2021a. [Parscap: The persian variant of the flickr8k dataset for image captioning](#).
- Reza Khanmohammadi, Mitra Sadat Mirshafiee, and Mehdi Allahyari. 2021b. [COPER: a query-adaptable semantics-based search engine for persian covid-19 articles](#). In *2021 7th International Conference on Web Research (ICWR)*, pages 64–70.
- Hamzeh Motahari Khansari and Mehrnoush Shamsfard. 2021. [Hmblogs: A big general persian corpus](#). *Preprint*, arXiv:2111.02362.
- Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozdeh 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.
- Hossein Khosravi and Ehsanollah Kabir. 2007. Introducing a very large dataset of handwritten farsi digits and a study on their varieties. *Pattern recognition letters*, 28(10):1133–1141.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. [Madlad-400: A multilingual and document-level large audited dataset](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc.
- Hynek Kydlíček, Guilherme Penedo, and Leandro von Werra. 2025. *Finepdfs*. <https://huggingface.co/datasets/HuggingFaceFW/finepdfs>.
- Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2025. [Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions](#). *Transactions of the Association for Computational Linguistics*, 13:1032–1055.
- language-ml-lab. 2024. [Kurdbert](#).
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. [Few-shot learning with multilingual generative language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. [SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages](#). In *Proceedings of the 2024 Conference on*

- Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.
- Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. [LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.
- Chunlan Ma, Ayyoob Imani, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schuetze. 2025. [Taxi1500: A dataset for multilingual text classification in 1500 languages](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 414–439, Albuquerque, New Mexico. Association for Computational Linguistics.
- Mann-E. 2024. [Hormoz-8B](#). Hugging Face.
- Mostafa Masumi, Seyed Soroush Majd, Mehrnosh Shamsfard, and Hamid Beigy. 2025. [FaBERT: Pre-training BERT on Persian blogs](#). In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 85–96, Albuquerque, New Mexico, USA. Association for Computational Linguistics.
- Mohammadali Memari, Soghra Mikaeyl Nejad, Amir Parsa Rabiei, Mehrshad Eisaei, and Saba Hesarakhi. 2024. [Bertcaps: Bert capsule for persian multi-domain sentiment analysis](#). *Preprint*, arXiv:2412.05591.
- Rayyan Merchant and Karine Megerdumian, editors. 2026. [The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family](#). Association for Computational Linguistics, Rabat, Morocco.
- Baqer M. Merzah, Tania Taami, Salman Asoudeh, Saeed Mirzaee, Amir reza Hossein pour, and Amir Ali Bengari. 2025. [Biopars: A pretrained biomedical large language model for persian biomedical text mining](#). *Preprint*, arXiv:2506.21567.
- Mohammad Reza Mirbagheri, Mohammad Mahdi Mirkamali, Zahra Motoshaker Arani, Ali Javeri, Amir Mahdi Sadeghzadeh, and Rasool Jalili. 2025. [Ept benchmark: Evaluation of persian trustworthiness in large language models](#). *Preprint*, arXiv:2509.06838.
- Sajad Mirzababaei, Amir Hossein Kargar, Hinrich Schütze, and Ehsaneddin Asgari. 2022. [Hengam: An adversarially trained transformer for Persian temporal tagging](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1013–1024, Online only. Association for Computational Linguistics.
- Hossein Mirzaee, Javad Peymanfard, Hamid Habibzadeh Moshtaghin, and Hossein Zeinali. 2025. [Armanemo: a persian dataset for text-based emotion detection: H. mirzaee et al.](#) *Language Resources and Evaluation*, 59(3):2565–2587.
- Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian, and Sauleh Eetemadi. [Parsvqa-caps: A benchmark for visual question answering and image captioning in persian](#).
- Omid Mohammad Nezami, Paria Jamshid Lou, and Mansoureh Karami. 2019. [Shemo: a large-scale validated database for persian speech emotion detection](#). *Language Resources and Evaluation*, 53(1):1–16.
- Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. [SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Salar Mohtaj, Behnam Roshanfekar, 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).
- Hojjat Mokhtarabadi, Ziba Zamani, Abbas Maazallahi, and Mohammad Hossein Manshaei. 2025. [Empowering Persian LLMs for instruction following: A novel dataset and training approach](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 31–67, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Erfan Moosavi Monazzah, Vahid Rahimzadeh, Yadollah Yaghoobzadeh, Azadeh Shakery, and Mohammad Taher Pilehvar. 2025. [PerCul: A story-driven cultural evaluation of LLMs in Persian](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12670–12687, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jamshid Mozafari, Arefeh Kazemi, Parham Moradi, and Mohammad Ali Nematbakhsh. 2022. [Peransel: A novel deep neural network-based system for persian question answering](#). *Computational Intelligence and Neuroscience*, 2022(1):3661286.
- 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.
- Kozhin muhealddin Awlla, Hadi Veisi, and Abdulhady Abas Abdullah. 2025. [Kubert: Central kurdish bert model and its application for sentiment analysis](#). *Preprint*, arXiv:2509.16804.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024a. [CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.
- Thuat Nguyen, Huu Nguyen, and Thien Nguyen. 2024b. [Culturay: A large cleaned multilingual dataset of 75 languages](#).
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, 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.
- Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2026. [Hplt 3.0: Very large-scale multilingual resources for llms and mt. mono- and bi-lingual data, multilingual evaluation, and pre-trained models](#). *Preprint*, arXiv:2511.01066.
- Romina Oji, Seyedeh Fatemeh Razavi, Sajjad Abdi Dehsorkh, Alireza Hariri, Hadi Asheri, and Reshad Hosseini. 2021a. [Parsinorm: A persian toolkit for speech processing normalization](#). In *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–5. IEEE.
- Romina Oji, Nasrin Taghizadeh, and Hesham Faili. 2021b. [Perspelldata: An exhaustive parallel spell dataset for Persian](#). In *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*, pages 8–14, Trento, Italy. Association for Computational Linguistics.
- Amir Panahandeh, Hanie Asemi, and Esmaeil Nourani. 2023. [Tppoet: Transformer-based persian poem generation using minimal data and advanced decoding techniques](#). *Preprint*, arXiv:2312.02125.
- Mahnaz Panahandeh Nigjeh and Shirin Ghanbari. 2024. [Leveraging parsbert for cross-domain polarity sentiment classification of persian social media comments](#). *Multimedia Tools and Applications*, 83(4):10677–10694.
- PartAI. 2024a. [Dorna](#). Hugging Face.
- PartAI. 2024b. [Dorna2](#). Hugging Face.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language](#). In *Second Conference on Language Modeling*.
- Javad Peymanfard, Samin Heydarian, Ali Lashini, Hossein Zeinali, Mohammad Reza Mohammadi, and Nasser Mozayani. 2024. [A multi-purpose audio-visual corpus for multi-modal persian speech recognition: The arman-av dataset](#). *Expert Systems with Applications*, 238:121648.
- Hanieh Poostchi, Ehsan Zare Borzeshi, Mohammad Abdous, and Massimo Piccardi. 2016. [PersoNER: Persian named-entity recognition](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3381–3389, Osaka, Japan. The COLING 2016 Organizing Committee.
- Hanieh Poostchi, Ehsan Zare Borzeshi, and Massimo Piccardi. 2018. [BiLSTM-CRF for Persian named-entity recognition ArmanPersoNERCorpus: the first entity-annotated Persian dataset](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Sampoorna Poria and Xiaolei Huang. 2025. [Bhaasha, bhāṣā, zaban: A survey for low-resourced languages in South Asia – current stage and challenges](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1386–1406, Suzhou, China. Association for Computational Linguistics.
- Zahra Pourbahman, Fatemeh Rajabi, Mohammadhossein Sadeghi, Omid Ghahroodi, Somayeh Bakhshaei, Arash Amini, Reza Kazemi, and Mahdieh Soleymani Baghshah. 2025. [ELAB: Extensive LLM alignment benchmark in Persian language](#). In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 458–470, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. [Scaling speech technology to 1,000+ languages](#). *Journal of Machine Learning Research*, 25(97):1–52.

- Targoman Intelligent Processing. 2024. [Tlpc \(revision ec03c70\)](#).
- Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R. Rabiee. 2025. [ManaTTS Persian: a recipe for creating TTS datasets for lower resource languages](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9177–9206, Albuquerque, New Mexico. Association for Computational Linguistics.
- Mahdi Rahmani, AmirHossein Saffari, and Reyhane Rahmani. 2025. [Megachat: A synthetic persian q&a dataset for high-quality sales chatbot evaluation](#). *Preprint*, arXiv:2511.23397.
- Z Raisi, VM Nazarzehi Had, E Sarani, and R Damani. 2025a. [Fatr: A comprehensive dataset and evaluation framework for persian text recognition in wild images](#). *Journal of Electrical and Computer Engineering Innovations (JECEI)*, pages 331–340.
- Zobeir Raisi, Valimohammad Nazarzehi Had, Esmail Sarani, and Raosul Damani. 2025b. [Ptldr: A real-world and synthetic benchmark dataset for persian scene and document text detection and recognition](#). *SN Computer Science*, 6(6):648.
- Surangika Ranathunga, En-Shiun Annie Lee, Marjana Priiti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. [Neural machine translation for low-resource languages: A survey](#). *ACM Comput. Surv.*, 55(11).
- Atefeh Ranjkesh Rashtehroudi, Alireza Akoushideh, and Asadollah Shahbahrani. 2023. [Pestd: a large-scale persian-english scene text dataset](#). *Multimedia Tools and Applications*, pages 1–16.
- Mohammad Sadegh Rasooli, Pegah Safari, Amirsaeid Moloodi, and Alireza Nourian. 2020. [The persian dependency treebank made universal](#). *Preprint*, arXiv:2009.10205.
- Mahdi Rasouli and Vahid Kiani. 2023. Investigating shallow and deep learning techniques for emotion classification in short persian texts. *Journal of AI and Data Mining*, 11(4):587–598.
- Roshan Research. 2026. [Hazm: Persian NLP toolkit](#). <https://github.com/roshan-research/hazm>.
- Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti. 2024. [Persianmind: A cross-lingual persian-english large language model](#). *Preprint*, arXiv:2401.06466.
- Behnam Sabeti, Hossein Abedi Firouzjaee, Ali Janalizadeh Choobbasti, S.H.E. Mortazavi Najafabadi, and Amir Vaheb. 2018. [MirasText: An automatically generated text corpus for Persian](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Sadra Sabouri, Elnaz Rahmati, Soroush Gooran, and Hossein Sameti. 2024. [naab: A ready-to-use plug-and-play corpus for farsi](#). *Journal of Artificial Intelligence Applications and Innovations*, 1(2).
- Nazanin Sabri, Reyhane Akhavan, and Behnam Bahrak. 2021. [EmoPars: A collection of 30K emotion-annotated Persian social media texts](#). In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 167–173, Online. INCOMA Ltd.
- Fatemeh sadat Hosseini, Shima Kashef, Elham Shabaninia, and Hossein Nezamabadi-pour. 2021. [Idpl-pfod: an image dataset of printed farsi text for ocr research](#). In *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLP 2021*, pages 22–31.
- Reyhaneh Sadeghi, Hamed Karbasi, and Ahmad Akbari. 2022. [Exappc: a large-scale persian paraphrase detection corpus](#). In *2022 8th International Conference on Web Research (ICWR)*, pages 168–175. IEEE.
- Seyed Mojtaba Sadjadi, Zeinab Rajabi, Leila Rabiei, and Mohammad-Shahram Moin. 2024. [Farssibert: A novel transformer-based model for semantic similarity measurement of persian social networks informal texts](#). *Preprint*, arXiv:2407.19173.
- Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoo-mian, 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, Suzhou, China. Association for Computational Linguistics.
- MohammadAli SadraeiJavaheri, Ali Moghaddaszadeh, Milad Molazadeh, Fariba Naeiji, Farnaz Aghababalo, Hamideh Rafiee, Zahra Amirmahani, Tohid Abedini, Fatemeh Zahra Sheikhi, and Amirmohammad Salehoof. 2024. [Tookabert: A step forward for persian nlu](#). *Preprint*, arXiv:2407.16382.
- Javad Sadri, Mohammad Reza Yeganehzad, and Javad Saghi. 2016. [A novel comprehensive database for off-line persian handwriting recognition](#). *Pattern Recognition*, 60:378–393.
- Ari M Saeed. 2024. Building a benchmark dataset for the kurdish news question answering. *Data in Brief*, 57:110916.
- Alireza Sakhaeirad, Ali Ma'manpoosh, and Arshia Hemmat. 2026. [Unmasking the factual-conceptual gap in Persian language models](#). In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*, pages 1–12, Rabat, Morocco. Association for Computational Linguistics.
- Mehran Sarmadi, Morteza Alikhani, Erfan Zinvandi, and Zahra Pourbahman. 2025. [Hakim: Farsi text embedding model](#). *Preprint*, arXiv:2505.08435.

- Hamid Jahad Sarvestani, Vida Ramezani, Saeed Saadat, Neda Taghizadeh Serajeh, Maryam Sadat Razavi Taheri, Shohreh Kasaei, MohammadAmin Fazli, and Ehsaneddin Asgari. 2026. [HarfoSokhan: A comprehensive parallel dataset for transitions between Persian colloquial and formal variations](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7380–7392, Rabat, Morocco. Association for Computational Linguistics.
- Alireza Sedghi. 2022. Virastar: A Persian text cleaner and normalizer for PHP. <https://github.com/AlirezaSedghi/Virastar>.
- 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](#). *Preprint*, arXiv:2505.21230.
- Mina Serajian, Saeed Najafzadeh Rahaghi, Hadi Veisi, and Saman Haratizadeh. 2025. Favc: A validated, transcribed, parallel farsi speech dataset for voice conversion. In *Proc. Interspeech 2025*, pages 4788–4792.
- Elham Shabaninia, Fatemeh sadat Eslami, Ali Afkari-Fahandari, and Hossein Nezamabadi-pour. 2023. [Sut: a new multi-purpose synthetic dataset for farsi document image analysis](#). In *2023 13th International Conference on Computer and Knowledge Engineering (ICCCKE)*, pages 253–258.
- Mahsa Sadat Shahshahani, Mahdi Mohseni, Azadeh Shakery, and Hesham Faily. 2019. Payma: A tagged corpus of persian named entities. *Signal and Data Processing*, 16(1):91–110.
- Mohammad Hasan Shamgholi, Vahid Saeedi, Javad Peymanfard, Leila Alhabib, and Hossein Zeinali. 2023. [Armannts single-speaker persian dataset](#). *Preprint*, arXiv:2304.03585.
- Mehrnoush Shamsfard, Zahra Saaberi, Mostafa Karimi manesh, Seyed Mohammad Hossein Hashemi, Zahra Vatankehah, Motahareh Ramezani, Niki Pourazin, Tara Zare, Maryam Azimi, Sarina Chit-saz, Sama Khoraminejad, Morteza Mahdavi Mortazavi, Mohammad Mahdi Chizari, Sahar Maleki, Seyed Soroush Majd, Mostafa Masumi, Sayed Ali Musavi Khoeini, Amir Mohseni, and Sogol Alipour. 2025. [Farseval-pkbets: A new diverse benchmark for evaluating persian large language models](#). *Preprint*, arXiv:2504.14690.
- Javad PourMostafa Roshan Sharami, Parsa Abbasi Sarabestani, and Seyed Abolghasem Mirroshandel. 2020. [Deepsentipers: Novel deep learning models trained over proposed augmented persian sentiment corpus](#). *Preprint*, arXiv:2004.05328.
- Kyumars Sheykh Esmaili and Shahin Salavati. 2013. [Sorani Kurdish versus Kurmanji Kurdish: An empirical comparison](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 300–305, Sofia, Bulgaria. Association for Computational Linguistics.
- Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. [mGPT: Few-shot learners go multilingual](#). *Transactions of the Association for Computational Linguistics*, 12:58–79.
- Kathleen Siminyu, Godson Kalipe, Davor Orlic, Jade Abbott, Vukosi Marivate, Sackey Freshia, Prateek Sibai, Bhanu Neupane, David I. Adelani, Amelia Taylor, Jamiil Toure ALI, Kevin Degila, Mombo-ladi Balogoun, Thierno Ibrahima DIOP, Davis David, Chayma Fourati, Hatem Haddad, and Malek Naski. 2021. [Ai4d – african language program](#). *Preprint*, arXiv:2104.02516.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2026. [OpenAI GPT-5 system card](#). *Preprint*, arXiv:2601.03267.
- Shivalika Singh, Freddie Vargus, Daniel D’souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O’Mahony, Mike Zhang, Ramith Het-tiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. [Aya dataset: An open-access collection for multilingual instruction tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.
- Snapp Incubator. 2024. Seperno: A customizable Persian text normalizer for search applications. <https://github.com/snapp-incubator/seperno>.
- Iman Souzanchi, Ramin Rahimi, Mohammad Ali Majidi Anvari, Atefeh Baniasadi, Ashkan Sadeghi, and Mohammad Reza Mohammadi. 2024. [A comprehensive dataset of real-scene images for text detection and recognition in persian](#). In *2024 14th International Conference on Computer and Knowledge Engineering (ICCCKE)*, pages 001–007.
- Mobina Taji, Arash Ghafouri, Hasan Naderi, and Behrouz Minaei-Bidgoli. 2025. [Persianmhqa: A dataset for open domain persian multi-hop question answering based on wikipedia encyclopedia](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(2).
- Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. [Crossmodal-3600: A massively multilingual multimodal evaluation dataset](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Martin Thoma. 2018. [The wili benchmark dataset for written language identification](#). *Preprint*, arXiv:1801.07779.
- Thomcles. 2025. [Persian-farsi-dataset](#). A cleaned and concatenated dataset for Persian text-to-speech tasks.
- David Uthus, Santiago Ontañón, Joshua Ainslie, and Mandy Guo. 2023. [mLongT5: A multilingual and efficient text-to-text transformer for longer sequences](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9380–9386, Singapore. Association for Computational Linguistics.
- Kianoosh Vadaei, Melika Shirian, Arshia Hemmat, Mohammad Hassan Heydari, Ali Mamanpoosh, and Afshaneh Fatemi. 2025. [From scenes to semantics: PersianCLEVR for bilingual 3d visual reasoning](#). In *1st Workshop on VLM4RWD @ NeurIPS 2025*.
- Amir Vaheb, Ali Janalizadeh Choobbasti, S.H.E. Mor-tazavi Najafabadi, Saeid Safavi, and Behnam Sabeti. 2018. [MirasVoice: A bilingual \(English-Persian\) speech corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Jörgen Valk and Tanel Alumäe. 2020. [Voxlingua107: a dataset for spoken language recognition](#). *Preprint*, arXiv:2011.12998.
- Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. [Writing system and speaker metadata for 2,800+ language varieties](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5035–5046, Marseille, France. European Language Resources Association.
- Hadi Veisi, Hawre Hosseini, Mohammad MohammadAmini, Wiryaa Fathy, and Aso Mahmudi. 2022. [Jira: a central kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon](#). *Language Resources and Evaluation*, 56(3):917–941.
- Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2019. [Toward kurdish language processing: Experiments in collecting and processing the asosoft text corpus](#). *Digital Scholarship in the Humanities*, 35(1):176–193.
- Hadi Veisi and Hamed Fakour Shandi. 2020. [A persian medical question answering system](#). *International Journal on Artificial Intelligence Tools*, 29(06):2050019.
- Yana Veitsman and Mareike Hartmann. 2025. [Recent advancements and challenges of Turkic Central Asian language processing](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 309–324, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. [ByT5: Towards a token-free future with pre-trained byte-to-byte models](#). *Transactions of the Association for Computational Linguistics*, 10:291–306.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Mohsen Yazdinejad and Marjan Kaedi. 2024. [Building a rich dataset to empower the persian question answering systems](#). *Preprint*, arXiv:2412.20212.
- Mohammad Yeghaneh Abkenar and Manfred Stede. 2024. [Neural mining of Persian short argumentative texts](#). In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 30–35, Torino, Italia. ELRA and ICCL.
- Hossein Zeinali, Lukáš Burget, and Jan Honza Černocký. 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.
- Abbas Zohrevand, Javad Sadri, and Zahra Imani. 2025. [A comprehensive persian offline handwritten database for investigating the effects of heritability and family relationships on handwriting](#). *Preprint*, arXiv:2509.03510.
- Saeedreza Zouashkiani. 2025. [Persianvox_nm: A 90+ hour single-speaker persian speech dataset](#). Derived from Nasle Mana Magazine, licensed under CC BY 4.0.

A Appendix

The tables below present all 101 ISO 639-3 coded Iranic varieties organized by branch and type.

Tables 2–7 enumerate the complete inventory of 101 ISO 639-3 codes assigned to the Iranic family, split by branch (Southwestern, Northwestern, Eastern, Ossetic, and extinct/historical) and by ISO type (macrolanguage, distinct language,

Name	ISO	Branch	Type	Speakers	Script(s)	Note
Persian (Farsi)	fas	SW	Macro	~110M	Perso-Arabic	Macrolang.; covers pes, prs
Western Farsi	pes	SW	Lang.	~70M	Perso-Arabic	Main spoken variety in Iran
Dari	prs	SW	Lang.	~20M	Perso-Arabic	Official language of Afghanistan
Tajik	tgk	SW	Lang.	~8M	Cyrillic	Official in Tajikistan
Northern Luri	lrc	SW	Lang.	~5M	Perso-Arabic	Part of Luri dialect continuum; W. Iran
Laki	lki	SW	Lang.	~1M	Perso-Arabic	Related to Luri; Kermanshah & Ilam, Iran
Judeo-Persian	jpr	SW	Lang.	~55K	Hebrew/Arabic	Persian in Hebrew script; Jewish communities
Kumzari	zum	SW	Lang.	~10K	Perso-Arabic	Outlier SW Iranian; Musandam, Oman
Judeo-Tat	jdt	SW	Lang.	~5K	Cyrillic/Hebrew	Endangered; Azerbaijan & Dagestan

Table 2: ISO 639-3 varieties — **Southwestern Iranian: Macrolanguages & Languages** (10 entries). Branch: SW = Southwestern, NW = Northwestern, E = Eastern, Oss. = Ossetic, Ext. = Extinct. Type: *Macro* = macrolanguage; *Lang.* = distinct language; *Dialect* = variety of a parent language; *Hist.* = no living speakers. Sorted by speaker count within each group.

dialect, or historical). Table 8 reports the digital resource and tool footprint of the major Indic languages, with Arabic and Turkish as regional reference points: pre-training corpus size (FineWeb-2), Wikipedia and Wiktionary article counts, UniMorph lemmata and forms, GlotLID V3 identification F1, parallel sentence totals, and availability of grammar, WordNet, spell-checker, and ASR tools. Table 10 compares speaker counts from two web-based estimation sources — the URL-NLP metadata of van Esch et al. (2022) and Wikipedia infoboxes — for the 33 Indic codes covered by URL-NLP. We emphasize that all reported counts are estimations aggregated from heterogeneous online sources, and the L1/L2 split is only meaningful for the few high-resource varieties where Wikipedia explicitly reports it (notably Persian and Pashto); for the rest, only a single rough figure is available and the two sources frequently disagree. For Indic varieties absent from both URL-NLP and Wikipedia, the only option is rougher estimation from secondary literature. Finally, Tables 11 and 12 compile publicly available NLP resources covering Indic languages: Table 11 catalogues datasets and benchmarks across all task families (NLU, MT, QA, knowledge benchmarks, NER, parsing, LID, summarization, sentiment/emotion, speech, vision-language, and pre-training corpora), while Table 12 covers models, toolkits, and plat-

forms (multilingual and Iranian-specific encoders, decoders, encoder-decoders, speech encoders, LID and script tools, and the major Persian preprocessing libraries).

Name	ISO	Branch	Type	Speakers	Script(s)	Note
Hazaragi	haz	SW	Dialect	~2.2M	Perso-Arabic	Dialect of Dari/Persian; Hazara people, Afghanistan
Aimaq	aiq	SW	Dialect	~1.8M	Perso-Arabic	Persian dialect cluster; NW Afghanistan
Bakhtiari	bqi	SW	Dialect	~1.3M	Perso-Arabic	Dialect of Luri; Khuzestan & Isfahan
Southern Luri	luz	SW	Dialect	~1M	Perso-Arabic	Dialect of Luri; Kohgiluyeh, Iran
Lari	lrl	SW	Dialect	~200K	Perso-Arabic	SW Iranian dialect; Fars Province, Iran
Fars Dialects	fay	SW	Dialect	~200K	Perso-Arabic	Dialect cluster; Fars Province, Iran
Dezfuli-Shushtari	def	SW	Dialect	~50K	Perso-Arabic	Dialect of Luri branch; Khuzestan, Iran
Dehwari	deh	SW	Dialect	~30K	Perso-Arabic	Dialect of Persian; Balochistan, Pakistan
Vafsi	vaf	SW	Dialect	~30K	Perso-Arabic	Endangered; Hamadan Province, Iran
Bashkardi	bsg	SW	Dialect	~20K	Perso-Arabic	Endangered; Hormozgan, Iran
Darwazi	drw	SW	Dialect	~15K	Arabic/Cyrillic	Dialect of Dari; Badakhshan
Bukharic	bhh	SW	Dialect	~10K	Hebrew/Cyrillic	Jewish Persian dialect; Bukhara, Uzbekistan
Dari (Zoroastrian)	gbz	SW	Dialect	~10K	Perso-Arabic	Zoroastrian Persian dialect; Yazd & Kerman
Parsi-Dari	prd	SW	Dialect	~10K	Perso-Arabic	Zoroastrian variety; may overlap with gbz
Pahlavani	phv	SW	Dialect	~10K	Perso-Arabic	Endangered; Gilan, Iran
Khunsaric	kfm	SW	Dialect	~10K	Perso-Arabic	Endangered; Isfahan Province, Iran
Gazic	gzi	SW	Dialect	~10K	Perso-Arabic	Endangered; Isfahan Province, Iran
Nayini	nyq	SW	Dialect	~10K	Perso-Arabic	Severely endangered; Isfahan, Iran
Soi	soj	SW	Dialect	~5K	Perso-Arabic	Endangered; Isfahan Province, Iran
Natanzic	ntz	SW	Dialect	~5K	Perso-Arabic	Severely endangered; Isfahan, Iran
Sivandi	siy	SW	Dialect	~3K	Perso-Arabic	Severely endangered; Fars, Iran
Lasgerdi	lsa	SW	Dialect	~2K	Perso-Arabic	Severely endangered; Semnan, Iran

Table 3: ISO 639-3 varieties — **Southwestern Iranian: Dialects** (22 entries). Branch: SW = Southwestern, NW = Northwestern, E = Eastern, Oss. = Ossetic, Ext. = Extinct. Type: *Macro* = macrolanguage; *Lang.* = distinct language; *Dialect* = variety of a parent language; *Hist.* = no living speakers. Sorted by speaker count within each group.

Name	ISO	Branch	Type	Speakers	Script(s)	Note
Kurdish (macro)	kur	NW	Macro	~30M	Lat./Arab./Cyr.	Macrolang.; covers kmr, ckb, sdh
Balochi (macro)	bal	NW	Macro	~9M	Arabic/Latin	Macrolang.; covers bcc, bgn, bgp
N. Kurdish (Kurmanji)	kmr	NW	Lang.	~15M	Latin/Arabic	Largest Kurdish variety; Turkey, Syria, Iraq
C. Kurdish (Sorani)	ckb	NW	Lang.	~7M	Arabic (Sorani)	Co-official in Iraqi Kurdistan
Mazanderani	mzn	NW	Lang.	~4M	Perso-Arabic	Caspian language; Mazandaran, Iran
Gilaki	glk	NW	Lang.	~3M	Perso-Arabic	Caspian language; Gilan Province, Iran
Southern Kurdish	sdh	NW	Lang.	~3M	Perso-Arabic	Iran & Iraq; sometimes treated as dialect
Western Balochi	bgn	NW	Lang.	~2M	Perso-Arabic	Pakistan, Iran, Afghanistan
Zazaki (Dimli)	diq	NW	Lang.	~2M	Latin	Zaza-Gorani branch; E. Turkey
Zaza	zza	NW	Lang.	~2M	Latin	E. Turkey; closely related to diq
Talysh	tly	NW	Lang.	~1M	Lat./Arab./Cyr.	Azerbaijan and Iran
Kirmanjki	kiu	NW	Lang.	~300K	Latin	Zaza-Gorani; debated relation to Kurdish
Gurani (Hawrami)	hac	NW	Lang.	~200K	Perso-Arabic	Zaza-Gorani branch; distinct from Kurdish
Muslim Tat	ttt	NW	Lang.	~30K	Cyrillic/Latin	Tatic branch; Azerbaijan & Dagestan
Shabaki	sdb	NW	Lang.	~10K	Perso-Arabic	Endangered; Nineveh, Iraq
Bajelani	bjm	NW	Lang.	~10K	Perso-Arabic	Endangered; Iraq and Iran

Table 4: ISO 639-3 varieties — **Northwestern Iranian: Macrolanguages & Languages** (17 entries). Branch: SW = Southwestern, NW = Northwestern, E = Eastern, Oss. = Ossetic, Ext. = Extinct. Type: *Macro* = macrolanguage; *Lang.* = distinct language; *Dialect* = variety of a parent language; *Hist.* = no living speakers. Sorted by speaker count within each group.

Name	ISO	Branch	Type	Speakers	Script(s)	Note
Southern Balochi	bcc	NW	Dialect	~3M	Perso-Arabic	Dialect of Balochi; Pakistan & Iran
Eastern Balochi	bgp	NW	Dialect	~2M	Perso-Arabic	Dialect of Balochi; eastern Pakistan
Takestani	tkr	NW	Dialect	~200K	Perso-Arabic	NW Iranian dialect; Zanjan, Iran
Shahmirzadi	srz	NW	Dialect	~30K	Perso-Arabic	Endangered; Mazandaran, Iran
Rudbari	rdb	NW	Dialect	~30K	Perso-Arabic	NW Iranian dialect; Gilan, Iran
Sangisari	sgr	NW	Dialect	~20K	Perso-Arabic	Endangered; Semnan, Iran
Koroshi	ktl	NW	Dialect	~15K	Perso-Arabic	Endangered; Hormozgan, Iran
Ashtiani	atn	NW	Dialect	~15K	Perso-Arabic	Endangered; Markazi Province, Iran
Semnani-Biyabuneki	smi	NW	Dialect	~15K	Perso-Arabic	Endangered; Semnan Province, Iran
Harzani-Kilit	hrz	NW	Dialect	~10K	Perso-Arabic	Endangered; East Azerbaijan, Iran
Karingani-Kalasuri	kgn	NW	Dialect	~10K	Perso-Arabic	Dialect cluster; East Azerbaijan, Iran
Maraghei	vmh	NW	Dialect	~10K	Perso-Arabic	Endangered; East Azerbaijan, Iran
Talysh (Shahrudi)	shm	NW	Dialect	~5K	Perso-Arabic	Dialect of Talysh; Semnan, Iran
Kho'ini	xkc	NW	Dialect	~5K	Perso-Arabic	Endangered; East Azerbaijan, Iran
Karanic	okh	NW	Dialect	~5K	Perso-Arabic	Endangered; Zanjan Province, Iran
Upper Taromi	toz	NW	Dialect	~5K	Perso-Arabic	Endangered; Zanjan Province, Iran
Eshtehardi	esh	NW	Dialect	~5K	Perso-Arabic	Endangered; Gilan, Iran
Alamuti	goz	NW	Dialect	~5K	Perso-Arabic	Endangered; Alamut, Iran
Alviri-Vidari	avd	NW	Dialect	~4K	Perso-Arabic	Endangered; Qazvin Province, Iran
Sorkhei-Aftari	sqr	NW	Dialect	~3K	Perso-Arabic	Endangered; Isfahan Province, Iran
Kajali	xkj	NW	Dialect	~3K	Perso-Arabic	Severely endangered; Zanjan, Iran
Kabatei	xkp	NW	Dialect	~3K	Perso-Arabic	Severely endangered; Zanjan, Iran
Sarli	sdf	NW	Dialect	~2K	Perso-Arabic	Severely endangered; Kerman-shah, Iran
Razajerdi	rat	NW	Dialect	~2K	Perso-Arabic	Severely endangered; Zanjan, Iran

Table 5: ISO 639-3 varieties — **Northwestern Iranian: Dialects** (24 entries). Branch: SW = Southwestern, NW = Northwestern, E = Eastern, Oss. = Ossetic, Ext. = Extinct. Type: *Macro* = macrolanguage; *Lang.* = distinct language; *Dialect* = variety of a parent language; *Hist.* = no living speakers. Sorted by speaker count within each group.

Name	ISO	Branch	Type	Speakers	Script(s)	Note
Pashto (macro)	pus	E	Macro	~60M	Perso-Arabic (ext.)	Macrolang.; covers pbt, pbu, pst
Northern Pashto	pbu	E	Lang.	~30M	Perso-Arabic (ext.)	Dominant variety in Afghanistan
Southern Pashto	pbt	E	Lang.	~20M	Perso-Arabic (ext.)	Dominant variety in Pakistan
Shughni	sgb	E	Lang.	~100K	Cyrillic/Latin	Pamiri; Gorno-Badakhshan, Tajikistan
Sarikoli	srh	E	Lang.	~100K	Cyrillic/Arabic	Pamiri; Tajikistan & Xinjiang, China
Wakhi	wbl	E	Lang.	~50K	Arabic/Cyrillic	Pamiri; Afghanistan, Pakistan, China
Waneci	wne	E	Lang.	~40K	Perso-Arabic	Eastern Iranian; Balochistan, Pakistan
Yaghnobi	yai	E	Lang.	~13K	Cyrillic	Descendant of Sogdian; Tajikistan
Yazgulyam	yah	E	Lang.	~7K	Cyrillic	Endangered; Yazgulyam Valley, Tajikistan
Ormuri	oru	E	Lang.	~7K	Unwritten	Endangered; Waziristan, Pakistan
Yidgha	ydg	E	Lang.	~6K	Unwritten	Endangered; Chitral, Pakistan
Parachi	prc	E	Lang.	~5K	Unwritten	Endangered; Parwan, Afghanistan
Munji	mnj	E	Lang.	~5K	Unwritten	Endangered; Munjan Valley, Afghanistan
Ishkashimi	isk	E	Lang.	~2K	Unwritten	Severely endangered; Afghanistan
Central Pashto	pst	E	Dialect	~10M	Perso-Arabic (ext.)	Dialect of Pashto; Pakistan
Sanglechi-Ishkashimi	sgl	E	Dialect	~3K	Unwritten	Dialect cluster related to isk
Sanglechi	sgy	E	Dialect	~1K	Unwritten	Dialect of sgl; Afghanistan

Table 6: ISO 639-3 varieties — **Eastern Iranian** (17 entries). Branch: SW = Southwestern, NW = Northwestern, E = Eastern, Oss. = Ossetic, Ext. = Extinct. Type: *Macro* = macrolanguage; *Lang.* = distinct language; *Dialect* = variety of a parent language; *Hist.* = no living speakers. Sorted by speaker count within each group.

Name	ISO	Branch	Type	Speakers	Script(s)	Note
<i>Ossetic</i>						
Ossetic	oss	Oss.	Lang.	~650K	Cyrillic	Only living Scythian language; N. & S. Ossetia
Digor Ossetic	osd	Oss.	Dialect	~450K	Cyrillic	Western dialect of Ossetic
<i>Extinct / Historical</i>						
Avestan	ave	Ext.	Hist.	0 (liturgical)	Avestan	Sacred lang. of Zoroastrianism; ~1000 BCE
Old Persian	peo	Ext.	Hist.	0	Cuneiform	Achaemenid inscriptions; 600–400 BCE
Mid. Persian (Pahlavi)	pal	Ext.	Hist.	0	Pahlavi	Official language of Sassanid Empire
Parthian	xpr	Ext.	Hist.	0	Aramaic/Pahlavi	NW Iranian; Parthian Empire
Sogdian	sog	Ext.	Hist.	0	Sogdian	Silk Road lingua franca; ancestor of Yaghnobi
Bactrian	xbc	Ext.	Hist.	0	Greek	Ancient Bactria; Afghanistan / Uzbekistan
Khwarezmian	xco	Ext.	Hist.	0	Khwarazmian	Ancient Khwarezm; Uzbekistan / Turkmenistan
Khotanese	kho	Ext.	Hist.	0	Brahmi	Eastern Iranian; ancient Khotan, Xinjiang
Tumshuqese	xtq	Ext.	Hist.	0	Brahmi	Eastern Iranian; Tumushuk, Xinjiang
Old Ossetic	oos	Ext.	Hist.	0	Georgian (hist.)	Ancient North Caucasus
Jassic	ysc	Ext.	Hist.	0	Latin (hist.)	Ossetic dialect in Hungary; 13th–15th c.

Table 7: ISO 639-3 varieties — **Ossetic** (2 entries) and **Extinct / Historical** (11 entries). Branch: SW = Southwestern, NW = Northwestern, E = Eastern, Oss. = Ossetic, Ext. = Extinct. Type: *Macro* = macrolanguage; *Lang.* = distinct language; *Dialect* = variety of a parent language; *Hist.* = no living speakers. Sorted by speaker count within each group.

ISO	Language	Resources								Tools			
		FineWeb2	Wikipedia	Wiktionary	Lemmata	Forms	LID F1	Speakers	Parallel	Grammar	WordNet	Spell	ASR
Regional neighbors (reference)													
arb	Arabic	106 GB	1,293K	74,093	4,134	140,003	0.845	481/488	—	✓	✓	✓	✓
tur	Turkish	135 GB	658K	43,253	3,579	275,460	0.972	91/82	—	✓	✓	✓	✓
Iranic languages													
fas	Persian	102 GB	1,075K	19,248	273	37,128	0.922	127/81	612M	✓	✓	✓	✓
tgk	Tajik	—	118K	430	0	0	0.996	—	—	✗	✗	✗	✗
kmr	N. Kurdish	979 MB	92K	8,254	15,083	216,370	0.997	29/35	848K	✓	✓	✓	✓
ckb	C. Kurdish	—	82K	—	274	24,317	0.969	—	—	✓	✓	✓	✓
sdh	S. Kurdish	—	15K	—	1	189	0.982	—	—	✓	✗	✗	✗
pus	Pashto	—	21K	1,100	395	6,945	0.990 [†]	—	—	✗	✗	✗	✗
mzn	Mazanderani	1.7 MB	65K	337	0	0	0.958	20/6	18.9K	✗	✗	✗	✗
glk	Gilaki	16.4 MB	48K	34	0	0	0.968	9.4/2	18.6K	✓	✗	✗	✗
bal	Balochi	—	1.2K	429	0	0	0.990 [‡]	8.8/12	175	✗	✗	✗	✗
tly	Talysh	3.0 MB	10K	302	0	0	0.993	1.1/0.9	8.7K	✓	✗	✗	✗
diq	Zazaki	—	43K	1,083	0	0	0.987	1.5/3	12.1K	✓	✗	✗	✗
hac	Gorani	38.7 MB	*	211	0	0	0.982	0.3/0.4	36.8K	✓	✗	✗	✗
lrc	N. Luri	82.7 KB	—	5	0	0	0.909	5/2.9	—	✓	✗	✗	✗
lki	Laki	57.7 KB	*	85	0	0	0.800	0.7/1	15.2K	✓	✗	✗	✗

Table 8: Digital resource and tool footprint for major Iranic languages, with regional neighbors as reference. (*) Wikimedia Incubator only; (—) no data. Lemmata/Forms: UniMorph coverage. LID F1: GlotLID V3. [†]Reported under pbt_Arab (Southern Pashto), used as proxy for the Pashto macro. [‡]Reported under bcc_Arab (Southern Balochi), used as proxy for the Balochi macro. Speakers: Wikipedia / URL-NLP estimate (millions). Parallel: total sentence pairs across all available pairs. ✓ = tool available; ✗ = absent.

Table 9: Parallel sentence counts across Iranic language pairs.

	en	fas	kur	glk	mzn	tly	lki	luz	hac	diq
fas	612M									
kur	511K	326K								
glk	9K	5.5K	488							
mzn	4.5K	5.1K	1.6K	1.1K						
tly	2.1K	2.1K	1.6K	54	851					
lki	4.6K	3.4K	994	437	3K	849				
luz	2K	2K	994	39	1.4K	785	1.2K			
hac	22K	7.8K	1.6K	2K	1.3K	409	723	934		
diq	4.4K	4.4K	3.3K	0	0	0	0	0	0	
bal	110	30	35	0	0	0	0	0	0	0

Name	ISO	Br.	Type	URL-NLP	L1 (Wiki)	L2 (Wiki)	L1+L2	Script(s)	Region
Southwestern Iranian									
Persian (macro)	fas	SW	Macro	81M	91M	35M	127M	Arab, Latn	IR, AF, TJ
Hazaragi	haz	SW	Dialect	3M	3M	—	—	Arab	AF, IR, PK
Aimaq	aiq	SW	Dialect	600K	600K	—	—	Arab	AF, IR
Bakhtiari	bqi	SW	Dialect	1M	1M	—	—	Arab	IR
N. Luri	lrc	SW	Lang.	2M	1.5M	—	—	Arab	IR, IQ
S. Luri	luz	SW	Dialect	900K	1M	—	—	Arab	IR
Larestani	lrl	SW	Lang.	200K	—	—	—	Arab	IR
Fars Dialects	fay	SW	Dialect	10–100K	—	—	—	Arab	IR
Zoroastrian Dari	gbz	SW	Dialect	10–100K	—	—	—	Arab	IR
Bukharic	bhh	SW	Dialect	100K	100K	—	—	Cyrl	UZ
Northwestern Iranian									
Kurdish (macro)	kur	NW	Macro	25M	26M	—	—	Arab, Latn, Cyrl	TR, IR, IQ, SY
C. Kurdish (Sorani)	ckb	NW	Lang.	7M	7M	—	—	Arab	IR, IQ
S. Kurdish	sdh	NW	Lang.	3M	3M	—	—	Arab	IR, IQ
Balochi (macro)	bal	NW	Macro	8M	8.8M	—	—	Arab, Latn	PK, IR, AF
W. Balochi	bgn	NW	Lang.	2M	2M	—	—	Arab	AF, IR, PK
E. Balochi	bgp	NW	Dialect	2M	2M	—	—	Arab, Latn	PK
Mazanderani	mzn	NW	Lang.	6M	1.35M	—	—	Arab	IR
Gilaki	glk	NW	Lang.	2M	3M	—	—	Arab	IR
Talysh	tly	NW	Lang.	900K	500–800K	—	—	Arab, Cyrl, Latn	AZ, IR
Gorani (Hawrami)	hac	NW	Lang.	400K	400K	—	—	Arab	IR, IQ
Zazaki (macro)	zza	NW	Macro	3M	2–4M	—	—	Latn	TR
Takestani	tko	NW	Dialect	200K	—	—	—	Arab	IR
Judeo-Tat	jdt	NW	Lang.	10–100K	70K	—	—	Cyrl	AZ
Muslim Tat	ttt	NW	Lang.	10–100K	30K	—	—	Arab, Latn	AZ, IR
Eastern Iranian									
Tajik	tgk	E	Lang.	8M	8M	—	—	Cyrl	TJ, UZ
Pashto (macro)	pus	E	Macro	13M	51M	4.9M	56M	Arab	AF, PK
Shughni	sgn	E	Lang.	10–100K	100K	—	—	Cyrl, Latn	TJ, AF
Wakhi	wbl	E	Lang.	10–100K	60K	—	—	Arab, Cyrl, Latn	AF, PK, CN, TJ
Yagnobi	yai	E	Lang.	10–100K	13K	—	—	Cyrl	TJ
Waneci	wne	E	Lang.	10–100K	—	—	—	Arab	PK
Ormuri	oru	E	Lang.	<10K	5K	—	—	Arab	AF, PK
Ossetic									
Iron Ossetian	oss	Oss.	Lang.	600K	600K	—	—	Cyrl	RU, GE

Table 10: Iranian language varieties present in the URL-NLP metadata dataset (van Esch et al., 2022), cross-referenced with Wikipedia speaker data. L1 = native speakers; L2 = second-language speakers; Total = combined. Branch: SW Southwestern, NW Northwestern, E Eastern, Oss. Ossetic. Type: Macro macrolanguage; Lang. distinct language; Dialect dialect/variety.

Name	Task	Number	Name	Task	Number
NLU & Semantics					
SIB-200 (Adelani et al., 2024)	Classification	1 (fas)	Taxi1500 (Ma et al., 2025)	Classification	1 (fas)
MASSIVE (FitzGerald et al., 2023)	Classification	1 (fas)	Persian-ATIS (Akbari et al., 2023)	Intent / Slot Filling	1 (fas)
FarsTail (Amirkhani et al., 2026)	Textual Entailment	1 (fas)	ExaPPC (Sadeghi et al., 2022)	Paraphrase Identification	1 (fas)
FarsSick (Ghasemi and Keyvanrad, 2021)	STS / NLI	1 (fas)	ParsiNLU (Khashabi et al., 2021)	Multi-task NLU	1 (fas)
PerCul (Moosavi Monazzah et al., 2025)	Cultural Reasoning / NLU	1 (fas)			
MT Tasks					
FLORES-200 (NLLB Team et al., 2022)	MT	6 (pes, prs, tgk, ckb, kmr, pbt)	NTREX-128 (Federmann et al., 2022)	MT	1 (fas)
TICO-19 (Anastasopoulos et al., 2020)	MT	2 (fas, prs)	SMOL (Caswell et al., 2025)	MT	3 (glk, hac, kmr)
PARME (Ahmadi et al., 2025b)	MT / LM	6 Iranian	Kurdish Par. Corpus (Ahmadi et al., 2022)	MT	1 (kur)
KMTP (Amini et al., 2021)	MT	1 (ckb)	Middle-East MT (Ahmadi et al., 2025a)	MT	4 (hac, kur, lki, glk)
Pewan (Sheykh Esmaili and Salavati, 2013)	MT / IR	2 (fas, ckb)	Script Norm. (Ahmadi and Anastasopoulos, 2023)	Translation / Norm.	5 Iranian
Question Answering					
Belebele (Bandarkar et al., 2024)	MRC	4 (pes, ckb, tgk, pbt)	Global-PIQA (Chang et al., 2025)	Commonsense MCQA	1 (fas)
PQuAD (Darvishi et al., 2023)	Reading Comprehension	1 (fas)	ParSQuAD (Abadani et al., 2021)	Reading Comprehension	1 (fas)
PersianQuAD (Kazemi et al., 2022)	Reading Comprehension	1 (fas)	PersianQA (Ayoubi, 2021)	Reading Comprehension	1 (fas)
PersianMHQA (Taji et al., 2025)	Multi-hop QA	1 (fas)	NextQuAD (Yazdinejad and Kaodi, 2024)	Open-domain QA	1 (fas)
FarsiQuAD (FontanRad et al., 2024)	Open-domain QA	1 (fas)	FarsNewsQA (Kazemi et al., 2023)	News QA	1 (fas)
PCoQA (Hemati et al., 2023)	Conversational QA	1 (fas)	PeCoQ (Etezadi and Shamsfard, 2020)	Complex KGQA	1 (fas)
PerCQA (Jamali et al., 2022)	Community QA	1 (fas)	PeQA (Arshia et al., 2022)	QA / Chatbot	1 (fas)
PerAnSel (Mozafari et al., 2022)	Answer Selection	1 (fas)	ParsSimpleQA (Babaei Giglou et al., 2022)	Simple QA / KG	1 (fas)
SynTran-fa (Farsi et al., 2024)	QA Generation	1 (fas)	MegaChat (Rahmani et al., 2025)	Synthetic Sales QA	1 (fas)
IslamicPCQA (Ghafouri et al., 2025)	Religious Multi-hop QA	1 (fas)	Religious QA (Boreshban et al., 2018)	Religious QA	1 (fas)
COPER (Khanmohammadi et al., 2021b)	QA / IR (COVID-19)	1 (fas)	Persian Med. QA (Veisi and Shandi, 2020)	Medical QA	1 (fas)
PersianMedQA (Kalahroodi et al., 2025b)	Medical MCQA	1 (fas)	Gaokerena Med. (Ghassabi et al., 2025)	Medical QA Suite	1 (fas)
Pers. Pharma QA (Hasannejad and Porbaha, 2024)	Pharmaceutical QA	1 (fas)	Pers. Med. Dialog (Hasannejad and Porbaha, 2023)	Medical Dialog QA	1 (fas)
PersianSciQA (Jolfaei et al., 2025)	Scientific QA	1 (fas)	KNQAD (Saeed, 2024)	News QA	1 (ckb)
Knowledge & LLM Evaluation Benchmarks					
Khayyam (PersianMMLU) (Ghahroodi et al., 2024)	Knowledge MC	1 (fas)	FarsEval-PKBETS (Shamsfard et al., 2025)	Comprehensive Bench	1 (fas)
PeKA (Hosseini et al., 2025b)	Knowledge Assessment	1 (fas)	PK-BETS (Hosseini et al., 2025b)	Bias / Ethics / Toxicity	1 (fas)
EPT (Mirbagheri et al., 2025)	Trustworthiness Eval	1 (fas)	ELAB (Pourbahman et al., 2025)	LLM Alignment Bench	1 (fas)
MELAC (Farsi et al., 2025a)	Culture Alignment Eval	1 (fas)	Persian-LLM Bench (Abaskoti et al., 2024)	Multi-task LLM Eval	1 (fas)
MEXA (Kargaran et al., 2025)	Cross-lingual Alignment	1 (fas)			
Information Extraction (NER)					
PEYMA (Shahshahani et al., 2019)	NER	1 (fas)	PersoNER (Poostchi et al., 2016)	NER	1 (fas)
ArmanPersoNER (Poostchi et al., 2018)	NER	1 (fas)	ARMAN (Farahani et al., 2021a)	NER	1 (fas)
Syntax: Parsing					
PerUDT (Rasooli et al., 2020)	POS / UD Parsing	1 (fas)	UD Kurmanji (Gökirmak and Tyers, 2017)	UD Parsing	1 (kmr)
CKB-Mukri (Asadpour, 2025)	UD Parsing	1 (ckb)			
Language / Dialect Identification					
GlottID (Kargaran et al., 2023)	LID	16 Iranian	WiLI-2018 (Thoma, 2018)	LID	4 Iranian
PALI (Ahmadi et al., 2023a)	LID (Perso-Arabic)	6 Iranian	Kurd/Zaza LID (Ahmadi et al., 2023b)	LID	4 Iranian
VoxLingua107 (Valk and Alumäe, 2020)	Spoken LID	3 (fas, pus, tgk)			
Text Generation & Instruction					
LSCP (Abdi Khojasteh et al., 2020)	Colloquial Generation	1 (fas)	Aya Dataset (Singh et al., 2024)	Instruction	2 (fas, kur)
xP3x (Muenighoff et al., 2023)	Instruction / Multi-task	2 (fas, kur)	MURI-IT (Köksal et al., 2025)	Instruction (multi-task)	1 (fas)
FarsInstruct (Mokhtarabadi et al., 2025)	Instruction	1 (fas)	Matina-IFT (Hosseini et al., 2025c)	Cultural Instruction	1 (fas)
Semi-Alpaca (Amiri, 2024)	Instruction	1 (fas)	Wergor (Ahmadi, 2019)	Transliteration	1 (kur)
Text Summarization					
CrossSum (Bhattacharjee et al., 2023)	Summarization	2 (fas, pus)	XL-Sum (Hasan et al., 2021)	Summarization	2 (fas, pus)
pn-summary (Farahani, 2020)	Summarization	1 (fas)	Wiki Summary (Farahani, 2020)	Summarization	1 (fas)
ParsBERT Summ. (Farahani et al., 2021b)	Summarization	1 (fas)			
Sentiment & Emotion					
SentiPers (Hosseini et al., 2021)	Sentiment Analysis	1 (fas)	DeepSentiPers (Sharami et al., 2020)	Sentiment Analysis	1 (fas)
MirasOpinion (Ashrafi Asli et al., 2020)	Sentiment Analysis	1 (fas)	Pars-ABSA (Ataei et al., 2019)	Aspect-based Sentiment	1 (fas)
ParsBERT Sent. (Panahandeh Nigjeh and Ghanbari, 2024)	Sentiment Analysis	1 (fas)	SnappFood (Farahani et al., 2021a)	Sentiment Analysis	1 (fas)
Digikala Sent. (Farahani et al., 2021a)	Sentiment Analysis	1 (fas)	ArmanEmo (Mirzaee et al., 2025)	Emotion Detection	1 (fas)
EmoPars (Sabri et al., 2021)	Emotion Detection	1 (fas)	ShortPersianEmo (Rasouli and Kiani, 2023)	Emotion Detection	1 (fas)
MirasIrony (Golizizian et al., 2020)	Irony Detection	1 (fas)	Iran COVID-19 (Hosseini et al., 2020)	Topic / Content Class.	1 (fas)
Other Text Classification					
Persian News (Farahani et al., 2021a)	News Classification	1 (fas)			
Spell Checking & Argumentation					
PerSpellData (Oji et al., 2021b)	Spell Checking	1 (fas)	Persian Typo (Dehghani and Faili, 2024)	Spell Checking	1 (fas)
PAMT (Yeghaneh Abkenar and Stede, 2024)	Argumentation Mining	1 (fas)			
Speech Tasks (ASR, TTS, Other)					
FLEURS (Conneau et al., 2023)	ASR	2 (fas, pbt)	XTREME-S (Conneau et al., 2022)	Multi-task Speech	1 (fas)
Common Voice (Ardila et al., 2020)	ASR / TTS	4 Iranian	Cordi (Ahmadi et al., 2024)	ASR / TTS / LID	2 (ckb, hac)
AsoSoft (Jira) (Veisi et al., 2022)	ASR	1 (ckb)	ParsVoice (Kalahroodi et al., 2025a)	TTS	1 (fas)
ManaTTS (Qharabagh et al., 2025)	TTS	1 (fas)	DeepMine-TTS (Adibian et al., 2025)	TTS	1 (fas)
DeepMine (Zeinali et al., 2019)	ASR / Speaker Rec	1 (fas)	MirasVoice (Vaheh et al., 2018)	ASR / TTS	1 (fas)
ArmanTTS (Shamgholi et al., 2023)	TTS	1 (fas)	PersianVox_NM (Zouashkiani, 2025)	ASR / TTS	1 (fas)
FARSDAT (Bijankhan et al., 1994)	ASR	1 (fas)	ShEMO (Mohamad Nezami et al., 2019)	Speech Emotion	1 (fas)
FaVC (Serajian et al., 2025)	Voice Conversion	1 (fas)	Arman-AV (Peymanfard et al., 2024)	Audio-Visual ASR	1 (fas)
PSRB (Sedghiyeh et al., 2025)	ASR Benchmark	1 (fas)	Persian-Farsi-Speech (Thomcles, 2025)	TTS	1 (fas)
DOLMA ASR (Hameed et al., 2025)	ASR	6 Iranian			
Vision-Language (VQA, Captioning)					
Crossmodal-3600 (Thapliyal et al., 2022)	Captioning	1 (fas)	ParsCap (Khanmohammadi et al., 2021a)	Captioning	1 (fas)
ParsVQA-Caps (Mobasher et al., 2025)	VQA / Captioning	1 (fas)	MEENA (Ghahroodi et al., 2026)	Multimodal QA	1 (fas)
Persian-VQA (Farsi et al., 2025b)	VQA / VLM Eval	1 (fas)	PersianCLEVR (Vadai et al., 2025)	3D Visual Reasoning	1 (fas)
Vision: OCR & Handwriting					
IDPL-PFOD (sadat Hosseini et al., 2021)	Printed OCR	1 (fas)	IDPL-PFOD2 (Asadi-zeydabadi et al., 2023)	Printed OCR	1 (fas)
Khosravi Digits (Khosravi and Kabir, 2007)	HW Digit Recog.	1 (fas)	Khayyam HW (Jafarzadeh et al., 2024)	HW Recognition	1 (fas)
SUT (Shabaninia et al., 2023)	Document Image Analysis	1 (fas)	CENPARMI (Haghighi et al., 2009)	HW Database	1 (fas)
Zohrevand HW (Zohrevand et al., 2025)	HW Database	1 (fas)	Sadri HW (Sadri et al., 2016)	HW Recognition	1 (fas)
Alaei Seg. (Alaei et al., 2011)	Line / Word Segm.	1 (fas)	Persian Scene (Souzanchi et al., 2024)	Scene Text Detection	1 (fas)
PTDR (Raisi et al., 2025b)	Scene / Doc Text	1 (fas)	PESTD (Rashtehroudi et al., 2023)	Bilingual Scene Text	1 (fas)
FATR (Raisi et al., 2025a)	Wild Image Text Recog.	1 (fas)			
Unlabeled Text / Pre-training Corpora					
GloVeCC (Kargaran et al., 2024a)	LM	16 Iranian	EMMA-500 (Ji et al., 2025)	LM	14 Iranian
FineWeb-2 (Penedo et al., 2025)	LM	8 Iranian	FinePDFs (Kydlicek et al., 2025)	LM	1 (fas)
MADLAD-400 (Kudugunta et al., 2023)	LM	8 Iranian	BLOOM (BigScience Workshop et al., 2023)	LM	1 (fas)
mC4 (Xue et al., 2021)	LM	1 (fas)	CulturaX (Nguyen et al., 2024a)	LM	2 (fas, pus)
CulturaY (Nguyen et al., 2024b)	LM	1 (fas)	HPLT (Oepen et al., 2026)	LM	2 (fas, pus)
OSCAR 23.01 (Jansen et al., 2022)	LM	5 Iranian	Naab (Sabouri et al., 2024)	LM	1 (fas)
Matina (Hosseini et al., 2025c)	LM	1 (fas)	hmBlogs (Khansari and Shamsfard, 2021)	LM	1 (fas)
MirasText (Sabati et al., 2018)	LM	1 (fas)	TLPC (Processing, 2024)	LM	1 (fas)
Glot500 Corpus (Imani et al., 2023)	LM	14 Iranian	AsoSoft Text (Veisi et al., 2019)	LM	1 (ckb)
Sorani Lyrics (Ahmadi et al., 2020)	LM	1 (ckb)	Kurd. Textbooks (Abdulahman et al., 2019)	LM	1 (kur)
Zaza-Gorani (Ahmadi, 2020)	LM	3 (zaz, kur, hac)			

Table 11: Publicly available datasets including Iranian languages. Counts give the number of Iranian languages covered; when only one or two, ISO codes are shown.

Name	Task	Number	Name	Task	Number
Translation Models					
NLLB-200 (NLLB Team et al., 2022)	MT	6 Iranian	MADLAD-MT (Kudugunta et al., 2023)	MT	8 Iranian
M2M-100 (Fan et al., 2021)	MT	3 (fas, kur, pus)	SMaLL-100 (Mohammadshahi et al., 2022)	MT	3 (fas, kur, pus)
Kurd. Text-to-Sign (Kamal and Hassani, 2020)	MT (sign)	1 (kur)			
Encoder Language Models (Multilingual)					
Glott500-m (Imani et al., 2023)	Encoder LM	14 Iranian	XML-R (Conneau et al., 2020)	Encoder LM	4 (fas, kur, pus, tsk)
mBERT (Devlin et al., 2019)	Encoder LM	3 (fas, kur, pus)	RemBERT (Chung et al., 2021)	Encoder LM	2 (fas, pus)
mDeBERTaV3 (He et al., 2023)	Encoder LM	1 (fas)			
Encoder Language Models (Persian)					
ParsBERT (Farahani et al., 2021a)	Encoder LM	1 (fas)	FaBERT (Masumi et al., 2025)	Encoder LM (formal+informal)	1 (fas)
TookaBERT (SadraeiJavaheri et al., 2024)	Encoder LM	1 (fas)	AriaBERT (Ghafouri et al., 2023)	Encoder LM (RoBERTa)	1 (fas)
FarSSiBERT (Sadjadi et al., 2024)	Encoder LM (social)	1 (fas)	Hakim (Sarmadi et al., 2025)	BERT + RetroMAE	1 (fas)
MatinaRoBERTa (Hosseinbeigi et al., 2025a)	Encoder LM (RoBERTa)	1 (fas)	BERTCaps (Memari et al., 2024)	BERT + Capsule hybrid	1 (fas)
Encoder Language Models (Other Iranian)					
KuBERT (muhealddin Awlla et al., 2025)	Encoder LM	1 (ckb)	KurdBERT (language-ml-lab, 2024)	Encoder LM	2 (ckb, kmr)
Ps-BERT (Haq et al., 2023)	Encoder LM	1 (pus)			
Decoder Language Models (Multilingual)					
MADLAD (Kudugunta et al., 2023)	Decoder LM	8 Iranian	Goldfish (Chang et al., 2026)	Decoder LM	6 Iranian
BLOOM (BigScience Workshop et al., 2023)	Decoder LM	1 (fas)	XGLM (Lin et al., 2022)	Decoder LM	1 (fas)
LLaMAX (Lu et al., 2024)	Decoder LM	2 (fas, pus)	mGPT (Shliazhko et al., 2024)	Decoder LM	3 (fas, pus, tsk)
Decoder Language Models (Persian)					
PersianMind (Rostami et al., 2024)	LLaMA-2 bilingual	1 (fas)	PersianLLaMA (Abbasi et al., 2023)	LLaMA 7B/13B	1 (fas)
LLaMA-2 Persian Adapt.	Continued pretrain	1 (fas)	Persian-Phi (Akhlaghi et al., 2025)	Phi-3 Mini + PEFT	1 (fas)
Dorna / Dorna2 (PartAI, 2024a,b)	LLaMA-3 fine-tune	1 (fas)	Maral / MaralGPT (Haghiri and Mohrechi, 2023)	Custom LLM	1 (fas)
Hormoz LLM (Mann-E, 2024)	8B LLM	1 (fas)	BioPars (Merzah et al., 2025)	Biomedical LLM	1 (fas)
FarsiInstruct Models (Mokhtarabadi et al., 2025)	Instruction-tuned	1 (fas)	TaaroF-aware FT (Sadr et al., 2025)	DPO / SFT cultural	1 (fas)
TPPoet (Panahandeh et al., 2023)	Poetry generation	1 (fas)			
Encoder-Decoder Language Models					
Aya-101 (Singh et al., 2024)	Enc-Dec LM	3 (fas, pus, kur)	mT0 (Muennighoff et al., 2023)	Enc-Dec LM	2 (fas, kur)
mT5 (Xue et al., 2021)	Enc-Dec LM	1 (fas)	ByT5 (Xue et al., 2022)	Enc-Dec LM	1 (fas)
mLongT5 (Uthus et al., 2023)	Enc-Dec LM	1 (fas)			
Speech Encoders					
XEUS (Chen et al., 2024)	Speech Rep.	n/a (4k+ langs)	MMS (Pratap et al., 2024)	Speech Rep.	n/a (1k+ langs)
mHuBERT-147 (Boito et al., 2024)	Speech Rep.	2 (fas, ckb)	XLSR-128 (Babu et al., 2021)	Speech Rep.	3 (fas, ckb, pus)
Toolkits: LID, Transliteration & Script					
GlottLID (Kargaran et al., 2023)	LID	16 Iranian	GlottScript (Kargaran et al., 2024c)	Script ID	—
Wergor (Ahmadi, 2019)	Transliteration	1 (kur)	Script Norm. (Ahmadi and Anastasopoulos, 2023)	Script Normalization	5 Iranian
Toolkits: Preprocessing & Morphology					
Hazm (Roshan Research, 2026)	Tokenise / POS / Lemma	1 (fas)	DadmaTools (Etezadi et al., 2022)	Tokenise / POS / Lemma	1 (fas)
Parsivar (Molhtaj et al., 2018)	Normalise / Tokenise / Stem	1 (fas)	Shekar (Amirvovjdan, 2025)	Normalise / POS / Lemma	1 (fas)
Khoshnevis (HamidReza Attar, 2022)	Text Cleaning / Tokenise	1 (fas)	Seperno (Snapp Incubator, 2024)	Normalisation (Go)	1 (fas)
FarsiNLPTools (Focly et al., 2014)	Normalisation	1 (fas)	Virastar (Sedghii, 2022)	Normalisation (PHP)	1 (fas)
Persstem (Dehdari and Lonsdale, 2008)	Stemming / Morphology	1 (fas)	ParsiAnalyzer (Esmailyfard, 2017)	Tokenisation (Java)	1 (fas)
ParsiNorm (Oji et al., 2021a)	Tokenise / Normalise	1 (fas)	Parstidx / HengamTagger (Mirzababaei et al., 2022)	Temporal Tagging	1 (fas)
Platforms					
Mozilla Common Voice (Ardila et al., 2020)	Speech	4 Iranian			

Table 12: Publicly available models, toolkits and platforms including Iranian languages. Counts give the number of Iranian languages covered; when only one or two, ISO codes are shown.