Evaluating Large Language Models on Urdu Idioms

Anonymous ACL submission

Abstract

We introduce a comprehensive benchmark for Urdu–English idiomatic translation, featuring a manually verified dataset of 4,000 aligned idiom-sentence pairs across Native and Roman Urdu. We evaluate translation, paraphrasing, idiom span detection, and back-translation using various prompting strategies, including cultural, idiomatic, and few-shot approaches. Our evaluation prioritizes semantic fidelity through neural metrics such as BERTScore, COMET, and XCOMET, alongside embedding-based similarities. Results demonstrate that frontier LLMs significantly outperform traditional NMT systems in preserving metaphorical intent. We find that cultural prompting and few-shot learning substantially enhance semantic preservation by activating deeper contextual associations. While performance remains relatively robust in Native Urdu, the lack of standardized orthography in Romanized input poses significant challenges for model consistency and span detection. This study establishes a high-quality foundation for cross-script idiomatic evaluation and highlights the critical role of prompt engineering in preserving figurative meaning.

1 Introduction

Idiomatic expressions pose a long-standing challenge in machine translation due to their non-compositional semantics, cultural grounding, and contextual dependence. For example, the Urdu idiom اونٹ کے منہ میں زیرہ ("a cumin seed in a camel's mouth") corresponds to the English idiom "a drop in the ocean," illustrating how figurative meaning is easily lost when systems rely on literal alignments. Accurate translation requires recognizing idiomatic spans, interpreting figurative intent, and generating culturally appropriate equivalents rather than surface-level mappings.

Urdu, the 10th most spoken language globally,[1] presents additional complexity due to its dual writing systems: the formal Perso–Arabic script (Native Urdu) and the widely used but orthographically inconsistent Roman Urdu. The lack of standardized spelling and phonetic ambiguity in Roman Urdu complicate idiom preservation across scripts. Despite the prevalence of idioms in Urdu communication, there is limited research on how modern LLMs or NMT systems handle idiomatic content in either script, and no unified evaluation spanning translation, paraphrasing, span detection, and back-translation.

To address these gaps, we introduce the first comprehensive benchmark for Urdu–English idiom understanding and translation. Our benchmark extends prior idiom-focused work by covering multiple subtasks, including direct and literal translation, culturally grounded translation, paraphrasing, idiom-span detection, idiomatic back-translation, and few-shot translation with 5-shot and 10-shot demonstrations. This design enables analysis of idiom recognition, cross-script transfer, prompt sensitivity, and robustness under varied conditions.

We evaluate a suite of open-source and closed-source LLMs and NMT models, including Gemini-3 Pro [2], GPT-OSS-20B (Kumar et al., 2025), Qwen 3–8B (Yang et al., 2025), DeepSeek R1-Distill-Llama-8B (Guo et al., 2025), Gemma 3–1B (DeepMind and Team, 2025), Mistral Instruct-7B (Jiang et al., 2023), LLama-3.2-3B-Instruct (Grattafiori et al., 2024), NLLB models (Team et al., 2022), and Google Translate [3] as a strong NMT baseline. We curated a total of 4,000 idiom–sentence pairs for evaluation, divided equally across scripts: 2,000 written in Native Urdu and 2,000 written in Roman Urdu. Each sentence contains at least one idiomatic expression and is paired with a verified English reference translation checked by native speakers.

[1] https://en.wikipedia.org/wiki/Urdu

[2] https://gemini.google.com/

[3] https://translate.google.com/

We assess model outputs using a diverse metric suite capturing surface similarity (BLEU, ChrF), semantic similarity (BERTScore, COMET, XCOMET), paraphrase consistency (ROUGE), and idiom-level variation (Levenshtein distance, unigram overlap, and multilingual embedding cosine scores). This multi-metric pipeline enables a holistic analysis of figurative meaning preservation across scripts, prompts, and architectures.

Our main contributions are:

- We release a unified benchmark for Urdu–English idiom understanding across translation, paraphrasing, idiom-span detection, and back-translation in Native and Roman Urdu.
- We provide a systematic analysis of prompt sensitivity in idiom translation for Urdu using eight prompting strategies.
- We present a cross-script comparison of idiomatic translation quality, revealing substantial performance gaps driven by orthographic representation.
- We introduce a comprehensive human and multi-metric evaluation combining BLEU, ChrF, BERTScore, COMET, XCOMET, ROUGE, Levenshtein distance, and embedding similarity.

2 Related Work

Idiomatic translation has been examined in several language pairs, with prior studies highlighting the difficulty of preserving figurative meaning in neural systems. Fadaee et al. (Fadaee et al., 2018) showed that data augmentation and semantic cues can improve idiom handling in NMT, while Liu et al. (Liu et al., 2023) explored cross-lingual idiom understanding using multilingual encoders. Recent work on Persian–English idiom translation demonstrates that retrieval-augmented prompting and semantic similarity models can enhance figurative meaning retention (Rezaeimanesh et al., 2025). However, these efforts remain limited to high- or mid-resource settings.

Research on Urdu NLP has focused primarily on low-resource challenges, including morphological complexity, script variation, and inconsistent orthography (Wahid et al., 2024; Butt et al., 2025). Despite the prevalence of idioms in Urdu communication, no prior work has investigated idiomatic translation across Native and Roman Urdu, and existing MT benchmarks do not include idiom-specific evaluation for Urdu.

3 Dataset

We construct a unified dataset of 4,000 Urdu–English idiom–sentence pairs, divided equally across writing systems. Each instance contains an Urdu sentence with at least one idiomatic expression, the extracted idiom, and a gold English translation verified by bilingual annotators. The dataset covers a broad range of figurative expressions spanning common proverbs, culture-dependent idioms, and semantically opaque phrases.

Table 1 shows an example idiom and its gold translation.

3.1 Native Urdu

The Native Urdu dataset contains 2,000 sentences written in the Perso–Arabic script, commonly used in formal communication, print media, and education. Idioms were collected from publicly available resources such as UrduPod101,[4] LingApp,[5] and EnglishUms[6], supplemented with manually curated examples from native speakers. Each idiom was paired with an English equivalent and contextual sentence. The initial English equivalents and contextual sentences were drafted with assistance from GPT-5; however, these outputs served only as preliminary templates. A panel of three bilingual postgraduate researchers (including the author) manually reviewed, edited, and corrected every entry to ensure absolute semantic fidelity and idiomatic resonance.

3.2 Roman Urdu

Roman Urdu refers to the representation of Urdu using the Latin script, commonly used in informal digital communication such as texting, social media, and online forums. Unlike Native Urdu, Roman Urdu lacks standardized orthography, resulting in substantial spelling variation and increased lexical diversity, which pose challenges for translation and evaluation systems.

For this work, the Roman Urdu portion of the dataset was constructed to mirror the Native Urdu split. Starting from the 2,000 Native Urdu idioms and their accompanying sentences, we utilized GPT-5[7] to produce the initial transliterated drafts.

[4] https://www.urdupod101.com/
[5] https://ling-app.com/
[6] https://englishums.com/
[7] https://chat.openai.com/

Idiom (Native Urdu): اونٹ کے منہ میں زیرہ

Idiom (Roman Urdu): Oont ke munh mein zeera

Idiom (English): A drop in the ocean

Sentence (Native Urdu): یہ تنخواہ تو میرے اخراجات کے مقابلے میں اونٹ کے منہ میں زیرہ ہے۔

Sentence (Roman Urdu): Ye tankhwah to mere kharchon ke muqablay mein oont ke munh mein zeera hai.

Gold Translation: This salary is just a drop in the ocean compared to my expenses.

Table 1: Example instance from the Native and Roman Urdu datasets.

To ensure the quality of this synthetic starting point, every entry was subjected to a thorough manual review by three native speakers proficient in both scripts.

These reviewers corrected any linguistic inaccuracies and adjusted the text to reflect naturalness and consistency with real-world usage, such as incorporating the informal phonetic spellings typical of digital Urdu. The final Roman Urdu dataset consists of 2,000 high-quality idiom–sentence pairs aligned with their Native Urdu and English counterparts, enabling controlled cross-script comparisons in downstream evaluation.

3.3 Dataset Statistics

Table 2 presents a quantitative comparison of the curated Native Urdu and Roman Urdu idiom–sentence datasets. Each split contains 2,000 idiom–sentence pairs, resulting in a total of 4,000 aligned instances with verified English translations.

Both datasets exhibit similar structural patterns: idioms average between six and seven words, while sentences typically range from nine to ten words. Roman Urdu shows a slightly larger vocabulary size (3,254 vs. 3,033 in Native Urdu) as well as a higher Type–Token Ratio (TTR), reflecting its greater orthographic variability and lack of standardized spelling conventions.

Figures 1 and 2 visualize the distributions of idiom and sentence lengths, respectively.

4 Experimental Design

This section outlines the prompting strategies, model configurations, and evaluation metrics used

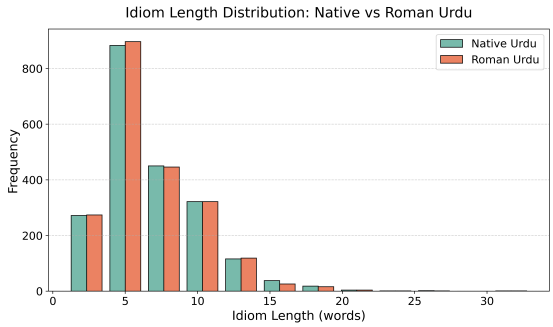


Figure 1: Distribution of idiom lengths in the Native and Roman Urdu datasets.

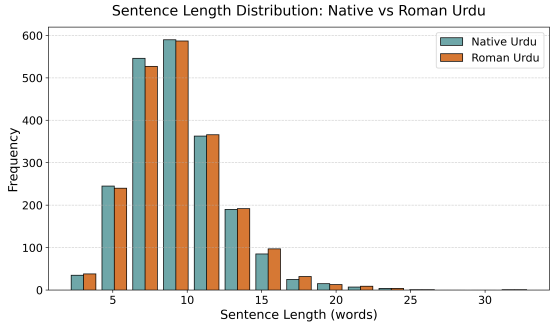


Figure 2: Distribution of sentence lengths in the Native and Roman Urdu datasets.

across all experiments, including idiom translation, idiom classification, English–Urdu back-translation, and Native vs Roman Urdu comparison.

4.1 Prompting Strategies

Table 3 shows the set of prompting strategies used in the experiments.

All large language model experiments were carried out using a chat-based interface in which the system role was explicitly set to *"You are an expert translator fluent in both Urdu and English."*

The Few-Shot translation prompt contains two variations: 5-Shot and 10-Shot with 5 and 10 Urdu-English translated examples provided in the prompts respectively.

4.2 Models

We evaluate a mix of closed-source and open-source LLMs and Neural Machine Translation (NMT) models. The LLMs include Gemini-3 [8], GPT-OSS-20B (Kumar et al., 2025), Qwen 3-8B (Yang et al., 2025), DeepSeek R1-Distill-Llama-8B (Guo et al., 2025), Gemma 3-1B (DeepMind and Team, 2025), Mistral Instruct-7B (Jiang et al.,

[8] https://gemini.google.com/

Dataset	# of Pairs	Avg. Idiom Length	Avg. Sentence Length	Vocab. Size	TTR
Native Urdu	2000	6.74	9.62	3033	0.213
Roman Urdu	2000	6.65	9.71	3254	0.232

Table 2: Linguistic statistics of the curated Urdu idiom translation datasets.

Prompt	Prompt Instruction
Literal Translation	Translate the following Urdu text to English. Do not include explanations, reasoning, or additional text.
Cultural Translation	Translate the following Urdu text into English. Use natural expressions and preserve the cultural meaning. Do not include explanations, reasoning, or additional text.
Paraphrase Translation	Paraphrase the following Urdu sentence into fluent English, keeping the meaning intact. Avoid word-for-word translations. Do not include explanations, reasoning, or additional text.
Idiomatic Translation	Identify the idiom in this Urdu sentence and produce an equivalent English idiom. Do not translate word-for-word. Do not include explanations or reasoning.
Idiomatic Back Translation	Translate the following Urdu idiom to an equivalent English idiom, and then back to Urdu. Provide only the final back-translated Urdu idiom.
Idiom Span Detection	Detect and extract the Urdu idiom in the sentence. Provide only the final detected Urdu idiom.
Few-Shot Translation	Translate the following Urdu text to English using the examples as references.

Table 3: Summary of full prompts used in the experiments.

2023), and LLama-3.2-3B-Instruct (Grattafiori et al., 2024). For NMT baselines, we include nllb-200-3.3B, nllb-200-distilled-600M (Team et al., 2022), and Google Translate [9].

The model temperature was set to 0.8 with a maximum output limit of 500 tokens. To account for models that generated conversational fillers, reasoning, or explanations alongside the translation, we employed regular expression (regex) patterns to programmatically isolate and extract the final translated sentences for evaluation.

4.3 Human Evaluation

To establish a gold-standard benchmark, we conducted a human evaluation of 1,000 translations (100 samples per model) generated from the *Cultural Translation* experiment. The dataset consisted of 500 Native Urdu and 500 Roman Urdu sentences. Two MSc students, who are native Urdu speakers and fluent in English as well, volunteered for the evaluation. Both annotators evaluated the 1,000 translations sample and each translation was evaluated on a 5-point Likert scale across three core dimensions:

- **Semantic Accuracy:** Evaluates whether the translation captures the underlying figurative and metaphorical meaning of the idiom rather than providing a literal, word-for-word rendering.
- **Fluency:** Assesses the grammatical correctness, naturalness, and readability of the output within an English linguistic context.
- **Cultural Equivalence:** Measures how well the translation resonates as a functional equivalent in English, determining if the output feels like a natural expression or idiom rather than an explained concept.

The final *Human Score* for each sample was calculated as the arithmetic mean of these three dimensions.

4.4 Automated Evaluation Metrics

We employ a combination of lexical, semantic, and idiomatic evaluation metrics to comprehensively assess translation quality. BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) measure surface-level n-gram overlap between the model output and

[9] https://translate.google.com/

reference translations. BERTScore (Zhang et al., 2020), COMET (Rei et al., 2020), and XCOMET (Guerreiro et al., 2024) capture semantic similarity and cross-lingual adequacy. ROUGE (Lin, 2004) is used for paraphrase evaluation, while Levenshtein distance and unigram overlap assess idiom-level variation and lexical preservation. Additionally, cosine similarity computed using multilingual embeddings (LASER, LaBSE, and USE-multilingual) quantifies semantic fidelity across scripts.

5 Results

5.1 Human Evaluation Results

To assess the reliability of the human evaluation, we computed inter-annotator agreement using Cohen's kappa (κ) on the sentence-level ratings provided by both annotators.

Agreement was calculated separately for each evaluation dimension, namely *Semantic Accuracy*, *Fluency*, and *Cultural Equivalence*, as well as for the aggregated *Human Score*. Table 4 reports the resulting κ values.

Evaluation Dimension	Cohen's κ
Semantic Accuracy	0.74
Fluency	0.81
Cultural Equivalence	0.73
Human Score	**0.76**

Table 4: Inter-annotator agreement measured using Cohen's kappa across evaluation dimensions.

The inter-annotator agreement results in Table 4 demonstrate substantial consensus across all dimensions. The overall Human Score achieved a Cohen's κ of 0.76, while Fluency reached the highest alignment at 0.81. Semantic Accuracy (0.74) and Cultural Equivalence (0.73) also showed consistent substantial agreement. These robust values validate the human evaluation as a reliable ground truth, confirming that the scoring rubric was applied consistently across evaluators.

5.2 Correlation Analysis

To assess the reliability of automated metrics for Urdu idiomatic translation, we calculated Spearman's rank correlation coefficient (ρ) between the automated scores and the averaged human ratings across 10 models.

As shown in Figure 3, neural-based metrics such as COMET and XCOMET consistently outperform traditional surface-matching metrics like BLEU and chrF. For instance, Gemini-3 achieves a ρ of 0.845 with COMET but only 0.166 with BLEU ($p > 0.05$), highlighting BLEU's inability to capture the figurative nuances of idioms.

The correlation matrix of the metrics themselves reveals a strong alignment between COMET and XCOMET, whereas BLEU shows a weak or negligible relationship with human judgment in several high-performing models (e.g., DeepSeek, $\rho = -0.04$). These results suggest that semantic-aware metrics are essential for evaluating culturally dense tasks like idiomatic translation.

5.3 Translation Performance and Strategy Analysis

The experimental results for idiom translation across Native and Roman Urdu scripts, alongside an ablation study of prompting strategies, are detailed in Tables 5 and 6. Analysis of these results reveals a consistent performance hierarchy where frontier LLMs significantly outperform specialized NMT models. This gap is most evident in neural semantic metrics; as established in the correlation analysis (Section 3), COMET and XCOMET serve as the primary indicators of quality due to their high alignment with human judgment ($\rho = 0.845$), whereas BLEU scores remain negligible (< 0.05) and statistically uncorrelated with idiomatic fidelity.

In the Native script subset, Gemini-3 established the performance ceiling with a peak XCOMET of 0.77 and COMET of 0.68. GPT-OSS-20B followed closely with an XCOMET of 0.74, suggesting that larger parameter counts and instruction tuning directly correlate with a model's capacity for metaphorical comprehension. Performance across the Roman Urdu dataset exhibited a general decline compared to the Native script due to the lack of standardized orthography. However, Gemini-3 maintained the strongest relative performance with an XCOMET of 0.70. Notably, traditional NMT models like NLLB 3.3B and Google Translate suffered the most significant drops; NLLB's chrF score fell from 21.14 to 9.65, highlighting the superior robustness of LLMs to phonetic variations in transliterated text.

The strategy analysis in Table 6 confirms that prompting methodology is as critical as model architecture. The Direct NMT approach struggled across both scripts, particularly in Roman Urdu where it recorded its lowest semantic score (XCOMET: 0.49). This reinforces the finding that standard

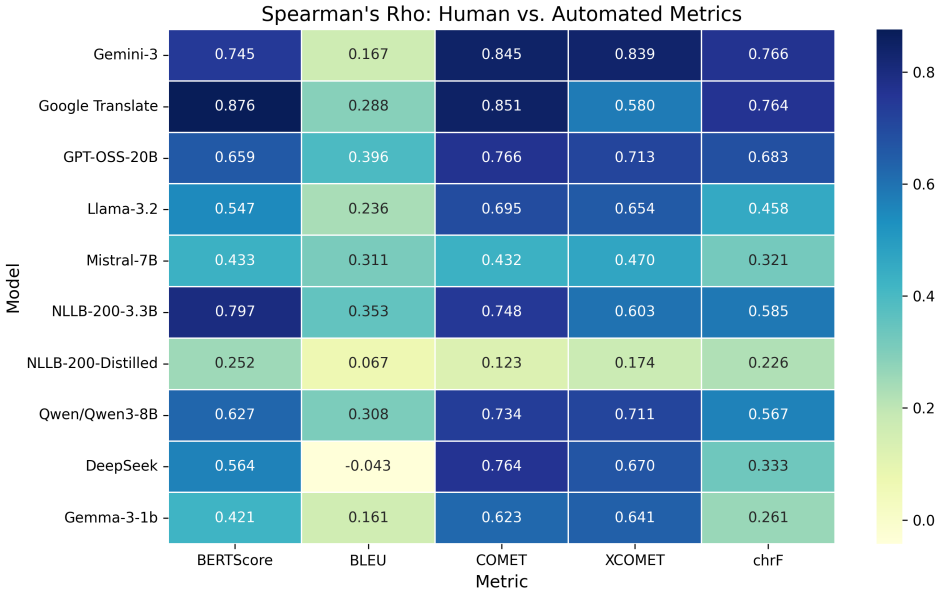


Figure 3: Heatmap of Spearman's Rank Correlation (ρ) between human evaluations and automated metrics across ten models. Darker colors indicate stronger alignment with human judgment.

	Native Urdu					Roman Urdu				
Model	**BERT**	**CMT**	**XCMT**	**BLEU**	**ChrF**	**BERT**	**CMT**	**XCMT**	**BLEU**	**ChrF**
Qwen 3-8B	0.88	0.65	0.71	0.03	25.38	0.83	0.60	0.62	0.04	23.79
DeepSeek-8B	0.87	0.59	0.59	0.02	21.82	0.81	0.52	0.46	0.02	20.08
NLLB 3.3B	0.87	0.63	0.61	0.02	21.14	0.82	0.41	0.34	0.00	9.65
NLLB 600M	0.87	0.64	0.62	0.02	21.33	0.78	0.42	0.56	0.00	10.25
Gemini-3	**0.91**	**0.68**	**0.77**	0.16	42.08	**0.86**	**0.67**	**0.70**	0.12	35.95
Gemma 3-1B	0.87	0.59	0.57	0.01	20.02	0.79	0.46	0.36	0.01	13.81
Google Trans.	0.86	0.63	0.64	0.03	21.70	0.80	0.44	0.56	0.00	12.26
LLaMA 3.2-3B	0.88	0.60	0.64	0.02	21.64	0.83	0.56	0.50	0.04	23.83
Mistral 7B	0.87	0.57	0.56	0.01	19.43	0.81	0.55	0.56	0.02	21.78
GPT-OSS-20B	0.89	0.66	0.74	0.04	26.33	0.84	0.62	0.66	0.06	28.37

Table 5: Comparative performance across Native and Roman Urdu scripts. The best scores for semantic metrics (COMET and XCOMET) are highlighted in bold.

pipelines are not equipped for the combined complexity of non-standardized scripts and idiomatic metaphors. Transitioning from Literal to Cultural prompting resulted in a consistent performance lift. In Native Urdu, Cultural Prompting achieved an XCOMET of 0.72, outperforming the Literal approach (0.66). By explicitly instructing models to seek cultural equivalents rather than literal translations, the models activate broader semantic associations that better preserve the idiom's intent. Even in the challenging Roman Urdu environment, the Cultural strategy maintained the lead (XCOMET: 0.57), proving that cultural instruction is a viable mechanism for mitigating semantic loss in transliterated Urdu.

5.4 Comparison of Few-Shot and Zero-Shot Translations

To evaluate the influence of in-context learning on idiomatic Urdu–English translation, we compare two few-shot settings (5-shot and 10-shot) with two representative zero-shot prompting strategies (Literal and Cultural). Table 7 reports BERTScore, COMET, XCOMET, BLEU, and ChrF scores for all settings.

Few-shot prompting consistently improves idiomatic translation quality over zero-shot settings, with semantic-aware metrics showing the most substantial gains. COMET and XCOMET, which correlate most strongly with human judgment, rise from zero-shot peaks of 0.64 and 0.67 to few-shot highs of 0.72 and 0.81, respectively. This progression indicates that in-context examples significantly

Experiment	Native Urdu					Roman Urdu				
	BERT	CMT	XCMT	BLEU	ChrF	BERT	CMT	XCMT	BLEU	ChrF
Direct (NMT)	0.87	0.63	0.62	0.02	25.79	0.80	0.43	0.49	0.00	10.72
Literal Prompt	0.89	0.64	0.66	0.02	26.39	0.88	0.58	0.56	0.03	24.04
Cultural Prompt	0.89	0.66	0.72	0.03	27.65	0.89	0.60	0.57	0.04	25.45

Table 6: Comparative performance of translation strategies across Native and Roman Urdu datasets: direct NMT, literal prompting, and cultural prompting.

enhance cross-lingual meaning preservation and idiomatic fidelity. Interestingly, while the move from 5-shot to 10-shot yields a notable increase in XCOMET, lexical metrics such as BLEU and ChrF remain stagnant or slightly decline. These results suggest that increasing the shot size primarily refines deep semantic nuance rather than surface-level lexical alignment, with diminishing returns for string-matching accuracy beyond five examples.

5.5 Paraphrasing Experiment

To evaluate how well models generate fluent English paraphrases of Urdu idioms and sentences while preserving meaning, we assess seven LLMs using semantic similarity, surface-level lexical overlap, Levenshtein distance, and multilingual embedding coherence. Table 8 presents the results.

Across evaluation metrics, Gemini-3 and GPT-OSS-20B consistently outperform smaller models. Gemini-3 achieves the highest lexical alignment with a ROUGE score of 0.32 and a unigram overlap of 35.94%, while maintaining top-tier semantic fidelity (BERTScore 0.90). Levenshtein distances between 39 and 43 across all systems suggest substantial rewriting behavior, with Gemini-3 and GPT-OSS-20B demonstrating the most precise paraphrasing capabilities. Cross-lingual embedding-based evaluations using LASER, LaBSE, and multilingual USE further confirm that Gemini-3 leads in preserving deep semantic structure (LaBSE 0.59), whereas smaller models such as Gemma-3 and Mistral show weaker alignment and reduced coherence.

5.6 Idiom Span Detection Results

In addition to translation and paraphrasing, we evaluated the models on their ability to detect idiomatic expressions within Urdu sentences. This experiment measures how accurately models can identify the span of an idiom, which is critical for subsequent translation or paraphrasing tasks.

For evaluation, we used ROUGE to measure n-gram overlap between predicted and reference spans, Levenshtein distance and unigram overlap for exact lexical matches, and cosine similarity using multilingual embeddings to capture semantic similarity across scripts.

Gemini-3 and GPT-OSS-20B show the highest BERTScore and cosine similarity, suggesting strong semantic alignment with reference idioms. Notably, Gemini-3 achieves the lowest Levenshtein distance (12.37) and highest unigram overlap (67.00), indicating superior boundary detection. LLama-3.2B also performs well in ROUGE-based overlap. Lower scores for Mistral-7B and Gemma-3-1B indicate significant challenges in identifying idiomatic spans accurately. These results highlight a clear performance gap, where high detection accuracy directly facilitates better downstream idiomatic translation quality.

5.7 Qualitative Analysis

To illustrate the behavior of our prompting strategies, we provide a representative example from the dataset where the Urdu sentence اچانک خبر سنتے ہی اس کے ہاتھ پیر پھول گئے۔ is paired with the golden English translation: “She completely panicked the moment she heard the sudden news.”

Table 10 demonstrates how different prompting strategies influence the model’s output, ranging from literal physical descriptions to culturally nuanced idiomatic equivalents. This comparison highlights the model’s ability to shift from surface-level translation to deep semantic interpretation when explicitly guided by task-specific instructions.

6 Conclusion

We present the first systematic benchmark for Urdu-English idiomatic translation across Native and Roman Urdu, covering translation, paraphrasing, idiom span detection, and back-translation. Central to this study is our curated dataset of 4,000 aligned idiom-sentence pairs, which provides a high-quality foundation for cross-script comparisons through manual verification by native speakers. Our findings indicate that large LLMs, particularly Gemini-3, GPT-OSS-20B, and Qwen 3-8B, significantly

Experiment	BERT	COMET	XCOMET	BLEU	ChrF
Few-Shot Experiments					
5-Shot	0.91	0.69	0.75	0.15	41.14
10-Shot	0.91	0.72	0.81	0.14	41.11
Zero-Shot Experiments					
Literal Translation	0.89	0.64	0.66	0.02	24.69
Cultural Translation	0.89	0.64	0.67	0.03	25.36

Table 7: Comparison of few-shot and zero-shot idiomatic translation performance across semantic and lexical metrics.

Model	BERT	ROUGE	Lev Dist	Uni Overlap	Cosine Similarity		
					LASER	USE	LaBSE
Qwen3-8B	0.90	0.30	39.36	30.42	0.56	0.53	0.55
DeepSeek-8B	0.89	0.26	42.34	27.57	0.49	0.48	0.51
Gemma-3-1B	0.89	0.21	42.67	21.21	0.45	0.43	0.46
Llama-3.2-3B	0.89	0.26	40.65	27.02	0.52	0.47	0.51
Mistral-7B	0.88	0.23	42.18	22.36	0.44	0.42	0.45
GPT-OSS-20B	0.90	0.31	40.36	32.45	0.57	0.54	0.57
Gemini-3	0.90	0.32	42.41	35.94	0.58	0.53	0.59

Table 8: Paraphrasing results using BERTScore, ROUGE, Levenshtein distance, unigram overlap, and embedding similarities.

Model	BERT	ROUGE	Lev Dist	Uni Overlap	Cosine Similarity		
					LASER	USE	LaBSE
Qwen3-8B	0.83	0.48	15.93	49.57	0.70	0.62	0.66
DeepSeek-8B	0.73	0.27	27.18	27.18	0.54	0.43	0.48
Gemma-3-1B	0.68	0.15	38.21	17.06	0.49	0.43	0.46
Llama-3.2-3B	0.82	0.55	19.62	57.35	0.69	0.61	0.66
Mistral-7B	0.66	0.05	36.64	5.83	0.45	0.40	0.45
GPT-OSS-20B	0.84	0.53	15.67	53.37	0.72	0.66	0.69
Gemini-3	0.88	0.52	12.37	67.00	0.71	0.68	0.65

Table 9: Idiom Span Detection results using ROUGE, Levenshtein distance, unigram overlap, and embedding similarities.

Prompt	Translation Result
Literal	Her hands and feet swelled up.
Cultural	She suddenly became extremely nervous.
Paraphrase	She lost her composure and couldn't think straight.
Idiomatic	She completely freaked out.
Back Translation	وہ گھبراہٹ میں آ گئی۔
Span Detection	ہاتھ پیر پھول جانا

Table 10: Sample translation results using Gemini-3 across different prompts.

outperform NMT systems in preserving idiomatic meaning.

We found that cultural and idiomatic prompts yield the highest semantic fidelity by activating broader associations that better preserve the idiom's intent. Few-shot prompting provides additional gains, confirming that in-context learning further improves idiomatic translation quality over zero-shot settings. Cross-script analysis reveals that while semantic metrics like COMET and XCOMET remain relatively stable across scripts, the lack of standardized orthography and phonetic ambiguity in Roman Urdu continues to pose challenges for model robustness. Furthermore, our idiom span detection results emphasize that the accurate identification of figurative expressions is a critical prerequisite for high-quality downstream translation. Overall, this study offers a unified multi-metric evaluation framework for idiomatic Urdu-English translation, providing insights for LLM development and the complexities of Romanized input.

Limitations

While this study provides a comprehensive benchmark for idiomatic translation, there are several limitations. First, the dataset is exclusively focused on the Urdu–English language pair. While this allows for a deep analysis of these specific languages, the findings may not directly generalize to other low-resource languages or different linguistic families. Second, the dataset covers 4,000 idiom-sentence pairs. Although verified by native speakers, this sample size may not fully represent the vast diversity of Urdu idioms found in real-world usage. Third, the lack of standardized orthography in Roman Urdu led to phonetic ambiguity and spelling inconsistencies, which posed significant challenges for both model detection and automated evaluation. Finally, the current scope of this study does not include expansion to other low-resource languages, which remains a key area for future investigation to validate the broader applicability of the proposed evaluation framework.

References

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, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Google DeepMind and Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, An-

dres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Deli Guo, Bo Zhang, Haoran Wei, Yufan Wang, and DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms. *Preprint*, arXiv:2501.12948.

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

Deepak Kumar, Divakar Yadav, and Yash Patel. 2025. Gpt-oss-20b: A comprehensive deployment-centric analysis of openai's open-weight mixture of experts model. *Preprint*, arXiv:2508.16700.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2025. Large language models for Persian-English idiom translation. 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 7974–7985, Albuquerque, New Mexico. Association for Computational Linguistics.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Abdul Wahid, Muhammad Arsalan, Adeel Mumtaz, Rafiullah Khan, and Saeed Anwar. 2024. Erupd: English to roman urdu parallel dataset. *arXiv preprint arXiv:2412.17562*.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

A Appendix

A.1 Resources

All experiments were performed on machines equipped with NVIDIA A100 40GB GPUs.

A.2 Roman Urdu Pitfalls

Roman Urdu often leads to lower translation quality compared to Native Urdu due to inconsistent spelling, phonetic ambiguity, and lack of standardization. For example, the Urdu idiom ہاتھ پیر پھول جانا can appear as *haath pair phool jana*, *hath pair phool jana*, or even *haath pair phool ja na*, causing models to misinterpret or fail to detect idiomatic meaning.

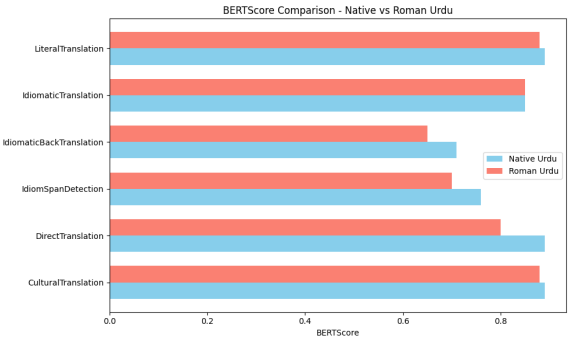
Other idioms show even greater variation in Romanization:

- آسمان سے باتیں کرنا (to daydream) → *aasman se baatein karna* / *asman sy baatein krna*
- نرگسیت میں ڈوب جانا (to become self-absorbed) → *nargisiyat mein doob jana* / *nargisiat me doob jana*

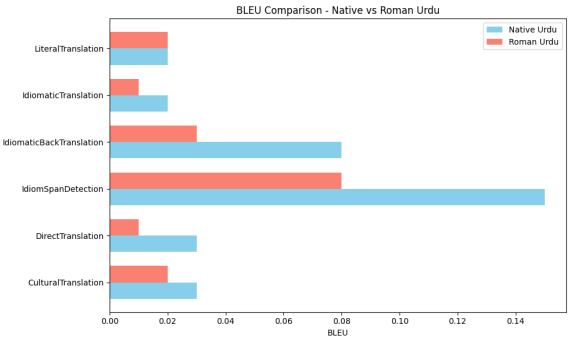
These variations increase lexical ambiguity and make it harder for models to match idioms with their English equivalents.

A.3 Comparison of Native and Roman Urdu

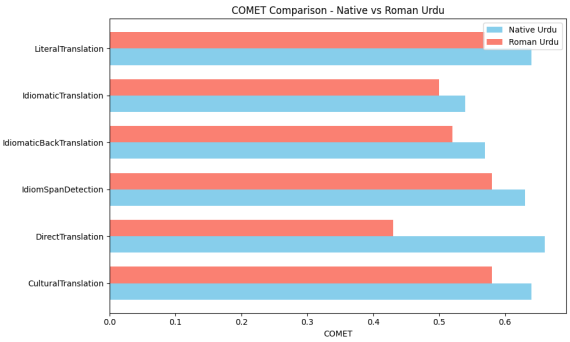
Figure 4 provides a detailed visualization of model performance across both script variations. These comparisons highlight the consistent score distribution shifts across lexical and neural metrics, further illustrating the impact of orthographic variance in Roman Urdu on automated evaluation.



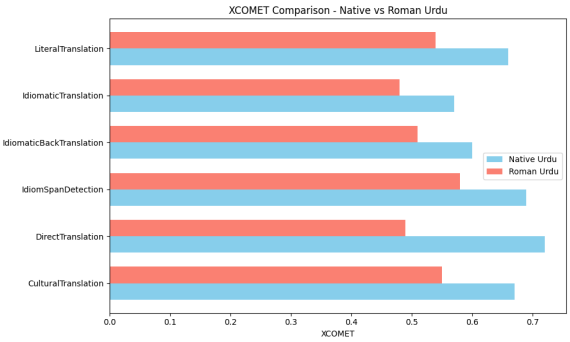
(a) BERTScore comparison



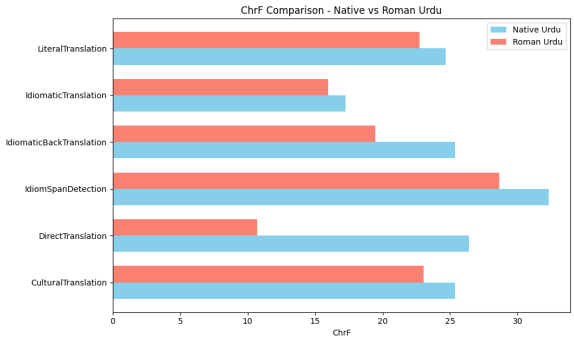
(b) BLEU comparison



(c) COMET comparison



(d) XCOMET comparison



(e) ChrF comparison

Figure 4: Model performance comparison across Urdu and Roman Urdu datasets using (a) BERTScore, (b) BLEU, (c) COMET, and (d) XCOMET.