Leveraging the Cross-Domain & Cross-Linguistic Corpus for Low Resource NMT: A Case Study On Bhili-Hindi-English Parallel Corpus

Pooja Singh[1] **Shashwat Bhardwaj**[3] **Vaibhav Sharma**[1] **Sandeep Kumar**[1,2,3]
[1]Department of Electrical Engineering, Indian Institute of Technology Delhi
[2]Bharti School of Telecommunication, Indian Institute of Technology Delhi
[3]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi
New Delhi, India
{eez228470, aiy237528, eey257541, ksandeep}@iitd.ac.in

Abstract

The linguistic diversity of India poses significant machine translation challenges, especially for underrepresented tribal languages like Bhili, which lack high-quality linguistic resources. This paper addresses the gap by introducing Bhili-Hindi-English Parallel Corpus (BHEPC), the first and largest parallel corpus worldwide comprising 110,000 meticulously curated sentences across Bhili, Hindi, and English. The corpus was created with the assistance of expert human translators. BHEPC spans critical domains such as education, administration, and news, establishing a valuable benchmark for research in low resource machine translation. To establish a comprehensive Bhili Machine Translation benchmark, we evaluated a wide range of proprietary and open-source Multilingual Large Language Models (MLLMs) on bidirectional translation tasks between English/Hindi and Bhili. Comprehensive evaluation demonstrates that the fine-tuned NLLB-200 distilled 600M variant model outperforms others, highlighting the potential of multilingual models in low resource scenarios. Furthermore, we investigated the generative translation capabilities of multilingual LLMs on BHEPC using in-context learning, assessing performance under cross-domain generalization and quantifying distributional divergence. This work bridges a critical resource gap and promotes inclusive natural language processing technologies for low-resource and marginalized languages globally.

1 Introduction

India is a linguistically diverse country, with over $1,369$ distinct mother tongues reported in the 2011 census and 22 constitutionally recognized languages under the 8^{th} Schedule. However, despite this linguistic richness, many indigenous and tribal languages remain critically endangered due to an acute scarcity of digitized linguistic resources and parallel corpora. Indigenous languages are deeply intertwined with cultural heritage, robust translation systems are imperative not only for effective communication and social inclusion but also for equitable access to government services (Nekoto et al., 2020). Effective translation supports crucial national activities ranging from the dissemination of policy directives and welfare schemes to judicial processes and educational initiatives, thereby reinforcing national cohesion within India's intricate multilingual landscape (Haddow et al., 2022).

Efforts to improve machine translation (MT) for Indic languages have accelerated, with early initiatives like the Hindi-English MT challenge at WMT'14 (Bojar et al., 2014) facilitating subsequent benchmarks for Gujarati-English (Barrault et al., 2019) and Tamil-English (Akhbardeh et al., 2021). Recent large-scale efforts such as Workshop on Asian Translation(WAT) (Nakazawa et al., 2021), FLORES 101 (Goyal et al., 2022) NLLB (Costa-jussà et al., 2022) and INDICGENBENCH (Singh et al., 2024) have expanded linguistic coverage significantly by incorporating multiple Indic languages into MT benchmarks. Parallel advances in Large Language Models (LLMs) (Achiam et al., 2023; Tay et al., 2022; Team et al., 2023) have revolutionized text generation and translation tasks. However, Contemporary multilingual models, predominantly trained on high-resource languages, inherently carry cultural biases, leading to suboptimal performance on underrepresented languages. In this context, the Bhili language, spoken by approximately 13 million people,[1] and written in the Devanagari script, is notably underserved. The lack of parallel corpora has severely impeded the development of effective MT systems for Bhili.

To address this gap, we present the Bhili-Hindi-English Parallel Corpus (BHEPC), a meticulously curated, community-driven, high-quality corpus

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

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9553–9579
November 4-9, 2025 ©2025 Association for Computational Linguistics

comprising 110,000 aligned sentences in Bhili, Hindi, and English. Leveraging richer linguistic resources available in Hindi and English, BHEPC enables the exploration of transfer learning and cross-lingual methodologies in an extremely low-resource setting. Although this study focuses on Bhili, the proposed corpus construction workflow, starting from a seed set and iteratively expanded through model-assisted generation with native speaker post-editing, offers a scalable methodology that can be adapted to other endangered languages with minimal digital presence. Building on this foundation, our primary contributions are:

- We introduce the Bhili-Hindi-English Parallel Corpus (BHEPC), the first large-scale, gold-standard parallel corpus for Bhili, comprising 110,000 sentences meticulously curated by native speakers through community-driven efforts.
- We extensively benchmark the multilingual translation capabilities of state-of-the-art open-source & proprietary models such as mT5, Qwen3, DeepSeek-V3, Gemma-2-9B, Mistral-7B-v0.1, BLOOMZ, Llama-2, Llama-3, Llama-4-Scout-17B-16E, Gemini 2.0 Flash, Gemini 2.5 Flash, GPT-3.5 Turbo, GPT-4o-0513, and GPT-4.5 across various model sizes across four translation directions: Hindi↔Bhili and English↔Bhili.
- We analyze cross-domain generalization and quantify distributional divergence across translation directions using Jensen-Shannon Divergence (JSD), while also proposing a hybrid seed-and-post-editing workflow that reduces manual translation effort and provides a scalable template for other endangered languages.

2 Background

Large Language models heavily rely on large annotated corpora, unintentionally favoring high-resource languages, leaving low-resource languages, particularly several regional languages of India, notably underrepresented in computational models and digital repositories. This discrepancy necessitates targeted initiatives for creating community-driven, gold-standard datasets to capture cultural nuances and ensure digital equity and cultural preservation.

The introduction of Transformer-based models (Vaswani et al., 2017), has brought about a paradigm change in neural machine translation (NMT). Despite these advances, translating extremely low-resource regional Indic languages such as Bhili, remains challenging due to inadequate parallel and monolingual data. While transfer learning (Zoph et al., 2016; Chen and Abdul-Mageed, 2023) and multilingual NMT (Johnson et al., 2017; Dabre et al., 2020) partly mitigate data scarcity, their effectiveness is still limited by reliance on parallel corpora.

Concurrent research has also focused on building natural language understanding models (Kakwani et al., 2020; Khanuja et al., 2021) and comprehensive evaluation datasets (Doddapaneni et al., 2023; Mhaske et al., 2023) for Indic languages, thus facilitating systematic benchmarking and comparison. Recent multilingual models like IndicTrans2 (Gala and Chitale, 2023) and MuRIL (Khanuja et al., 2021) offer improved translation and natural language understanding capabilities, leveraging synthetic data and transliteration to compensate for resource constraints. Meanwhile, large language models (Costa-jussà et al., 2022; Arivazhagan et al., 2019; Maurya et al., 2024) demonstrate significant potential for cross-lingual transfer, yet their applicability to translation tasks without direct parallel supervision remains underexplored. Collectively, these advancements highlight the necessity of developing robust community-driven gold standard datasets, evaluation benchmarks, and specialized architectures to enhance translation for severely underrepresented languages like Bhili.

3 Dataset

The acute scarcity of publicly available parallel corpora continues to hinder the development of neural MT models for languages such as Bhili. Given Bhili's linguistic and cultural significance, we address this gap by curating a large-scale, gold-standard parallel corpus translated by native speakers through community-driven efforts, constructing robust evaluation benchmarks, and leveraging multilingual models to exploit linguistic similarities across Indic languages.

3.1 Corpus Details

We introduce the Bhili-Hindi-English Parallel Corpus (BHEPC), the first large-scale, human-curated,

	Lang.	Train	Test	Dev
#Sent.		1,08,000	1,000	1,000
#Tokens	eng	22,21,303	20,413	20,976
	hin	23,57,588	24,062	24,546
	bhb	23,83,485	26.017	26,326
#Types	eng	1,08,012	6,478	6,589
	hin	1,06,749	5,324	5,834
	bhb	2,00,248	7,648	7,851

Table 1: Statistics of Datasets

gold-standard dataset developed through community efforts to support Bhili language translation. The corpus comprises 1,08,000 training, 1000 validation, and 1,000 test sentences, across various domains, including education, administration, and mass media. Hindi source sentences were derived from established resources and meticulously translated into Bhili by language experts, thereby ensuring precise and contextually appropriate renderings. Hindi source sentences were primarily drawn from the Bharat Parallel Corpus Collection (BPCC) (Gala and Chitale, 2023), supplemented by publicly available government documents from Legislative Assembly Speeches (Siripragada et al., 2020), PMIndia corpus (Haddow and Kirefu, 2020), and NCERT Textbooks[2]. Data collection and manual translation efforts, spanning May 2024 to March 2025, involved an average 10 professional translators contributing a total of 27,000 hours. The dataset includes Bhili-Hindi-English tripartite parallel sentences, systematically screened to remove personally identifiable information, hate speech, and redundancy before segmentation into sentence pairs, more details are provided in Appendix A.11.

In addition to structural details, BHEPC was deliberately curated to encode cultural and social dimensions; Appendix A.10.1 outlines how orthography, idioms, and community practices were preserved to make the dataset both a linguistic resource and a cultural benchmark. Table 1 presents detailed corpus statistics, including sentence counts, token distributions, and vocabulary diversity across language pairs.

3.2 Data Preprocessing

We applied extensive preprocessing to improve data quality and linguistic consistency. This included normalizing Bhili homophones, removing extraneous characters, converting English text to lowercase, and enforcing strict de-duplication to avoid the repetition of nearly identical sentences, keeping only one example from sets of very similar sentences. We also set limits on sentence length, rejecting those with fewer than 6 or more than 80 words. Furthermore, to eliminate redundant sentence pairs, we applied cosine similarity-based filtering to those with nearly identical source and target segments. Tokenization was performed using the SentencePiece model, ensuring uniform segmentation across language pairs. The English dataset was generated by translating Hindi sentences from the specified resources using the IndicTrans2 model. Further details on preprocessing steps, quality control, and validation procedures are provided in Appendix A.11.

3.3 Evaluation Dataset & Metrics

To assess multilingual language models for the Bhili language effectively across all the translation directions, we curated a high-quality evaluation dataset. A stratified sampling approach was employed to extract a representative set of sentences, ensuring proportional coverage across different domains while avoiding redundancy. Sentences from various source domains were combined to construct a balanced evaluation dataset. The remaining corpus was partitioned into training and validation sets in a 99:1 ratio while preserving domain distribution. We also curated a domain-specific evaluation set to establish robust benchmarks for evaluating cross-domain adaptability. This included 288 sentences from the NCERT domain, 487 from the Govt/PMI domain, and 1,063 from the mass media domain. Expert translators provided gold-standard translations to ensure high-quality reference data. This benchmark dataset serves as a foundation for evaluating fine-tuning strategies, cross-lingual transfer learning, and domain generalization on Bhili MT. For evaluating translation performance in low-resource language (LRL) settings, we report the chrF++ and a sentence-level variant of BLEU, spBLEU which is more robust than corpus-level metrics in scenarios with limited reference translations following prior work (Khiu et al., 2024). We complement these automatic evaluations with detailed human judgments and inter-annotator agreement analysis, presented in Section 5.

[2] https://ncert.nic.in/textbook.php

Model (LLM)	hin-bhb			bhb-hin			eng-bhb			bhb-eng		
In-Context Ex.	0	5	10	0	5	10	0	5	10	0	5	10
Llama -2-7B	5.61	12.87	14.89	20.07	19.57	15.06	0.43	12.92	11.67	11.45	9.25	15.15
Llama -3.2-1B	4.78	9.50	11.64	19.11	13.50	12.14	0.36	7.08	11.26	10.76	14.93	12.82
Llama -3-8B	7.50	13.50	15.65	21.23	20.50	16.80	0.55	13.50	13.80	12.00	16.50	17.50
Llama-4-Scout-17B-16E	13.89	15.76	17.41	22.45	25.64	17.32	6.35	16.78	23.64	24.68	25.67	22.29
BLOOMZ-560M	5.67	17.67	18.76	23.07	23.89	25.37	0.37	9.88	7.50	13.70	18.35	13.87
BLOOMZ-3.1B	6.35	12.19	14.00	23.87	25.42	27.04	0.64	10.13	12.45	19.01	18.65	20.12
BLOOMZ-7B1	12.58	13.54	15.24	25.21	26.33	26.15	0.81	11.12	15.54	21.34	22.98	23.87
Mistral-7B-v0.1	4.61	14.86	12.87	13.89	12.50	17.23	0.76	10.87	11.43	12.89	13.25	14.21
Gemma-2-9B	9.89	19.03	16.34	17.51	19.64	15.03	5.27	13.53	11.90	30.50	23.25	24.02
Gemini 2.0 Flash	21.17	23.44	24.56	33.01	34.78	36.01	16.82	18.60	19.13	35.62	37.56	38.37
Gemini 2.5 Flash	24.19	25.05	26.74	34.09	35.17	35.89	19.35	21.30	23.54	37.56	38.89	39.12
GPT-3.5 Turbo	21.53	26.31	28.32	39.12	40.32	42.13	20.39	21.78	23.01	38.03	40.15	41.91
GPT-4o-0513	24.01	28.33	28.89	43.09	44.19	45.32	22.19	22.76	24.89	40.51	41.36	43.57
GPT-4.5	27.47	29.75	31.23	45.72	46.68	49.65	23.35	25.67	27.12	42.16	43.78	45.16

Table 2: chrF++ scores with varying in-context examples across four different translation directions. For open-source models, the best scores are in green and second-best in grey. Proprietary models (GPT-4.5 & Gemini) dominate overall, while smaller open source models remain competitive in low-resource settings.

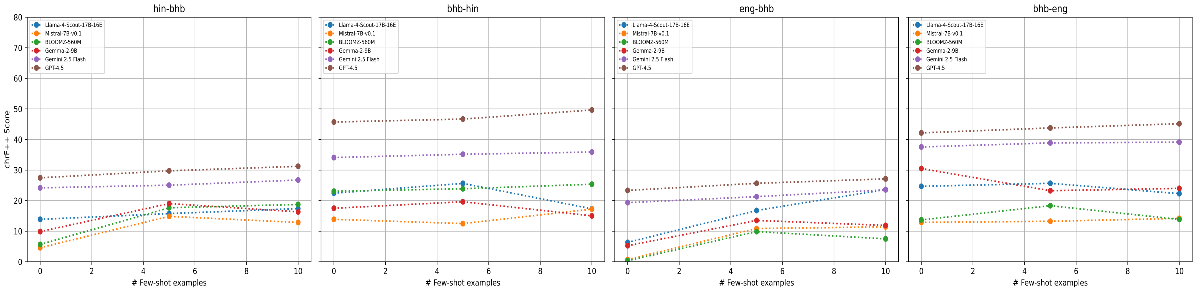


Figure 1: chrF++ performance trends of LLMs across in-context examples (0, 5, 10) shots for four translation directions: Hindi↔Bhili and English↔Bhili direction.

4 Experimental Results & Analysis

4.1 Baseline Models

Despite the proliferation of multilingual language models, significant gaps persist for underrepresented and endangered languages (Protasov et al., 2024; Costa-jussà et al., 2022). Even state-of-the-art multilingual transformers such as mT5, NLLB-200, XLM-R (Conneau et al., 2020), and decoder-based architectures such as BLOOMZ often underperform on low-resource languages, focusing primarily on well-represented ones. To establish strong benchmarks for Bhili translation, we leverage the Bhili-Hindi-English Parallel Corpus (BHEPC) and evaluate a range of open-source & proprietary language models, including IndicTrans2, NLLB 200, mT5 (Xue et al., 2021), Qwen3-8B (Yang et al., 2025), DeepSeek-V3 (DeepSeek-AI et al., 2025), Gemma-2-9B (Team et al., 2024), Mistral-7B-v0.1 (Jiang et al., 2023), BLOOMZ-7B1 (Muennighoff et al., 2023), Llama-2-7B (Touvron et al., 2023), Llama-3-8B (Grattafiori et al., 2024), Llama-4-Scout-17B-16E (Meta AI, 2025), Gemini 2.0 Flash (Google Cloud, 2025a), Gemini 2.5 Flash (Google Cloud, 2025b), GPT-3.5 Turbo (OpenAI, 20253), GPT-4o-0513 (OpenAI et al., 2024), and GPT-4.5 (OpenAI, 2025) across various model sizes. Models are assessed under diverse paradigms, including in-context learning (0,5 and 10 shots), fine-tuning, and cross-domain generalization.

4.2 In-context learning on BHEPC

Large language models (LLMs) exhibit strong few-shot learning capabilities, effectively performing tasks by leveraging a limited number of exemplars. However, their generative capabilities remain inadequate for underrepresented languages due to imbalances in pretraining data. In this section, we empirically examine the impact of varying in-context examples on LLM translation performance across Hindi (hin), Bhili (bhb), and English (eng). Experiments with 0, 5, and 10-shot prompting were conducted across all translation directions, and chrF++ scores are reported in Table 2. We benchmarked both open-source and proprietary models,

Model (LLM)	**hin-bhb**		**bhb-hin**		**eng-bhb**		**bhb-eng**	
Eval. Metric	spBLEU	chrF++	spBLEU	chrF++	spBLEU	chrF++	spBLEU	chrF++
IndicTrans2	9.29	35.67	27.21	53.12	5.21	31.45	15.32	42.00
NLLB-200 (600M)	11.30	42.27	**37.59**	**60.62**	**7.85**	**35.18**	**27.00**	**53.00**
mT5 small	11.00	42.66	29.64	54.54	5.50	27.82	16.29	41.34
mT5 base	**11.68**	**42.83**	34.67	58.67	7.30	33.63	19.76	45.45
Llama-3-8B	10.78	27.89	19.35	35.21	4.32	21.23	11.70	30.59
BLOOMZ-7B1	9.34	23.67	14.32	32.56	3.25	19.43	8.76	29.21
Mixtral-7B-v0.1	7.97	21.30	12.30	27.89	1.80	13.50	3.56	21.65
Gemma-2-9B	8.34	25.45	13.21	29.45	3.36	17.89	7.87	28.67
DeepSeek-V3	3.56	15.34	6.12	18.46	3.87	13.23	7.45	16.34
Qwen3-8B	2.12	18.79	4.86	17.96	1.32	19.42	9.97	22.27

Table 3: spBLEU and chrF++ scores for fine-tuned LLMs. Best scores are highlighted in cyan, with NLLB-200 and mT5-base leading across directions.

including Mistral-7B-v01, Gemma-2-9B, Llama-2-7B, Llama-3.2-1B, Llama-3-8B, Llama-4-Scout-17B-16E, BLOOMZ (560M, 3.1B, 7B1), Gemini 2.0/2.5 Flash, GPT-3.5 Turbo, GPT-4o, and GPT-4.5 over the BHEPC dataset.

Results indicate that proprietary models consistently outperform open-source models across all directions, with improvements correlated to larger model sizes and more in-context examples (Figure. 6). Notably, in Figure. 1 translations from Bhili to English and Bhili to Hindi show sharper gains compared to translations into Bhili. While open-source models show significant improvement from 0 to 5 shots, but gains from 5 to 10 shots are marginal, particularly in hin to bhb and eng to bhb directions.

Among open-source models, Llama-4-Scout-17B-16E performs best for eng→bhb across all shots and for hin→bhb at 0 shots. Gemma leads at 5 shots, and BLOOMZ-560M shows better results at 10 shots. However, Mistral-7B and smaller Llama variants perform poorly across directions. In bhb→hin, BLOOMZ 3.1B and 7B outperform others, while bhb→eng shows degraded performance for Gemma and BLOOMZ-560M beyond 5 shots. Overall, no consistent trend emerges across models, but chrF++ scores are higher when Bhili is the source language. This is most evident in bhb to eng translations, where Gemma achieves high scores (e.g., 30.50 at 0 shots), likely benefiting from extensive English pretraining. In contrast, translations into Bhili remain challenging, with Llama-4-Scout scoring only 6.35 at 0 shots and other models scoring even lower for eng to bhili direction.

These results highlight that translations into high-resource languages are handled more effectively than translations into low-resource languages like Bhili. This disparity reflects the complex linguistic structure of Bhili and the lack of sufficient resources. Larger models demonstrate greater robustness and benefit more from additional context, but translation quality remains highly dependent on language pair, translation direction, data availability, and model architecture.

4.3 Fine-tuning LLMs on BHEPC & Comparison with In-Context Learning

In this section, we evaluated the performance of fine-tuned large language models (LLMs) on the BHEPC dataset. Except for IndicTrans2, which was pre-trained exclusively in 22 Indian languages, the majority of the models underwent intensive multilingual pretraining encompassing 100–200 languages. However, none of these models have been pre-trained on the Bhili language that we explore in this work.

Table 3 reports fine-tuning results across all translation directions. Each model is fine-tuned on the training data, with early stopping based on the validation set, and performance is reported on the test set. The NLLB-200 distilled 600M variant consistently outperforms other models, achieving the highest chrF++ scores, followed by mT5 small and base variants, which also demonstrate strong generalization. In contrast, IndicTrans2 performs poorly

Finetuning Corpus	**Size**	**Testing Corpus**					
		NCERT	Gov/PMI	Mass Media	NCERT	Gov/PMI	Mass Media
			hin-bhb			*bhb-hin*	
NCERT	10k	**30.38**	24.39	30.95	**54.20**	36.14	44.33
Gov/PMI	34k	19.44	**38.75**	37.29	33.40	**60.09**	58.04
Mass Media	64k	20.50	34.68	**42.08**	32.26	50.98	**61.40**
			eng-bhb			*bhb-eng*	
NCERT	10k	**87.35**	34.51	24.77	**70.37**	28.51	38.18
Gov/PMI	34k	17.68	**43.65**	31.16	25.22	**47.69**	45.58
Mass Media	64k	20.18	29.40	**37.31**	25.57	51.11	**52.35**

Table 4: chrF++ scores of the fine-tuned NLLB-200 distilled (600M) model under cross-domain generalization for Hindi↔Bhili and English↔Bhili translation directions. Bold values denote the highest performance per column.

across all directions, most likely as a result of its specialization and subsequent overfitting on its pre-trained set of 22 Indian languages, resulting in poor generalization to unseen languages like Bhili. For instance, in the Hindi-to-Bhili translation direction, IndicTrans2 achieves a spBLEU score of 9.29 and a chrF++ score of 35.67 only, highlighting constraints due to its limited pre-training scope. On the other hand, the NLLB 600M and mT5 base models perform well, with spBLEU scores of 11.30 and 11.68 and chrF++ values of 42.27 and 42.83, respectively, suggesting their superior ability to handle the intricacies of this language pair.

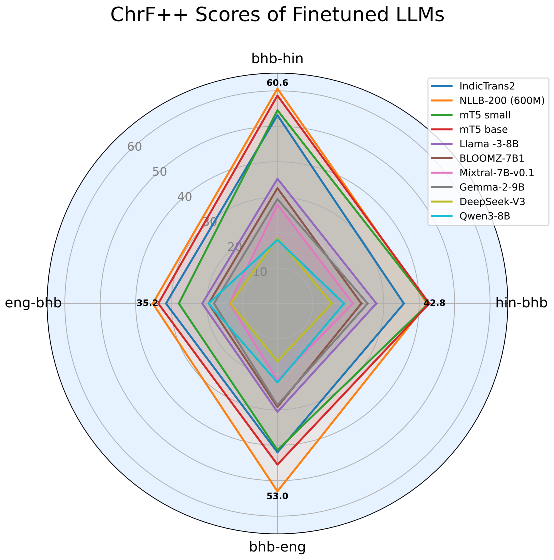


Figure 2: Radar plot showing chrF++ scores of fine-tuned LLMs across four translation directions. NLLB-200 (600M) excel in low-resource scenarios, while other models exhibit balanced performance.

In addition, the Mixtral-7B-v0.1, DeepSeek-V3, Qwen3-8B and Gemma-2-9B models show limited performance, suggesting that increased model capacity alone is insufficient without adequate low-resource language exposure. LLaMA-3-8B and BLOOMZ-7B1 achieve relatively better results, emphasizing the importance of multilingual pre-training quality over sheer model size.

Figure. 2 shows chrF++ scores across four translation directions, with each curve representing a model. Models perform better in Bhili→English and Bhili→Hindi directions due to richer target language resources, while generating Bhili translations remains challenging. We verified these performance differences using paired bootstrap resampling (Koehn, 2004), and report detailed significance tests in Appendix A.10.5. Except in the hin→bhb direction, NLLB-200 significantly outperforms all other models ($p < 0.005$). Comparatively, In-Context Learning (ICL) approaches achieve competitive performance, particularly with larger open-source models like LLaMA-4-Scout-17B-16E, BLOOMZ-7B1, Gemma-9B, and proprietary models such as Gemini 2.5 Flash and GPT-4.5. ICL proves especially effective in low-to-high resource translation directions, benefiting from extensive pretraining and richer contextual examples.

Our findings highlight that fine-tuning remains highly effective for low-resource translation, particularly with models like NLLB-200 and mT5. However, ICL offers a competitive alternative for larger models, reducing the need for expensive fine-tuning while achieving comparable results. This suggests that a hybrid strategy combining fine-tuning for smaller models and ICL for larger ones can effectively optimize translation performance in multilingual, low-resource settings.

4.4 Assessing Cross-Domain Generalization in Machine Translation

Domain adaptation remains a critical yet often underexplored challenge in MT for low-resource languages (LRLs). Prior studies show that performance significantly degrades when language models encounter unfamiliar vocabulary and writing styles (Blitzer, 2008; Elsahar and Gallé, 2019). In this section, we investigate two key factors influencing MT performance: (1) the domain similarity between the fine-tuning and testing corpora, and (2) the effect of domain divergence on translation quality. To analyze cross-domain generalization, we consider three distinct fine-tuning corpora: NCERT textbooks (educational domain), Govt/PMI (administrative/speeches), and Mass Media (News). Performance is assessed across four translation directions: Hindi↔Bhili and English↔Bhili. We define in-domain experiments as cases where the fine-tuning and testing corpora originate from the same domain, while cross-domain experiments occur when the domains differ.

As shown in Table 4 in-domain fine-tuning consistently yields higher chrF++ scores compared to cross-domain settings, reinforcing the negative impact of domain shift on translation accuracy. For instance, in the Bhili to Hindi direction, the model fine-tuned on Govt/PMI achieves 60.09 chrF++ scores on the same domain but drops significantly to 33.40 when evaluated on NCERT, highlighting a severe loss in translation quality when faced with cross-domain data. We quantify domain similarity using Jensen-Shannon Divergence (JSD) (Menéndez et al., 1997). The heatmap in Figure. 3 visualizes JSD scores across different training and testing domain pairs. Darker cells represent lower JSD values, indicating higher domain similarity, while lighter cells signify greater divergence. The results reveal that models fine-tuned on domains with lower JSD (higher similarity) achieve better transaltion quality, whereas those trained on highly divergent corpora struggle to adapt. These findings emphasize the necessity of domain-aware training for building robust, cross-domain generalizable MT models for low-resource language pairs like Bhili. For detailed JSD computations and additional analysis, refer to Appendix A.6 and A.7.

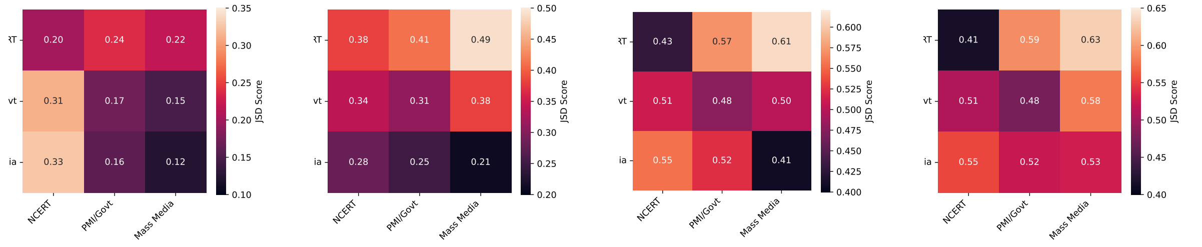
5 Human Evaluation and Qualitative Analysis

Evaluating MT models for low-resource languages like Bhili necessitates a comprehensive assessment framework that combines both quantitative human judgments and qualitative error analysis. This dual perspective ensures a deeper understanding of translation quality beyond surface-level automatic metrics, addressing the linguistic and cultural complexities inherent in low-resource settings.

5.1 Quantitative Human Evaluation: Alignment with Automatic Metrics

To systematically assess the correlation between automatic metrics and human judgments of the translation quality, we conducted a large-scale annotation study following the Multidimensional Quality Metric (MQM) (Sai B et al., 2023; Lommel et al., 2013) and Direct Assessment (DA) guidelines. Candidate translations were generated by eight state-of-the-art multilingual models such as IndicTrans2, NLLB-200, mT5-base, Llama-3-8B, BLOOMZ-7B1, Gemma-2-9B, Mixtral-7B-v0.1, and DeepSeek-V3 across four translation directions: Hindi↔Bhili and English↔Bhili where the segments drawn from the test set.

After all eight models translated each segment, these source translation pairs were presented to language experts in randomized order without revealing the model identities. For each translation direction, we selected 250 segments and employed two bilingual experts, each a native speaker of the target language and fluent in Hindi and English, to perform evaluations. Annotators highlighted error spans, assigned error categories and severity ratings, and provided DA scores on a 1-5 scale. To ensure annotation consistency, experts initially evaluated 50 shared segments, resolving minor disagreements through discussion. This process yielded a total of 1,000 annotated segments, forming a robust foundation for metric evaluation. These human-curated annotations subsequently served as test data for spBLEU and chrF++ metric evaluations and underpinned our MQM score computations. The final Inter-Annotator Agreement (IAA) coefficients were 0.60 for Hindi→Bhili, 0.66 for Bhili→Hindi, 0.53 for English→Bhili, and 0.57 for Bhili→English directions, confirming high annotation reliability across language pairs. In addition, we conducted an inter-annotator agreement (IAA) study on our manually curated Hindi→Bhili



(a) Hindi to Bhili

(b) Bhili to Hindi

(c) English to Bhili

(d) Bhili to English

Figure 3: Jensen-Shannon Divergence (JSD) heatmap for cross-domain generalization evaluation. JSD is computed between in-domain and cross-domain data to quantify distributional divergence between fine-tuning and testing corpora across four translation directions. The results demonstrate that domain shifts significantly impact translation performance, affecting model generalization.

gold references data: two translators independently rendered 250 sentences into Bhili and rated each other's outputs on the MQM scale, yielding an IAA of 0.59. Since only the Hindi→Bhili side was human-translated (with English generated via IndicTrans2), we report gold-data IAA only for this direction. Table 5 presents the segment-level Kendall's τ and Pearson's ρ correlations between spBLEU, chrF++, and human MQM scores. The results demonstrate that chrF++ consistently exhibits stronger alignment with human judgments across all translation directions. Notably, the correlation is weakest when Bhili is the target language (hin→bhb: $\tau = 0.18$, $\rho = 0.25$; eng→bhb: $\tau = 0.14$, $\rho = 0.20$), reflecting the challenges models face in accurately generating Bhili translations. These errors often involve critical lexical mistranslations and cultural inaccuracies, which receive the most severe MQM penalties. Conversely, when Bhili serves as the source language (bhb→hin: $\tau = 0.28$, $\rho = 0.36$; bhb→eng: $\tau = 0.24$, $\rho = 0.31$), models exhibit fewer critical lexical errors, leading to higher metric correlations and stronger alignment with human adequacy and fluency assessments.

5.2 Qualitative Error Analysis: Linguistic and Cultural Insights

To complement the quantitative analysis, we conducted a detailed qualitative error analysis focusing on the fine-tuned NLLB-200 distilled 600M variant, identified as the best-performing model in our experiments. We randomly selected 100 sentences from four translation directions. Native speakers of Bhili, with 1 to 20 years of linguistic expertise, reviewed these translations to identify and categorize prevalent error patterns.

Key issues observed include:

- **Language Mixing:** Due to Bhili's high lexical overlap with Gujarati, the model frequently introduced Gujarati words or inappropriate verb inflections, particularly in administrative and mass media domains.
- **Hallucination and Omission:** Across all directions, the model exhibited a tendency to hallucinate content not present in the source or omit essential information, severely affecting translation fidelity.
- **Polysemy and Lexical Ambiguity:** The model frequently misinterpreted words with multiple meanings, particularly in English-to-Bhili translations, leading to contextually inappropriate outputs.
- **Domain-Specific Translation Failures:** The model struggled with specialized terminology and formal registers, producing inconsistent translations in the education and administrative domains.

Representative examples of these errors are visualized in Figures. 7 and 8. Such qualitative insights underscore the need for culturally grounded gold standard datasets and fine-tuning strategies that better capture the linguistic richness and structural characteristics of Bhili.

6 Conclusion

In this work, we present the first large-scale Bhili-Hindi-English Parallel Corpus (BHEPC) and establish strong benchmarks for Bhili translation. Our data acquisition methodology was both resource-intensive and community-engaged, reflecting the complexities of collecting high-quality linguistic data for an under-documented language. To assess whether LLMs trained on high-resource Devanagari-script languages generalize to Bhili, we conducted extensive benchmarking. While Bhili shares the Devanagari script with Hindi, our findings revealed significant performance degradation particularly in generation tasks challenging the assumption of positive transfer. This observation underscores a pivotal insight script-level similarity alone is insufficient to guarantee semantic transfer or effective representation learning. Bhili's distinct linguistic and cultural characteristics remain underrepresented in existing models, underscoring the need for dedicated LLMs tailored to low-resource languages like Bhili.

Limitations & Future Work

Although our study introduces a high-utility parallel corpus and benchmarks for Bhili, it is not without constraints. The scarcity of monolingual Bhili data limited us to supervised fine-tuning, restricting the use of unsupervised or semi-supervised approaches. Similarly, we did not experiment with augmentation techniques such as back-translation or pivot-based transfer, which have proven effective in other low-resource NMT contexts. Furthermore, while the creation of a 110k, sentence parallel corpus is a substantial contribution, the manual effort required raises concerns about scalability to the thousands of other low-resource languages worldwide. To mitigate this, we adopted a hybrid workflow that begins with a modest seed corpus and iteratively expands it through model-assisted generation and native speaker post-editing. This approach reduces reliance on exhaustive manual translation, yet broader generalizability and cross-domain robustness remain open challenges for future work.

Ethical Statements

We are committed to upholding the highest ethical standards throughout this work. All human translations and annotations were performed by professional language experts with verified proficiency in the target languages and relevant domain expertise. Annotators were fairly compensated with competitive monthly remuneration, aligned with prevailing government standards and reflective of their linguistic skills and the time and effort invested. Additionally, we utilized publicly available Hindi datasets from the BPCC corpus, released under the CC-0 and CC-BY-4.0 licenses. All external resources were used strictly in accordance with their intended research purposes, and the resulting BHEPC dataset is intended solely for academic research and not for commercial purposes. We have obtained the consent from all the language experts.

Acknowledgements

We gratefully acknowledge the support of the Ministry of Tribal Affairs, Government of India, and in particular Shri Vibhu Nayar, Secretary, whose encouragement, visionary leadership, and pivotal guidance in shaping the research direction have been essential to advancing this work. Shri Nayar's foresight in conceptualizing and initiating the Adi Vaani Project has laid the foundation for the preservation and technological empowerment of India's tribal languages. His ongoing support and proactive efforts have facilitated the allocation of critical resources and helped create an enabling ecosystem for this research. The generous support of the Adi Vaani Project, along with the Tribal Research Institute, Bhopal, has enabled us to recruit essential human resources, procure the cloud infrastructure required for large-scale computational work, and engage expert human translators for dataset creation. We sincerely thank all officials from the Ministry of Tribal Affairs and the Tribal Research Institute, Bhopal, for their assistance in coordinating data collection. We also express our deep gratitude to all translators for their dedicated efforts in creating the BHEPC benchmark, and to Prof. Parag Singla, Prof. Vivek Kumar, and Dr. Vipul Rathore for their guidance throughout this project.

Dataset Availability

The Bhili-Hindi-English Parallel Corpus (BHEPC) presented in this study constitutes the first large-scale, high-quality resource for the extremely low-resource Bhili language. Curated through community-driven efforts, BHEPC aims to facilitate research in low-resource machine translation and promote the digital inclusion of marginalized tribal communities. Given the cultural sensitivity of the Bhili language, its endangered status, and the ongoing research objectives associated with this work, access to the BHEPC dataset will be provided upon request to researchers and institutions for academic and non-commercial purposes.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, et al. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the sixth conference on machine translation*, pages 1–88. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

John Blitzer. 2008. *Domain adaptation of natural language processing systems*. Ph.D. thesis, University of Pennsylvania.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Wei-Rui Chen and Muhammad Abdul-Mageed. 2023. Improving neural machine translation of indigenous languages with multilingual transfer learning. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 73–85.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop

Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426.

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173.

Jay Gala and Pranjal A Chitale. 2023. Aswanth kumar m, janki atul nawale, anupama sujatha, ratish puduppully, vivek raghavan, pratyush kumar, mitesh m khapra, raj dabre, and anoop kunchukuttan. 2023. indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Google Cloud. 2025a. Gemini 2.0 flash model - google cloud vertex ai.

Google Cloud. 2025b. Gemini 2.5 flash model - google cloud vertex ai.

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.

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, Andres 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.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Eric Khiu, Hasti Toossi, Jinyu Liu, Jiaxu Li, David Anugraha, Juan Flores, Leandro Roman, A Seza

Doğruöz, and En-Shiun Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

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*.

Kaushal Maurya, Rahul Kejriwal, Maunendra Desarkar, and Anoop Kunchukuttan. 2024. Charspan: Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 294–310.

María Luisa Menéndez, JA Pardo, L Pardo, and MC Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Meta AI. 2025. Llama-4-scout-17b-16e. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A large-scale named entity annotated data for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Toshiaki Nakazawa, Hideki Nakayama, Isao Goto, Hideya Mino, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Shohei Higashiyama, Hiroshi Manabe, Win Pa Pa, et al. 2021. Proceedings of the 8th workshop on asian translation (wat2021). In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

OpenAI. 2025. Gpt-4.5 turbo. https://platform.openai.com/docs/models/gpt-4-turbo.

OpenAI. 20253. Gpt-3.5 turbo. https://platform.openai.com/docs/models/gpt-3-5-turbo.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Vitaly Protasov, Elisei Stakovskii, Ekaterina Voloshina, Tatiana Shavrina, and Alexander Panchenko. 2024. Super donors and super recipients: Studying cross-lingual transfer between high-resource and low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 94–108.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. *arXiv preprint arXiv:2404.16816*.

Shashank Siripragada, Jerin Philip, Vinay P Namboodiri, and CV Jawahar. 2020. A multilingual parallel corpora collection effort for indian languages. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, volume 1, page 8.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, et al. 2022. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray

Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. advances in neural information processing systems. *Advances in neural information processing systems*, 30(2017).

Yuk Wah Wong. 2005. *Learning for semantic parsing using statistical machine translation techniques*. Computer Science Department, University of Texas at Austin.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.

A Appendix

A.1 Extremely low-resource languages

Languages with extremely low resources are characterized by a severe scarcity of accessible data and documentation. In the context of Indian regional languages, many fall into this category, where available resources are minimal compared to more widely studied languages. Many of these languages are either not published or have very little data available, and they are often said to be under-documented, under-resourced, or under-digitized. Therefore, massive obstacles exist when trying to gather and process raw textual data in these languages.

A.2 Bhili Language

Approximately 13 million people across the Indian states of Rajasthan, Gujarat, Maharashtra, and Madhya Pradesh speak Bhili, a Western Indo-Aryan language written in the Devanagari script and deeply rooted in Bhil culture. The dataset we present cover Bhili dialect spoken in the Madhya Pradesh Jhabua region. Despite its significance, due to the lack of publicly available parallel corpora, Bhili has been mostly unexplored in the domains of NLP and machine translation. Given its large speaker base and close lexical ties to Gujarati and Marathi, developing a robust MT system for Bhili particularly for Hindi-Bhili and English-Bhili language pairs has the potential to bridge critical communication gaps and enhance digital inclusion. The growing need for effective digital communication in Bhili-speaking regions emphasizes the potential impact of such a system, making the consolidation of existing resources and the creation of new parallel corpora a critical step toward enabling seamless interaction between Bhili speakers and the broader global community.

A.3 Pre-trained Multilingual LLMs

Pretrained multilingual models have revolutionized the field of natural language processing (NLP) by making substantial advances in machine translation (MT) and cross-lingual transfer learning. Despite the growing number of large-scale Multilingual language models such as IndicTrans2 (supports 22 Indian languages), NLLB (covering 200 languages), and mT5 (spanning 101 languages), Bhili remains largely overlooked. This highlights a critical gap in multilingual MT frameworks. Similarly, even the latest large language models, such as Gemma, Mixtral, DeepSeek,Qwen3, the Llama , and BLOOMZ family series, have expanded multilingual representation but still do not include Bhili, making it less accessible for computational applications. Beyond academic research, commercial MT platforms like Google Translate [3] and Microsoft Translator [4] also exclude Bhili, limiting its digital presence and practical usability. This lack of representation both in research-driven and commercial models makes it even harder to preserve the language, improving accessibility, and integrating Bhili into modern NLP applications.

A.4 Training Details

We evaluated a wide range of pre-trained open-source models for low-resource language translation, including both encoder-decoder and decoder-only architectures. For all models, we maintain a consistent experimental setup to ensure fair comparison. We explore both full fine-tuning and Parameter-Efficient Fine-Tuning (PEFT) using LoRA (Low-Rank Adaptation) (Hu et al., 2022). Hyperparameters are selected through an extensive grid search over batch sizes (8,16, 32) and learning rates (5e-3, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5), and final selections are based on the chrF++ validation scores for both in-domain and cross-domain datasets. The complete hyperparameter configurations are provided in Table 6. In ICL experiments, exemplars are randomly sampled from the training set. For all decoder-based models, we set the decoding temperature to 0.1 to avoid degenerate outputs. For all translation directions, we applied a uniform prompt across all models, as shown in Table 7. Given the high computational demands of fine-tuning large models, we performed only a single run per fine-tuning experiment rather than averaging results across multiple runs.

A.5 Computing Infrastructure

All experiments are performed on a High Performance Computing Cluster having NVIDIA A100 GPUs. Model training times range from 6 to 48 hours, depending on model size and dataset scale. The Hugging Face Transformers library is used for model implementation and fine-tuning, while evaluation metrics are computed using NLTK (Bird and Loper, 2004) and SacréBLEU (Post, 2018). LoRA experiments leverage the PEFT library for efficient adaptation.

A.6 Jensen-Shannon Divergence (JSD) for Cross-Domain Generalization Analysis

Jensen-Shannon Divergence (JSD) quantifies the similarity between two probability distributions, A and B, and is defined as:

$$JSD(A \parallel B) = \frac{1}{2}KL(A \parallel M) + \frac{1}{2}KL(B \parallel M) \tag{1}$$

where M is the mean distribution, and $KL(\cdot \parallel \cdot)$ represents the Kullback-Leibler divergence.

As shown in Equation (1), the JSD (Menéndez et al., 1997) is a symmetrized version of the KL divergence. To compute JSD across domains, we tokenize text using `bert/base/multilingual-/cased`, process batches efficiently, normalize numeric and temporal expressions, and construct token frequency distributions. Missing tokens across corpora are assigned zero probability for proper alignment. We evaluate JSD for all translation directions across the NCERT, Govt/PMI, and Mass Media domains. Heatmap visualizations in Figure 3 show domain shifts, where lower JSD values indicate higher similarity. These insights inform cross-domain adaptation strategies in neural machine translation (NMT), helping mitigate distributional disparities and enhance model robustness.

A.7 Cross Domain Impact Analysis on Machine Translation: Correlating spBLEU, chrF++, and JSD Scores

To further analyze the impact of domain shift on machine translation performance, we evaluate spBLEU scores across different translation directions and examine the relationship between domain similarity (JSD scores) and translation quality metrics (spBLEU and chrF++).

Figure. 4 presents bar plots of spBLEU scores for four translation directions: Hindi to Bhili, Bhili to Hindi, English to Bhili, and Bhili to English, across three domain-specific fine-tuning settings (NCERT, Govt/PMI, Mass Media). Each bar indicates the translation performance when a model trained on one domain is tested on another. Higher bars signal closer alignment between training and testing domains, while lower bars highlight the effects of domain mismatch. The results indicate that in-domain fine-tuning leads to consistently higher

[3] Google Cloud Translation API Reference: https://cloud.google.com/translate/docs/reference/rest.

[4] Microsoft Translator API Reference: https://www.microsoft.com/en-us/translator/business/translator-api/.

Metric	hin→bhb		bhb→hin		eng→bhb		bhb→eng	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ
spBLEU	0.18	0.25	0.28	0.36	0.14	0.20	0.24	0.31
chrF++	0.20	0.30	0.30	0.45	0.15	0.22	0.25	0.38

Table 5: Segment-level Pearson ρ and Kendall's τ correlations of spBLEU and chrF++ with human MQM judgments. chrF++ shows stronger alignment with human evaluations across all translation directions.

Hyperparameters	Values Used
Optimizer	Adam
Beta Values (β_1, β_2)	(0.9, 0.98)
Learning Rate	5e-4
Scheduler	Inverse Sqrt
Loss Criterion	Cross-Entropy
Max Gradient Norm	1.0
Weight Decay	0.01
Batch Size	16
Gradient Accumulation Steps	4
Patience (Early Stopping)	10
Mixed Precision Training	FP16
LoRA Rank (r)	16 (LoRA FT only)
LoRA Alpha	32 (LoRA FT only)
LoRA Dropout	0.1 (LoRA FT only)
Decoding Temperature	0.7

Table 6: Unified hyperparameter configuration across all models and experiments

spBLEU scores, whereas cross-domain settings exhibit performance degradation. Notably, in the Bhili to English direction, models fine-tuned on Mass Media outperform those trained on NCERT and Govt/PMI, suggesting that domain alignment plays a crucial role in translation effectiveness.

Figure. 5 examines the relationship between Jensen-Shannon Divergence (JSD) and spBLEU scores using scatter plots with regression curves. The plot highlights how domain divergence (JSD) impacts translation quality (spBLEU), with trends varying across fine-tuning corpora. A weak or negative correlation is observed in the NCERT setting, suggesting that higher domain divergence leads to lower translation quality, whereas Govt/PMI and Mass Media show a slight positive correlation. The shaded confidence intervals indicate variability, emphasizing the influence of domain adaptation on model performance.

A.8 Translation Guidelines

To ensure consistency and semantic fidelity in translation, we developed a comprehensive set of guidelines that balances linguistic rigor with practical limitations (notably the lack of Bhili-specific glossaries, literature and linguistic resources). Translators are instructed to preserve the source content's meaning and stylistic register without introducing or omitting content, while handling named entities, numerals, dates, and technical vocabulary in accordance with target language's conventions. Specifically:

- **General Principles:** Faithfully reproduce the source text's meaning, tone, style, and register whether formal, colloquial, or emphatic without additions or deletions; correct minor typos or grammatical slips while preserving any factual inconsistencies; and ensure the translation reads fluently and naturally.
- **Named Entities:** Use established conventional translations where available; otherwise, transliterate entities accurately into the target script; and strictly follow language-specific norms without inventing alternative renderings.
- **Numbers & Units:** Mirror the source's numeric format exactly (spelled out or in digits);

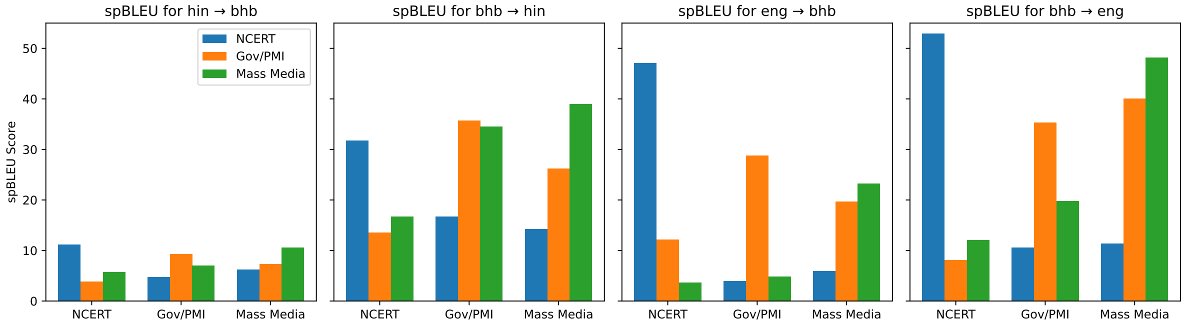


Figure 4: Bar plots showing spBLEU scores across four translation directions: (1) hin → bhb, (2) bhb → hin, (3) eng → bhb, and (4) bhb → eng. The NLLB model fine-tuned on domain-specific datasets: NCERT, Govt/PMI, and Mass Media. The evaluation is conducted on both in-domain and cross-domain data. Each bar represents the translation quality achieved for a given direction and training corpus.

Translation Direction	**Prompt**
English to Bhili	Translate the following English sentence to Bhili: **Input: [Sentence in Source Language]** **Output:**
Bhili to English	Translate the following Bhili sentence to English: **Input: [Sentence in Source Language]** **Output:**
Hindi to Bhili	Translate the following Hindi sentence to Bhili: **Input: [Sentence in Source Language]** **Output:**
Bhili to Hindi	Translate the following Bhili sentence to Hindi: **Input: [Sentence in Source Language]** **Output:**

Table 7: Prompt templates used for different translation directions for all the multilingual LLMs. N-shot examples follow the same format as the last test example given to the model.

apply local counting conventions for large values while retaining “billion”/ “trillion” in English or accepted local terms; and preserve the original units of measurement.

- **Dates:** Maintain the exact date format, whether fully spelled or numeric, and keep the same digit length for years, avoiding any expansion or contraction.

A.9 Annotation Guidelines Based on MQM: Error Categories and Severities

Annotators evaluate translations at the segment level, a segment can consist of a single sentence or multiple sentences by aligning each translated unit with its original source and presenting both sides by side. In Table 8, we describe each error type with a clear hierarchy, illustrating how errors

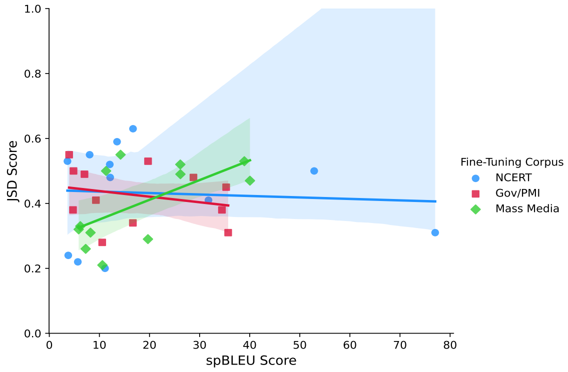


Figure 5: Plot showing the relationship between JSD and spBLEU scores for the NLLB model across three domains. Data points are domain color-coded, with regression lines and confidence intervals highlighting domain-specific trends. NCERT shows little correlation, while Govt/PMI and Mass Media exhibit slight positive correlations, suggesting a trade-off between JSD and spBLEU scores.

are structured across categories. Each category is then assigned a severity rating on a five-point scale (Very low, Low, Medium, High, and Very high) thereby enabling fine-grained distinctions in error impact. Table 9 presents the descriptors at the scale's endpoints as shown to the annotators. To translate their judgments into numeric scores, we assign a weight of 1 for very low, 2 for low, 3 for medium, 4 for high, and 5 for very high. Furthermore, each subcategory such as Accuracy, Fluency, Terminology Inappropriate, and Style, is paired with its own severity marking, so we treat them all with equal significance. Errors unrelated to translation automatically receive a zero score, and any sentence flagged for a source error is omitted from the evaluation.

The following instructions were shared with the annotators:

- Annotators were instructed to scrutinize each translated segment and pinpoint every error present, with a strict limit of five errors per segment. Whenever a segment contains more than five mistakes, they should then select and report only the five most consequential errors.

- Initially, mark the exact span of text by applying color coding; then select the appropriate category/sub category and assign a severity level from the available options. If the error stems from the original content or represents an omission, the highlighted fragment may instead reside within the source segment to ensure the correct context is captured.

- Identify errors at the finest possible granularity. For example, if two words in a sentence are mistranslated, log two separate mistranslation errors.

- In instances where multiple errors overlap within the same text segment, record only the single most severe error; if their severity levels are equal, choose the first matching category in the error typology (e.g., Accuracy, then Fluency, then Terminology).

- Treat Source error and Non-translation as special cases: annotate Source errors by highlighting the relevant span in the source segment (such sentences are exempt from scoring, though the source error must still be marked).

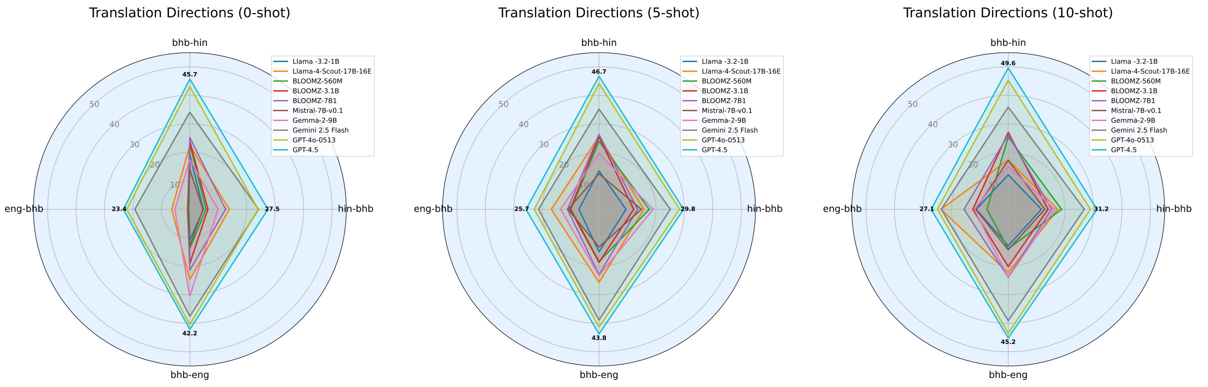
- If the translation is so heavily distorted or entirely unrelated that discrete errors cannot be reliably distinguished, flag a single Non-translation error spanning the entire segment no other errors should be noted when this category is selected.

- Finally, after annotating all errors, assign each translation a score out of 5 and record this value in the final score column.

A.10 Additional Analysis

Supplementary bar plots provide a more detailed, metric-wise comparison across translation directions, reinforcing trends observed in Figure. 9 &10. The bhb → hin direction consistently outperforms others, highlighting the advantages of linguistic similarity and better Hindi representation in pretrained models. While NLLB-200 and mT5-base remain the strongest performers, other LLMs struggle, particularly in the low-resource eng → bhb direction.

Additionally, chrF++ scores consistently surpass spBLEU, indicating that character-level metrics are more effective for evaluating translations involving morphologically rich, low-resource languages like Bhili. These findings underscore the importance of targeted multilingual pretraining and appropriate evaluation metrics in low-resource machine translation.



(a) 0-shot (b) 5-shot (c) 10-shot

Figure 6: Comparison of model performance across four translation directions (hin→bhb, bhb→hin, eng→bhb, bhb→eng) under varying few-shot scenarios. The radar plots highlight that larger models, particularly GPT-4.5 and Llama-4-Scout-17B-16E, consistently outperform smaller models across all settings, with noticeable performance gains as the number of shots increases.

Error Category	**Explanation**
Accuracy	**Addition** The translation injects information absent from the original source, constituting extraneous content. **Omission** Translation is missing content from the source. **Mistranslation** The target text fails to faithfully render the semantic intent of the source. **Untranslated text** Source text has been left untranslated
Fluency	**Orthographic Inconsistency** Spelling or capitalization deviates from standard conventions. **Syntactic Inaccuracy** Grammatical constructions are erroneous, excluding orthographic faults. **Register** The level of formality or pronoun usage is contextually inappropriate. **Character Encoding** Character corruption arises from improper encoding
Terminology Inappropriate	Term selection is non-standard or ill-suited to the domain context.
Style Awkward	The tone or sentence structure is discordant with the genre or unduly verbose. (Example: 1. The source sentence feels formal like in a newspaper, but the translation doesn't. 2. Sentences are correct, but simply too long, etc..)
Transliteration	If it transliterates instead of translating words/phrases, where it should not.
Other	Any issue not encompassed by the specified categories.
Source Error	An error residing in the original source that requires annotation.
Non Translation	The segment is so garbled or unrelated reliably characterize the 5 most severe errors.

Table 8: Hierarchy of errors accompanied by the corresponding explanations provided to the annotators

A.10.1 Cultural Aspects & Dataset Representation

One of the defining challenges of documenting tribal and indigenous languages lies in capturing their cultural specificity alongside linguistic content. The Bhili-Hindi-English Parallel Corpus (BHEPC) was deliberately curated to reflect not only sentence-level alignments but also the social

Translation from **English to Bhili** :

Input: He is survived by his wife Pilar and their two sons, Michael and Bobby.
Model Output: उन परिवार मा घेरवाली फिलर ऐवु उने बे बेटा माइकल ऐ बाबी है |
Ground Truth: तीनी बयेर पिलार अन तिहुना बेहु सोरां माईकल अन बॉबी रिहया से।

Input : This format is not approved by any official body.
Model Ouput: इना फॉर्मेट ने कइनी पण ओपिसियल बडी यें मंजूर नी से।
Ground Truth: यो प्रारूप कईना अधिकारिक निकाय थी स्वीकृत नी मले से |

Input : Children play more games on mobile phones .
Model Ouput: सोरां मोबाईल फोन पर वदु रमे से।
Ground Truth: पुरिया मुबेल मा खेल घणु खेले से।

(a) In the first two examples, the model output primarily has more content from high-resource languages, Hindi and English, like परिवार मा घेरवाली, format, and official, whereas the third output shows the children are occupied with their mobile devices, rather than explicitly stating that they are playing games on mobile phones.

Translation from **Bhili to English** :

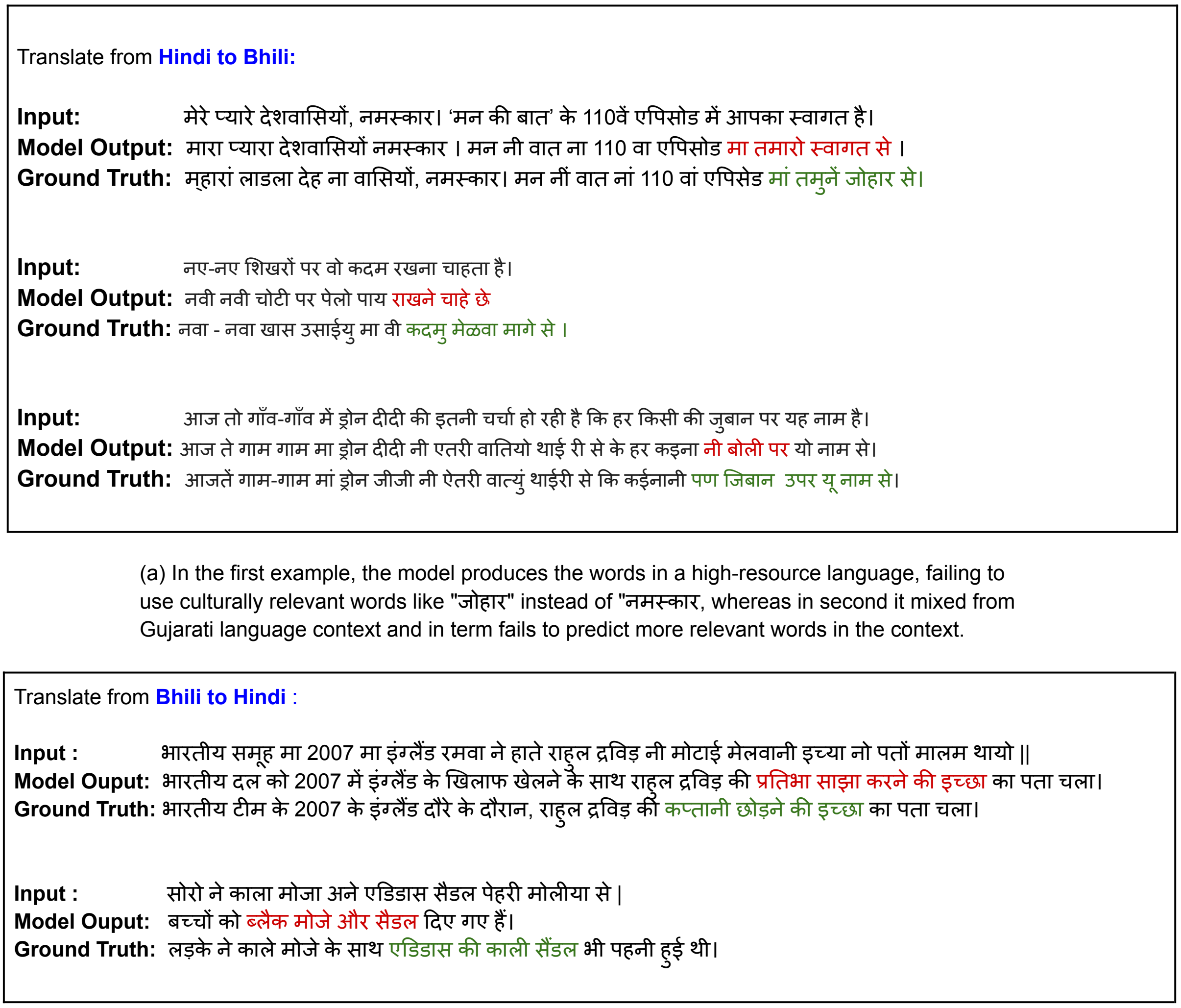
Input : मारा दातु मा दुखावो थाई रियो छे |
Model Ouput: My teeth are **not working.**
Ground Truth: I have pain in my teeth.

Input : हमू काजे कोसरोत कोरुणु चाहे |
Model Ouput: Shall we just do **the math work**?
Ground Truth: We need to do **exercise.**

Input : इना लेदे हमारी सरकार आदिवासी इलाकु मा एकलव्य आवासीय स्कूल बणावा ना अभियान मा घणी मीनत करी यीं से ।
Model Ouput: Therefore, our government is **making a tribal society in tribal areas** a huge effort under the campaign to build a Eklavya Residential School.
Ground Truth: For this, our government is working **very hard in the campaign to build Eklavya residential schools** in tribal areas.

(b) In the first examples,, the actual intent is to convey pain in the teeth, not that they've failed to function, whereas in the second example, the model confuses the Bhili word for "exercise" with "math work," completely changing the subject matter, and in the third example, it fails to capture the correct context of the sentences.

Figure 7: Error analysis in the predictions from the fine-tuned NLLB-200 distilled 600M variant model on the evaluation dataset highlighting errors in the model prediction from English to Bhili and Bhili to English direction.



Translate from **Hindi to Bhili:**

Input: मेरे प्यारे देशवासियों, नमस्कार। ‘मन की बात’ के 110वें एपिसोड में आपका स्वागत है।
Model Output: मारा प्यारा देशवासियों नमस्कार । मन नी वात ना 110 वा एपिसोड मा तमारो स्वागत से ।
Ground Truth: म्हारां लाडला देह ना वासियों, नमस्कार। मन नीं वात नां 110 वां एपिसेड मां तमुनें जोहार से।

Input: नए-नए शिखरों पर वो कदम रखना चाहता है।
Model Output: नवी नवी चोटी पर पेलो पाय राखने चाहे छे
Ground Truth: नवा - नवा खास उसाईयु मा वी कदमु मेळवा मागे से ।

Input: आज तो गाँव-गाँव में ड्रोन दीदी की इतनी चर्चा हो रही है कि हर किसी की जुबान पर यह नाम है।
Model Output: आज ते गाम गाम मा ड्रोन दीदी नी एतरी वातियो थाई री से के हर कइना नी बोली पर यो नाम से।
Ground Truth: आजतें गाम-गाम मां ड्रोन जीजी नी ऐतरी वात्युं थाईरी से कि कईनानी पण जिबान उपर यू नाम से।

(a) In the first example, the model produces the words in a high-resource language, failing to use culturally relevant words like "जोहार" instead of "नमस्कार, whereas in second it mixed from Gujarati language context and in term fails to predict more relevant words in the context.

Translate from **Bhili to Hindi** :

Input : भारतीय समूह मा 2007 मा इंग्लैंड रमवा ने हाते राहुल द्रविड़ नी मोटाई मेलवानी इच्या नो पतों मालम थायो ||
Model Ouput: भारतीय दल को 2007 में इंग्लैंड के खिलाफ खेलने के साथ राहुल द्रविड़ की प्रतिभा साझा करने की इच्छा का पता चला।
Ground Truth: भारतीय टीम के 2007 के इंग्लैंड दौरे के दौरान, राहुल द्रविड़ की कप्तानी छोड़ने की इच्छा का पता चला।

Input : सोरो ने काला मोजा अने एडिडास सैडल पेहरी मोलीया से |
Model Ouput: बच्चों को ब्लैक मोजे और सैडल दिए गए हैं।
Ground Truth: लड़के ने काले मोजे के साथ एडिडास की काली सैंडल भी पहनी हुई थी।

(b) In the first example, the model misinterpreted "कप्तानी छोड़ने" (resigning from captaincy) as "प्रतिभा साझा करने" (sharing talent). This indicates a semantic error where the meaning of the phrase is incorrectly conveyed whereas in second case the model ignored the brand name "Adidas" and mistranslated the context. Omission of key details like "Adidas" in the translation.

Figure 8: Error analysis in the predictions from the fine-tuned NLLB-200 distilled 600M variant model on the evaluation dataset highlighting errors in the model prediction from Hindi to Bhili and Bhili to Hindi direction.

Error Severity	Description
Very High	Errors that fundamentally alter or obscure the original semantic content, especially in pivotal passages, thereby risking substantial misinterpretation by the reader.
Very Low	Minor blemishes that preserve the core meaning yet introduce subtle stylistic or grammatical inconsistencies, marginally affecting fluency or reader engagement.

Table 9: Definitions of error severity end-points based on impact on meaning and readability

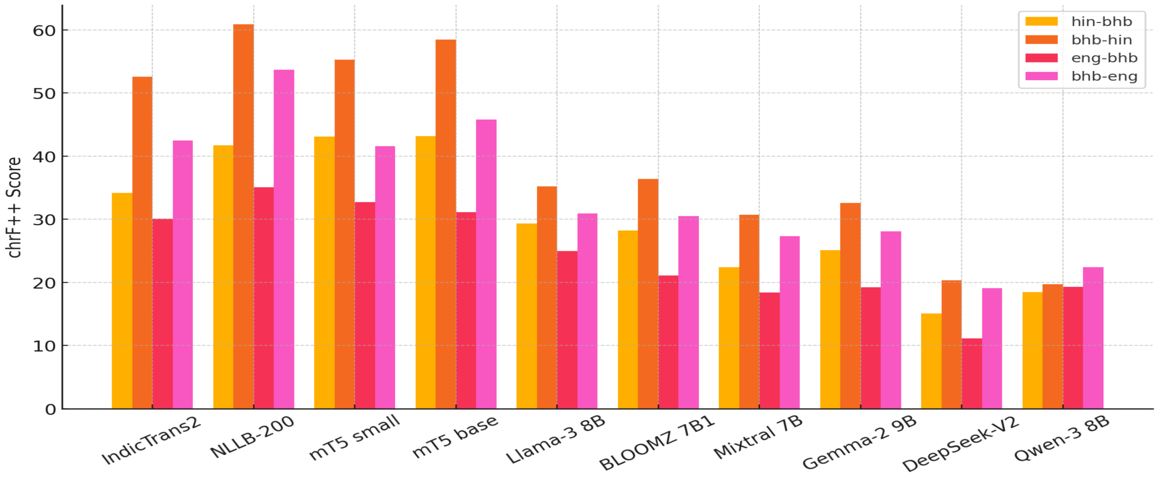


Figure 9: chrF++ scores across models and translation directions

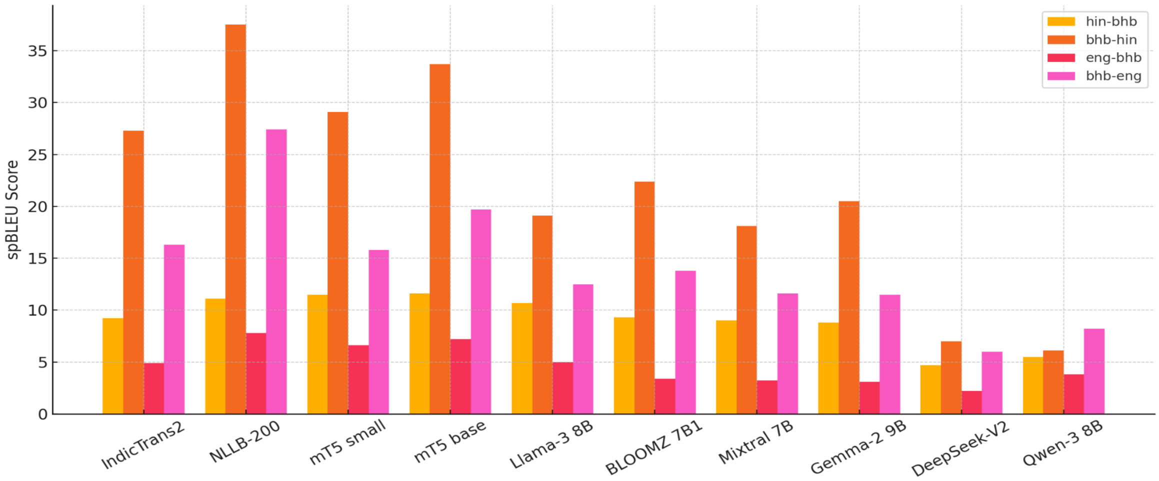
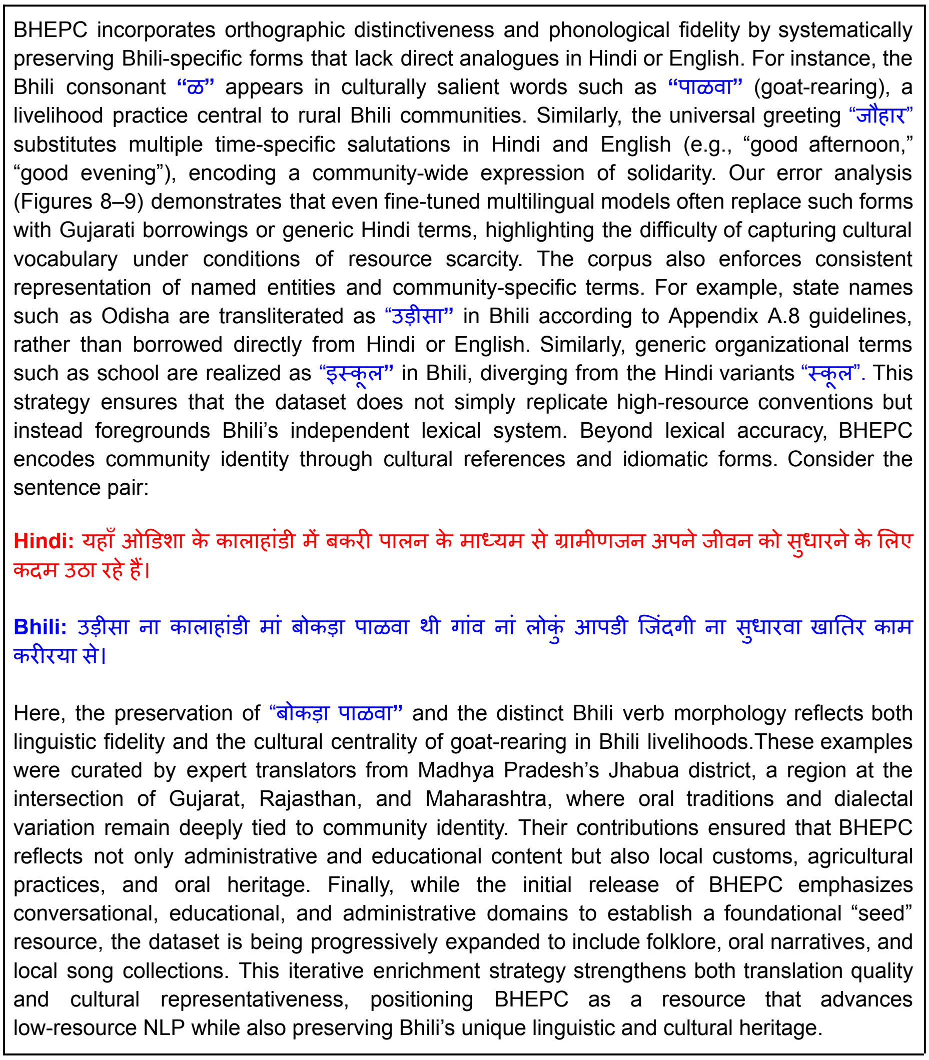


Figure 10: spBLEU scores across models and language pairs

practices, idiomatic expressions, and orthographic features that characterize Bhili usage in real contexts. As noted in Section 5.2, model errors frequently arise from the substitution of Bhili-specific lexical forms with Gujarati borrowings or standard Hindi constructions, underscoring the importance of explicitly encoding cultural vocabulary in the dataset. Further details are provided in Figure 11.

A.10.2 Scalability and Generalizability

Achieving global linguistic inclusivity requires solutions that are both scalable and resource-efficient, particularly for the thousands of languages with minimal or no digital presence. In this work, we systematically evaluated open-source and proprietary multilingual models ranging from 300M to 17B parameters to assess their adaptability to Bhili, an extremely low-resource language that shares its script with Hindi. While these models perform



BHEPC incorporates orthographic distinctiveness and phonological fidelity by systematically preserving Bhili-specific forms that lack direct analogues in Hindi or English. For instance, the Bhili consonant **“ळ”** appears in culturally salient words such as **“पाळवा”** (goat-rearing), a livelihood practice central to rural Bhili communities. Similarly, the universal greeting “जौहार” substitutes multiple time-specific salutations in Hindi and English (e.g., “good afternoon,” “good evening”), encoding a community-wide expression of solidarity. Our error analysis (Figures 8–9) demonstrates that even fine-tuned multilingual models often replace such forms with Gujarati borrowings or generic Hindi terms, highlighting the difficulty of capturing cultural vocabulary under conditions of resource scarcity. The corpus also enforces consistent representation of named entities and community-specific terms. For example, state names such as Odisha are transliterated as **“उड़ीसा”** in Bhili according to Appendix A.8 guidelines, rather than borrowed directly from Hindi or English. Similarly, generic organizational terms such as school are realized as **“इस्कूल”** in Bhili, diverging from the Hindi variants “स्कूल”. This strategy ensures that the dataset does not simply replicate high-resource conventions but instead foregrounds Bhili’s independent lexical system. Beyond lexical accuracy, BHEPC encodes community identity through cultural references and idiomatic forms. Consider the sentence pair:

Hindi: यहाँ ओडिशा के कालाहांडी में बकरी पालन के माध्यम से ग्रामीणजन अपने जीवन को सुधारने के लिए कदम उठा रहे हैं।

Bhili: उड़ीसा ना कालाहांडी मां बोकड़ा पाळवा थी गांव नां लोकुं आपडी जिंदगी ना सुधारवा खातिर काम करीरया से।

Here, the preservation of **“बोकड़ा पाळवा”** and the distinct Bhili verb morphology reflects both linguistic fidelity and the cultural centrality of goat-rearing in Bhili livelihoods.These examples were curated by expert translators from Madhya Pradesh’s Jhabua district, a region at the intersection of Gujarat, Rajasthan, and Maharashtra, where oral traditions and dialectal variation remain deeply tied to community identity. Their contributions ensured that BHEPC reflects not only administrative and educational content but also local customs, agricultural practices, and oral heritage. Finally, while the initial release of BHEPC emphasizes conversational, educational, and administrative domains to establish a foundational “seed” resource, the dataset is being progressively expanded to include folklore, oral narratives, and local song collections. This iterative enrichment strategy strengthens both translation quality and cultural representativeness, positioning BHEPC as a resource that advances low-resource NLP while also preserving Bhili’s unique linguistic and cultural heritage.

Figure 11: Examples illustrate cultural and orthographic features preserved by the corpus, such as community greetings, livelihood terms (e.g., goat-rearing), and standardized transliteration of named entities. The paired sentence examples show how BHEPC captures community-specific vocabulary and morphology across domains (rural life, education, administration) while maintaining clean parallel alignment.

well on high-resource languages, they struggle to generate culturally accurate and linguistically rich translations for Bhili, as demonstrated in our error analyses (Figures. 7 & 8) and performance comparisons (Tables. 2 & 3). This performance gap arises because high-resource languages benefit

Model	Hin-Eng (0-shot)	Hin-Eng (10-shot)	Eng-Hin (0-shot)	Eng-Hin (10-shot)
Llama-2-7B	6.78 / 38.26	10.84 / 41.21	6.53 / 22.06	7.50 / 24.00
Llama-3-8B	40.18 / 65.72	40.57 / 65.88	7.91 / 26.79	9.50 / 30.00
BLOOMZ-560M	3.92 / 15.21	3.96 / 17.28	2.50 / 12.00	3.50 / 15.00
BLOOMZ-3.1B	12.70 / 30.30	18.54 / 42.27	10.00 / 25.00	11.00 / 35.00
BLOOMZ-7B1	28.18 / 53.10	30.45 / 50.82	19.70 / 39.85	22.00 / 41.00
Mistral-7B-v0.1	0.41 / 3.06	0.42 / 14.38	0.30 / 2.00	0.40 / 3.00
Gemma-2-9B	28.36 / 64.65	35.60 / 66.65	13.37 / 43.54	33.99 / 52.73
GPT-3.5 Turbo	41.48 / 66.89	43.62 / 68.28	30.00 / 55.00	33.00 / 58.00
GPT-4o-0513	50.94 / 73.22	53.37 / 74.31	35.00 / 60.00	38.00 / 63.00
GPT-4.5	52.94 / 75.22	54.61 / 77.66	38.00 / 65.00	42.00 / 68.00

Table 10: Hindi–English bidirectional results on the test set: spBLEU/chrF++ (↑) for zero-shot and 10-shot in Hin→Eng and Eng→Hin.

from large and diverse corpora, whereas Bhili suffers from the absence of standardized orthography, lack of monolingual corpora, and minimal prior digital representation. For instance, certain Bhili-specific words contain unique orthographic forms that do not exist in Hindi or other high-resource languages, which often leads to systematic mistranslations as shown in Figure 11. Producing fluent Bhili thus requires handling complex morphology and culturally grounded vocabulary that current models underrepresent. Consistent with prior low-resource MT literature, the first critical step toward improving translation for such languages is the creation of a representative parallel corpus. Our dataset is the first publicly available resource for Bhili, developed through community-driven efforts. To balance quality with scalability, we adopted a hybrid workflow: (i) curating a seed corpus of 80,000 sentences with professional translators, (ii) fine-tuning models to reach a reliable level of accuracy, and (iii) using the fine-tuned models to generate additional Bhili sentence pairs, which were then post-edited by native speakers. This iterative pipeline substantially reduces human effort compared to fully manual translation while preserving linguistic and cultural fidelity.

We believe that this hybrid pipeline combining modest manual seeding, model-assisted generation, and post-editing, offers a scalable methodology that can be extended to other low-resource and endangered languages. While it does not eliminate the inherent costs of corpus creation, it provides a practical pathway for bootstrapping translation resources across multiple languages, thereby contributing to more inclusive global language technologies.

A.10.3 Baseline Performance of Hindi–English Bidirectional MT

To contextualize the difficulty of Bhili translation, we also evaluated our model suite on the high-resource Hindi↔English pair. Table 10 presents results under both zero-shot and 10-shot in-context settings, evaluated with spBLEU and chrF++. Across all models, Hindi↔English achieves substantially higher performance than Bhili↔(English/Hindi), with differences of approximately 20–30 spBLEU and 30–40 chrF++ points. For example, GPT-4.5 obtains a 10-shot chrF++ of 77.66 on Hindi→English, compared to only 25.67 on English→Bhili. This stark gap highlights the effect of severe data scarcity and cultural specificity in Bhili, even for advanced multilingual models.

A.10.4 Scaling Open-Source Models

We further examined the effect of model scaling by evaluating larger variants of two best-performing open-source baselines: mT5-large (1.2B) and NLLB-1.3B. The results show that NLLB-1.3B performs comparably to its smaller 600M counterpart while consistently outperforming mT5-(Base & large) in both spBLEU and chrF++ across all four translation directions. These findings reinforce our earlier observation (Section 4.3, Table 3) that the NLLB architecture is particularly well-suited for low-resource translation. Nevertheless, even at this scale, a substantial gap remains between open-source models and proprietary systems such as GPT-4.5, indicating that architectural design and domain adaptation are as crucial as model size in advancing low-resource machine translation.

Model	ΔchrF++ [95 % CI]	p
mT5-base	+0.56 [0.32, 0.80]	0.003
IndicTrans2	–7.16 [–7.55, –6.78]	<0.001
BLOOMZ-7B1	–6.12 [–6.50, –5.75]	<0.001
Gemma-2-9B	–4.88 [–5.20, –4.55]	<0.001
Llama-3-8B	–7.42 [–7.80, –7.10]	<0.001
Mixtral-7B-v0.1	–8.10 [–8.50, –7.70]	<0.001

Table 11: ΔchrF++ (95% CI, p-values) vs. NLLB-200 for Hindi→Bhili.

A.10.5 Statistical Significance Testing

To evaluate whether observed performance differences between finetuned models are statistically significant, we applied paired bootstrap resampling with 1,000 iterations, following established practice in MT evaluation (Wong, 2005). All tests were conducted at the segment level using chrF++, which we found to correlate most closely with human judgments (see Section 5.1). For each system pair (e.g., NLLB-200 vs. another model), we repeatedly resampled the test set with replacement and computed the mean chrF++ difference. The resulting distribution of 1,000 differences was used to estimate the 95% confidence interval and the two-tailed p-value. Tables 11–14 report ΔchrF++ scores with 95% confidence intervals and p-values relative to NLLB-200. The only exception is the Hindi→Bhili direction, where mT5-base achieves a small but significant advantage (+0.56 chrF++, $p = 0.003$). In all other directions (bhb→hin, eng→bhb, bhb→eng), NLLB-200 significantly outperforms all open-source baselines ($p < 0.005$). For example, in the English→Bhili setting, the mean gain of NLLB-200 over BLOOMZ-7B1 is –5.75 chrF++ with a 95% confidence interval excluding zero ($p \ll 0.001$). These findings confirm that the reported improvements are statistically robust and consistent with both automatic metrics and human evaluations. Bootstrap significance testing demonstrates that NLLB-200's improvements are not only numerically higher but also statistically reliable, thereby strengthening the validity of our conclusions.

A.11 Data Preprocessing Details

To ensure the reliability and consistency of the Bhili–Hindi–English Parallel Corpus (BHEPC), we adopted a multi-stage preprocessing pipeline that combined automated filtering with manual validation.

Model	ΔchrF++ [95 % CI]	p
mT5-base	–0.62 [–0.85, –0.39]	0.002
IndicTrans2	–6.85 [–7.20, –6.50]	<0.001
BLOOMZ-7B1	–5.90 [–6.25, –5.55]	<0.001
Gemma-2-9B	–4.75 [–5.10, –4.40]	<0.001
Llama-3-8B	–7.05 [–7.40, –6.70]	<0.001
Mixtral-7B-v0.1	–7.80 [–8.15, –7.45]	<0.001

Table 12: ΔchrF++ (95% CI, p-values) vs. NLLB-200 for Bhili→Hindi.

Model	ΔchrF++ [95 % CI]	p
mT5-base	–0.48 [–0.70, –0.26]	0.004
IndicTrans2	–6.50 [–6.85, –6.15]	<0.001
BLOOMZ-7B1	–5.75 [–6.10, –5.40]	<0.001
Gemma-2-9B	–4.62 [–4.95, –4.30]	<0.001
Llama-3-8B	–6.88 [–7.22, –6.54]	<0.001
Mixtral-7B-v0.1	–7.25 [–7.60, –6.90]	<0.001

Table 13: ΔchrF++ (95% CI, p-values) vs. NLLB-200 for English→Bhili.

Model	ΔchrF++ [95 % CI]	p
mT5-base	–0.51 [–0.75, –0.27]	0.003
IndicTrans2	–6.65 [–6.99, –6.31]	<0.001
BLOOMZ-7B1	–5.80 [–6.15, –5.45]	<0.001
Gemma-2-9B	–4.70 [–5.05, –4.35]	<0.001
Llama-3-8B	–7.12 [–7.46, –6.78]	<0.001
Mixtral-7B-v0.1	–7.55 [–7.90, –7.20]	<0.001

Table 14: ΔchrF++ (95% CI, p-values) vs. NLLB-200 for Bhili→English.

Length Filtering: Approximately 4.3% of sentences were removed based on length. Sentences shorter than 6 words were often repetitive or contextually uninformative (e.g., "Thank you," "Yes, sir"), while sentences longer than 80 words introduced alignment and tokenization difficulties. These thresholds follow common heuristics in multilingual NMT datasets such as FLORES-200 and BPCC.

Near-Duplicate Removal: We excluded 1,867 sentence pairs with cosine similarity above 0.95 to avoid redundancy and preserve content diversity.

Normalization: Over 1,200 lexical and orthographic variants in Bhili were standardized using a phonetic lexicon curated by native speakers. Additional script normalization was applied across all three languages to reduce variation and ensure consistency.

Screening for PII, Hate Speech, and Redundancy: The screening process combined automated and manual checks. Automated steps included script normalization, strict de-duplication, and cosine similarity based redundancy filtering. Human reviewers inspected flagged cases to ensure accuracy. Hindi sentences were sourced from vetted public corpora (BPCC, PMIndia, NCERT, Legislative Assembly proceedings), which are inherently low-risk for PII or offensive content. Bhili translations were produced by native speakers following translation guidelines, while English translations were validated for semantic fidelity.

English Translations: The English portion of the corpus was generated using the IndicTrans2 model. To mitigate potential noise, ten bilingual experts reviewed a stratified subset of outputs over two weeks, flagging and post-editing sentences with critical errors. Approximately 1.6% of sentence pairs were removed due to hallucination, misalignment, or semantic mismatch. Only translations that passed this validation were retained.

By combining automated preprocessing with manual oversight, BHEPC adheres to established corpus construction practices while maintaining a high standard of linguistic and cultural accuracy. The resulting dataset offers a reliable foundation for both training and evaluation in low-resource machine translation.