

Linguistic Proximity Enables Pivot-Based Machine Translation for an Under-Resourced Tribal Language

Pooja Singh^{1*} M Kaab Bin Shahid^{2*} Atai Waris Khan³

Aryan Kumar Jha⁴ Sandeep Kumar¹

¹Indian Institute of Technology Delhi ²University of Stuttgart

³NIT Jalandhar ⁴Indian Institute of Technology Delhi, Abu Dhabi, UAE

{eez228470,ksandeep}@ee.iitd.ac.in st200141@stud.uni-stuttgart.de

ataiwbk.ai.24@nitj.ac.in aryankumarjha2007@gmail.com

Abstract

For more than 10.4 million speakers of Bhili, an under-resourced tribal Indo-Aryan language of central and western India, limited machine translation support restricts digital accessibility and language inclusion. Beyond data scarcity, this challenge is intensified by the linguistic distance between English and Bhili. We examined whether this gap can be reduced through linguistically motivated pivot translation via Hindi and Marathi, two better-resourced languages that are more closely related to Bhili. Computational analysis shows substantially higher character-level overlap with Hindi and Marathi (67.6% and 57.5%) than with English (1.9%), providing an empirical basis for pivot selection. In supervised experiments, cascaded fine-tuning via pivot languages improves the strongest direct translation of NLLB-200 baseline from 7.87 to 9.46 BLEU for English→Bhili and from 25.23 to 29.84 BLEU for Bhili→English, with corresponding ChrF++ gains of +1.91 and +5.18. Pivot-based prompting further shows that LLMs can benefit from related-language cues: Gemini 2.5 Flash improves by +20.09 ChrF++ for English→Bhili using Hindi as the pivot, while GPT-4.5 reaches 46.87 ChrF++ for Bhili→English. We release **BhilSetu**, a five-language parallel corpus of 1,33,000 sentences and the first Bhili FLORES-200-aligned benchmark. Our results suggest that leveraging linguistic proximity can improve translation quality while supporting broader efforts toward digital inclusion and preservation of under-resourced indigenous languages.

1 Introduction

Neural Machine Translation (NMT) has advanced rapidly for high-resource languages through large-scale parallel corpora and multilingual pre-trained

models (Devlin et al., 2019; Faheem et al., 2024). However, this progress remains uneven: many under-resourced languages, including Bhili, Santali, and Bodo, are still largely absent from translation systems that mediate access to education, governance, healthcare, and digital communication. Even commercial systems cover only a small fraction of the world’s 7,000+ languages (Haddow et al., 2022), leaving many linguistic communities without basic translation support (Fan et al., 2021).

Bhili occupies a particularly urgent position among these excluded languages. Spoken by more than 10.4 million¹ people across Madhya Pradesh, Gujarat, and Maharashtra, it is one of India’s largest tribal languages, yet it has limited presence in formal education, scarce digitized resources, and almost no dedicated translation infrastructure. Since many Bhili speakers belong to Scheduled Tribe communities facing longstanding marginalization, the absence of translation technologies for Bhili represents not merely a technical gap but a question of linguistic justice and cultural survival. It also affects access to public services and mother-tongue information, while language shift toward dominant regional languages threatens indigenous knowledge embedded in agricultural practices, medicinal taxonomies, kinship systems, ecological memory, and oral traditions (Joshi et al., 2020).

However, developing NMT for Bhili remains challenging even with the available parallel data. Although the BHEPC corpus (Singh et al., 2025) provides Bhili–English–Hindi parallel data, models fine-tuned directly on English↔Bhili produce substantially degraded translations (Section 6.2), while LLMs also perform poorly in zero-shot settings (Table 3), despite stronger performance reported for other low-resource languages (Elsner

*Equal contribution.

¹https://en.wikipedia.org/wiki/Bhil_languages

et al., 2024). One factor contributing to this difficulty is the linguistic distance between Germanic English and Indo-Aryan Bhili: models must learn correspondences between structurally different linguistic systems (Littell et al., 2017), a challenge compounded by Bhili’s extremely limited representation in model pre-training.

Pivot-based translation provides a promising alternative (Utiyama and Isahara, 2007; Wu and Wang, 2009): rather than directly bridging this typological gap, translation is routed through intermediate languages that are structurally closer to Bhili (Zhu et al., 2025). Its effectiveness, however, depends on selecting pivots that are both linguistically close to Bhili and sufficiently well-resourced to support reliable translation. Hindi, Marathi, and Gujarati are natural candidates given their shared Indo-Aryan ancestry (Masica, 1991), geographical proximity, and long-standing contact with Bhili-speaking communities. Rather than relying solely on these intuitions, we evaluate their suitability empirically by quantifying morphological, lexical, and syntactic proximity in Section 3.

Building on this analysis, we present the first comprehensive analysis of pivot-based Bhili translation for English↔Bhili, covering both cascaded fine-tuning of multilingual models (mT5 (Xue et al., 2021), NLLB (Costa-Jussà et al., 2022), Llama-3.1-8B (Grattafiori et al., 2024), and BLOOMZ-7B1 (Muennighoff et al., 2023)) and zero-shot pivot prompting with six contemporary LLMs (GPT-4.5 (Achiam et al., 2023), Llama-3.1-8B, Gemma-2-9B (Team et al., 2024), BLOOMZ-7B1, Mistral-7B-v0.1 (Jiang et al., 2023), and Gemini-2.5 Flash (Comanici et al., 2025)).

Our contributions are:

1. **Linguistic Analysis:** First systematic computational analysis of Bhili’s structural relationship to surrounding Indo-Aryan languages, quantifying morphological, lexical, and syntactic proximity to Hindi, Gujarati, and Marathi, establishing theoretical foundations for pivot selection.
2. **Multilingual Corpus:** We introduce **BhilSetu**, the largest 1,33,000-sentence, five-language parallel corpus covering Bhili, English, Hindi, Gujarati, and Marathi.
3. **Bhili FLORES-200 benchmark:** We create the first Bhili-aligned FLORES-200 test set with 997 sentences across all five languages.

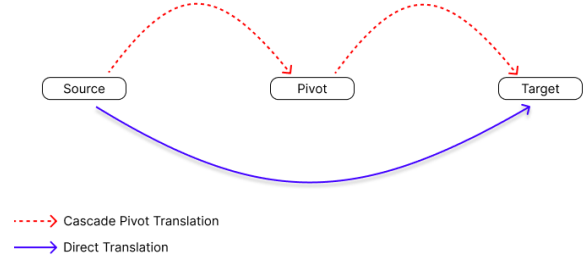


Figure 1: Direct and cascade pivot translation pathways. The blue arrow denotes direct source-to-target translation, while the red dashed arrows denote cascade translation through a pivot language.

4. **Cascaded Fine-tuning:** Cascaded fine-tuning strategies with sequential source-to-pivot and pivot-to-target training, demonstrating substantial improvements over direct translation.
5. **Pivot-based prompting:** We systematically evaluate whether LLMs benefit from translating through intermediate Indo-Aryan languages under zero-shot prompting.

2 Related Work & Background

Translation between high resource and low-resource endangered languages has remained a persistent challenge in NMT (Zhu et al., 2024). Several frameworks have been proposed to adapt existing NMT models for low-resource settings (Sennrich et al., 2016), utilizing data augmentation (Elmadani and Buys, 2024), transfer learning (Zoph et al., 2016), and pre-training strategies (Elmadani and Buys, 2024). Multilingual transformer-based architectures such as mBERT (Pires et al., 2019), and IndicTrans 2 (Gala et al., 2023) have further enhanced cross-lingual transfer capabilities for Indian languages. However, despite notable progress for major Indian languages, extremely low-resource languages like Bhili remain largely absent from these multilingual models.

Pivot-based translation strategies offer an effective alternative when the linguistic distance between the high- and low-resource language is large (as quantified in Section 3). Existing work on Indic pivoting, however, has two recurring limitations. First, it predominantly relies on *high-resource* pivots such as English to bridge gaps between Indic languages (Zhang et al., 2024), and where multiple pivots are combined, they are typically ensembled to boost coverage rather than selected for structural proximity to the target (Gaik-

wad et al., 2024). Second, pivot choice is rarely justified empirically; linguistically related *regional* languages are seldom evaluated as candidates. For instance, Gujarati and Marathi share substantial overlap with Bhili but have not been systematically explored in the literature. Our work departs from both practices: rather than defaulting to English or to convenience-driven selection, we choose pivots (Hindi, Marathi) through a quantitative linguistic-proximity analysis and route through related Indo-Aryan intermediaries.

In parallel, LLMs have introduced new paradigms in low-resource translation through in-context learning (ICL) and multi-lingual prompting (Chitale et al., 2024; Brown et al., 2020). Studies have shown that LLMs can effectively generalize translation patterns even from limited examples, leveraging linguistic cues embedded in prompts (Cahyawijaya et al., 2024; Zhang et al., 2023). However, their effectiveness for tribal and endangered languages remains underexplored. Linguistic studies suggest that translation accuracy improves with greater relatedness between source and pivot languages (Siripragada et al., 2020), but empirical evidence for this in the Indian tribal context is scarce.

3 Linguistic Analysis and Pivot Selection

Motivation and Analytical Framework. The selection of pivot languages in low-resource machine translation is not a neutral decision. It is a linguistically consequential choice whose implications compound across both legs of the translation pipeline. A poorly chosen pivot, one that is typologically distant from the target language, amplifies errors at each stage (Dabre et al., 2017), ultimately producing translations that fail both lexically and structurally. For Bhili, identifying optimal pivot languages requires systematic analysis of its structural relationships to candidate intermediaries.

We evaluate three candidate pivot languages: Hindi, Marathi, and Gujarati. Yet demographic proximity does not guarantee linguistic commensurability, and the question of which language constitutes the most reliable structural bridge between English and Bhili demands empirical rather than intuitive answers.

To ground this choice quantitatively, we conducted a comprehensive computational analysis

of the BhilSetu parallel corpus (1,33,000 sentence pairs), examining six complementary metrics across lexical, morphological, and syntactic dimensions (Lin et al., 2019). Specifically, we measured: *character-level overlap* (Jaccard similarity over character sets, capturing orthographic affinity); *word-level overlap* (Jaccard similarity over token sets, capturing direct lexical sharing); *sentence length correlation* (Pearson r , capturing syntactic complexity parallelism); *normalized edit distance* (mean Levenshtein distance normalized by maximum sentence length, capturing surface form divergence); *character trigram similarity* (Jaccard over character 3-grams, capturing sub-word morphological patterns); and *heuristic cognate density* (proportion of word pairs with normalized Levenshtein similarity ≥ 0.6 , serving as a proxy for shared etymological vocabulary (Inkpen et al., 2005)). Together, these metrics capture complementary aspects of cross-linguistic proximity that are theoretically linked to pivot translation quality: lexical overlap (Mohammadshahi et al., 2024) predicts vocabulary preservation, morphological similarity facilitates inflection transfer, and syntactic correlation reduces reordering errors across the two translation stages, computed from the BhilSetu corpus using the six metrics defined above (Table 1).

Hindi and Marathi: Devanagari-Normalized Evidence. Hindi and Marathi, sharing Devanagari script with Bhili, show consistent proximity across all six metrics as shown in Table 1. Hindi is the closest computational neighbor (character overlap 0.676, word overlap 0.150, length correlation $r = 0.945$, normalized edit distance 0.472, trigram similarity 0.247, cognate density 2.71%), reflecting both genetic relatedness and parallel orthography and syntax. Marathi ranks second (character overlap 0.575, word overlap 0.042, $r = 0.848$), maintaining strong structural similarity despite more divergent vocabulary due to Sanskrit *tatsama* influence and southern contact.

The Gujarati Paradox: Sociolinguistic Proximity vs. Computational Distance. Gujarati illustrates a gap between sociolinguistic reality and computational evidence. In Bhili speech communities, Gujarati is omnipresent: speakers in the eastern Gujarat tribal belt (Dahod, Panchmahal, Chhota Udaipur) code-switch fluidly with Gujarati. Yet our analysis yields very low Gujarati scores

Table 1: Computational linguistic similarity metrics between Bhili and candidate pivot languages. $\uparrow$ = higher is more similar; $\downarrow$ = lower is more similar.

Language	Char. $\uparrow$	Word $\uparrow$	Len. Corr. $\uparrow$	Edit Dist. $\downarrow$	3-gram $\uparrow$	Cognate% $\uparrow$
Hindi	0.676	0.150	0.945	0.472	0.247	2.71%
Marathi	0.575	0.042	0.848	0.689	0.114	1.23%
Gujarati	0.012	0.003	0.886	0.891	0.003	0.04%

across all character-level metrics. This is because, for example, *ghar* (house) is written घर in Bhili, Hindi, and Marathi (Devanagari) but ઘર in Gujarati script, thus distinct Unicode code points encoding the same phoneme. Although Gujarati shares syntactic complexity patterns with Bhili ($r = 0.886$), two languages may share deep lexical and morphological ancestry yet exhibit near-zero character overlap simply because they employ different writing systems.

Pivot Selection. Based on the computational evidence and sociolinguistic context (Table 1), we select **Hindi and Marathi** as pivot languages. Hindi is justified by its dominance across all valid metrics and its role as the primary contact language in Bhili speakers’ educational and civic lives. Marathi is justified by its consistent secondary proximity and its geographic overlap with Bhili communities in northern Maharashtra.

4 Dataset

4.1 Corpus Construction & Processing

We extend the Bhili–Hindi–English Parallel Corpus (BHEPC) by adding parallel Marathi and Gujarati sentences and also added 23,000 human-translated high-quality sentence pairs across the five languages, resulting in a final corpus **BhilSetu** of 1,33,000 parallel sentences. This makes **BhilSetu** the first large-scale, human-curated, gold-standard dataset for Bhili MT across five languages. The additional 23,000 Hindi source sentences were primarily drawn from the Bharat Parallel Corpus Collection (BPCC) (Gala et al., 2023), government speeches (Ramesh et al., 2022; Kunchukuttan et al., 2018), and NCERT textbooks². All translations were created and validated by native-speaking linguists to ensure contextual fidelity and naturalness. The dataset statistics are summarized in Table 2.

²<https://ncert.nic.in/textbook.php>

Table 2: Statistics of the corpus and FLORES-200 test set.

Dataset	Train	Test (FLORES)	Langs
BHEPC	1,10,000	-	3
Bhili-FLORES	-	997	5
BhilSetu	1,33,000	997	5

To ensure data quality, we applied preprocessing steps, including normalization of Bhili homophones, removal of extraneous symbols, lowercasing of English text, and strict de-duplication using cosine-similarity filtering. Sentences shorter than 6 or longer than 80 words were discarded. Throughout this work, we use bhb to denote Bhili, following its ISO 639-3 language identifier.

4.2 Evaluation Datasets & Metrics

For evaluation, we use the *FLORES-200* benchmark dataset (Costa-Jussà et al., 2022), a multilingual corpus created to assess translation quality across 200 languages. We rely on its *devtest* split, which contains 997 parallel sentences per language pair, allowing standardized evaluation of multilingual and low-resource MT systems. *FLORES devtest* is translated into Bhili with the help of language experts, and the pivot languages, Hindi, Gujarati and Marathi to assess the reliability of pivot-based translation across two directions. Performance is measured using BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2017), which evaluate n-gram precision and character-level similarity respectively, both suitable for morphologically rich Indian languages. All scores are computed with SacreBLEU (Post, 2018) and NLTK (Bird et al., 2009) for reproducibility.

4.3 Model Training & Hyperparameters

All models are trained under a consistent setup to enable fair comparison. For Llama-3.1-8B and Bloomz-7B1 we experiment with Parameter-Efficient Fine-Tuning (PEFT) using LoRA (Hu et al., 2022), rank 8. Key hyperparameters include

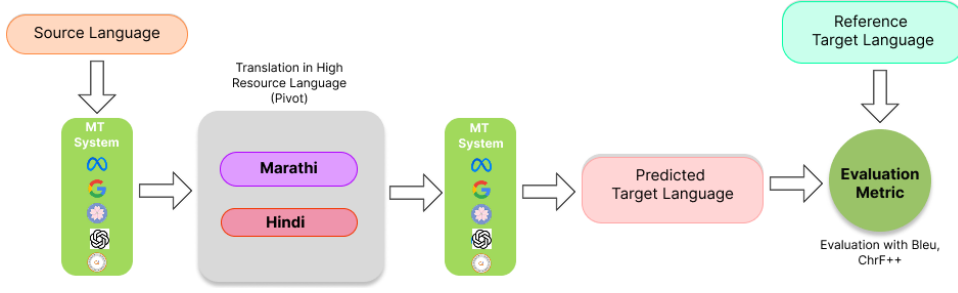


Figure 2: Illustration of the Pivot-Based Translation Framework.

Adam optimizer, learning rate $5e-5$, batch size 16, weight decay 0.01, and early stopping patience of 10 based on validation chrF++ scores. For prompting, all models are evaluated with temperature 0 and the same fixed prompt template. Five-shot examples are sampled once from the training set and kept identical across all models. Experiments are conducted on a NVIDIA A100 GPUs, with training times ranging from 10 to 72 hours.

5 Methodology

Pivot-based translation addresses the fundamental challenge of translating between a high-resource and an extremely low-resource language by routing through one or more intermediate languages that maintain closer linguistic affinity to the target. As established in Section 3, direct translation between English and Bhili is hampered not only by data scarcity but by the wide typological gap separating a Germanic language from a tribal Indo-Aryan one. Even with the 1,10,000-sentence BHEPC corpus, models trained directly on English–Bhili pairs produce degraded output, motivating the pivot architectures described here (see Table 3). Figures 1 and 2 illustrate the direct and cascaded pivot pipelines.

Within this work, we explore two distinct pivoting strategies, **1. Cascading Pivoting** and **2. Pivot Prompting**, evaluated across two translation directions: English→Bhili and Bhili→English, using Hindi and Marathi as pivot languages (see Section 3 for linguistic justification).

5.1 Cascading Pivoting

The cascading approach decomposes translation into two sequential, independently trained stages: source-to-pivot ($S \rightarrow P$) and pivot-to-target ($P \rightarrow T$). Each stage leverages the same underlying multilingual model architecture, ensuring consistent vocabulary representations and embedding spaces

across both directions. The shared model parameters promote stable lexical and syntactic mappings, reducing the distributional drift that commonly afflicts cascade pipelines operating with heterogeneous model pairs. Let $\mathbf{x}$, $\mathbf{z}$, and $\mathbf{y}$ denote the source, pivot, and target sequences, respectively. Ideally, the final translation can be obtained by marginalizing over all possible pivot sequences:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} \sum_{\mathbf{z}} P(\mathbf{y} | \mathbf{z}) P(\mathbf{z} | \mathbf{x}). \quad (1)$$

Since exact marginalization over all pivot sequences is computationally intractable, we approximate this process using a standard two-step cascade. First, the source sentence is decoded into the most likely pivot sentence:

$$\hat{\mathbf{z}} = \arg \max_{\mathbf{z}} P(\mathbf{z} | \mathbf{x}), \quad (2)$$

and the resulting pivot sentence is then decoded into the final target sentence:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} P(\mathbf{y} | \hat{\mathbf{z}}). \quad (3)$$

For training, the $S \rightarrow P$ and $P \rightarrow T$ models are optimized separately using parallel source–pivot and pivot–target sentence pairs. Given triplets $(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}, \mathbf{y}^{(i)})$, the training objective is the sum of the two negative log-likelihood losses:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \left[\log P(\mathbf{z}^{(i)} | \mathbf{x}^{(i)}) + \log P(\mathbf{y}^{(i)} | \mathbf{z}^{(i)}) \right] \quad (4)$$

This sequential decomposition provides modularity: the $S \rightarrow P$ and $P \rightarrow T$ components can be trained, evaluated, and updated independently, and existing high-resource translation models can be reused for the first stage without re-training from scratch. Concretely, we fine-tune a single multilingual checkpoint on both the source–pivot and

pivot–target directions (Eq. 4). At inference, the model is used twice: first to generate the pivot hypothesis $\hat{\mathbf{z}}$ from the source, and then to generate the target $\hat{\mathbf{y}}$ from $\hat{\mathbf{z}}$ (Eqs. 2–3). Using the same model for both steps keeps the vocabulary and embedding space consistent and helps reduce the distributional drift that can arise when different models are used for the two stages of the cascade.

5.2 Pivot Prompting

While cascading relies on separately fine-tuned models for each translation stage, the pivot prompting approach leverages a single multilingual LLM to perform zero-shot translation across pivot pathways within a unified prompt. Rather than producing an explicit intermediate output, the model receives the source sentence alongside a pivot language in its context and is instructed to translate directly into the target while implicitly reasoning through the specified intermediate steps.

This design allows the model to internally synthesize lexical and structural cues from pivot languages simultaneously, without committing to a single intermediate output as cascading requires. Conceptually, the pivot formulation models the conditional distribution:

$$P(\mathbf{y} \mid \mathbf{x}, P_1) \quad (5)$$

where $\{P_1\}$ is the auxiliary pivot language incorporated through the prompt context. Unlike cascading, which explicitly produces a pivot sequence $\mathbf{z}$ as a discrete intermediate artifact, pivot prompting treats pivot-level reasoning as an implicit latent process encoded within the LLM’s internal cross-lingual representations. The model is not required to externalize pivot-language text; instead, it leverages its multilingual pretraining to bridge linguistic gaps internally.

This helps particularly in zero-shot settings where fine-tuning data is unavailable, and where the target language (Bhili) may have minimal direct representation in the model’s pretraining corpus. By prompting through linguistically proximate pivots; Hindi and Marathi, the model is guided toward a region of its representation space closer to Bhili, improving the fidelity of the final output even without Bhili-specific supervision.

5.3 Experimental Setup

We compare three settings across English→Bhili (eng→bhb) and Bhili→English (bhb→eng):

Table 3: Direct translation baseline results for English↔Bhili without pivot languages. Fine-tuned models report BLEU / ChrF++. Direct prompting models report ChrF++ for 0-shot / 5-shot prompting. Best results within each block and metric are shown in **blue**.

Model	eng→bhb	bhb→eng
<i>Fine-tuned models: BLEU / ChrF++</i>		
mT5 Base	7.13 / 35.63	19.76 / 46.45
NLLB-200	7.87 / 37.18	25.23 / 54.34
Llama-3.1-8B	5.12 / 22.23	12.10 / 30.59
BLOOMZ-7B1	4.28 / 21.43	9.12 / 29.78
<i>Direct prompting: 0-shot / 5-shot ChrF++</i>		
Llama-3.1-8B	0.55 / 13.50	12.00 / 16.50
BLOOMZ-7B1	0.81 / 11.12	19.34 / 21.02
Mistral-7B-v0.1	0.76 / 10.87	12.89 / 13.25
Gemma-2-9B	5.27 / 13.53	30.50 / 23.25
Gemini 2.5 Flash	19.35 / 21.30	37.56 / 38.89
GPT-4.5	23.35 / 25.67	42.16 / 43.78

- **Direct Baseline:** the source sentence is translated directly into the target language without any pivot.
- **Cascaded Pivot:** the source sentence is translated into Hindi or Marathi, and the pivot output is then translated into the target language.
- **Pivot Prompting:** a single LLM is prompted to translate the source sentence into the target language using Hindi or Marathi as an implicit pivot cue.

6 Experimental Results and Analysis

6.1 Baseline Models

We evaluate encoder-decoder and decoder-only models to benchmark direct and pivot-based Bhili translation. For supervised experiments, we fine-tune **mT5 Base** (Xue et al., 2021), **NLLB-200 (600M)** (Costa-Jussà et al., 2022), **Llama-3.1-8B** (Grattafiori et al., 2024), and **BLOOMZ-7B1** (Muennighoff et al., 2023) under both direct English↔Bhili translation and cascaded pivot translation through Hindi and Marathi. For zero-shot prompting, we evaluate **Llama-3.1-8B**, **BLOOMZ-7B1**, **Mistral-7B-v0.1** (Jiang et al., 2023), **Gemma-2-9B** (Team et al., 2024), **Gemini 2.5 Flash** (Comanici et al., 2025), and **GPT-4.5** (Achiam et al., 2023) under both direct and pivot-based prompting. These models cover multilingual pre-training, instruction tuning, and different translation regimes.

Table 4: Cascading pivot results for English↔Bhili translation on the Flores-200 benchmark. Scores are BLEU / ChrF++. Δ reports the difference compared with the corresponding direct fine-tuned baseline for the same model. Best pivot result for each model and direction is shown in **blue**.

Model	Pivot	eng→bhb	Δ	bhb→eng	Δ
mT5 Base	Hindi	8.71 / 37.85	+1.58 / +2.22	22.57 / 49.91	+2.81 / +3.46
	Marathi	7.57 / 36.84	+0.44 / +1.21	21.22 / 47.14	+1.46 / +0.69
NLLB-200 (600M)	Hindi	9.46 / 39.09	+1.59 / +1.91	29.84 / 59.52	+4.61 / +5.18
	Marathi	8.51 / 37.97	+0.64 / +0.79	27.58 / 55.17	+2.35 / +0.83
Llama-3.1-8B	Hindi	6.12 / 26.20	+1.00 / +3.97	14.54 / 33.05	+2.44 / +2.46
	Marathi	5.96 / 23.02	+0.84 / +0.79	15.63 / 34.01	+3.53 / +3.42
BLOOMZ-7B1	Hindi	7.58 / 27.96	+3.30 / +6.53	10.85 / 31.03	+1.73 / +2.82
	Marathi	5.38 / 22.86	+1.10 / +1.43	9.37 / 28.42	+0.25 / +0.21

6.2 Direct Translation Baseline

Before assessing the contribution of pivot strategies, we establish direct translation baselines for eng→bhb and bhb→eng without any intermediate pivot language. These results quantify what supervised and zero-shot models can achieve when forced to bridge the full English–Bhili typological gap in a single step, providing the reference point against which all pivot-based improvements are measured. Table 3 shows that direct translation performance is consistently low across model types, confirming the central difficulty of the English–Bhili translation problem. Among fine-tuned models, NLLB-200 achieves the strongest direct baseline in both directions, while direct prompting shows that larger instruction-tuned models perform better but still struggle substantially when Bhili is the target language.

6.3 Cascading Pivot Results

Table 4 presents the results of cascaded pivot translation using Hindi and Marathi as intermediate languages. Pivot-based routing consistently improves over direct translation across all fine-tuned models and both translation directions. For English→Bhili, the strongest result is obtained by NLLB-200 using Hindi as the pivot, improving from 7.87 / 37.18 to 9.46 / 39.09 (+1.59 BLEU, +1.91 ChrF++). Similar gains are observed for mT5 Base, Llama-3.1-8B, and BLOOMZ-7B1, with Hindi generally outperforming Marathi. For Bhili→English, improvements are even more pronounced, with NLLB-200 reaching 29.84 / 59.52 using Hindi as the pivot and Llama-3.1-8B achieving its best performance through Marathi (15.63 / 34.01). Overall, results indicate that linguistically related pivots reduce translation difficulty and provide consistent benefits over direct fine-tuning.

Pivot Effectiveness: Across models, Hindi emerges as the most reliable pivot language, yielding the best performance in the majority of settings. Marathi also provides positive gains, although its improvements are generally smaller and less consistent. These findings align with the linguistic similarity analysis in Section 3, where Hindi exhibits the strongest structural and lexical proximity to Bhili.

6.4 Pivot Prompting Results

Table 5 evaluates pivot prompting under a zero-shot setting. Unlike cascaded fine-tuning, pivot prompting produces more modest gains and exhibits a clear directional asymmetry. For English→Bhili, improvements are generally small for most models, suggesting that current LLMs still struggle to generate fluent Bhili without supervised exposure. The largest gain is observed for Gemini 2.5 Flash, which improves from 19.35 to 39.44 ChrF++ using Hindi as the pivot. In contrast, Bhili→English benefits more consistently from pivot prompting, with GPT-4.5 improving from 42.16 to 46.87 ChrF++ and Gemini 2.5 Flash reaching 44.75 ChrF++. Overall, pivot prompting helps LLMs interpret Bhili inputs but cannot reliably generate fluent Bhili outputs, whereas routing through Hindi or Marathi markedly improves English generation.

6.5 Pivot Contribution: Direct vs. Pivot Comparison

Tables 4 and 5 show that pivot-based translation consistently improves over direct translation (Table 3), but the gains depend strongly on direction, model type, and strategy. Cascaded fine-tuning delivers stable improvements in both directions for all fine-tuned models, with Hindi providing the

Table 5: Full pivot-prompting results for English $\leftrightarrow$ Bhili translation. Scores are ChrF++. Δ reports the gain of the better pivot over the corresponding direct zero-shot baseline. Best pivot result within each model and direction is highlighted in **blue**.

Model	English $\rightarrow$ Bhili				Bhili $\rightarrow$ English			
	Direct	Hindi	Marathi	Δ	Direct	Hindi	Marathi	Δ
Llama-3.1-8B	0.55	1.44	1.39	+0.89	12.00	13.82	10.99	+1.82
BLOOMZ-7B1	0.81	3.76	2.00	+2.95	19.34	19.56	19.76	+0.42
Mistral-7B-v0.1	0.76	1.94	1.65	+1.18	12.89	14.40	13.68	+1.51
Gemma-2-9B	5.27	6.31	4.51	+1.04	30.50	33.86	32.31	+3.36
Gemini 2.5 Flash	19.35	39.44	38.39	+20.09	37.56	44.75	42.05	+7.19
GPT-4.5	23.35	24.56	23.76	+1.21	42.16	46.87	43.24	+4.71

strongest overall pivot in most settings. Translating through a well-represented Indo-Aryan intermediate language improves source interpretability and reduces the structural mismatch before Bhili generation, leading to consistent gains over direct English $\leftrightarrow$ Bhili translation.

In contrast, pivot prompting remains asymmetric and more model-dependent: gains are clearer for bhb $\rightarrow$ eng, where models can generate English more reliably, but are smaller for eng $\rightarrow$ bhb, where limited Bhili generation capacity remains a bottleneck in zero-shot LLMs. While pivot cues improve source understanding, they do not fully compensate for the absence of supervised Bhili training signals. Thus, supervised cascading is preferable when target-language generation must be reliable, whereas pivot prompting mainly benefits translation *from* low-resource Bhili into English.

7 Conclusion

This work presents a linguistically grounded pivot-based approach for English $\leftrightarrow$ Bhili machine translation, targeting an extremely under-resourced tribal language of India. Our results show that direct translation remains difficult across both fine-tuned and prompted models, especially when Bhili is the target language, while routing through related Indo-Aryan pivots provides consistent improvements over direct baselines. Hindi emerges as the most reliable pivot in most settings, with Marathi also offering useful gains for selected models and directions. Our computational linguistic analysis quantifies lexical, morphological, and structural proximity between Bhili and candidate pivot languages, providing an empirical basis for pivot selection. We also release **BhilSetu**, a 1,33,000-sentence five-language parallel corpus covering Bhili, English, Hindi, Gujarati,

and Marathi, along with the first Bhili-aligned FLORES-200 benchmark of 997 sentences. By grounding pivot selection in linguistic proximity rather than resource availability alone, this work offers a practical and reproducible path toward extending machine translation to other marginalized indigenous languages.

Limitations

Although pivot-based training improves Bhili translation performance, error propagation through intermediate pivot languages may still affect final outputs, particularly in morphologically rich and low-resource settings. Future work will explore extending the approach to additional under-resourced languages, improving robustness through synthetic and monolingual data augmentation, and combining Hindi and Marathi through multi-source pivot ensembling and adaptive pivot selection.

Ethical Considerations

All translations and annotations were carried out by qualified native speakers and linguists under informed consent procedures and with compensation provided in accordance with government norms. Publicly available datasets and the newly created **BhilSetu** corpus were used and curated following research ethics and open-data practices.

Acknowledgments

We gratefully acknowledge the Ministry of Tribal Affairs, Government of India, and the Adi Vaani Project, together with the Tribal Research Institute, Bhopal, which enabled the human resources, cloud infrastructure, and expert translators essential to this work. We also thank all translators for their efforts in creating the BhilSetu.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433.
- Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. An empirical study of in-context learning in LLMs for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- 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, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia conference on language, information and computation*, pages 282–286.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Khalid N Elmadani and Jan Buys. 2024. Neural machine translation between low-resource languages with synthetic pivoting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12144–12158.
- Micha Elsner et al. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354.
- Mohamed Atta Faheem, Khaled Tawfik Wassif, Hanaa Bayomi, and Sherif Mahdy Abdou. 2024. Improving neural machine translation for low resource languages through non-parallel corpora: a case study of egyptian dialect to modern standard arabic translation. *Scientific Reports*, 14(1):2265.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.
- Pranav Gaikwad, Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. How effective is multi-source pivoting for translation of low resource indian languages? *arXiv preprint arXiv:2406.13332*.
- Jay P Gala, Pranjal A Chitale, Varun Gumma, Sumanth Doddapaneni, Kumar M Aswanth, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, et al. 2023. Indictans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv 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.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel,

- Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6282–6293.
- Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhat-tacharyya. 2018. The iit bombay english-hindi parallel corpus. In *proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.
- Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.
- Colin P Masica. 1991. *The indo-aryan languages*. Cambridge University Press.
- Alireza Mohammadshahi, Jannis Vamvas, and Rico Sennrich. 2024. Investigating multi-pivot ensembling with massively multilingual machine translation models. In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 169–180.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4996–5001.
- Maja Popovi  . 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.
- 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.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 86–96.
- Pooja Singh, Shashwat Bhardwaj, Vaibhav Sharma, and Sandeep Kumar. 2025. Leveraging the cross-domain & cross-linguistic corpus for low resource nmt: A case study on bhili-hindi-english parallel corpus. *arXiv preprint arXiv:2511.00486*.
- Shashank Siripragada, Jerin Philip, Vinay P Namboodiri, and CV Jawahar. 2020. A multilingual parallel corpora collection effort for indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3743–3751.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L  onard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram  , et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491.
- Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 154–162.
- 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.

- Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International conference on machine learning*, pages 41092–41110. PMLR.
- Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024. Plug: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046.
- Enchang Zhu, Yuxin Huang, Yantuan Xian, Junguo Zhu, Minghu Gao, and Zhiqiang Yu. 2025. Enhancing distant low-resource neural machine translation with semantic pivot. *Alexandria Engineering Journal*, 116:633–643.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.
- 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.