PSMT-eval: A Benchmark Dataset for Machine Translation Evaluation in Punjabi Shahmukhi

Abdul Basit and Abdul Hameed Azeemi
Lahore University of Management Sciences, Lahore, Pakistan
{a_basit, abdul.azeemi}@lums.edu.pk

Abstract

This paper introduces PSMT-eval, a new dataset for machine translation (MT) evaluation in Punjabi Shahmukhi, a low-resource language spoken by over 80 million people. We constructed an English–Punjabi Shahmukhi parallel corpus through pivot-based human translation, where the Urdu side of existing English–Urdu datasets was manually translated into Punjabi Shahmukhi. Four widely used English–Urdu datasets were selected: tatoeba-test.eng-urd, FLORES101, IN22, and MKB, with total count of 5182 sentences. To assess the dataset, we evaluated translations generated by state-of-the-art large language models (LLMs), including GPT-5, GPT-OSS-120B, Claude 4 Sonnet, LLaMA-3.3 70B Instruct, and Qwen-3 32B. Results show that only Claude 4 Sonnet achieved competitive performance, with a maximum **SacreBLEU score of 17.51** on IN22. GPT-5 followed with a maximum **SacreBLEU score of 10**, while other models performed poorly. Further analysis highlights major challenges in Punjabi Shahmukhi MT. To the best of our knowledge, this is the first benchmark dataset specifically for MT evaluation in Punjabi Shahmukhi. PSMT-eval is a valuable resource for advancing low-resource MT research and is publicly [1] available to support future work.

1 Introduction

There has been tremendous growth in natural language processing (NLP) in terms of tasks, models, and datasets for high-resource languages. However, progress in low-resource languages remains limited, even for fundamental tasks such as machine translation (MT). Among these low-resource languages is Punjabi Shahmukhi, spoken by over 80 million people[2], yet highly underrepresented in the NLP domain.

Punjabi is widely spoken in both Pakistan and India, but it is written in two distinct scripts: Shahmukhi,is a right-to-left script derived from the Perso-Arabic script and primarily used in Pakistan, and Gurmukhi, derived from Indic scripts and used mainly in India (Saini and Lehal, 2011). Existing benchmarks such as FLORES-101 (Goyal et al., 2022) provide valuable resources for MT evaluation in many low-resource languages, including Punjabi, but only in the Gurmukhi script. Extending these resources to Shahmukhi may appear straightforward, yet it presents unique challenges due to script differences, lack of tools, and limited prior resources. As a preliminary experiment, we translated the first 100 sentences from the Tatoeba Translation Challenge (tatoeba-test.eng-urd) dataset into Punjabi Shahmukhi using a pivot-based human translation approach and evaluated it on GPT-4. The system achieved a **SacreBLEU score of only 9**, indicating very poor translation quality. Table 1 illustrates representative cases, where sentences intended to be translated into Punjabi Shahmukhi were instead rendered in Urdu. This outcome highlights the difficulty of the task and motivates the development of a dedicated dataset specifically for machine translation evaluation in Punjabi Shahmukhi. Establishing standardized evaluation datasets accelerates progress in model development by providing consistent and challenging benchmarks. The proposed Punjabi Shahmukhi test set **PSMT-eval** serves as a foundational resource to facilitate systematic evaluation and guide future improvements in model performance.

2 Related Work

Parallel corpora are the foundation of modern machine translation, enabling the training

[1] https://huggingface.co/datasets/Niazi/PSMT_EVAL

[2] https://www.pbs.gov.pk/sites/default/files/population/2017/tables/pakistan/Table11n.pdf

Proceedings of the 13th Workshop on Asian Translation (WAT 2026), pages 1–8
November 9, 2026 ©2026 Association for Computational Linguistics

Source (English)	Reference Translation (Punjabi Shahmukhi)	GPT-4 Translation
How rude of you!.	اتوں بڑا بدتمیز ایں	کتنے بد تمیز ہو تم!
Science can be used for good or evil purposes.	سائنس نوں چنگے یا ماڑے مقصد لئی ورتیا جا سکدا اے۔	سائنس اچھے یا برے مقاصد کے لئے استعمال کیا جا سکتا ہے۔
The animal died from hunger	اوہ جانور بُھک نال مر جاوے گا۔	وہ جانور بھوک سے مر گیا۔
My best friend is a book.	کتاب میری چنگی ساتھی اے۔	کتاب میری بہترین دوست ہے
Urdu is her mother tongue	اردو اوہدی ماں بولی اے۔	اُردو اس کی مادری زبان ہے۔

Table 1: Comparison between reference translations and GPT-4 outputs in Punjabi Shahmukhi.

and evaluation of data-driven models across languages. A number of large-scale resources have been developed for high-resource languages, creating standardized benchmarks for fair system comparison. The Europarl corpus (Koehn, 2005) is one of the earliest and most widely used resources, containing proceedings of the European Parliament. It provides a standardized test set of approximately 40,000 sentence pairs (about one million words) from the last quarter of 2000, with subsets aligned across 11 European languages. Similarly, the United Nations corpus Ziemski et al. (2016) provides a large, multilingual parallel resource for the six official UN languages, comprising test sets of 4,000 unique, one-to-one aligned sentences across all languages. For low-resource languages, Tiedemann (2020) provides training and test sets for machine translation across more than 500 languages. For each language, each test set contains at least 200 sentences. For multilingual and low-resource MT, several benchmark resources have been introduced in recent years. The FLORES evaluation datasets (Guzmán et al., 2019) were specifically designed for low-resource languages such as Nepali and Sinhala. The test set comprises 2,924 sentences for Nepali–English and 2,905 sentences for Sinhala–English. This effort was later expanded into FLORES-101 (Goyal et al., 2022), which covers over 100 languages and provides a unified evaluation framework with the test set of 997 per language pair. However, for Punjabi, FLORES-101 only includes the Gurmukhi script, used primarily in India, while excluding Shahmukhi, the Perso-Arabic script variant widely used in Pakistan. For Indian languages, the IN22 test set was introduced by Gala et al. (2023). It covers 22 Indian languages, with each test set containing 1,503 sentence pairs. While it includes Punjabi, the data is available only in the Gurmukhi script.

In summary, although a wide range of benchmark datasets exist for high-resource and several low-resource languages, Punjabi Shahmukhi remains critically underrepresented. Existing resources either focus on Punjabi Gurmukhi or exclude Shahmukhi altogether. To the best of our knowledge, no standardized dataset has been developed for English–Punjabi Shahmukhi MT evaluation. Our work addresses this gap by introducing PSMT-eval, a parallel corpus designed specifically for evaluating MT systems on Punjabi Shahmukhi, thereby contributing a novel benchmark to the NLP community.

3 Methodology and Experiments

Since Punjabi Shahmukhi is written in the same Perso-Arabic script as Urdu, the two languages share orthography and some vocabulary. However, they remain distinct, with important cultural and linguistic differences (Saini and Lehal, 2011). To construct the dataset, we first identified suitable Urdu resources to serve as the source for translation into Punjabi Shahmukhi. We then engaged a linguistic expert proficient who is a native speaker in both Urdu and Punjabi Shahmukhi and published work in Punjabi to carry out high-quality human translations. Each dataset was translated using a segment-by-segment approach. To validate the translations, we randomly selected ten percent of the sentences from each dataset and asked an independent linguistic expert to verify their accuracy and faithfulness. If needed they provide the alternate translation. The linguist experts were compensated with market-competitive rates.

3.1 Data Creation

To create PSMT-eval, we selected four publicly available test datasets. First, we utilized the tatoeba-test.eng-urd test set (Tiedemann, 2020),

Dataset	Total Sentences	Avg. words per sentence	Avg. length per sentence	Avg. word ratio	Avg. char ratio
Flores101	997	25.89	121.99	0.82	1.04
MKB	1019	20.05	93.25	0.96	1.18
tatoeba-test	1663	6.94	31.84	0.91	0.99
IN22	1503	11.56	53.02	0.88	1.01

Table 2: Statistics of Punjabi Shahmukhi datasets used in this study, including the total number of sentences, average words per sentence, average sentence length in characters, average Word ratio (EN/PNB), and average char ratio (EN/PNB).

Model	**tatoeba-test.eng-urd**	**MKB**	**IN22**	**FLORES101**
GPT-5	7.69	6.64	10.00	8.89
GPT-OSS-120B	7.32	5.72	9.58	7.81
Claude 4 Sonnet	**14.49**	**11.40**	**17.51**	**14.93**
Qwen 3 32B	2.00	3.12	3.01	3.62
LLaMA-3.3 70B Inst.	7.80	5.21	6.54	7.14

Table 3: SacreBLEU scores of five large language models on PSMT-eval

Model	**tatoeba-test.eng-urd**	**MKB**	**IN22**	**FLORES101**
GPT-5	35.42	34.92	39.01	38.26
GPT-OSS-120B	33.88	33.37	38.03	37.7
Claude 4 Sonnet	**41.97**	**40.06**	**45.18**	**44.42**
Qwen 3 32B	19.19	21.30	21.42	24.46
LLaMA-3.3 70B Inst.	28.68	21.27	24.41	23.69

Table 4: Chrf++ scores of five large language models on PSMT-eval

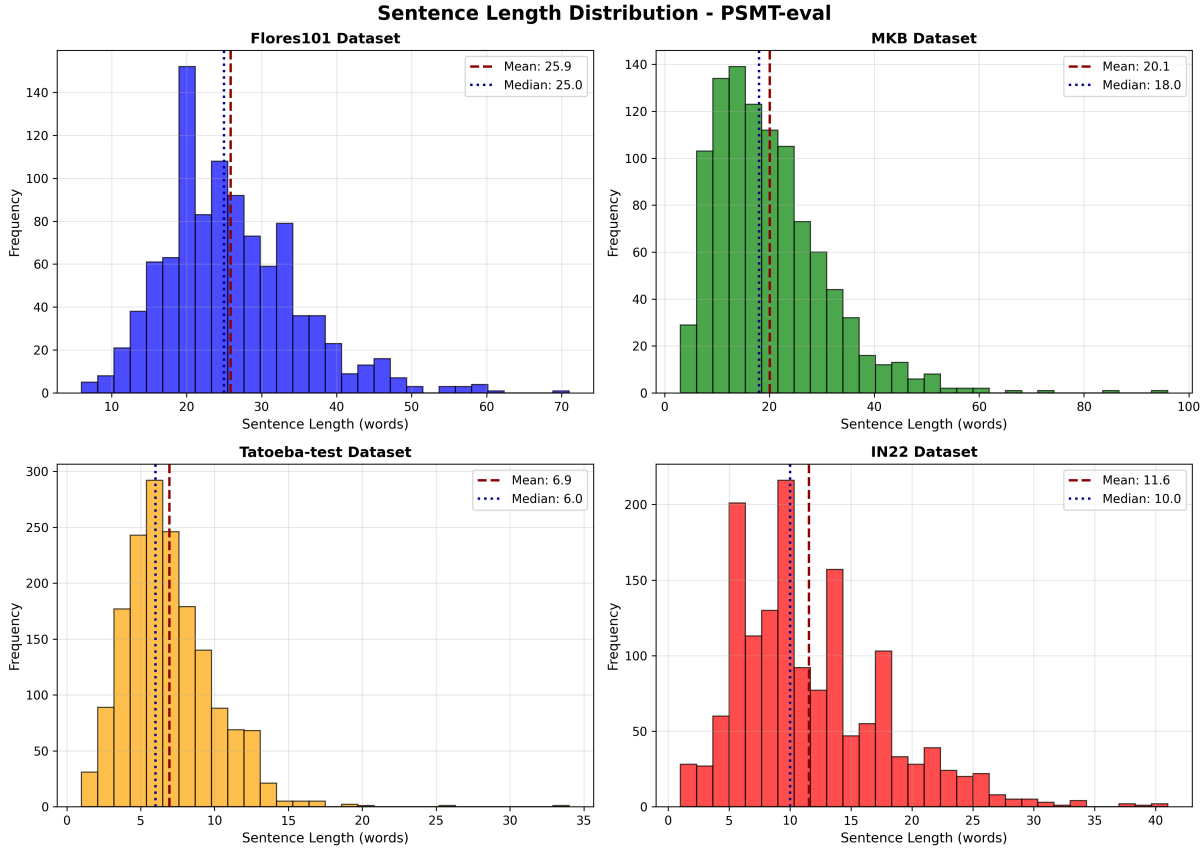


Figure 1: Sentence Length Distribution across Punjabi Language Datasets. This figure illustrates the distribution of sentence lengths (in words) across four Punjabi datasets: Flores101, MKB,tatoeba-test.eng-urd, and IN22. Each subplot shows the mean (red dashed line) and median (blue dotted line) sentence lengths, highlighting variation in dataset characteristics and text complexity.

which is part of the Tatoeba Translation Challenge and includes test sets for over 500 languages. Our study specifically focuses on the Urdu subset. It consists of simple, informal, short sentences. Second, we employed the FLORES-101 dataset (Goyal et al., 2022), a valuable benchmark for low-resource languages covering 101 languages, which is multidomain, from which we used the Urdu portion. Third, we incorporated the Mann Ki Baat test dataset (Siripragada et al., 2020), which contains Urdu translations of speeches delivered by the Indian Prime Minister in multiple Indian languages; its domain is political. Finally, we used the IN22 test set (Gala et al., 2023), which provides a parallel corpus for 22 Indian languages, selecting the Urdu subset for our work. It spans 13 different domains, including Culture, Economy, Education, Entertainment, Geography, Government, Health, Industry, Legal, News, Religion, Sports, and Tourism.

3.2 Inter-Annotator Agreement

Evaluation was conducted by two evaluators holding a Master's degree in Punjabi (Shahmukhi) and fluent in Urdu, who independently rated the Punjabi Shahmukhi translation of each sampled sentence against its Urdu source using a 5-point adequacy scale (1 = Unacceptable to 5 = Perfect).Inter-annotator agreement was measured using quadratic-weighted Cohen's Kappa(Cohen, 1960, 1968) on a 10% sample (518 sentences) drawn independently from each of the four datasets. Combined agreement was moderate (Landis and Koch, 1977) (k = 0.549, 65.1% raw agreement), with substantial agreement on Indic (k = 0.785) and MKB (K = 0.606), moderate agreement on Flores (k = 0.432), and slight agreement on OPUS (k = 0.141) despite 63.3% raw agreement there. The lower scores reflect a systematic one-point scoring offset between evaluators on Flores and a skewed score distribution on OPUS a known cause of deflated kappa despite high observed agreement(Feinstein and Cicchetti, 1990), rather

than random disagreement.

3.3 Data distribution

The figure 1 illustrates clear differences in sentence length distributions across the four Punjabi datasets. The Flores101 dataset exhibits a moderately right-skewed distribution, with most sentences ranging between 15–30 words and a close alignment between the mean (25.9) and median (25), reflecting a balanced structure typical of translation benchmarks. The MKB dataset shows a slightly longer tail, with a mean of 20.1 and a median of 18, indicating moderate variability and the presence of longer sentences. In contrast, the tatoeba-test.eng-urd dataset displays a sharp peak between 5–10 words, with a mean of 6.9 and a median of 6, suggesting shorter and simpler sentence constructions characteristic of conversational or instructional content. The IN22 dataset demonstrates a wider spread, with a mean of 11.6 and a median of 10, reflecting relatively short yet diverse sentence lengths and greater heterogeneity within the corpus.

Table 2 summarizes key statistics for these datasets. Flores101 contains 997 sentences with the highest average of 25.89 words and 121.99 characters per sentence, indicating more complex and information-dense text. tatoeba-test.eng-urd, despite being the largest corpus with 1,663 sentences, exhibits the shortest average sentence length 6.94 words and 31.84 characters, implying simpler or fragmented linguistic structures. MKB and IN22 fall between these extremes, with MKB comprising 1,019 sentences averaging 20.05 words and IN22 including 1,503 sentences with an average of 11.56 words per sentence, representing moderate complexity levels across these corpora.

3.4 Models

To evaluate the effectiveness of PSMT-eval, we selected a diverse set of state-of-the-art large language models (LLMs), encompassing both commercial black-box systems and publicly available open-weight models. Specifically, we assessed the performance of proprietary models GPT-5 (OpenAI) and Claude 4 Sonnet (Anthropic) alongside open-weight models, including GPT-OSS-120B, LLaMA-3.3 70B Instruct (Meta), and Qwen-3 32B (Alibaba). We employed large language models in this study due to the absence of any publicly available machine translation systems capable of translating from English to Punjabi Shahmukhi.

For GPT-5, we configured the reasoning effort parameter to minimal and set verbosity to low, while for the open-weight models we fixed the temperature at 0.1. These settings were chosen to minimize stochastic variation and ensure outputs suitable for machine translation tasks.

To standardize the evaluation, all models were prompted with the instruction: “Please translate the English sentence into Punjabi Shahmukhi script. Only return the translation.” To enhance task clarity, we supplemented this with the contextual cue ”You are an expert English-to-Punjabi translator. Translate the given English text to Punjabi using Shahmukhi script (Perso-Arabic-based script). Provide only the translation without any explanations, notes, or additional text. Maintain the original meaning and context while ensuring natural Punjabi expression.”

In open-source LLMs, GPT-OSS-120B[3] and LLaMA-3.3 70B[4] Instruct were accessed via the Fireworks API , while Qwen-3 32B[5] was deployed using the Groq API. This setup enabled efficient experimentation under consistent conditions, ensuring comparability across both proprietary and open-weight systems.

3.5 Metrics

We evaluate translation performance using the SacreBLEU metric with default settings (Post, 2018), which provides standardized tokenization and scoring for detokenized output, ensuring fair and reproducible comparisons across models. In addition, we report Chrf++ scores (Popović, 2017), a character n-gram–based metric that is particularly effective for languages with rich morphology and complex sentence structures.

3.6 Results

The evaluation results reveal a clear performance gap between proprietary and open-weight models. Among all systems, Claude 4 Sonnet consistently outperforms others, achieving the highest SacreBLEU3 and Chrf++4 scores across all datasets. GPT-5 ranks second overall, producing competitive results particularly on the IN22 and

[3] https://fireworks.ai/models/fireworks/gpt-oss-120b

[4] https://fireworks.ai/models/fireworks/llama-v3p3-70b-instruct

[5] https://console.groq.com/docs/model/qwen/qwen3-32b

Issue	Source	Reference Translation (Punjabi Shahmukhi)	Model Translation	Issue Detail
Language Error	The hostel's guests were mostly citizens of the United Arab Emirates.	ابوسٹل دے زیادہ تر پروہنے متحدہ عرب امارات دے شہری سن	ہاسٹل کے مہمان زیادہ تر متحدہ عرب امارات کے شہری تھے۔	All sentence is translated in Urdu Rather then Punjabi Shahmukhi
Mistranslation	A hostel collapsed in Mecca, the holy city of Islam at about 10 o'clock this **morning** local time.	اسلام دے مقدس شہر مکہ وچ اج **سویرے** مقامی ویلے دے مطابق سویرے 10 وجے ہاسٹل ڈِگ گیا۔	مکہ وچ اک ہوسٹل ڈگ گیا، اسلام دا مقدس شہر، سہ پہر تقریبن 10 بجے مقامی وقت۔	**سویرے** (Morning) is replaced by **سہ پہر** (after noon) in the translation.
Word-Omission	its second claw was larger, giving rise to the name Hesperonychus which means "western claw."	نسبتاً وڈا سی، جس دی وجہ توں ہیسپرونکس دے ناں دی مقبولیت ہوئی، جس دا مطلب اے مغربی پنجہ۔	اُس دی دوجی نکھ وڈی سی، جس کر کے اس دا نام ہیسپیرون	Missing the translation of *which means "western claw."* .
Different language	Pittman suggested that conditions wouldn't improve until sometime next week .	یہ پٹمین نیں اشارہ دتا کہ اگلے ہفتے تک حالات چنگے نئیں ہون گے۔	ਪਿਟਮੈਨ ਨੇ ਸੁਝਾਇਆ ਕਿ ਸਥਿਤੀਆਂ ਸੁਧਾਰ ਨਹੀਂ ਹੋਵੇਗਾ ਅਗਲੀ ਹਫ਼ਤੇ ਦੇ ਕੁਝ ਵੇਲੇ ਤੱਕ।	The whole sentence is translated in Gurmukhi rather than in Punjabi Shahmukhi
Transliteration	Following the **race**, Keselowski remains the Drivers' Championship leader with 2,250 points.	دوڑ دے بعد، 2,250 پوائنٹاں دے نال کیسیالوسکی ڈرائیوراں دی چیمپئن شپ چ برتری رہی۔	ریس کے بعد، کیزلاؤسکی 2,250 پوائنٹس کے ساتھ ڈرائیورز چیمپئن شپ کا لیڈربرقرار رہتا ہے۔	**ریس** used instead of **دوڑ** in translation.

Table 5: Examples of different Punjabi translation issues observed in Neural Machine Translation outputs.

Flores101 datasets though it still lags behind Claude.

In contrast, open-weight models demonstrate mixed and generally weaker performance. GPT-OSS-120B shows promise, achieving relatively strong Chrf++ scores (close to GPT-5 in some cases). However, LLaMA-3.3 70B Instruct and especially Qwen-3 32B perform poorly, with very low SacreBLEU and ChrF++ scores. Their weak results highlight the limitations of current open-weight models when applied to low-resource language translation without domain-specific adaptation.

The performance patterns remain consistent dataset-wise: Claude dominates across tatoeba-test. eng-urd MKB, IN22, and Flores101, followed by GPT-5, while open-weight models trail behind. The stronger ChrF++ scores for proprietary models indicate their robustness in capturing complex linguistic features. This performance indicates that there is a large room for improvement in the open weight models and also in the proprietary models for low-resource languages like Punjabi Shahmukhi.

4 Challenges

Our research has unveiled several challenges associated with Punjabi Shahmukhi translation. These challenges vary across models—some models exhibit a higher frequency of specific errors, while others demonstrate comparatively fewer issues. All issues fall under the domain of accuracy (Lommel et al., 2013). The challenges are summarized below:

- Machine translation between closely related languages tends to be less challenging and generally produces fewer errors compared to translation between distant languages (Popović et al., 2016). Claude Sonnet 4 demonstrates a recurring issue in which it translates the target sentence from English into Urdu instead of Punjabi Shahmukhi. This challenge persists across all evaluated models, likely due to the high degree of orthographic and linguistic similarity between Urdu and Punjabi Shahmukhi. This issue is illustrated in the first row of Table 5.

- When a translation fails to accurately convey the meaning of the source text, it is referred to as a mistranslation (Freitag et al., 2021). This issue was observed across all evaluated models. For instance, the LLaMA-3 70B Instruct model demonstrates this challenge, as illustrated in the second row of Table 5.

- Neural Machine Translation (NMT) systems often omit important words from the source text, which noticeably reduces the adequacy and completeness of the translation (Yang et al., 2019). This challenge is primarily observed in GPT-OSS-120B and, to a lesser extent, in other models. For example, in the third row of Table 5

- Qwen 3 (32B) and LLaMA 3.3 (70B Instruct) frequently produce translations in Punjabi Gurmukhi rather than Punjabi Shahmukhi, despite the contextual instruction clearly specifying that the output should be in the Perso-Arabic (Shahmukhi) script. In contrast, GPT-5 and GPT-OSS-120B exhibit

Error Type	Qwen 3 32B	GPT-OSS-120B	LLaMA-3.3 70B	GPT-5	Claude 4 Sonnet
Language Error	19.0	18.8	20.0	19.2	**33.3**
Mistranslation	14.3	12.5	**20.0**	15.4	16.7
Word Omission	14.3	**31.3**	13.3	15.4	11.1
Different Language	**38.1**	9.4	33.3	11.5	22.2
Transliteration	14.3	28.1	13.3	**38.5**	16.7

Table 6: Error type distribution (%) across models on all dataset

fewer occurrences of this issue. This error is categorized as **Major Severity** because it substantially compromises the clarity, reliability, and practical usability of the translated text for its intended purpose (Alemayehu et al., 2024). This issue is demonstrated in the fourth row of Table 5.

- Transliteration errors may occur due to unclear transliteration choices or inconsistent segmentation between the source and target text (Sennrich et al., 2015). This issue is observed across models, with an example provided of GPT-5 in the last row of Table 5.

- To quantify translation error patterns, a stratified random sample of 270 sentences was drawn proportionally across all four corpora — Flores101 (n=51), MKB (n=54), Tatoeba-test (n=86), and IN22 (n=79) — with each stratum's sample size reflecting its relative proportion of the total 5,182 sentences. Table 6 presents the resulting percentage distribution of each error type across all five models on all four datasets, providing a comprehensive cross-model and cross-corpus view of translation error patterns in Punjabi Shahmukhi.

4.1 Reasons

We observed that the primary challenges arise from models translating sentences either into Urdu or Gurmukhi rather than Punjabi Shahmukhi. Although Urdu shares significant orthographic similarity with Punjabi Shahmukhi, the two languages differ culturally and linguistically. We hypothesize that the limited availability of Punjabi Shahmukhi data contributes to this performance limitation across models.

5 Limitation and Conclusion

In this study, we introduced PSMT-eval, a benchmark dataset designed for evaluating machine translation performance on Punjabi Shahmukhi, and assessed several state-of-the-art large language models (LLMs) using it. Our evaluation revealed that existing LLMs exhibit notable limitations when handling low-resource languages, particularly in accurately distinguishing between closely related scripts such as Urdu, Gurmukhi, and Shahmukhi. These findings underscore the urgent need for more robust translation systems and fine-tuned models that can effectively handle the linguistic and cultural nuances of underrepresented languages.

One key limitation of this work lies in the restricted availability of high-quality Punjabi Shahmukhi data, which may have influenced the models' overall performance. Additionally, the evaluation primarily focused on general-domain text, leaving domain-specific translation (such as news, literature, or social media) an open area for further exploration.

For future research, we plan to expand the PSMT-eval dataset by incorporating diverse domains and larger parallel corpora for Punjabi Shahmukhi. Furthermore, this work can be extended to other low-resource languages of Pakistan, including Sindhi, Balochi, and Pashto, to establish a broader evaluation framework for multilingual translation in the region. Ultimately, we aim to explore the development of specialized large language models optimized for low-resource languages, ensuring inclusivity and equitable language technology advancement across diverse linguistic communities.

References

Hizkiel Mitiku Alemayehu, Hamada M Zahera, and Axel-Cyrille Ngonga Ngomo. 2024. Error

analysis of multilingual language models in machine translation: A case study of english-amharic translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19758–19768.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit Proceedings of Conference*, pages 79–86. International Association for Machine Translation.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Maja Popović, Mihael Arcan, and Filip Klubička. 2016. Language related issues for machine translation between closely related south slavic languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 43–52.

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

Tejinder Singh Saini and Gurpreet Singh Lehal. 2011. Word disambiguation in shahmukhi to gurmukhi transliteration. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 79–87.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shashank Siripragada, Jerin Philip, Vinay P Namboodiri, and CV Jawahar. 2020. A multilingual parallel corpora collection effort for indian languages. *arXiv preprint arXiv:2007.07691*.

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

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

A AI Usage Statement

AI-based tools were used to assist with drafting portions of the text during preparation of this resubmission. All experimental design, data analysis, interpretation of results, and final scientific content were reviewed and are the sole responsibility of the authors.