# **Evaluating Large Language Models on Urdu Idiom Translation**

# Muhammad Farmal Khan, Mousumi Akter

Technical University Dortmund, Germany {farmal.khan, mousumi.akter}@tu-dortmund.de

#### **Abstract**

Idiomatic translation remains a significant challenge in machine translation, especially for low-resource languages such as Urdu, and has received limited prior attention. To advance research in this area, we introduce the first evaluation datasets for Urdu–English idiomatic translation, covering both Native Urdu and Roman Urdu scripts and annotated with gold-standard English equivalents. We evaluate multiple open-source Large Language Models (LLMs) and Neural Machine Translation (NMT) systems on this task, focusing on their ability to preserve idiomatic and cultural meaning. Automatic metrics—including BLEU, BERTScore, COMET, and XCOMET—are used to assess translation quality. Our findings indicate that prompt engineering enhances idiomatic translation compared to direct translation, though performance differences among prompt types are relatively minor. Moreover, cross-script comparisons reveal that text representation substantially affects translation quality, with Native Urdu inputs producing more accurate idiomatic translations than Roman Urdu.

Keywords: Machine Translation, Idiomatic Expressions, Urdu, Large Language Models, Evaluation

#### 1. Introduction

Idiomatic expressions are difficult to translate because their meanings often cannot be inferred from the literal meanings of individual words. Their interpretation depends heavily on cultural and contextual understanding. For instance, the Urdu idiom اونٹ کے منہ میں زیرہ ("Oont ke munh mein zeera") literally means "a cumin seed in a camel's mouth," but it corresponds to the English idiom "a drop in the ocean." Such expressions highlight the challenge of preserving figurative meaning in translation rather than relying on word-for-word equivalence.

Previous studies have explored various strategies to improve idiomatic translation, including semantic alignment, retrieval-augmented methods, and specialized idiom datasets (Fadaee et al., 2018; Rezaeimanesh et al., 2025; Liu et al., 2023). For instance, recent work on Persian–English idiom translation shows that using semantic similarity models and retrieval-based examples can improve figurative meaning retention (Rezaeimanesh et al., 2025).

Urdu, the 10th most spoken language in the world, <sup>1</sup> lacks sufficient linguistic resources and annotated datasets, hindering progress in large language model development. Its dual writing systems—the formal Perso-Arabic script and the informal Roman Urdu—further complicate idiomatic translation. Roman Urdu's inconsistent spelling and phonetic variation make it difficult for models to capture idiomatic meaning across forms (Wahid et al., 2024; Butt et al., 2025). Despite its widespread use, little research has examined how translation models handle idiomatic expressions

This study evaluates the idiomatic translation capability of several open-source LLMs and Neural Machine Translation (NMT) models. The models include 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), and Mistral Instruct-7B (Jiang et al., 2023), nllb-200-3.3B, nllb-200-distilled-600M (Team et al., 2022). To support this evaluation, we curated two datasets:

- Native Urdu Dataset: 1100 idiom-sentence pairs written in standard Urdu, paired with English reference translations verified by native speakers.
- Roman Urdu Dataset: 460 idiom-sentence pairs written in Roman Urdu (Urdu written with the Latin alphabet) with gold-standard English translations.

The translation outputs are evaluated using automatic metrics including BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), COMET (Rei et al., 2020), and XCOMET (Guerreiro et al., 2024), which together assess both surface-level and semantic similarity. We also analyze the role of prompt design in guiding LLMs toward idiomatic translation, using three tailored prompt types: *Literal, Cultural*, and *Paraphrase*, to examine their influence on model behavior.

Our main contributions are as follows:

 We introduce two new Urdu–English idiom translation datasets: one in native Urdu and

across these two scripts. Addressing this gap is vital for building translation systems that preserve natural and idiomatic language use in both formal and informal Urdu.

https://en.wikipedia.org/wiki/Urdu

another in Roman Urdu, annotated with goldstandard English equivalents.

- We benchmark several open-source LLMs on the task of idiomatic translation, studying their ability to handle figurative meaning and cultural context.
- We examine how prompt phrasing affects translation quality through three distinct prompting strategies designed to elicit idiomatic understanding.
- We compare translation quality across writing systems using BLEU, BERTScore, COMET, and XCOMET, providing a reproducible baseline for future idiomatic translation research in low-resource languages.

#### 2. Dataset

#### 2.1. Native Urdu

Our first dataset consists of Urdu idioms written in native Urdu, along with their meanings and example sentences. Native Urdu refers to the standard form of the Urdu language that uses the Perso-Arabic script, which is commonly employed in formal writing, print media, and educational contexts. For example, the idiom النكه كا تارا ("aankh ka tara") literally means "the apple of the eye" and is used to describe someone who is very dear or beloved.

The idioms in this dataset were compiled and created by native Urdu speakers based on their own linguistic knowledge and a variety of public resources (e.g., websites and printed collections of idioms). Public online sources included Urdu-Pod101², LingApp³ and englishums⁴ which provide comprehensive lists and explanations of Urdu idioms and proverbs. We then used an LLM (Chat-GPT) to generate corresponding English idioms and example sentences, which were manually reviewed and corrected (if necessary) by a group of university students, fluent in Urdu and English, to ensure translation quality and idiomatic equivalence. The resulting dataset contains 1,100 idiom—sentence pairs. An example is shown in Table 1.

#### 2.2. Roman Urdu

Roman Urdu refers to the representation of the Urdu language using the Latin (English) script instead of the traditional Perso-Arabic script. It is commonly used in informal digital communication such as social media, texting, and online forums. For example, the Urdu idiom (3,2) (4) (4)

اونٹ کے منہ میں زیرہ:

Meaning (Urdu): انتہائی کم چیز

Idiom (English): A drop in the ocean

Sentence (Urdu):
یہ تنخواہ تو میرے اخراجات کے مقابلے میں اونٹ کے منہ میں

Gold Translation: This salary is just a drop

Table 1: Example from the Urdu idiom translation dataset.

in the ocean compared to my expenses.

drop in the ocean") can be written in Roman Urdu as *Oont ke munh mein zeera*. However, because Roman Urdu lacks standardized spelling and orthography, it poses additional challenges for machine translation and evaluation models.

For the second dataset, we collected 100 idioms from an online website, scribd<sup>5</sup>. In addition, we created another 360 Urdu-English idiom pairs that were reviewed and annotated by native speakers to ensure correctness. This resulted in 460 idioms in Roman Urdu.

#### 2.3. Dataset Statistics

Table 2 provides a quantitative comparison of the Native and Roman Urdu idiom datasets. The Native Urdu dataset contains 1,100 idioms with an average idiom length of 5.82 words and a vocabulary size of 1,911 unique tokens. In contrast, the Roman Urdu dataset includes 460 idioms, averaging 5.59 words in length with a vocabulary size of 937.

Figure 1 illustrates the distribution of idiom lengths in both datasets. Most idioms in both scripts fall between 4 and 7 words, indicating similar structural patterns across writing systems. However, the Native Urdu dataset exhibits slightly greater lexical diversity and a longer tail of idioms exceeding 10 words.

Dataset	Num of Idioms	Average Idiom Length	Vocabulary Size
Native Urdu	1100	5.82	1911
Roman Urdu	460	5.59	937

Table 2: Statistics of the Urdu idiom datasets.

## 3. Experimental Setup

# 3.1. Prompt Design

For the LLMs, three different types of prompts were curated, as shown in Table 4. These prompts were created to test whether the models can distinguish between literal and contextual translations

<sup>&</sup>lt;sup>2</sup>https://www.urdupod101.com/

<sup>3</sup>https://ling-app.com/

<sup>4</sup>https://englishums.com/

<sup>5</sup>https://www.scribd.com/document/
382070512

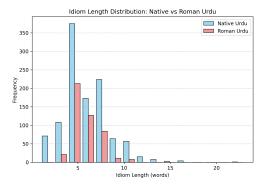


Figure 1: Distribution of idiom lengths in the Native and Roman Urdu datasets.

when given further instructions. Each prompt had to include instructions for not providing any other text except the translation to avoid ambiguity.

For the Neural Machine Translation (NMT) models, no prompts were needed as they just directly translate the text given.

#### 3.2. Models and Data

We generate translations using various opensource LLMs and NMT models which include 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), and Mistral Instruct-7B (Jiang et al., 2023), nllb-200-3.3B, nllb-200-distilled-600M (Team et al., 2022).

The models were executed with their default configurations and the output tokens were allowed to reach a maximum 500 tokens. However some of the models were including explanations and reasoning in their output, due to which the translations were extracted using regular expressions for final output.

For both datasets, 100 samples were randomly selected for translation using the seven models.

### 3.3. Evaluation Metrics

To assess the quality of idiomatic translations, we employ a combination of lexical and semantic metrics, including BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), COMET (Rei et al., 2020), and XCOMET (Guerreiro et al., 2024). This mix captures both surface-level overlap and deeper semantic equivalence between model outputs and gold-standard translations. All four metrics are computed for each translation across different prompt configurations.

### 4. Results

Each translation in our dataset was annotated with the four evaluation metric scores and the mean of each score was calculated for each experiment (prompts and direct translations) and model. The detailed results are shown in Table 3.

### 4.1. Model Performance

Across both Native and Roman Urdu datasets, GPT-OSS-20B consistently achieves the highest performance. It scores best in BERTScore, COMET, and XCOMET, indicating strong semantic alignment, fluency, and idiomatic understanding. BLEU scores are low for all models, reflecting the difficulty in capturing lexical variability of Urdu, but GPT-OSS-20B still outperforms other models in nearly all settings.

The NMT models (3.3B and Distilled 600M) under Direct Translation perform well on BERTScore but show low COMET and XCOMET, especially for Roman Urdu, indicating challenges in capturing idiomatic meaning or written form variations.

**Findings:** GPT-OSS-20B demonstrates superior idiomatic translation capability across all metrics, confirming that large LLMs outperform traditional NMT models in handling linguistic nuance and semantic alignment.

# 4.2. Prompt Performance

Prompt design significantly influences translation quality for instruction-tuned models.

Large models benefit more from instructive prompts. For GPT-OSS-20B, Cultural and Paraphrase prompts consistently outperform the Literal Prompt in semantic metrics, indicating that explicit guidance improves idiomatic translation.

Smaller models like Deepseek R1-Distill-Llama-8B, Gemma 3-1B, and Mistral Instruct-v0.3-7B show modest improvements with more instructive prompts, highlighting their lower sensitivity to prompt variations. NMT models do not use prompts and represent direct translation baselines.

Overall, using instructive prompts (Cultural and Paraphrase) is especially effective for large, reasoning-capable models.

**Findings:** Instructive, context-aware prompts significantly enhance idiomatic translation quality especially for larger instruction-tuned models like GPT-OSS-20B.

# 4.3. Native vs Roman

To analyze the effect of using different writing forms on translation quality, we compared model performance on the Native Urdu and Roman Urdu

Model / Prompt	Native Urdu				Roman Urdu			
	BERTScore	BLEU	COMET	XCOMET	BERTScore	BLEU	COMET	XCOMET
Qwen 3-8B								
Literal Prompt	89.6	2.5	66.4	72.8	87.2	3.3	55.9	64.3
Cultural Prompt	89.5	2.6	65.7	72.7	87.4	3.3	56.9	63.3
Paraphrase Prompt	89.9	3.5	66.7	72.8	87.8	3.3	57.7	64.3
Deepseek R1-Distill-Llama-8B								
Literal Prompt	88.1	0.6	59.3	59.1	85.8	1.5	48.5	52.0
Cultural Prompt	88.5	1.3	60.8	61.0	85.9	1.3	49.9	52.5
Paraphrase Prompt	88.6	1.3	61.0	61.0	86.2	8.0	49.9	52.6
Gemma 3-1B								
Literal Prompt	88.1	1.2	60.5	55.2	85.6	0.0	45.8	45.4
Cultural Prompt	88.1	8.0	60.7	55.2	86.0	0.7	46.8	44.7
Paraphrase Prompt	88.4	8.0	60.3	55.7	85.9	0.3	47.0	45.8
Mistral Instruct-v0.3-7B								
Literal Prompt	88.1	0.5	58.4	54.7	86.8	3.4	55.2	61.2
Cultural Prompt	88.2	0.6	59.1	55.5	86.6	3.2	56.0	62.2
Paraphrase Prompt	88.1	0.5	58.2	54.1	86.9	3.2	55.4	61.6
GPT OSS-20B								
Literal Prompt	90.2	3.9	68.4	77.7	88.1	4.6	59.5	67.4
Cultural Prompt	89.9	3.3	68.9	77.8	88.0	6.5	60.8	69.6
Paraphrase Prompt	90.2	3.9	77.8	78.5	88.1	6.2	59.7	68.3
NLLB-200-3.3B								
Direct Translation	89.2	2.8	65.1	69.97	80.3	0.0	39.5	53.6
NLLB-200-Distilled-600M								
Direct Translation	89.4	2.2	64.7	69.97	77.6	0.0	38.5	69.6

Table 3: Performance comparison of different prompt strategies across translation directions using automatic evaluation metrics. Bold values indicate the best-performing prompt per model.

**Literal Prompt:** Translate this Urdu text to English. Do not include explanations, reasoning, or additional text. Only provide the English translation, nothing else.

**Cultural Prompt:** Translate the following Urdu text into English. Use natural expressions and preserve the cultural meaning. Do not include explanations, reasoning, or additional text. Only provide the English translation, nothing else.

Paraphrase Prompt: Translate the following Urdu text into English. Avoid word-for-word translations. Do not include explanations, reasoning, or additional text. Only provide the English translation, nothing else.

Table 4: Prompt templates used for idiomatic translation experiments.

datasets. Figure 2 shows a side-by-side comparison of BERTScore, BLEU, COMET, and XCOMET for all evaluated models.

As illustrated in the figures, models generally perform better on the Native Urdu dataset compared to Roman Urdu. This is especially visible in COMET and XCOMET scores, indicating that semantic alignment is stronger when the input is in standard Urdu. BLEU scores remain low for both writing systems due to the variability in idiomatic expressions and morphological differences, but a similar trend is observed, with Native Urdu outperforming Roman Urdu.

**Findings:** Models generally perform better on Native Urdu than Roman Urdu, suggesting that inconsistent Romanization reduces translation accuracy and semantic consistency.

### 5. Conclusion

In this research, We curated the first Urdu-English idiomatic translation dataset and evaluated multiple open-source LLMs and NMT models on translating idioms from both Native and Roman Urdu. Our experiments demonstrate that instruction-tuned large models, particularly GPT-

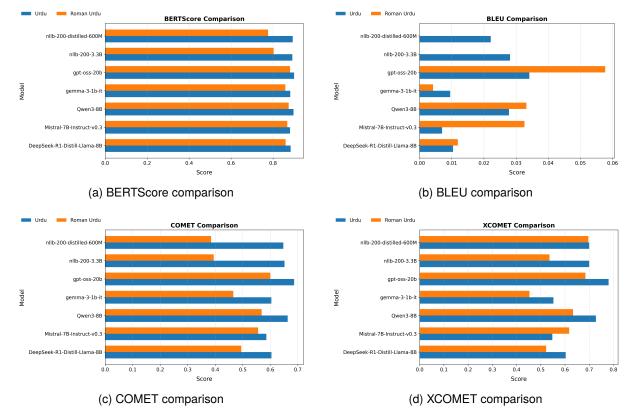


Figure 2: Model performance comparison across Urdu and Roman Urdu datasets using (a) BERTScore, (b) BLEU, (c) COMET, and (d) XCOMET.

OSS-20B, consistently outperform other models across semantic metrics (BERTScore, COMET, XCOMET), indicating strong fluency and cultural understanding in idiomatic translation. Prompt design plays a significant role in translation quality for large models: Cultural and Paraphrase prompts generally improve semantic alignment over Literal prompts, while smaller models show limited sensitivity to prompt variations.

Across the two writing systems, Native Urdu inputs consistently yield higher translation quality compared to Roman Urdu, highlighting the importance of orthographic representation. Overall, our findings provide a benchmark for low-resource idiomatic translation in Urdu and demonstrate the effectiveness of prompt engineering combined with powerful instruction-tuned LLMs, as well as the value of carefully curated evaluation datasets.

# 6. Limitations

While this study establishes a benchmark for Urdu idiom translation using both Native Urdu and Roman Urdu datasets, several limitations must be acknowledged. The datasets are relatively small—comprising 1,100 Native Urdu idioms and 460 Roman Urdu idioms—which may restrict the generalizability of the findings to broader idiomatic or

sentence-level contexts. Furthermore, the lack of standardized spelling and orthography in Roman Urdu introduces considerable variability, potentially influencing translation quality and evaluation outcomes. The evaluation itself relies exclusively on automatic metrics such as BERTScore, BLEU, COMET, and XCOMET, which may not fully capture idiomatic meaning or cultural nuance. Finally, although multiple open-source LLMs and NMT models were assessed, the study excludes proprietary or larger-scale models that might yield stronger results but are not publicly available.

#### 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.

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).
- 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-Al. 2025. Deepseek-r1: Incentivizing reasoning capability in Ilms.
- 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.
- Deepak Kumar, Divakar Yadav, and Yash Patel. 2025. Gpt-oss-20b: A comprehensive deployment-centric analysis of openai's openweight mixture of experts model.
- 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.
- 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 Linquistics.
- 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.
- 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.
- 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*).