Building and Evaluating a High Quality Parallel Corpus for English Urdu Low Resource Machine Translation

Munief Hassan Tahir, Hunain Azam, Sana Shams, Sarmad Hussain
Center for Language Engineering, Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology, Lahore, Pakistan
munief.hassan@kics.edu.pk, l216202@lhr.nu.edu.pk,
sana.shams@kics.edu.pk, sarmad.hussain@kics.edu.pk

Abstract

Low-resource languages like Urdu suffer from limited high quality parallel data for machine translation. We introduce a curated English–Urdu corpus of 80,749 high-fidelity sentence pairs across 18 diverse domains, built via ethical collection, manual alignment, deduplication, and strict length-based filtering (AWCD $\leq$ 5). The corpus is converted into a bidirectional SFT dataset with bilingual (English/Urdu) instructions to enhance prompt-language robustness. Fine-tuning Llama-3.1-8B-Instruct (Llama-FT) and UrduLlama 1.1 (UrduLlama-FT) yields major gains over the baseline. sacreBLEU scores reach 24.65–25.24 (En→Ur) and 76.14–77.97 (Ur→En) for Llama-FT, with minimal sensitivity to prompt language. Blind human evaluation on 90 sentences per direction confirms substantial perceptual improvements. Results demonstrate the value of clean parallel data and bilingual instruction tuning, revealing complementary benefits of general SFT versus Urdu specific pretraining. This work provides a reproducible resource and pipeline to advance Urdu machine translation and similar low-resource languages.

1 Introduction

In an increasingly interconnected world, machine translation (MT) has become a fundamental component of natural language processing (NLP), facilitating smooth cross-lingual communication. Low-resource languages like Urdu, one of the most widely spoken languages in the world with over 230 million speakers (SIL International, 2022), remain woefully underserved, whereas high-resource language pairs like English–French or English–Chinese have profited from decades of research and enormous parallel corpora. Modern neural machine translation systems face significant obstacles due to Urdu's intricate morphology, right-to-left Nastaliq script, rich code-mixing with Persian and Arabic loanwords, and scarcity of high-quality parallel data. The public resources currently available for translating between English and Urdu are either domain-restricted, noisily aligned, or insufficient in scale, which leads to models that are culturally misaligned, have poor generalization, and have lexical sparsity. This study fills these gaps by presenting a large-scale curated English–Urdu parallel corpus that includes 80,749 high-fidelity sentence pairs from 18 standardized domains. The dataset guarantees linguistic equivalency, traceability, and suitability for neural MT architecture training and evaluation. It is constructed through a rigorous pipeline of ethical data acquisition, structural and manual sentence alignment, and multi-stage preprocessing. In addition to being larger and cleaner than previous English–Urdu resources, this corpus creates a repeatable framework for creating domain-balanced, ethically sourced datasets for other low-resource languages.

2 Related Work

FConv-NN introduces a simple yet effective method to improve Urdu–English (UR-EN) neural machine translation in low-resource settings (Israr et al., 2024). The model was trained and evaluated using FAIRSEQ, an open-source PyTorch toolkit, on a system with an Intel Core i9-9900K CPU (3.60 GHz) and an NVIDIA GeForce GTX 1650 SUPER GPU. Experiments on a 100K Urdu–English corpus (90K train, 5K validation, 5K test) showed the FConv-NN model achieved a BLEU score of 40.22, a 18.42 point gain over baseline CNN and a BLEURT score of 0.565, outperforming CNN and CNN-ADL across all metrics. It also produced more fluent translations than Google and Bing. The model enables efficient, real-time translation in low-resource settings but still trails RNN-based models, highlighting the need for improved hybrid architectures.

The English–Urdu NMT model employs a three-

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 102–110

layer LSTM encoder-decoder with a sequence-to-sequence architecture, integrating preprocessing steps (noise removal, tokenization, POS-tagging) and Word2Vec Skip-Gram embeddings to manage morphological variations and out-of-vocabulary words without segmentation (Andrabi and Wahid, 2022). Trained with the Adam optimizer and Soft-Max loss on a cleaned parallel corpus prepared via Python tokenization and Selenium web scraping, the model used a fixed sequence length of 15, achieving optimal performance at sequence length 10. It attained BLEU scores of 50.86 (training) and 47.06 (testing), showing strong context retention for short sequences. Human evaluations rated translations 53% excellent and 38% good, confirming the model's fluency and reliability for English–Urdu translation.

The English–Urdu NMT system uses a six-layer LSTM encoder-decoder with Bahdanau attention and GloVe embeddings to enhance context retention and translation quality (Kumhar et al., 2022). Extensive preprocessing—truecasing (English), Unicode normalization, non-printable character removal, and sentence padding—addresses Urdu's complex morphology and script alignment. Trained on Google Colab using the Adam optimizer and categorical cross-entropy, the system utilized a parallel corpus of 1,083,734 tokens (542,810 English, 540,924 Urdu) from religious texts, web-scraped news, and daily-use phrases, split 70:30 for training and testing. It achieved an average BLEU score of 45.83, though accuracy remains limited in specialized domains (e.g., health, tourism, business) and the system supports text-only translation, lacking speech input.

Expectation Maximization (EM) based transliteration is an unsupervised, language-independent technique proposed to enhance Urdu-to-English translation by learning transliteration patterns and handling out-of-vocabulary (OOV) words without a separate dataset (Mohy ud Din, 2019). Using the UMC005 Quran-based parallel corpus of 6,414 sentence pairs, the approach achieved BLEU score improvements of 0.63–0.91 for SMT and 1.28–2.05 for NMT, with the Transformer model outperforming LSTM (11.61 vs. 9.08). Tokenization and preprocessing further improved results by +3.5 BLEU points. However, the study was limited by the lack of a large, high-quality Urdu–English parallel corpus, emphasizing the need for better data resources and exploration of unsupervised translation methods for low-resource languages like Urdu.

A model for English to Urdu and Hindi machine translation using translation rules and artificial neural networks presents a hybrid multilingual translation framework that integrates rule-based methods with artificial neural networks (ANN) to translate English text into Urdu and Hindi (Khan and Usman, 2019). The system employs translation rules and bilingual dictionaries within an encoder-decoder architecture implemented in Java and MATLAB, where linguistic rules and tokens are numerically encoded for neural processing. Trained on approximately 465 grammar rule pairs and 9,000 bilingual word entries for each language pair, the model achieved strong results, with BLEU 0.6054, METEOR 0.8083, and F-score 0.8250 for Urdu translations. The study emphasizes that expanding grammar rules, enriching the bilingual dictionary, and refining case marking are key to enhancing accuracy in Urdu and Hindi translation.

The study "Linguistics Knowledge-Driven Multi-Task (LKMT) Neural Machine Translation for Urdu and English" (Hassan et al., 2024) presents a novel pre-trained model that enhances Urdu–English translation by integrating linguistic knowledge such as part-of-speech (POS) and dependency (DEP) features into a Transformer-based architecture. Trained on a large monolingual corpus of 5,464,575 sentences and fine-tuned using the Tanzil and religious Urdu–English parallel corpora, the model effectively captures deeper lexical and syntactic relationships. Experimental results show BLEU score improvements of +1.97 for Urdu→English and +2.42 for English→Urdu over previous models like XLM and mBART. While demonstrating strong performance for low-resource languages, the study notes that limited high-quality Urdu parallel data remains a key challenge. Future work aims to incorporate richer linguistic and semantic features to further enhance translation quality and domain adaptability.

The paper "Enriching Source for English-to-Urdu Machine Translation" (Jawaid et al., 2016) proposes adding artificial case markers (pseudo-words) to the English source text to improve phrase-based SMT performance when translating into Urdu, a morphologically rich and free word-order language. Using the Moses framework, GIZA++, and a 5-gram SRILM model, the system was trained on the "ALL" parallel corpus and supported with Urdu monolingual data. By preordering English sentences to match Urdu syntax and inserting pseudo-markers to represent grammatical roles, the

approach improved alignments and translation accuracy, achieving up to +1 BLEU gain over the baseline. However, occasional over-generation of markers caused inconsistent results, indicating a need for further refinement.

The paper "Advancing Roman Urdu to Urdu Transliteration using Machine Learning Techniques" (Ahmad and Ahmad, 2024) uses a diverse dataset of 6.5 million sentences from social media, messaging, and poetry to train and evaluate transliteration models. The authors compared Seq2Seq, RNN+LSTM, and Tensor2Tensor attention-based models, finding the Transformer-based approach achieved the best performance with a BLEU score of approximately 75. Training was conducted on an Alienware 15R2 laptop with a Core i7 processor, 32 GB RAM, and 8 GB GPU memory, using four attention layers, dropout, and subword tokenization for effective contextual understanding. While the Transformer model accurately handled complex, compound, and long sentences, limitations included occasional errors with rare or unseen words and reliance on high computational resources, highlighting the importance of data diversity, model architecture, and context-aware design in Roman Urdu transliteration.

The paper "Low-Resource Transliteration for Roman-Urdu and Urdu Using Transformer-Based Models" (Butt et al., 2025) addresses transliteration challenges between Roman-Urdu and Urdu using a transformer-based m2m100 model. The authors employ two datasets: the large-scale Roman-Urdu-Parl (RUP) corpus with 6.3 million sentence pairs and the smaller, domain-specific Dakshina dataset with 10,000 sentences. Their methodology includes Masked Language Modeling (MLM) pretraining on monolingual Roman-Urdu and Urdu text to improve subword-level generalization, followed by fine-tuning for direct transliteration with language-specific tokens. Models were trained using standard transformer settings, including batch sizes of 64–128, learning rate 1e-5, and gradient accumulation, on hardware provided by DFKI. Results show that the fine-tuned m2m100 model achieves Char-BLEU scores up to 97.44 for Roman-Urdu $\rightarrow$ Urdu and 96.37 for Urdu $\rightarrow$ Roman-Urdu, outperforming previous RNN baselines and GPT-4o Mini in zero-shot evaluation. Limitations include inherent ambiguity in Roman-Urdu spelling, trade-offs between domain adaptation and retention of previously learned patterns, and reliance on parallel corpora, highlighting potential avenues for multi-reference evaluation or semi-supervised approaches.

The paper "Urdu-to-English-Based Unsupervised Machine Translation" (Raza et al., 2024) addresses the challenge of translating between a low-resource language pair with significant structural differences, where Urdu follows a Subject-Object-Verb (SOV) order and English uses Subject-Verb-Object (SVO). To overcome the scarcity of parallel corpora, the authors proposed an unsupervised neural machine translation model trained solely on monolingual data from the Tanzil Corpus. The system employs a Transformer-based architecture with a shared encoder, leveraging cross-lingual embeddings, denoising autoencoders, and on-the-fly backtranslation to optimize translation quality. Experimental results showed that the unsupervised approach achieved a BLEU score of 5.21, while a supervised variant reached 13.85, indicating improvements but also highlighting the challenges posed by structural differences, idiomatic expressions, and morphological complexity.

The Transtech system (Masroor et al., 2019) is a rule-based Roman Urdu to English translator designed to handle variable spelling and grammar. It operates in three phases: a scanner for tokenization and spell checking, a POS tagger using an LL(1) parser and context-free grammar, and a translator module for semantic analysis and sentence generation. The system uses a self-constructed corpus of over 3,000 words and 2,000 sentences, supported by a knowledge base storing linguistic and syntactic information. Evaluation against Google Translator showed Transtech provides more accurate translations, especially for word order, tense, and pronouns. Limitations include rare/unseen words, complex sentences, and dependency on the manually curated knowledge base, with future improvements suggested through corpus expansion and machine learning integration.

3 Data Collection and Preprocessing

This section outlines the methodology for constructing a high-quality English-Urdu parallel corpus through systematic data collection, sentence-level alignment, and rigorous preprocessing. The process ensures linguistic equivalence, data integrity, and suitability for downstream NLP tasks, particularly machine translation.

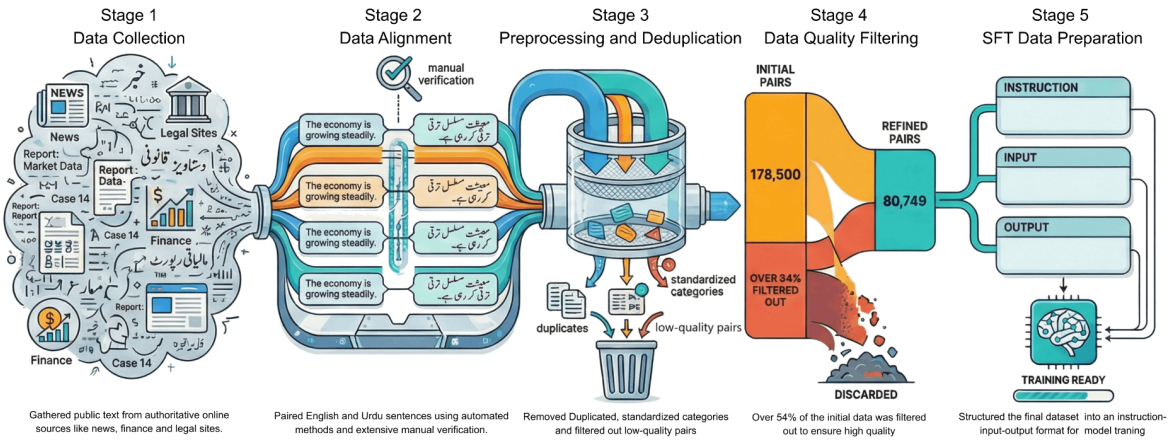

Figure 1: Data Processing Pipeline

3.1 Data Collection

We collected bilingual textual content from publicly accessible, authoritative online sources spanning diverse domains, including news outlets, agricultural extensions, financial institutions, culinary archives, telecommunications providers, legal repositories, and historical records. Priority was given to reputable platforms offering verifiable parallel content.

Using automated, non-intrusive pipelines, we extracted only publicly available textual material such as articles, reports, guidelines, and descriptive passages while avoiding restricted or login-protected sections. All retrieved content was cataloged with source metadata and stored in a structured repository, adhering to ethical standards for open-domain data acquisition.

3.2 Data Alignment

Parallel documents were pre-paired with explicit filename correspondence. Sentence-level alignment began automatically by exploiting structural cues (e.g., punctuation, paragraph breaks, and formatting markers) to generate initial English-Urdu sentence pairs, which were consolidated into a central corpus.

Subsequently, a designated annotator conducted manual pairwise verification to resolve alignment ambiguities caused by structural variations, cultural adaptations, or content omissions/additions in Urdu or English translations. A final exhaustive validation was then performed on the consolidated corpus to ensure strict one to one sentence correspondence, eliminating discrepancies and yielding high fidelity parallel pairs.

3.3 Data Preprocessing and Deduplication

The aligned corpus initially comprising 178,500 sentence pairs from verified sources underwent systematic preprocessing to eliminate noise, standardize formatting, and enhance translation quality. This step ensured the final dataset was clean, consistent, and optimized for bilingual modeling tasks.

3.3.1 Merging and Initial Analysis

The aligned sentence pairs from all sources were consolidated into a single corpus containing 178,500 parallel entries, with fields `English`, `Urdu`, and `Category`. Basic descriptive statistics including total rows, column structure, and non-null counts per field were computed. Empty rows (where every field was missing) arose occasionally during source ingestion and were removed, yielding a negligible reduction in size.

3.3.2 Category Normalization

The `Category` field, originally assigned by the data collector for each source, exhibited inconsistencies such as duplicate labels (e.g., “News” vs. “News ”), trailing whitespace, and typographical errors (e.g., “Relgion”). Normalization was essential to

Table 1: Distribution of Standardized Categories in the Final Dataset

Category Group	Count
National News	49,516
International News	8,898
Law & Judiciary	6,484
Food & Recipes	4,941
Banking	2,576
Agriculture	2,277
Sports	2,042
Telecom	1,766
Religion	615
Health	521
Business	326
Science	268
Corporate & Business	173
Anti-Drug Education	154
Showbiz	73
Daily Life & Communication	54
Weather	41
History & Arts	24

enforce a uniform taxonomy across the entire corpus, thereby enabling reliable domain-based analysis and stratified sampling. Domain identification was performed on sentences belonging to the "News" category using the Domain Identification API.[1] Similar categories were collapsed into 18 standardized groups while preserving thematic integrity. The resulting distribution, after all subsequent preprocessing steps, is reported in Table 1.

3.3.3 Identifier Assignment

To maintain traceability throughout the pipeline, unique identifiers were introduced at two stages. A `Raw_Sentence_ID` (RSID000003–RSID178450) was assigned to every entry in the merged raw corpus. After completion of the cleaning stages described in the following subsections (deduplication, filtering, and quality assurance), a `Cleaned_Sentence_ID` (CSID000003–CSID118482) was allocated to each retained high-quality pair, ensuring unambiguous reference in downstream modeling and evaluation.

3.3.4 Deduplication and Filtering

Duplicates were removed in two stages: (1) exact matches on both `English` and `Urdu` (removing 56,434 rows), and (2) duplicates on `English` with varying `Urdu` (removing 1,703 rows, retaining the first occurrence). Sentences with two or fewer words in English were filtered out (1,880 rows), as they typically represent fragments unsuitable for translation tasks.

3.4 Data Quality Filtering

Word counts were calculated for each sentence: `EWC` (English Word Count) and `UWC` (Urdu Word Count), using whitespace tokenization. The absolute difference, $\texttt{AWCD} = |\texttt{EWC} - \texttt{UWC}|$, was computed to gauge translation fidelity, assuming well-aligned translations exhibit similar lengths.

Statistics on the post-filtering dataset (80,749 pairs) include: average EWC of 16.20, average UWC of 17.83, and average AWCD of 2.24 (Table 2).

Table 2: Key Statistics of Final Dataset

Metric	Value
Total Sentence Pairs	80,749
Unique Categories	18
Average EWC	16.20
Average UWC	17.83
Average AWCD	2.24

To ensure the highest possible translation quality for downstream modeling, we adopted a conservative filtering strategy based on AWCD. Sentences with AWCD ≤ 5 were classified as high-fidelity pairs and retained without modification, as this threshold captures translations with closely matching lengths, which strongly correlates with accurate semantic alignment in parallel corpora.

Sentences with AWCD > 5 were excluded from the final dataset. While some pairs in this range may still represent valid translations (e.g., due to legitimate structural differences between English and Urdu), higher AWCD values introduce greater risk of misalignment, partial translations, or added/omitted content. Given the priority of producing a clean, high-precision corpus suitable for training and evaluating machine translation systems where even moderate misalignment can degrade performance. We opted to prioritize precision by retaining only the most reliable pairs.

This conservative approach yielded 80,749 high-quality sentence pairs for the final dataset. The distribution of retained and excluded pairs is shown in Table 3.

[1] https://tech.cle.org.pk/domainidentification

Table 3: Distribution of Translation Quality Groups in the Final Dataset

Status	Count	Percentage
Retained (AWCD ≤ 5)	80,749	60.9%
Excluded (AWCD > 5)	51,920	39.1%

This preprocessing pipeline ensures a clean, structured corpus ready for linguistic review and NLP applications.

3.5 Supervised Fine-Tuning Data Preparation

To prepare the corpus for supervised fine-tuning (SFT) of instruction tuned large language models, we transformed the high quality parallel sentence pairs into an instruction-following format consisting of three fields: `instruction`, `input`, and `output`.

We first split the retained high fidelity pairs into training and testing sets using a 90:10 ratio, ensuring stratified sampling across the standardized categories to preserve domain diversity in both splits.

To fully exploit the bidirectional nature of the parallel data, we augmented the dataset by creating separate examples for each translation direction. For each parallel pair, we generated two SFT instances sharing the same `Cleaned_Sentence_ID` (CSID) but differing in direction: one with the English sentence as source and Urdu as target, and one with Urdu as source and English as target. This effectively doubled the number of training examples while maintaining perfect alignment.

A key aspect of our preparation strategy was the use of bilingual instructions to enhance robustness and mitigate potential biases arising from instruction language. Prior work has demonstrated that the language of the instruction can significantly influence performance in multilingual LLMs (Zhu et al., 2024). To address this and promote balanced improvement across both directions, we constructed four distinct types of instruction examples:

- English instruction + English input $\rightarrow$ Urdu output
- English instruction + Urdu input $\rightarrow$ English output
- Urdu instruction + English input $\rightarrow$ Urdu output
- Urdu instruction + Urdu input $\rightarrow$ English output

For each type, we curated separate pools of manually crafted prompt templates (approximately 10-15 varied phrasings per pool) that explicitly indicate the required translation direction while varying in stylistic nuance (e.g., "Translate the following English text into Urdu:", "Provide an accurate Urdu translation of this English sentence:", and the Urdu equivalents). During dataset construction, a template was randomly sampled from the appropriate pool based on the desired instruction language and translation direction. This random selection ensures lexical and structural diversity in the instructions, which has been shown to improve generalization in instruction tuning (Wang et al., 2023; Zhang et al., 2023).

By incorporating instructions in both languages and covering all direction combinations, the resulting SFT dataset encourages the model to perform reliably regardless of whether the prompt is presented in English or Urdu. The final SFT training set comprises twice the number of original training pairs providing a rich, diverse resource for bidirectional machine translation fine-tuning.

4 Experiments

In this section, we describe the finetuning experiments conducted using the supervised fine-tuning (SFT) dataset prepared in Section 3.5. We evaluate the translation performance of the finetuned models against appropriate baselines on established benchmarks, focusing on the impact of prompt language and translation direction.

4.1 Models

We experiment with the following models:

- **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024): The base instruction-tuned model from Meta, serving as our primary baseline (referred to as **Base**).
- **UrduLlama 1.1**: An Urdu adapted variant built on Llama-3.1-8B-Instruct through continued pretraining (CPT) on approximately 800 million Urdu tokens followed by instruction finetuning on 432K Urdu instructions (referred to as **UrduLlama 1.1**).

Both UrduLlama 1.1 model and Llama-3.1-8B-Instruct were finetuned on the SFT Data prepared in Section 3.5.

Table 4: BLEU scores on the held-out test set from our corpus.

Prompt Language	Direction	Base	Llama-FT	UrduLlama-FT
Urdu	En → Ur	14.22	**25.24**	23.11
English	En → Ur	13.65	**24.65**	23.82
Urdu	Ur → En	19.79	**76.14**	29.14
English	Ur → En	16.99	**77.97**	29.79

4.2 Fine-Tuning Setup

We fine-tuned Llama-3.1-8B-Instruct and UrduLlama 1.1 using the `torchtune` library (torchtune maintainers and contributors, 2024) with LoRA (Hu et al., 2021) for parameter efficiency.

LoRA adapters (rank 8, $\alpha = 16$) were applied to attention projections and MLP layers. Training used AdamW (learning rate 3×10^{-4}, weight decay 0.01), cosine scheduling with 100 warmup steps, and `bfloat16` precision. The per-device batch size was 1 with gradient accumulation of 8 steps (effective batch size 8). Activation checkpointing was enabled to reduce memory usage.

Using Nvidia A100 40GB GPU, models were trained for 3 epochs on the full SFT training set with shuffling. The same hyperparameters were used for both models to ensure a fair comparison.

4.3 Automatic Evaluation

We automatically evaluated translation quality using sacreBLEU (Post, 2018) on a held-out test set from the split of our corpus. Evaluations were conducted across the four combinations of prompt language (English or Urdu) and translation direction.

Table 4 reports BLEU scores for three models: the original Llama-3.1-8B-Instruct without further fine-tuning (Base), Llama-3.1-8B-Instruct fine-tuned on our corpus (Llama-FT), and UrduLlama 1.1 fine-tuned on the same corpus (UrduLlama-FT).

Fine-tuning on our high-quality parallel corpus yields large improvements over the Base model in both directions. For English → Urdu, Llama-FT achieves the highest scores (24.65–25.24), outperforming UrduLlama-FT by 1–2 BLEU points. In the Urdu → English direction, the gains are particularly striking: Llama-FT reaches BLEU scores above 76, far exceeding both the Base (17) and UrduLlama-FT (29). This demonstrates the value of supervised finetuning on clean, domain diverse parallel data, especially when translating from the lower-resource language.

A key observation is the robustness of Llama-FT to prompt language. The performance difference between English and Urdu prompts is negligible (e.g., 24.65 vs. 25.24 for En → Ur; 77.97 vs. 76.14 for Ur → En). In contrast, the Base and UrduLlama-FT show more noticeable drops when prompted in Urdu (up to 3 points in some cases). This stability can be attributed to our bilingual instruction strategy during SFT preparation, which exposed the model to instructions in both languages and reduced sensitivity to the prompt's language which is a common issue in multilingual LLMs. (Zhu et al., 2024)

4.4 Human Evaluation

To complement automatic metrics, we conducted a blind human evaluation on translations generated using English prompts in both directions.

We randomly sampled 90 sentences from the validation split, covering all categories. For each sentence, the three models (Llama 3.1-8B-Instruct, UrduLlama, and Llama 3.1-8B-Instruct-finetuned) produced translations, which were pooled, shuffled, and presented anonymously to two bilingual annotators fluent in English and Urdu.

Annotators assigned a single overall quality score on a 5-point Likert scale (5 = excellent, 1 = poor), assessing the translation's combined adequacy and fluency. Final scores per model and direction were obtained by averaging the ratings from both annotators. Table 5 reports the average human scores.

The human judgments align closely with the automatic BLEU scores while providing additional nuance on perceived quality. In the English → Urdu direction, the Base model receives an extremely low rating, confirming its limited ability to generate coherent Urdu text. Both fine-tuned models show dramatic improvements, with UrduLlama-FT slightly edging out Llama-FT (3.80 vs. 3.68). This small advantage for UrduLlama-FT likely stems from its prior continued pre-training on massive monolingual Urdu data, which enhances Urdu

generation capability.

Table 5: Human evaluation results

Direction	Base	UrduLlama	Fine-tuned
English → Urdu	1.05	**3.80**	3.68
Urdu → English	3.36	3.57	**3.86**

Conversely, in the Urdu → English direction, Llama-FT achieves the highest rating (3.86), outperforming UrduLlama-FT (3.57) and substantially surpassing the Base model (3.36). This mirrors the large BLEU gains observed for Llama-FT in this direction and underscores the value of high quality bidirectional parallel data for improving comprehension and translation from the lower resource language.

Overall, human evaluation corroborates the automatic metrics: supervised finetuning on clean, domain diverse parallel sentences yields major perceptual quality improvements, particularly when starting from a general-purpose base model. The complementary strengths UrduLlama-FT's edge in Urdu generation and Llama FT's superiority in Urdu to English translation highlight how different adaptation strategies (monolingual continued pretraining vs. parallel SFT) benefit distinct aspects of bidirectional performance.

5 Conclusion

In this work, we presented a carefully constructed English-Urdu parallel corpus comprising 80,749 high-quality sentence pairs across 18 diverse domains. Through systematic collection from authoritative sources, rigorous sentence level alignment, extensive preprocessing, and conservative length-based filtering (AWCD ≤ 5), we produced a clean and traceable resource suitable for low resource machine translation. By transforming this corpus into a bidirectional supervised finetuning (SFT) dataset augmented with bilingual instructions (both English and Urdu prompts), we effectively doubled the training examples while promoting robustness to prompt language.

Fine-tuning Llama-3.1-8B-Instruct (Llama-FT) and UrduLlama 1.1 (UrduLlama-FT) on this dataset resulted in substantial improvements over the unmodified Llama-3.1-8B-Instruct baseline. Automatic evaluation using sacreBLEU showed Llama-FT achieving BLEU scores of 24.65–25.24 (En→Ur) and 76.14–77.97 (Ur→En), far outperforming the baseline and demonstrating near-complete insensitivity to prompt language. UrduLlama-FT also delivered strong gains, particularly in Urdu generation. Blind human evaluation on 90 sentences per direction corroborated these findings: fine-tuned models scored 3.68–3.80 (En→Ur) and 3.57–3.86 (Ur→En) on a 5-point scale, compared to the baseline's 1.05 and 3.36.

The results underscore the value of high quality parallel data for bidirectional translation in low resource settings and show that bilingual instruction tuning effectively reduces prompt language bias. Complementary strengths emerged: Llama-FT excelled in Urdu-to-English translation, while UrduLlama-FT retained an edge in English-to-Urdu generation due to its prior Urdu specific adaptation. Together, these contributions advance open research on Urdu machine translation and provide a reproducible pipeline for similar low resource languages.

6 Limitations and Future Work

The corpus, while diverse, is heavily weighted toward news domains, potentially limiting generalization. The evaluation was primarily internal on held-out data from the same corpus, which ensures consistency but restricts the evaluation of out-of-domain performance.

Future directions include publicly releasing the full cleaned corpus. We also intend to evaluate the fine-tuned models on external benchmarks such as FLORES-200 and WAT Urdu tasks for broader comparability. To enable practical deployment on resource-constrained devices common in Urdu-speaking regions, we plan to apply the same fine-tuning recipe to smaller, efficient base models to further reduce memory footprint and inference latency while preserving translation quality.

References

Ahsan Ahmad and Mohsin Ali Ahmad. 2024. Advancing Roman Urdu to Urdu transliteration using machine learning techniques. *Asian Journal of Multidisciplinary Research & Review*, 5(2):108–127.

Syed Abdul Basit Andrabi and Abdul Wahid. 2022. Machine translation system using deep learning for English to Urdu. *Computational Intelligence and Neuroscience*, 2022:7873012.

Umer Butt, Stalin Varanasi, and Günter Neumann. 2025. Low-resource transliteration for Roman-Urdu and Urdu using transformer-based models. In *Proceedings of the Eighth Workshop on Technologies for*

Machine Translation of Low-Resource Languages (LoResMT 2025), pages 144–153. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Muhammad Naeem Ul Hassan, Zhengtao Yu, Jian Wang, Ying Li, Shengxiang Gao, Shuwan Yang, and Cunli Mao. 2024. LKMT: Linguistics knowledge-driven multi-task neural machine translation for Urdu and English. *Computers, Materials & Continua*, 81(1):1901–1923.

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

Huma Israr, Muhammad Khuram Shahzad, and Shahid Anwar. 2024. Improved Urdu–English neural machine translation with a fully convolutional neural network encoder. *International Journal of Mathematical, Engineering and Management Sciences*, 9(5):1067–1088.

Bushra Jawaid, Amir Kamran, and Ondřej Bojar. 2016. Enriching source for English-to-Urdu machine translation. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP@COLING)*, pages 54–63, Osaka, Japan. Association for Computational Linguistics.

Shahnawaz Khan and Imran Usman. 2019. A model for English to Urdu and Hindi machine translation system using translation rules and artificial neural network. *The International Arab Journal of Information Technology*, 16(1):125–131.

Sajadul Hassan Kumhar, Syed Immamul Ansarullah, Akber Abid Gardezi, Shafiq Ahmad, Abdelaty Edrees Sayed, and Muhammad Shafiq. 2022. Translation of english language into urdu language using lstm model. *Computers, Materials & Continua*, 74(2):3899–3912.

Hafsa Masroor, Muhammad Saeed, Maryam Feroz, Kamran Ahsan, and Khawar Islam. 2019. Transtech: development of a novel translator for roman urdu to english. *Heliyon*, 5(5):e01780.

Usman Mohy ud Din. 2019. Urdu–English machine transliteration using neural networks. Master's thesis, COMSATS University Islamabad, Lahore Campus.

Matt Post. 2018. A call for clarity in reporting bleu scores. *Preprint*, arXiv:1804.08771.

Ahmed Raza, Usama Ahmed, Kainat Saleem, Muhammad Azam Hussain, and Amna Sarwar. 2024. Urdu-to-english-based unsupervised machine translation. *Journal of Computer Science and Applications*, 1(2):1–10.

SIL International. 2022. Urdu. `https://www.ethnologue.com/language/urd`. Ethnologue: Languages of the World (25th edition, accessed 2025).

torchtune maintainers and contributors. 2024. torchtune: Pytorch's finetuning library.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and 1 others. 2023. Instruction tuning for large language models: A survey. *ACM Computing Surveys*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.