One Language, Three of Its Voices: Evaluating Multilingual LLMs Across Persian, Dari, and Tajiki on Translation and Understanding Tasks

Noor Mairukh Khan Arnob[1], **Abu Bakar Siddique Mahi** [2]
[1]Research & Innovation Lab, University of Asia Pacific
[2]BUBT Research Graduate School, Bangladesh University of Business and Technology
Correspondence: arnob@uap-bd.edu

Abstract

The Iranian linguistic family is pluricentric, encompassing Iranian Persian, Dari (Afghanistan), and Tajiki (Tajikistan). While Multilingual Large Language Models (MLLMs) claim broad coverage, their robustness across these regional variants and script differences (Perso-Arabic vs. Cyrillic) remains under-explored, particularly in the open-weight landscape. We evaluate five open-weight models from the Qwen, Bloomz, and Gemma families across four downstream tasks: Sentiment Analysis, Machine Translation (MT), NLI, and QA. Utilizing a dataset of over 240,000 processed samples, we observe severe performance disparities. While the fine-tuned `gemma-3-4b-persian` achieves promising results on Iranian Persian (77.3% accuracy in Sentiment), almost all tested models appear to suffer catastrophic degradation on Tajiki script (dropping to <1.0 BLEU). These findings highlight a critical "script barrier" in current open-weight MLLM development for Central Asian languages. Code and data available here.

1 Introduction

The Persian language (*Farsi*) serves as a lingua franca for millions across the Silk Road region, officially recognized as Persian in Iran, Dari in Afghanistan, and Tajiki in Tajikistan. While these varieties share a high degree of mutual intelligibility in their spoken forms, they diverge significantly in the written domain. Iranian Persian and Dari utilize the Perso-Arabic script with subtle lexical and morphological differences, whereas Tajiki adopted the Cyrillic script during the Soviet era (Windfuhr, 2009).

This pluricentric nature presents a unique challenge for Natural Language Processing (NLP). Current Multilingual Large Language Models (MLLMs) often treat Persian as a monolith, predominantly trained on data scraped from the Iranian webspace. This creates a representational bias (Blasi et al., 2022) that may marginalize Dari and Tajiki speakers.

In this work, we present a focused evaluation of five open-weight Large Language Models, ranging from 1.5B to 4B parameters. By focusing on models suitable for edge deployment rather than proprietary giants, we aim to understand the accessibility of Persian NLP. We focus on the following Research Questions (RQs):

- **RQ1 (Task Performance):** How do general-purpose multilingual models compare to language-specific fine-tunes on standard Persian tasks?
- **RQ2 (Script Robustness):** Does the change from Perso-Arabic to Cyrillic script (Tajiki) cause catastrophic generalization failure in models that claim Persian support?
- **RQ3 (Dialectal Variance):** How do models perform on Dari, which shares the script but differs in vocabulary and grammar?

We evaluate these questions across four distinct tasks: Sentiment Analysis, Natural Language Inference (NLI), Question Answering (QA), and Machine Translation (MT).

2 Related Work

2.1 Persian NLP Resources

The landscape of Persian NLP has expanded significantly in recent years, though it remains skewed towards the Iranian standard. The SentiPers corpus (Hosseini et al., 2018) established a baseline for polarity detection. However, modern evaluations require larger, diverse sources. An openly available dataset (Khayati, 2021) for persian sentiment analysis was collected from the popular iranian eccormerce website, Digikala. Another valuable resource provides comments from the food delivery service, Snappfood for sentiment analysis. The introduction of FarsTail (Amirkhani et al., 2023) for NLI and PQuAD (Darvishi et al., 2023) for

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 98–104

reading comprehension provided the first standardized benchmarks for reasoning in Persian. These datasets, however, are exclusively in Iranian Persian, limiting their utility for cross-variant assessment.

2.2 Multilingual LLM Evaluation

Evaluations of MLLMs such as BLOOM (Workshop et al., 2022) and Qwen (Bai et al., 2023) often report aggregate metrics on massive benchmarks (e.g., MMLU). While these models include Persian in their training data, detailed breakdown by dialect is rarely provided. Recent works have highlighted the "curse of multilinguality," (Chang et al., 2024) where increasing the number of languages can dilute performance on low-resource variants. Fine-tuned approaches, such as the *Gemma-Persian* model (Shojaei, 2025) adapted from Google's Gemma family, attempt to mitigate this by focusing on specific language families. Our work contextualizes these models within the specific linguistic constraints of the Iranian family.

3 Experimental Setup

3.1 Models

We selected five open-weight models based on their accessibility and claimed multilingual support. We chose moderate sized LLMs for future deployability in edge-devices. All models were loaded using 4-bit quantization (NF4) on an NVIDIA RTX 4090 GPU to simulate resource-constrained environments, with the exception of Gemma-3, which utilized bfloat16 precision based on architecture requirements.

1. **Qwen2.5-Instruct (1.5B & 3B):** Alibaba's diverse multilingual models (Team, 2024), known for strong instruction following.
2. **Bloomz (1.7B & 3B):** The instruction-tuned variants (Muennighoff et al., 2023) of the BigScience BLOOM model, trained on xP3.
3. **Gemma-3-4b-persian:** A specialized fine-tune of Google's Gemma-3, specifically optimized for Persian instruction following (Shojaei, 2025).

3.2 Data Processing and Prompts

We utilized a unified data loading pipeline to process raw datasets into a standardized JSONL format.

Preprocessing: Text normalization was applied to handle common Persian orthographic variations (e.g., unifying Arabic/Persian *Ye* (ی) and *Kaf* (ک), normalizing zero-width non-joiners). For Sentiment Analysis, we mapped diverse label spaces (e.g., 5-star ratings, -2 to +2 scales) to a unified {Negative, Neutral, Positive} schema.

Prompt Engineering: We utilized English-language zero-shot prompts for all tasks to maintain consistency across the multilingual models. However, for the Machine Translation task, we explicitly conditioned the models on the specific variant. The prompts followed the template: *"Translate the following [Source Language] text to English"*, where [Source Language] was dynamically populated as "Dari", "Tajik", or "pes" (Standard Persian) based on the FLORES-200 subset. This ensures that performance degradation on variants (e.g., Tajiki) is due to model capability gaps rather than ambiguous instructions. English prompts were chosen for cross-model consistency since all models are instruction-tuned in English, but Persian-language prompts could yield different results. (See Appendix A for full templates).

3.3 Tasks and Datasets

We construct an evaluation benchmark spanning over 240,000 samples across four tasks, from which 1,000 test instances per task are sampled for model evaluation. Details of the datasets, their sources, and the specific variants covered are detailed in Table 1. For deterministic evaluation considering time and computational constraints, we sampled a stratified test set of $N = 1000$ per task (Seed=42).

Task	Dataset	Domain	Total Size	Variant
Sentiment	SentiPers (Hosseini et al., 2018)	Digital Reviews	15.6k	Iranian
	Digikala (Khayati, 2021)	E-commerce	98.4k	
	SnappFood (Farahani et al., 2021)	Food Delivery	70.0k	
MT	FLORES-200 (Team et al., 2022)	Wiki/News	3.0k	Pes/Prs/Tgk
	Tatoeba (Tiedemann, 2020)	General	13.7k	Iranian
NLI	FarsTail (Amirkhani et al., 2023)	General	10.3k	Iranian
QA	PQuAD (Darvishi et al., 2023)	Wikipedia	60.3k	Iranian

Table 1: Summary of datasets used in this study. "Pes", "Prs", and "Tgk" refer to Iranian Persian, Dari, and Tajiki respectively. Total Size refers to available samples before test split sampling.

3.4 Evaluation Metrics

For classification tasks, we use Accuracy and Macro-F1. We incorporated a strict parsing logic: model outputs that did not strictly match the label space were categorized as "unknown," penalizing the accuracy score. For QA, we report Exact Match (EM) and F1. For MT, we report BLEU and chrF scores using `sacrebleu`. chrF is particularly im-

portant for Persian due to its agglutinative morphology, where token-based BLEU may penalize valid variations.

4 Results & Analysis

4.1 Overall Performance (RQ1)

Table 2 summarizes the performance across all tasks. Addressing **RQ1**, the fine-tuned **Gemma-3-4b-persian** demonstrates superior performance compared to general-purpose baselines, achieving the highest scores in Sentiment (77.3%), QA (41.0 EM), and MT (23.3 BLEU). This suggests that for languages with high morphological richness like Persian, general multilingual pre-training is perhaps less effective than targeted fine-tuning.

4.2 Sentiment and NLI Analysis

In Sentiment Analysis (Figure 1), the Qwen and Gemma models performed consistently well. Notably, a source breakdown revealed that models performed best on SnappFood (short, informal reviews) compared to SentiPers (formal). For example, `Qwen2.5-3B` achieved 78.7% on SnappFood but only 61.8% on SentiPers, suggesting modern LLMs are better aligned with web-style informal text.

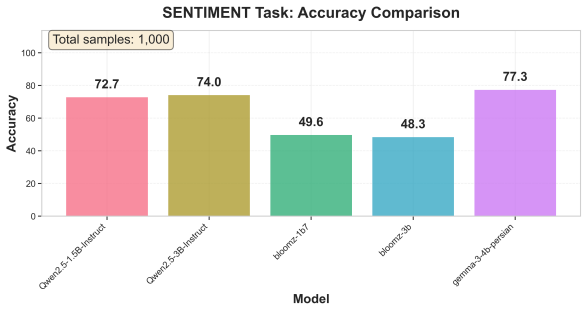

Figure 1: Sentiment Accuracy by Model. Gemma and Qwen significantly outperform the Bloomz baselines.

To better understand the failure modes of these models, we analyzed the specific error types in the sentiment task, as shown in Figure 2. The stacked bar chart reveals a distinct pattern: while the `gemma-3-4b-persian` model maintains a relatively low error count overall, the baseline models (particularly Bloomz) exhibit a massive disparity. The high "parsing failure" rate in older models is absent in Qwen and Gemma, yet the latter still struggle with distinguishing between *Neutral* and *Negative* sentiment. This typically occurs in Persian reviews where polite formalities (Taarof) often mask negative feedback, confusing models that lack deep cultural fine-tuning.

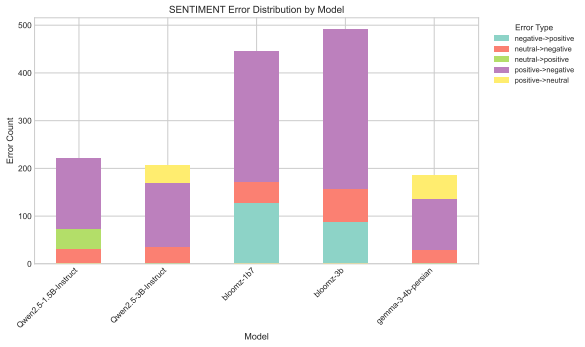

Figure 2: Distribution of top error types in Sentiment Analysis. While newer models avoid formatting errors, conflation of Neutral and Negative classes remains a persistent linguistic challenge.

The Bloomz family struggled significantly with NLI, yielding accuracy scores near the random baseline (33%). Analysis of the logs reveals a high parsing failure rate for Bloomz (up to 1.8%), indicating that the models often failed to generate the specific labels requested, instead hallucinating continuations of the premise.

4.3 The Script Barrier: Tajiki Robustness (RQ2)

Addressing **RQ2**, our findings indicate a severe cross-script degradation of models on the Tajiki variant. As illustrated in Figure 3, there is a stark disparity between performance on Perso-Arabic scripts (Persian, Dari) and Cyrillic (Tajiki).

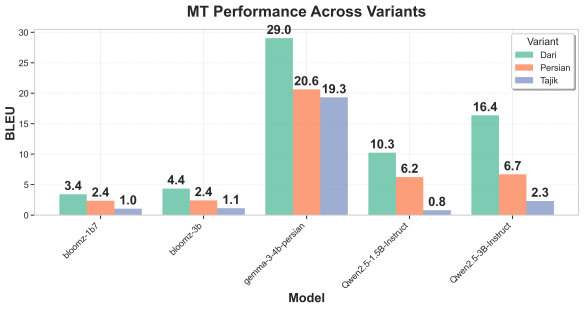

Figure 3: Performance across Persian variants (MT Task). Note the near-zero performance on Tajiki for Qwen and Bloomz.

For the `Qwen2.5-1.5B` model, performance drops from 6.2 BLEU on Persian to 0.8 BLEU on Tajiki. This implies that the model probably treats Tajiki as an entirely foreign language, despite it being linguistically identical to Persian in syntax and morphology. However, it is to be noted that we did

Model	Sentiment		NLI		MT (Avg)		QA	
	Acc	F1	Acc	F1	BLEU	chrF	EM	F1
`Bloomz-1b7`	49.6	35.3	33.9	20.3	2.4	21.2	10.1	23.6
`Bloomz-3b`	48.3	33.8	35.4	24.8	2.7	22.6	17.5	32.5
`Qwen2.5-1.5B`	72.7	57.9	56.2	51.0	4.6	32.0	10.8	35.2
`Qwen2.5-3B`	74.0	62.9	**63.9**	**59.3**	7.6	37.4	6.1	37.4
`gemma-3-4b-persian`	**77.3**	**67.8**	49.4	44.2	**23.3**	**50.8**	**41.0**	**68.4**

Table 2: Main results on Persian evaluation tasks. `gemma-3-4b-persian` dominates in generation tasks (MT, QA) and Sentiment, while `Qwen2.5-3B` shows strong reasoning capabilities in NLI.

not conduct a controlled transliteration experiment to disentangle whether the observed performance gap is primarily due to script mismatch (Cyrillic tokenization) or genuine dialectal divergence. The `gemma-3-4b-persian` model is the outlier, maintaining a score of 19.3 BLEU on Tajiki. This suggests that the base Gemma model's pre-training may have included Cyrillic data (perhaps Russian or Central Asian text) that allows for cross-script transfer, a capability preserved during fine-tuning.

4.4 Question Answering Capabilities

The QA task (PQuAD) proved most difficult. The Qwen models, despite good instruction following, struggled with the extractive nature of the task. They often paraphrased the answer rather than extracting the span, leading to low Exact Match scores.

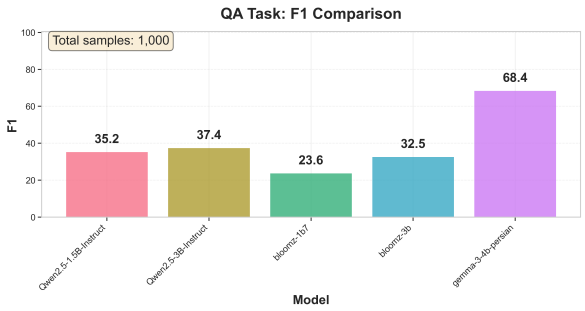

Figure 4: F1 Scores on PQuAD Question Answering. The fine-tuned Gemma model shows superior span extraction capabilities compared to general multilingual baselines.

As shown in Figure 4, textttgemma-3-4b-persian achieved a dominating F1 score of 68.4, nearly double that of the nearest competitor. This indicates that the fine-tuning process for Gemma likely included Persian reading comprehension tasks, aligning it better with the cultural context and structural requirements of PQuAD.

4.5 The Dari Divergence: Lexical vs. Structural (RQ3)

Addressing **RQ3**, qualitative analysis reveals that Dari divergence is primarily lexical rather than structural, stemming from English and Pashto loanwords (e.g., نوتهوپ (*Pohantoon*) for University) versus Iranian Persian's French influence. As detailed in Table 3, weaker models like **Bloomz** struggled with these dialectal markers, frequently hallucinating (e.g., misinterpreting *Pohantoon* as unrelated entities). In contrast, **Qwen2.5** and `gemma-3-4b-persian` demonstrated robust zero-shot generalization. Additionally, while the fine-tuned Gemma model respected Dari orthography (e.g., *mekonand*), base models often over-corrected text to the Tehrani standard during generation, highlighting a persistent training bias despite high aggregate BLEU scores.

Concept	Iranian Persian	Dari (Target)	Model Handling
University	Dāneshgāh (هاگشناد)	Pohantoon (پوهنتون)	**Gemma:** ✓ | **Bloomz:** ×
Company	Sherkat (شرکت)	Kompani (کمپنی)	**All Models:** ✓
Technology	Fanāvari (فناوری)	Teknālozhi (تکنالوژی)	**Qwen:** ✓ | **Bloomz:** ×
Policy	Siāsat (سیاست)	Pālisi (پالیسی)	**Gemma:** ✓ | **Qwen:** ✓

Table 3: Qualitative analysis of lexical divergence in Dari. Stronger models (Gemma/Qwen) successfully bridge the lexical gap, while weaker baselines (Bloomz) suffer from hallucinations when encountering non-Iranian specific vocabulary (e.g., Pashto loanwords like *Pohantoon*).

5 Reproducibility

To encourage future work and enable other researchers to replicate our results, we release the code for this manuscript in this link.

6 Limitations and Future Work

This study has several limitations that should be considered when interpreting our findings. All experiments used English zero-shot prompts for consistency; Persian-language prompts may yield

different results. Our model selection was limited to small open-weight models (1.5B–4B), and the absence of larger or proprietary baselines limits its generalizability. The `gemma-3-4b-persian` model used bfloat16 precision while others used NF4 quantization, which may partially explain its performance advantage. Our Tajiki results conflate script mismatch with dialectal divergence; a controlled transliteration experiment would be needed to disentangle these factors, and our findings should be read as indicating insufficient Cyrillic coverage rather than a definitive failure to understand Tajiki itself. Finally, results are from a single 1,000-sample split without confidence intervals, and bootstrap resampling would strengthen statistical reliability.

7 Acknowledgment

We would like to extend our gratitude to the Research and Innovation Lab, Department of Computer Science and Engineering, University of Asia Pacific for graciously providing ample computational resources for this research work.

8 Conclusion

Our investigation reveals a stark dichotomy in the landscape of Persian NLP within the open-weight model ecosystem: while fine-tuning has successfully adapted LLMs to the Iranian standard, a formidable "script barrier" appears to isolate Tajiki speakers, and dialectal nuances in Dari remain under-utilized. Since this study focuses on smaller open-weight LLMs, further work is required to determine if these limitations persist in larger, proprietary models. The superior performance of `gemma-3-4b-persian` underscores that generic multilingual pre-training is likely insufficient for complex tasks like QA; rather, achieving true linguistic inclusivity across the Iranian continuum requires curated, multi-script instruction tuning that treats these variants not as separate low-resource languages, but as a unified linguistic heritage.

References

Hossein Amirkhani, Mohammad AzariJafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, Zohreh Pourjafari, and Azadeh Amirak. 2023. Farstail: A persian natural language inference dataset. *Soft Computing*, pages 1–13.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. 2023. Pquad: A persian question answering dataset. *Computer Speech & Language*, 80:101486.

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

Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. 2018. Sentipers: A sentiment analysis corpus for persian. *arXiv preprint arXiv:1801.07737*.

Armin Khayati. 2021. Digikala-comments-sentiment-analysis. `https://github.com/Arminkhayati/Digikala-comments-sentiment-analysis`. GitHub repository, accessed 14 January 2026.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Mohammad Shojaei. 2025. Gemma 3-4B Persian (v0) Model. `https://huggingface.co/mshojaei77/gemma-3-4b-persian-v0`. Hugging Face model repository, accessed 15 January 2026.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jörg Tiedemann. 2020. The tatoeba translation challenge—realistic data sets for low-resource and multilingual mt. *arXiv preprint arXiv:2010.06354*.

Gernot Windfuhr. 2009. *The Iranian Languages*, volume 2009. Routledge London.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

A Prompt Templates

We utilized the following zero-shot prompts for our experiments. For the Machine Translation task, the prompt was dynamically adjusted based on the specific variant (Persian/Dari/Tajik) in the FLORES-200 dataset to ensure models were not penalized for script confusion. It is to be noted that English prompts bias toward instruction-tuned models.

Task	Template
Sentiment	Classify the sentiment of the following Persian text as one of: negative, neutral, positive. Text: {text} Respond with only the sentiment label, nothing else. Sentiment:
MT	Translate the following {source_lang} text to {target_lang}. {source_lang} text: {text} {target_lang} translation:
NLI	Given the premise and hypothesis below, determine the relationship between them. Choose one of: entailment, neutral, contradiction. Premise: {premise} Hypothesis: {hypothesis} Respond with only the label, nothing else. Relationship:
QA	Answer the question based on the given context. Extract the answer directly from the context. Context: {context} Question: {question} Answer:

Table 4: Zero-shot prompt templates used for evaluation.

B Reproducibility Details

Experiments were conducted using 4-bit quantization (NF4) for Qwen and Bloomz models to simulate consumer-grade hardware constraints. We utilized python 3.10.19 for running our scripts. The Gemma-3 model was loaded in bfloat16 precision as it was an architectural requirement for loading Gemma. We utilized the `bitsandbytes` python package for quantization and the `transformers` python pacakge for inference. Deterministic generation was enforced by setting the temperature to 0.0 and `do_sample=False` for all tasks. We acknowledge that a single 1,000-sample split without confidence intervals is a limitation, especially for noisy MT metrics. Bootstrap confidence intervals should be adopted in future works.

C Detailed Experimental Results

C.1 Sentiment Analysis by Domain

Table 5 presents the performance breakdown across the three source datasets used in the aggregated Sentiment Analysis task.

- **SnappFood:** Short, informal food delivery reviews.
- **Digikala:** Product reviews, semi-formal/colloquial.
- **SentiPers:** Formal/Academic corpus.

The results indicate that general-purpose multilingual models (Qwen, Bloomz) struggle significantly with the formal register of SentiPers (e.g., Qwen2.5-1.5B drops from 76.6% on SnappFood to 48.0% on SentiPers). In contrast, the fine-tuned **gemma-3-4b-persian** maintains high robustness (74.5% on SentiPers), suggesting successful alignment across different registers of Iranian Persian. A visual depiction of the LLMs' performance across different datasets is shown in Figure 5.

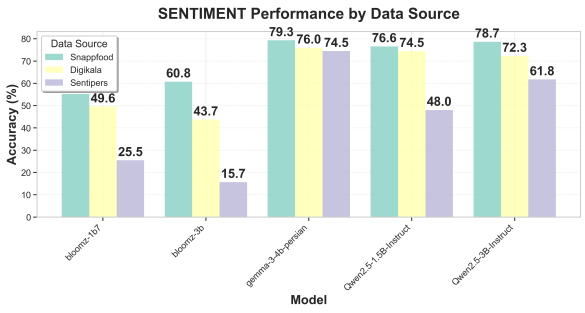

Figure 5: Performance of Multilingual LLMs across datasets of different nature, indicating more recent LLMs are better adept to handle web data.

C.2 Machine Translation by Variant and Source

Table 6 presents a granular breakdown of BLEU scores across language variants and data sources.

Model	**SnappFood** (Informal)		**Digikala** (Semi-Formal)		**SentiPers** (Formal)	
	Acc	F1	Acc	F1	Acc	F1
`Bloomz-1b7`	55.3	37.1	49.6	34.6	25.5	19.8
`Bloomz-3b`	60.8	39.7	43.7	30.7	15.7	12.9
`Qwen2.5-1.5B`	76.6	51.4	74.5	56.8	48.0	41.0
`Qwen2.5-3B`	78.7	52.6	72.3	55.6	61.8	55.7
`gemma-3-4b-persian`	**79.3**	**53.4**	**76.0**	**62.1**	**74.5**	**70.0**

Table 5: Detailed performance breakdown on Sentiment Analysis datasets. The *SentiPers* dataset proved most challenging for base models, while the fine-tuned Gemma model demonstrated cross-domain robustness.

Model	**Variant Performance (BLEU)**			**Source Breakdown (BLEU)**	
	Persian	**Dari**	**Tajiki**	**FLORES-200**	**Tatoeba**
`Bloomz-1b7`	2.4	3.4	1.0	2.5	0.2
`Bloomz-3b`	2.4	4.4	1.1	2.9	0.1
`Qwen2.5-1.5B`	6.2	10.3	0.8	5.0	0.1
`Qwen2.5-3B`	6.7	16.4	2.3	9.1	0.1
`gemma-3-4b-persian`	**20.6**	**29.0**	**19.3**	**25.0**	**0.2**

Table 6: Detailed Machine Translation results. The left section highlights the "Script Barrier" where performance collapses on Tajiki for non-fine-tuned models. The right section shows the model performance averaged across variants on specific datasets.

Three critical patterns emerge from this data:

The Script Barrier (Tajiki): The most salient result is the catastrophic failure on Tajiki. Although Tajiki and Persian are highly mutually intelligible and share core syntax and morphology, the Cyrillic script poses a major barrier for base multilingual models. Both `Qwen2.5-1.5B` and Bloomz collapse to near-zero performance (< 1.1 BLEU), effectively treating Tajiki as unseen. In contrast, the fine-tuned gemma-3-4b-persian remains usable at 19.3 BLEU, indicating that targeted instruction tuning can bridge the Perso-Arabic–Cyrillic script gap where general multilingual pre-training fails.

The Dari Inversion: Contrary to expectations, models often perform better on Dari than on Standard Persian (e.g., Gemma-3: 29.0 vs. 20.6 BLEU). Qualitative analysis suggests this is not due to better dialectal modeling, but to test-set artifacts: the Dari subset of FLORES-200 contains simpler, less metaphorical sentences that models translate more literally and accurately.

Domain Brittleness (Tatoeba vs. FLORES): We observe a variation between data sources. While models achieve respectable performance on FLORES-200 (Wiki/News domain), they fail almost completely on the Tatoeba dataset (averaging < 0.2 BLEU) (Figure 6). Tatoeba consists largely of short, context-free, and idiomatic sentences. The models' inability to handle these samples indicates a high sensitivity to domain and sentence length; they struggle to ground short, ambiguous text without the extensive context provided in Wikipedia-style paragraphs.

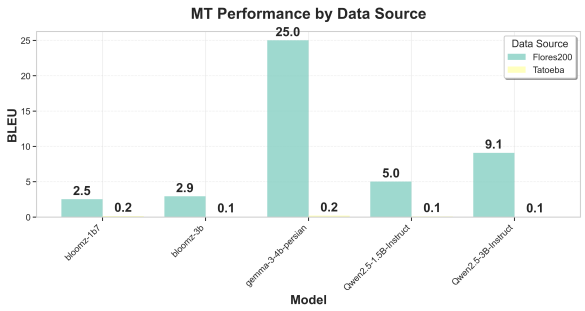

Figure 6: Comparison of Machine Translation performance (BLEU) across data sources. While models achieve reasonable scores on the standard FLORES-200 benchmark, they exhibit near-total failure on the Tatoeba dataset, highlighting a lack of robustness to diverse sentence structures and community-sourced data.