# Ensemble Decoding Across Pivot-Transfer Mechanisms for Low-Resource Indic Neural Machine Translation

**Anonymous ACL submission**

## Abstract

Neural machine translation (NMT) has made significant progress but performance on low resource languages (LRL) has been limited due to scarcity of parallel corpora and divergence between language pairs. Multilingual models like IndicTrans2 and NLLB-200 can translate between many languages with a single model but have relatively low performance on LRL. Previous work has shown that pivot-based translation is an effective method to enhance translation quality in low-resource scenarios. We apply this approach to Kashmiri-Manipuri, an Indic language pair from distinct language families (Indo-Aryan and Tibeto-Burman) and scripts (Perso-Arabic and Meitei). We propose a corpus-driven pivot language selection methodology based on word-order divergence and sentence-length metrics computed across 22 scheduled Indian languages, identifying Urdu as a structurally compatible pivot language. Due to limited parallel data, we construct synthetic triangulated corpora for this setting. We evaluate four Urdu-based pivot-transfer strategies and ensemble decoding across these heterogeneous transfer mechanisms. The best ensemble achieves 28.25 BLEU and 49.81 chrF++, statistically outperforming the best individual pivot model by 1.26 BLEU and 2.85 chrF++. Our results demonstrate that ensemble decoding across diverse pivot-transfer mechanisms can improve translation quality for low-resource and structurally distant Indic language pairs, even when the pivot language is fixed.

## 1 Introduction

India is linguistically diverse with over 1.4 billion speakers belonging to four major language families namely, Indo-Aryan, Dravidian, Tibeto-Burman, and Austroasiatic. Among the 22 constitutionally recognized scheduled languages, several are considered as low-resource by the natural language processing community. Low resource languages (LRL) lack sufficient digital resources, parallel corpora, and robust computational infrastructure, leading to poor translation quality (Qumar et al., 2024; Tan et al., 2020). Although neural machine translation (NMT) has made significant progress, performance for LRL remains limited due to scarcity of parallel data (Tafa et al., 2025). Translation between low-resource Indic languages, particularly those belonging to different language families and written in different scripts, presents an additional challenge due to structural divergence and morphological complexity. Kashmiri and Manipuri exemplify this challenge. Kashmiri (Indo-Aryan, Perso-Arabic script) and Manipuri (Tibeto-Burman, Meitei script) are among the lowest-resourced scheduled languages, with declining speaker populations and a critical need for language technology to preserve cultural and literary heritage (Raja and Vats, 2025; Qumar et al., 2025).

Although multilingual models increasingly support Indic translation, low-resource languages may remain underrepresented. Models such as IndicTrans2 (Gala et al., 2023) provide reasonable performance but require large-scale multilingual pretraining and substantial computation for downstream finetuning. Recent work has shown pivot-based translation to be an effective strategy for LRL translation (Kim et al., 2019; Sulistyo et al., 2025b,a), where translation is decomposed through a higher-resource intermediate language. Unlike cascaded source-pivot-target translation, pivot-based transfer learning enables direct source-to-target inference using a single model, reducing inference complexity and error propagation. Mohammadshahi et al. (2024); Fan et al. (2021); Firat et al. (2016) have also explored ensemble decoding across pivot models for improved translation quality, typically using identical architectures with different pivot languages. In contrast, different pivot-transfer strategies introduce distinct induc-

tive biases and error characteristics, motivating the exploration of ensemble decoding across heterogeneous pivot mechanisms.

In this work, we address the translation from Kashmiri to Manipuri with pivot-based strategies and propose a novel framework for ensemble decoding. Although English is commonly used as the pivot language in LRL translation, recent work suggests that linguistically related pivots may improve quality of translation for certain Indic languages (Bala Das et al., 2024). We therefore apply a systematic methodology and select Urdu as a linguistically compatible pivot language. We implement four pivot-based transfer learning strategies and systematically evaluate them against direct translation and cascade baselines using BLEU and chrF++ metrics. We further propose an ensemble combining the distinct pivot strategies to produce the final Manipuri translations. Finally, we provide quantitative and qualitative results comparing our ensemble approach against baselines and individual pivot-based models.

Our contributions are as follows: (1) We propose a corpus-driven pivot language selection methodology and identify Urdu as a suitable pivot for Kashmiri-Manipuri translation. (2) We construct triangulated Kashmiri-Urdu, Urdu-Manipuri, and Kashmiri-Manipuri corpora from Bharat Parallel Corpus Collection (BPCC). (3) We evaluate four pivot-transfer learning strategies for low-resource Indic NMT. (4) We propose an ensemble framework combining heterogeneous pivot-transfer mechanisms and demonstrate improved translation performance over individual pivot models and direct baselines.

## 2 Related Work

**Multilingual NMT for Low Resource Indic Languages.** Recent progress in Indic machine translation is driven by multilingual models such as IndicTrans2-M2M (Gala et al., 2023) and NLLB-200 (Costa-Jussà et al., 2022). These models support many Indic languages through large-scale multilingual training, but translation performance on LRL Indic pairs remains limited due to lack of direct parallel corpora. Further, some languages such as Kashmiri in Perso-Arabic script are currently not supported in commercial systems such as Google Translate (Google, 2026).

**Pivot-based Translation Strategies.** Pivot-based translation through an intermediate language is a widely studied approach for zero-resource and low-resource machine translation. Gispert and Mariño (2006), Wu and Wang (2007), Sulistyo et al. (2025b) propose a cascaded approach that combines pretrained source-pivot and pivot-target models. Utiyama and Isahara (2007) studied selecting the best among top-K predictions to reduce the propagation error in cascade systems. Later work explored fine-tuning (Wu and Wang, 2007) and transferring model components across languages after pretraining with a pivot language (Zoph et al., 2016; Kim et al., 2019). These approaches differ in how the pivot language is incorporated during model pretraining and parameter transfer.

**Ensemble Decoding in NMT.** Ensemble decoding has been used to improve translation quality in NMT by aggregating predictions from multiple models during inference (Firat et al., 2016). Wu and Wang (2007) show that a combination of multiple pivot-based systems can outperform individual pivoting strategies. Mohammadshahi et al. (2024); Garmash and Monz (2016) propose confidence-based and weighted aggregation methods for better translation quality. Existing pivot-based ensemble approaches combine models trained using similar training configurations with different pivot languages. In contrast, we explore an ensemble with heterogeneous pivot mechanisms.

**Pivot Language Selection.** The choice of pivot language is a critical design decision in pivot-based NMT. Lin et al. (2019); Isozaki et al. (2010) demonstrate that selection of pivot language guided by linguistic and corpus-level similarity can improve quality of translation. Zhu et al. (2025) apply Kendall's $\tau$ to measure structural similarity for pivot selection in low-resource NMT, while Ramasethu et al. (2026) investigate linguistically related pivot languages in low-resource settings. Despite this growing body of evidence, systematic corpus-driven pivot selection remains underexplored in low-resource Indic NMT.

## 3 Methodology

Figure 1 presents an overview of the proposed framework. We first perform corpus-driven pivot language selection, followed by triangulated corpus construction, pivot-based transfer learning, and ensemble decoding across heterogeneous pivot strategies.

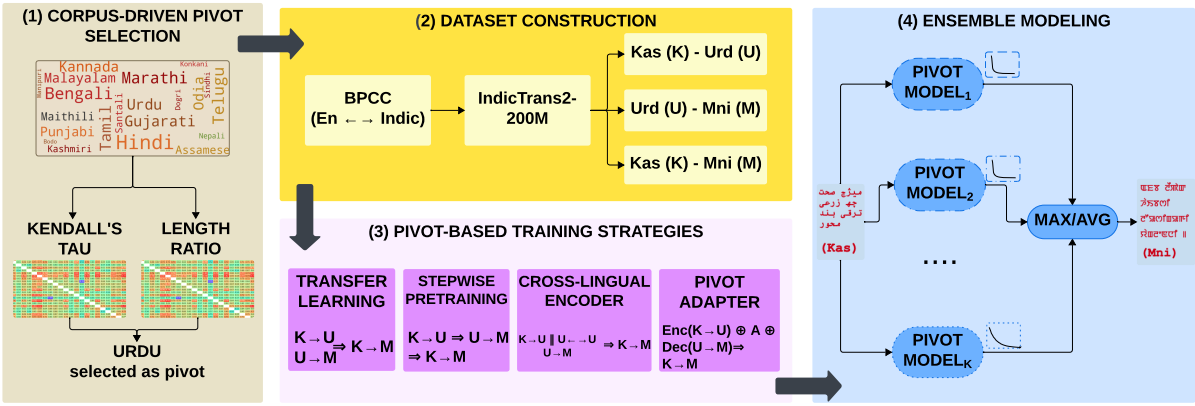


Figure 1: Overview of the proposed framework comprising corpus-driven pivot selection, triangulated dataset construction, pivot-based transfer learning, and ensemble decoding across heterogeneous pivot strategies.

### 3.1 Corpus-Driven Pivot Language Selection

Motivated by recent work on linguistic similarity and pivot language selection, we adopt a corpus-driven pivot selection strategy using two metrics: Kendall's $\tau_b$ for structural divergence and sentence length ratio for token-level length divergence. These metrics are intended to capture reordering complexity and token-level variation arising from morphological and syntactic differences between languages.

**Kendall's Tau $\tau_b$ for Structural Divergence** Following previous work (Isozaki et al., 2010; Birch et al., 2010), we quantify structural similarity using Kendall's rank correlation coefficient ($\tau_b$) computed over word-aligned sentence pairs.

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}$$

where $C$ and $D$ denote the number of concordant and discordant alignment pairs, respectively, and $T_x$ and $T_y$ denote ties in the source and target sequences. Word alignments were obtained using Awesome-Align (Dou and Neubig, 2021) with `bert-base-multilingual-cased`. Higher values of $\tau_b$ indicate lower structural divergence and reduced reordering requirements between languages.

**Sentence Length Ratio.** We use sentence length ratio as a corpus-level proxy for token-level expansion and compression behavior between languages arising from morphological and segmentation differences. To estimate token-level length divergence between languages from $N$ sentences, we compute the mean source-to-target sentence length ratio:

$$R(A, B) = \frac{1}{N} \sum_{i=1}^{N} \frac{|s_i(A)|}{|s_i(B)|}$$

where $|s_i(A)|$ and $|s_i(B)|$ denote the whitespace-tokenized length of the $i$-th sentence pair in languages $A$ and $B$, respectively. A ratio close to unity indicates that the two languages encode information at comparable token-length scales, while large deviations indicate systematic expansion or compression between languages.

Together, Kendall's Tau $\tau_b$ and sentence length ratio capture two corpus-observable dimensions relevant to pivot-based translation: structural agreement and token-level length divergence. Additional implementation details are provided in Appendix A. Both metrics were computed over sentence pairs from the *bpcc-seed-latest* BPCC subset.

### 3.2 Pivot Language Selection for Kashmiri-Manipuri

**Source and Target Languages.** Kashmiri and Manipuri are among the most typologically distant language pairs across the 22 Indic languages. They have a Kendall's Tau of 0.35, which is the lowest score in our pairwise analysis across all $22 \times 22$ language combinations, and a sentence length ratio of about 0.4, indicating divergence in both word order and morphology. This makes it a challenging test case under low resource conditions.

**Pivot Language.** Urdu is identified as the most compatible pivot for this pair under both metrics (Figure 2). It has a Kendall's Tau of 0.74 with Kashmiri, consistent with their shared Indo-Aryan family and Perso-Arabic script, and a comparatively

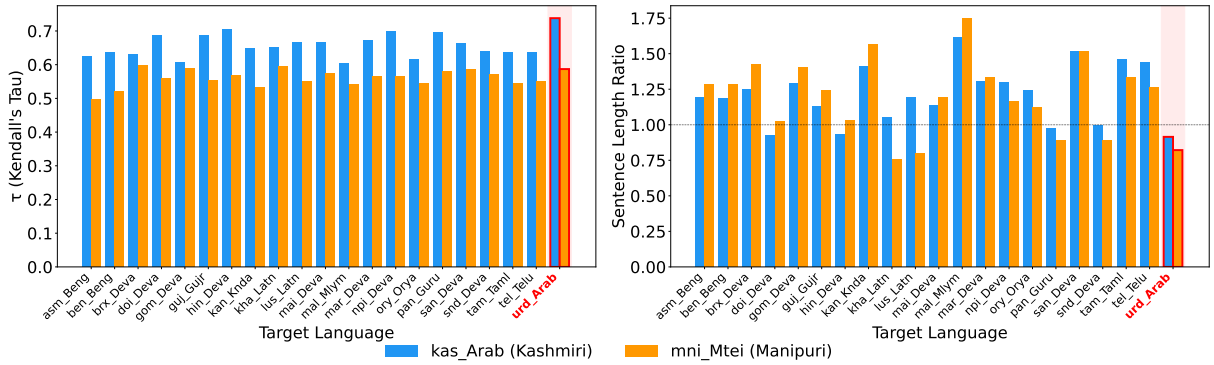


Figure 2: Pairwise structural similarity (Kendall's $\tau_b$) and sentence length ratio between Kashmiri, Manipuri, and other Indic languages. Urdu exhibits high structural similarity and moderate length divergence with both languages, motivating its selection as the pivot language.

high score of 0.59 with Manipuri. The sentence length ratio of Urdu with Kashmiri and Manipuri is around 0.8. Furthermore, the large coverage of Urdu compared to other Indic languages makes it a strong pivot candidate. Results for all 22x22 Indic language pairs are in Figure 3.

### 3.3 Pivot-Based Transfer Mechanisms

In this section, we present four pivot-based transfer learning strategies adapted from Kim et al. (2019). Pivot-based learning improves source-target translation in low-resource settings by leveraging a higher-resource, linguistically similar pivot language. All pivot-transfer strategies consist of two pretraining stages on the source-pivot and pivot-target language pairs, respectively, followed by finetuning on the low-resource source-target corpus. The methods differ in how knowledge is transferred across the intermediate pivot-training stages.

**Plain Transfer Learning (TL).** TL transfers knowledge through encoder-decoder recombination between independently pretrained source-pivot and pivot-target models.

**Stepwise Pretraining (SW).** SW preserves source-language encoder representations during sequential pivot-target pretraining through encoder freezing.

**Cross-Lingual Encoder (XL).** XL learns shared latent representations of the source and pivot languages using joint translation and autoencoding objectives.

**Pivot Adapter (PA).** PA performs explicit representation alignment by mapping source encoder representations to the pivot representation space.

Although the four pivot strategies use the same pivot language, they facilitate cross-lingual knowledge transfer through different mechanisms: encoder-decoder recombination (TL), sequential transfer (SW), shared latent representation learning (XL), and explicit representation alignment (PA). Kudugunta et al. (2019) showed that training configurations affect how cross-lingual representations are learned and transferred in multilingual NMT. We therefore hypothesize that different pivot-transfer strategies may produce complementary translation behaviors that can be leveraged in an ensemble setup. Additional implementation details are provided in Appendix B.

### 3.4 Ensemble Modeling

Diversity is a fundamental principle in an ensemble. Ensemble methods are most effective when the constituent models produce diverse predictions. Recent work has shown that diversity arising from pivot-based translation can improve ensemble-based NMT systems (Oh et al., 2025). Existing pivot-based ensemble approaches introduce diversity by combining models trained using different pivot languages. In contrast, our study focuses on diversity across pivot-transfer mechanisms while keeping the pivot language fixed.

Let:

$$M = \{m_1, m_2, \ldots, m_K\}$$

denote a collection of $K$ source-target translation models trained using different pivot-transfer strategies. At decoding step $t$, each model predicts a probability distribution:

$$p_k(y_t \mid y_{<t}, x)$$

where $x$ is the source sentence and $y_{<t}$ denotes previously generated target tokens.

During beam-search decoding, token probabilities from the constituent pivot-transfer models are aggregated at each decoding step. We evaluate two ensemble aggregation configurations.

**Average Ensemble.** The average ensemble computes the mean probability distribution across all models:

$$p_{avg}(y_t) = \frac{1}{K} \sum_{k=1}^{K} p_k(y_t)$$

Average probability aggregation is widely used in NMT because it smooths individual model variance while preserving consensus predictions across models.

**Maximum Ensemble.** For token $v \in$ vocab $\mathcal{V}$, the maximum ensemble selects the normalized maximum probability across models:

$$p_{\max}(y_t = v) = \frac{\max_k p_k(y_t = v)}{\sum_{v' \in \mathcal{V}} \max_k p_k(y_t = v')},$$

This configuration favors highly confident predictions from individual models.

The aggregated ensemble distribution is used to select the next token candidates during beam-search decoding.

# 4 Experiments

## 4.1 Data Construction via Triangulation

Direct Kashmiri-to-Manipuri (kas2mni) parallel data is extremely limited. We therefore construct Kashmiri-to-Urdu (kas2urd), Urdu-to-Manipuri (urd2mni), and kas2mni corpora through triangulation using English-Indic parallel data from the Bharat Parallel Corpus Collection (BPCC) (Gala et al., 2023). Synthetic Indic-Indic alignments are generated through the English bridge using IndicTrans2-200M (Gala et al., 2023).

Each synthetic corpus is constructed using the half-half strategy proposed in previous work (Elmadani and Buys, 2024) to maximize human-curated coverage on both sides while reducing compounded translation errors. For a language pair $(X, Y)$, the first half retains all human-curated $X$ sentences from BPCC and generates synthetic $Y$ translations using the paired English sentences, yielding $(X_{\text{gold}}, Y_{\text{syn}})$ pairs. The second half reverses this process to produce $(X_{\text{syn}}, Y_{\text{gold}})$ pairs.

| Dataset | Train | Val | Test | Total |
|---|---|---|---|---|
| kas2urd | 221,918 | 4,990 | 4,994 | 231,902 |
| urd2mni | 208,842 | 4,714 | 4,689 | 218,245 |
| kas2mni | 117,087 | 2,647 | 2,641 | 122,375 |

Table 1: Corpus sizes after deduplication and train/validation/test splitting.

The two halves are then concatenated after deduplication. Additional details are provided in Appendix D.

To create train, validation, and test splits, unique English bridge sentences are compiled across all corpora and randomly assigned to a single split. Each Indic-Indic sentence pair inherits the split of its associated English sentence, ensuring that no validation or test English bridge sentence is observed during training in any corpus. Validation and test sets each contain up to 5,000 unique English sentences, with the remaining data used for training (Table 1). To the best of our knowledge, no publicly available human-curated kas2mni parallel corpus currently exists. Consequently, both training and evaluation are performed on triangulated synthetic sentence pairs.

## 4.2 Implementation

Following previous work (Gala et al., 2023), we train a SentencePiece BPE model with a shared vocabulary size of 32,000. All pivot and direct models share a single, optimized Transformer architecture consisting of 4 encoder and decoder layers, hidden dimension of 768, feed-forward dimension of 2048 and 4 attention heads. The resulting model has approximately 137M parameters. Each model was optimized with Adam with an initial learning rate of 0.0005 and the inverse-square root scheduler with a warm-up ratio of 4%. Models were trained on a single L4 GPU with a batch size of 75 sentences and gradient accumulation over 2 steps. The individual pivot models were trained for up to 30 epochs, followed by 5 epochs of source-target finetuning. Early stopping was applied based on the validation BLEU score. Additional hyperparameter optimization details and complete training configurations are provided in Appendix E.

## 4.3 Baselines

We compare our models trained on the synthetic corpus against the following:

**(1) Multilingual baselines.** IndicTrans2-1B (IT2-1B) and IndicTrans2-320M (IT2-320M)

| Model | BLEU↑ | chrF++↑ |
|---|---|---|
| IT2-320M[1] | 27.10 ± 0.52 | 53.33 ± 0.40 |
| IT2-1B[1] | 28.79 ± 0.51 | 54.84 ± 0.37 |
| NLLB-distilled-600M | 20.97 ± 0.42 | 45.60 ± 0.33 |
| direct-scratch | 24.91 ± 0.51 | 47.90 ± 0.41 |

Table 2: Performance of direct translation models on the kas2mni benchmark. ± indicates bootstrap-estimated standard deviation.

| Model | BLEU↑ | chrF++↑ |
|---|---|---|
| direct-scratch | 24.91 ± 0.51 | **47.90 ± 0.41** |
| Cascade | 21.86 ± 0.48 | 45.78 ± 0.42 |
| TL | 26.95† ± 0.51 | 46.24 ± 0.48 |
| SW | 26.65† ± 0.53 | 47.27 ± 0.47 |
| XL | **26.99** † **± 0.51** | 46.96 ± 0.47 |
| PA | 23.02 ± 0.48 | 43.24 ± 0.50 |

Table 3: Pivot-based transfer performance on the kas2mni benchmark. TL, SW, XL pivot-transfer systems perform significantly better than Cascade and direct-scratch in BLEU(†: $p < 0.05$, paired bootstrap resampling).

(Gala et al., 2023) are multilingual encoder-decoder models trained on the BPCC corpus with a many-to-many translation objective. We additionally finetune NLLB-distilled-600M (Costa-Jussà et al., 2022) after vocabulary augmentation on the synthetic kas2mni corpus using the recommended training configuration from the original work (details in Appendix F).

Because our synthetic kas2mni benchmark is constructed from BPCC subsets, the evaluation data may overlap with the multilingual pretraining data used by IT2-1B and IT2-320M. Consequently, the IT2 results are reported only as a reference multilingual system and should not be interpreted as a strictly comparable baseline.

**(2) Direct** Transformer model with the same architecture as our pivot models trained on the synthetic kas2mni corpus (direct-scratch).

**(3) Cascade pivot baseline.** In this pivot-based approach, a source-to-pivot and a pivot-to-target model are connected in a series. The first model translates the source sentence into the pivot language, which the second model translates into the target language. Both models were configured with the optimized pivot architecture and were trained on the corresponding synthetic corpus. We use this (Cascade) as our pivot baseline since it is trained on the same corpora without any pivot-based training procedure.

### 4.4 Evaluation metrics

During inference, we utilize beam search decoding with a beam size of 5 and a length penalty of 1.0. We evaluate translation quality using BLEU (Papineni et al., 2002), which measures the geometric mean of modified n-gram precision between the reference and system outputs. Following Gala et al. (2023), we additionally use chrF++ (Popović, 2017), which captures character-level n-gram similarity and is particularly informative for morphologically rich Indic languages. We therefore report both metrics to capture complementary aspects of translation quality. All scores are computed using SacreBLEU (Post, 2018) to ensure standardized tokenization and fair comparison across systems. Standard deviations are estimated using bootstrap resampling of the test set. We also perform paired bootstrap resampling-based statistical significance tests (Koehn, 2004) for both metrics, with differences being considered statistically significant at $p < 0.05$.

## 5 Results

### 5.1 Direct Translation Baselines

Table 2 reports the performance of the direct translation models on the synthetic kas2mni benchmark. IT2-1B achieves the highest performance with a BLEU score of 28.79 and chrF++ score of 54.84, followed by IT2-320M. The direct-scratch model trained only on the kas2mni corpus achieves a BLEU of 24.91 and a chrF++ of 47.90. After vocabulary augmentation and finetuning, NLLB-distilled-600M achieves 20.97 BLEU and 45.60 chrF++, underperforming direct-scratch. This is likely because Manipuri in Meitei script is not adequately represented in the multilingual pretraining data. The IT2 results should be interpreted cautiously because the synthetic kas2mni benchmark is constructed from BPCC subsets that overlap with the multilingual pretraining corpus used by IT2. Consequently, these scores are reported primarily as reference multilingual system performance rather than as strictly comparable baselines.

### 5.2 Pivot-Based Translation

Table 3 compares our Urdu-based pivot-transfer strategies against direct-scratch and Cascade. Cas-

[1]Evaluation data may overlap with IT2 multilingual pretraining data due to BPCC-based synthetic corpus construction; results are therefore reported only as a reference multilingual system.

| **Example 1:** "The declaration was given legal basis under the Statute of Westminster of 1931." | |
|---|---|
| XL | "The agreement was legally enforced under the **1931 West Bengal law**." |
| TL | "Ablation was legalized under the **Constitution and the 1931 Constitution of the United States**." |
| SW | "**West Bengal is a state in eastern India**." |
| PA | "The **Elbmennt and the 1931 Westrumannt** gave it a legal basis." |
| Ensemble | "The proclamation was enacted into law under the **1931 Statute of Westminster**." |
| *Observation* | *Recovery of institutional and temporal information.* |
| **Example 2:** "Alcohol, even in small amounts, affects the brain." | |
| XL | "**It is not known whether the infection has spread to the brain**." |
| TL | "**Brennoshta is all-encompassing and all-encompassing**." |
| SW | "**The alphabet and the alphabet are not the same**." |
| PA | "**It is the official website of the Government of India**." |
| Ensemble | "Alcohol affects the brain, even in small amounts." |
| *Observation* | *Recovery from semantic hallucination.* |
| **Example 3:** "They live in northwestern North America, Canada, and Alaska." | |
| XL | "**M.V.S.M.U.S., KenaNaDaDaDaDaDaDaDaDaDaDaDa**." |
| TL | "The main countries of origin are **North America, Canada, Peru, and Alaska**." |
| SW | "The main languages are **Kenaestra, Kenaetaxa, Khusu, and Elsekada**." |
| PA | "The main distribution is in northwestern North America, Canada, and Alaska." |
| Ensemble | "They live in northwestern North America, Canada, and Alaska." |
| *Observation* | *Reduction of repetition and entity corruption.* |

Table 4: Qualitative comparison between ensemble and individual pivot systems. Bold spans indicate major translation errors. English shown for readability only.

| **Model** | **BLEU↑** | **chrF++↑** |
|---|---|---|
| direct-scratch | $24.91 \pm 0.51$ | $47.90 \pm 0.41$ |
| Best Pivot (XL) | $26.99 \pm 0.51$ | $46.96 \pm 0.47$ |
| Avg(TL,SW,XL,PA) | $28.18^{\dagger} \pm 0.53$ | $\mathbf{50.17}^{\dagger} \pm \mathbf{0.46}$ |
| Avg(TL,SW,XL) | $27.97 \pm 0.50$ | $49.49^{\dagger} \pm 0.46$ |
| Max(TL,SW,XL,PA) | $\mathbf{28.25}^{\dagger} \pm \mathbf{0.52}$ | $49.81^{\dagger} \pm 0.45$ |
| Max(TL,SW,XL) | $27.57 \pm 0.50$ | $49.11^{\dagger} \pm 0.44$ |

Table 5: Ensemble performance on kas2mni. †: significant improvement over best pivot ($p < 0.05$)

cade achieves a BLEU score of 21.86 which is more than 3 BLEU lower than direct-scratch. This is consistent with previous work showing that cascaded pivot systems can accumulate translation errors across sequential decoding steps (Utiyama and Isahara, 2007).

Among the pivot-transfer strategies, TL, SW and XL achieve statistically significant BLEU score improvements over both direct-scratch and Cascade, while PA underperforms the other pivot-based models. The chrF++ performance of TL, SW and XL pivot-transfer methodologies are statistically comparable to Cascade, but remain comparable to or lower than direct-scratch.

Table 3 shows that the relative ranking of pivot systems differs between BLEU and chrF++. This can be attributed to the metrics capturing complementary properties of translation quality. BLEU estimates word-level n-gram precision while chrF++ focuses on character-level n-gram similarity. Although differences between pivot mechanisms are modest, the variation in performance suggests that different strategies may learn partially complementary translation behaviors, motivating exploration of more sophisticated ensemble methods.

### 5.3 Ensemble Decoding

We evaluate ensemble decoding across the pivot-transfer models (Table 5). We additionally evaluate three-model ensemble configurations excluding PA because PA underperformed the other pivot-transfer strategies in both BLEU and chrF++.

Overall, the ensemble models achieve significant improvements over direct-scratch and the best individual pivot system. The best maximum ensemble achieves 28.25 BLEU and 49.81 chrF++, corresponding to improvements of 3.34 BLEU and 1.91 chrF++ over direct-scratch, and 1.26 BLEU and 2.85 chrF++ over the best individual pivot model. Similar trends are observed for the average ensemble configurations. However, unlike Mohammadshahi et al. (2024), the maximum ensemble produces comparable results to the average ensemble. In both the average and maximum aggregation settings, the four-model ensembles slightly outperform the corresponding three-model ensembles, despite inclusion of the weaker PA model. This suggest that even weaker pivot-transfer models may contribute to useful model diversity during ensemble decoding. Across all ensemble configurations, chrF++ improvements over direct-scratch and the best pivot are statistically significant un-

| Source (English) | Cascade | Ensemble | Observation |
|---|---|---|---|
| "The architecture of Amaravati also influenced the architecture of Southeast Asia." | "**Amaravati was influenced by this architecture**, as well as the idols of Southeast Asia." | "The architecture of Amaravati also influenced Southeast Asian sculpture." | Preservation of semantic relation and causality. |
| "It is notable for having restored its flow after being dry for 60 years." | "**There is a 60 km long and narrow road leading to the airport**." | "It is famous for reviving its stream after drying up for 60 years." | Reduced semantic hallucination. |
| "In total, he played 20 times for his country." | "In total, he **lost 20 caps for his country and lost them all**." | "In total, he played 20 times for his country." | Preservation of predicate semantics. |

Table 6: Qualitative comparison between ensemble decoding and Cascade. Bold spans indicate translation errors in Cascade outputs. English shown for readability only.

der paired bootstrap resampling ($p < 0.05$), while BLEU improvements are not consistently statistically significant. chrF++ improvements are particularly relevant for morphologically rich Indic languages like Manipuri (Gala et al., 2023) because the metric evaluates character-level similarity and is less sensitive to exact word matches than BLEU.

Overall, the results indicate that the individual pivot-transfer models are not redundant, consistent with the pairwise BLEU analysis in Appendix G. Instead, they appear to produce partially complementary predictions that can be effectively combined in an ensemble setting, leading to better performance. The improved results also suggest that diversity across pivot-transfer mechanisms may be beneficial for LRL translation.

### 5.4 Qualitative Analysis of Ensemble Decoding

**Comparison with Pivot Strategies** We qualitatively compare the ensemble outputs against the individual pivot-transfer systems to better understand the observed BLEU and chrF++ improvements. In many cases, the ensemble recovers the correct sentence-level meaning despite substantial degradation in one or more constituent models.

We observe three common qualitative trends. First, the ensemble preserves institutional, geographic, and named entities more accurately than the individual pivot systems (Example 1 in Table 4). Second, it is more robust to semantic drift and incorrect semantic relations (Example 2). Third, it reduces repetition and lexical corruption observed in several pivot-transfer outputs (Example 3). Additional examples are provided in Section H.

**Comparison with Cascade** When qualitatively compared with ensemble decoding, the cascade system frequently exhibits semantic distortion, incorrect semantic relations and predicate corruption. In contrast, the ensemble more consistently preserves sentence-level meaning and semantic relations across diverse examples. Representative examples are shown in Table 6.

## 6 Conclusions

In this work, we studied pivot-based transfer learning for low-resource NMT using Kashmiri-Manipuri, a structurally distant Indic language pair spanning distinct language families and scripts. We proposed a corpus-driven pivot selection methodology using Kendall's Tau $\tau_b$ and sentence length ratio, identifying Urdu as a structurally compatible pivot language. We systematically evaluated four pivot-transfer strategies under low-resource conditions. The best results were obtained through ensemble decoding across the heterogeneous pivot-transfer mechanisms. The proposed ensemble consistently outperformed the individual pivot models, as well as direct and cascade baselines, with statistically significant chrF++ improvements. Qualitative analysis showed that complementary behaviors across pivot-transfer strategies allow the ensemble to better preserve semantic relations and sentence-level meaning. Unlike existing pivot-based ensemble approaches that rely on diversity across pivot languages, our work demonstrates that diversity across pivot-transfer mechanisms can provide effective ensemble performance with a fixed pivot language. These findings suggest that heterogeneous pivot-transfer ensembles are a promising direction for low-resource translation of structurally distant Indic language pairs. Future work could investigate adaptive ensemble methods that dynamically weight pivot-transfer models during decoding.

## Limitations

We restrict our work to a single Kashmiri to Manipuri setting, chosen because the two languages

are among the most structurally distant of the Indic languages. Controlled comparison across alternative pivots under BPCC-derived triangulation is challenging. Changing the pivot language also alters the triangulated corpus distribution and potential overlap with multilingual pretraining corpus. Future work can extend the framework to additional language pairs and independently curated multi-pivot settings. Construction of the synthetic corpora using the existing half-half approach required translating over English as an intermediate language. Although performance of the direct and ensemble translation methods suggests that the synthetic corpora are sufficient for effective model training, the intermediate English triangulation step may still introduce translation errors relative to a fully human-curated corpus. Additionally, ensemble decoding increases inference cost because multiple pivot-transfer models must be evaluated simultaneously during decoding. Future work could investigate distillation or single-model ensemble approaches to reduce inference overhead while retaining ensemble gains.

## References


Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435.

Sudhansu Bala Das, Divyajyoti Panda, Tapas Kumar Mishra, Bidyut Kr. Patra, and Asif Ekbal. 2024. Multilingual neural machine translation for indic to indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(5).

Alexandra Birch, Miles Osborne, and Phil Blunsom. 2010. Metrics for mt evaluation: evaluating reordering. *Machine Translation*, 24(1):15–26.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

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, Torino, Italia. ELRA and ICCL.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki 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. *Preprint*, arXiv:2305.16307.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.

Adrià Gispert and José Mariño. 2006. Catalan-english statistical machine translation without parallel corpus: Bridging through spanish.

Google. 2026. Google translate. `https://translate.google.com/`. Accessed: 2026-05-17.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 944–952.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. 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 866–876, Hong Kong, China. Association for Computational Linguistics.

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, Barcelona, Spain. Association for Computational Linguistics.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual nmt representations at scale. 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 1565–1575.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, and 1 others. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

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, Mexico City, Mexico. Association for Computational Linguistics.

Seokjin Oh, Keonwoong Noh, and Woohwan Jung. 2025. A single model ensemble framework for neural machine translation using pivot translation. *arXiv preprint arXiv:2502.01182*.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Syed Matla Ul Qumar, Muzaffar Azim, and S. M. K. Quadri. 2024. Emerging resources, enduring challenges: a comprehensive study of kashmiri parallel corpus. *AI & Society*.

Syed Matla Ul Qumar, Muzaffar Azim, S. M. K. Quadri, Mohannad Alkanan, Mohammad Shuaib Mir, and Yonis Gulzar. 2025. Deep neural architectures for kashmiri-english machine translation. *Scientific Reports*, 15(1):30014.

Rahul Raja and Arpita Vats. 2025. Parallel corpora for machine translation in low-resource Indic languages: A comprehensive review. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 129–143, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Aishwarya Ramasethu, Rohin Garg, Niyathi Allu, Harshwardhan Fartale, and Dun Li Chan. 2026. Can linguistically related languages guide LLM translation in low-resource settings? In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 168–185, Rabat, Morocco. Association for Computational Linguistics.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *International conference on machine learning*, pages 5719–5728. PMLR.

Danang Arbian Sulistyo, Didik Dwi Prasetya, Fadhli Almu'iini Ahda, and Aji Prasetya Wibawa. 2025a. Pivoted low resource multilingual translation with ner optimization. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(5).

Danang Arbian Sulistyo, Aji Prasetya Wibawa, Didik Dwi Prasetya, and Fadhli Almuíini Ahda. 2025b. An enhanced pivot-based neural machine translation for low-resource languages. *International Journal of Advances in Intelligent Informatics*, 11(2):258–274.

Taofik O. Tafa, Siti Zaiton Mohd Hashim, Mohd Shahizan Bin Othman, Hitham Alhussian, Maged Nasser, Said Jadid Abdulkadir, Sharin Hazlin Huspi, Sarafa O. Adeyemo, and Yunusa Adamu Bena. 2025. Machine translation performance for low-resource languages: A systematic literature review. *IEEE Access*, 13:72486–72505.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1:5–21.

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, Rochester, New York. Association for Computational Linguistics.

Elan van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation. *Preprint*, arXiv:2004.04418.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara

Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic. Association for Computational Linguistics.

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.

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, Austin, Texas. Association for Computational Linguistics.

# A Pivot Selection Methodology Details

Word alignments were obtained using Awesome-Align (Dou and Neubig, 2021) with the pretrained `bert-base-multilingual-cased` model and the softmax extraction strategy.

For Kendall's Tau $\tau_b$ computation, alignment links were first sorted according to source-side token order to construct a permutation of target-side token indices. $\tau_b$ was then computed over the resulting aligned token permutations following Isozaki et al. (2010). A value of $\tau_b = 1$ indicates perfectly preserved word order, $\tau_b = -1$ indicates complete reversal, and $\tau_b = 0$ indicates no rank correlation. In pivot-based translation, higher values therefore indicate lower reordering complexity between languages.

Unlike approaches requiring external typological databases or language-specific lexical resources, both metrics can be estimated directly from the corpora.

Pairwise sentence length ratio and $\tau_b$ are shown in Figure 3. All reported $\tau$ and sentence length ratio values were computed over sentence pairs per language pair from the *bpcc-seed-latest* BPCC subset.

# B Pivot-Based Transfer Learning Strategies

This section explains the methodology for the four pivot-based transfer learning strategies introduced in section 3.3. The four pivot-transfer strategies are adapted from Kim et al. (2019).

## B.1 Plain Transfer Learning

The encoder from the pretrained source-pivot model and the decoder from the pretrained pivot-target model are combined to initialize the final source-target model prior to finetuning. This is depicted in Figure 4.

## B.2 Stepwise Pretraining

During pivot-target pretraining (Figure 5), the pretrained source-pivot encoder is frozen to preserve source-language representations while adapting the decoder to target-language generation.

## B.3 Cross-Lingual Encoder

Figure 6 shows the steps involved in training a cross-lingual encoder pivot model. The source-pivot encoder is jointly trained using translation and denoising autoencoding objectives. The denoising objective is applied to span-corrupted pivot-language sentences to encourage the encoder to learn shared cross-lingual representations of the source and pivot languages.

## B.4 Pivot Adapter

In PA, an adapter maps source encoder and pivot encoder representations to a shared latent space. Sentence representations are obtained by average pooling the final encoder hidden states from parallel source-pivot sentences. Let $S$ and $P$ denote the pooled source and pivot representations respectively. The adapter matrix $A$ is estimated by minimizing:

$$\hat{A} = \arg\min_{A} \|AS - P\|^2$$

under an orthogonality constraint, yielding the standard orthogonal Procrustes solution. The learned adapter is inserted between the source encoder and target decoder during source-target finetuning. Figure 7 illustrates the pretraining steps and how the Kashmiri-Urdu adapter is used in the final model.

| Source → Target Corpus | Kashmiri → Urdu | Urdu → Manipuri | Kashmiri → Manipuri |
|---|---|---|---|
| Source corpus before dedup | 125598 | 299245 | 125598 |
| Source corpus after dedup | 91447 | 211450 | 91447 |
| Target corpus before dedup | 299245 | 134005 | 134005 |
| Target corpus after dedup | 140455 | 6795 | 30928 |
| Final corpus after triangulation | 231902 | 218245 | 122375 |

Table 7: Corpus size before and after deduplication (dedup), and triangulation.

| Hyperparameter | Value |
|---|---|
| Vocabulary size | **32000**, 20000 |
| Encoder/Decoder layers | **4**, 5 |
| Label smoothing | **0.2**, 0.3 |
| Warm-up steps (% of total optimization steps) | **4%**, 8% |

Table 8: Hyperparameter combinations evaluated during grid search. Bold denotes the optimal value.

## C Ensemble of multiple pivot-based transfer learning strategies

A detailed description of the ensemble setup is in Section 3.4. The ensemble decoding process is illustrated in Figure 8.

## D Data Processing

We use the Bharat Parallel Corpus Collection (BPCC) (Gala et al., 2023), a publicly available large-scale multilingual corpus containing sentence pairs across all 22 scheduled Indian languages, including Kashmiri, Urdu and Manipuri. BPCC is organized into subsets based on the source domain of the data. Each subset contains English-Indic data corresponding to different language pairs. Since direct Indic-Indic aligned sentences are not available, we construct Indic-Indic parallel corpora following previous work (Elmadani and Buys, 2024) through triangulation using shared English sentences between two English-Indic BPCC subsets. To reduce noise from the underlying English-Indic sentences, we only consider data from human-curated BPCC subsets. These include *bpcc-seed-latest*, *bpcc-seed-v1*, *ilci*, *nllb-seed*, and *massive*. Because the underlying data are drawn from human-curated subsets, we do not perform additional filtering to the original BPCC sentences; however, we apply the following deduplication pipeline prior to triangulation to ensure unique sentence pairs in the final Indic-Indic corpus.

- Stage 1 - Intra-corpus deduplication: Duplicate English sentences within each English-Indic corpus are removed.
- Stage 2 - Inter-subset deduplication: While merging the selected BPCC subsets for a language pair, duplicate English sentences across subsets are removed.
- Stage 3 - Cross-corpus deduplication: Duplicate English sentences shared between the source-language and target-language English-Indic corpora are removed from the target-language corpus. The choice of removal side is arbitrary and does not affect the experimental results.

At the end of this multi-stage deduplication process, the source-language and target-language corpora contain no overlapping English sentences. The corpus is then split into train, validation and test splits as described in 4.1. The corpora sizes used in our experiments are shown in Table 7

Because the third stage of deduplication removes duplicate sentences from only the target side of a language pair, the number of retained sentences for a language can vary depending on whether the language appears as the source or target in a triangulated pair, and on the amount of overlap with the paired language. Consequently, the retained contributions differ across language pairs even when the same language is involved in multiple triangulation settings.

## E Hyperparameter Optimization for Pivot Model Pretraining

Transformer performance in low-resource NMT is highly sensitive to hyperparameter selection (Araabi and Monz, 2020). We therefore performed a grid search on the Kashmiri-Urdu (source-pivot) model and following Araabi and Monz (2020), applied the resulting optimal configuration to all downstream models, since the parallel corpora used in our experiments are of comparable scale.

Based on previous work relating dataset scale to vocabulary and model size (van Biljon et al., 2020; Gowda and May, 2020), we explored vocabulary sizes of 20K and 32K. Motivated by recent findings that shallower transformers are effective in low-resource settings (van Biljon et al., 2020), we evaluated architectures with 4 and 5 encoder/decoder layers.

The number of attention heads was fixed at 4, the feed-forward dimension at 2048, and the embedding dimension at 768. We additionally explored label smoothing values of 0.2 and 0.3, and warm-up ratios of 4% and 8% of the total optimization steps using the inverse square-root learning-rate scheduler.

The optimal configuration consisted of 4 encoder/decoder layers, a 32K vocabulary, label smoothing of 0.2, and a warm-up ratio of 4%.

Table 8 contains the explored hyperparameter combinations, while the final training configuration used in pretraining is provided in Table 9.

## F NLLB-200 Finetuning Details

We finetuned the pretrained NLLB-200 distilled 600M model (Costa-Jussà et al., 2022) on the synthetic Kashmiri–Manipuri corpus using HuggingFace Transformers (Wolf et al., 2020). Because the original SentencePiece vocabulary lacked coverage of the Meitei script, Manipuri text tokenized predominantly as <unk>. To address this, we added 84 Meitei characters observed in the training corpus to the tokenizer vocabulary and resized the model embeddings accordingly. Model outputs were normalized by replacing zero-width spaces with standard whitespace prior to evaluation. We used the recommended training configuration for finetuning (Costa-Jussà et al., 2022) with a beam size 5.

## G Diversity Across Pivot-Transfer Mechanisms

To quantify diversity across the pivot-transfer mechanisms, we compute average pairwise corpus BLEU between the hypotheses generated by XL, TL, SW, and PA on the kas2mni test set. Based on previous work in diversity evaluation in machine translation (Shen et al., 2019), the hypothesis from each model is scored against another model's output using SacreBLEU. Lower pairwise BLEU indicates greater diversity between model predictions.

| Training parameter | Value |
|---|---|
| Optimizer | Adam |
| Beta values ($\beta_1$, $\beta_2$) | 0.90, 0.98 |
| Learning rate | 0.0005 |
| Scheduler | Inverse square-root |
| Criterion | Cross-entropy |
| Encoder/Decoder layer dropout | 0.0/0.1 |
| Embedding dropout | 0.3 |
| Effective minibatch size | 150 sentences ($\sim$4000 tokens) |
| Number of epochs | 30 (training) / 5 (finetuning) |
| Encoder/decoder vocabulary size | 32000 |
| Label smoothing | 0.2 |
| Maximum source/target sequence length | 192 |
| Evaluation metric | Perplexity |

Table 9: Complete set of parameters used in the final model configuration.

| Model | XL | TL | SW | PA |
|---|---|---|---|---|
| XL | 100.00 | 66.05 | 62.04 | 56.29 |
| TL | 66.05 | 100.00 | 62.12 | 56.44 |
| SW | 62.03 | 62.11 | 100.00 | 57.31 |
| PA | 56.28 | 56.44 | 57.30 | 100.00 |

Table 10: Pairwise corpus BLEU between hypotheses generated by the pivot-transfer models on the kas2mni test set. Lower values indicate greater diversity between model predictions.

Table 10 shows that TL and XL produce the most similar hypotheses, while PA consistently exhibits the lowest pairwise BLEU relative to the other models. Although PA underperforms as an individual system, its lower pairwise similarity suggests more distinct prediction behavior, which may contribute to the observed ensemble improvements.

## H Qualitative Comparison of Ensemble with Pivot-Transfer Mechanisms

Examples in Table 11 further illustrate the complementary behavior of the pivot-transfer mechanisms. Individual pivot systems exhibit hallucination, repetition, lexical corruption, and incorrect entity translation. In many cases, different pivot models preserve different parts of the sentence semantics, allowing the ensemble to recover a substantially more accurate translation. Overall, the qualitative results suggest that the diverse pivot-transfer mechanisms produce partially complementary predictions that can be effectively combined through ensemble decoding.

# I Pivot-Transfer and Ensemble Implementation

All the models are based on the transformer encoder-decoder architecture. We reuse and adapt some parts of the IndicTrans2 model code released under the permissive MIT license (Gala et al., 2023). It includes the use of an Indic pre-processor to pre-process Indic sentences before applying to our models and an Indic post-processor to post-process the model outputs. We use the Hugging Face transformers and dataset libraries for model training and dataset handling. The code developed as part of this work is intended for use in the research of low-resource NMT systems.

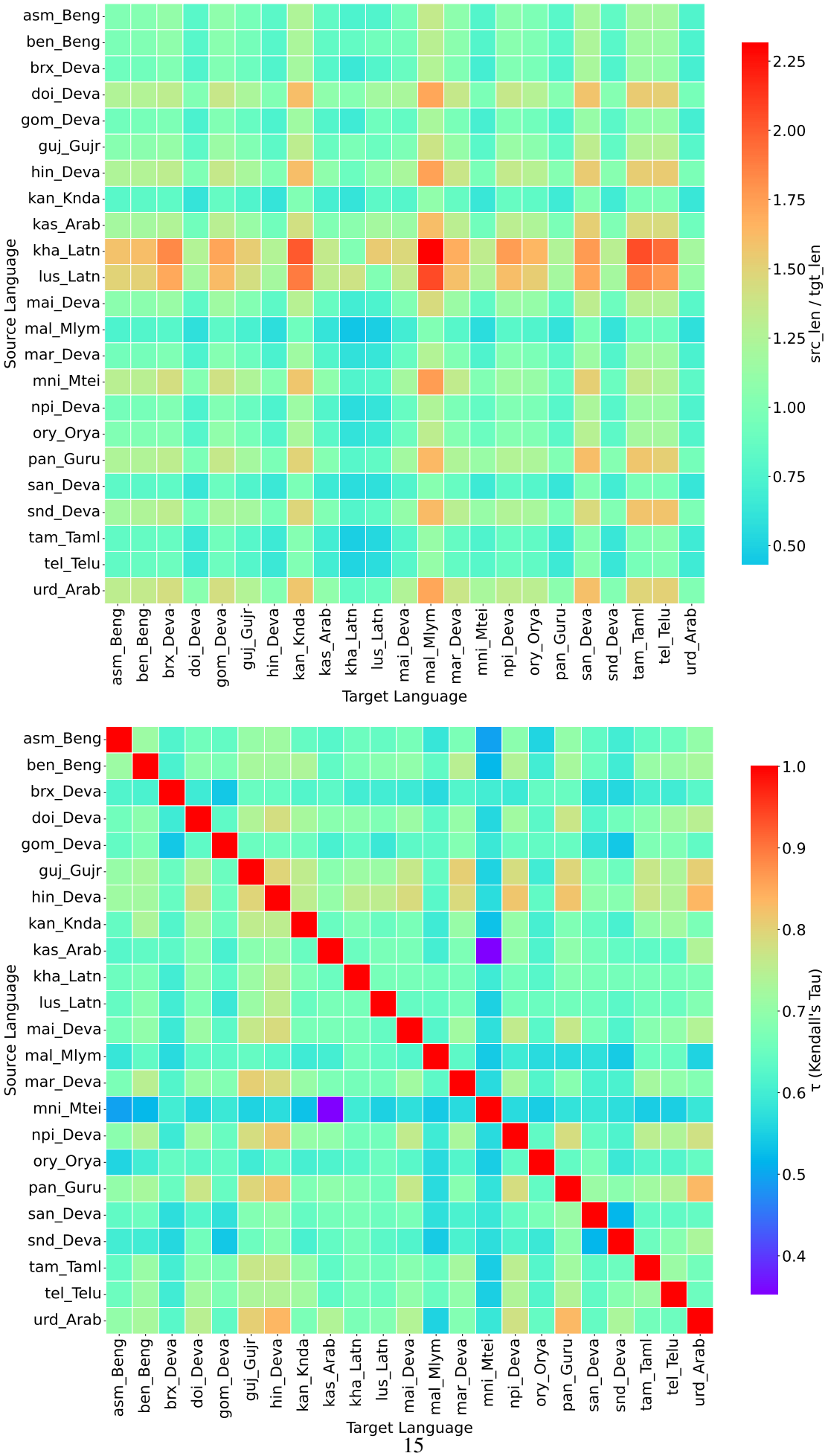


Figure 3: Sentence length ratio and Kendall's Tau $\tau_b$ across 22 $\times$ 22 Indic language pairs. Top: sentence length ratio heatmap. Bottom: Kendall's Tau $\tau_b$ heatmap.

| Source (English) | XL / TL / SW / PA | Ensemble | Observation |
|---|---|---|---|
| "Jones was eight years old when his father died." | XL: "Jones was eight years old when his father died."<br>TL: "**Jones died at the age of eight**, when his father died."<br>SW: "**Jones died at the age of eight**, when his father died."<br>PA: "Jones was eight years old when his father died." | "Jones was eight years old when his father died." | Recovery of semantic role. |
| "The dynamic information that is used needs to be reviewed periodically." | XL: "**There are three ways to do this : either by yourself or through a social network**."<br>TL: "**There are also three types of insurance policies for employees**."<br>SW: "**Ancestry is a living organism and a living organism**."<br>PA: "**It is also the seat of the Arunachal Pradesh Legislative Assembly**." | "The dynamic addresses used need to be reviewed periodically." | Recovery from semantic hallucination. |
| "The territory of Ladakh was divided into the Kargil and Leh districts in 1979." | XL: "The Ladakh region was divided into Kargil and **eleven districts** in 1979."<br>TL: "Ladakh region was divided into Kargil and 13 districts in 1979."<br>SW: "The Ladakh region was divided into Kargil and **three districts** in 1979."<br>PA: "The region of Ladakh was divided into Kargil and **14 districts** in 1979." | "The Ladakh region was divided into the Kargil and Leh districts in 1979." | Recovery of geographic entity. |
| "Other uses include sprinklers, toilets, and water filtration." | XL: "**Autobiography also includes tika, taiyolashan, dhusu, and nalakshashan**."<br>TL: "**They are also known by the names Talash, Talash, and Avishka**."<br>SW: "Other uses include sanitation, toilets, and irrigation."<br>PA: "**Automotive industry, including manufacturing, shipping, and distribution**." | "Other uses include sanitation, toilets, and irrigation." | Recovery from hallucination and lexical corruption. |
| "Three of these places were for men, and one for women." | XL: "The current President of Bangladesh is Abdul Hamid, **the Prime Minister is Sheikh Hasina, and the Prime Minister is Sheikh Hasina**."<br>TL: "**He is the father of the Prime Minister, the Vice President, and the Prime Minister**."<br>SW: "The current WMF is **the WMF, the WMF, and the WMF**."<br>PA: "The main characters are **M.F.M., Ahum Yatris, Khushbu, and Kanupi**." | "Three of these places were for men, and one was for a woman." | Recovery from repetition and entity corruption. |

Table 11: Additional qualitative comparison between individual pivot-transfer strategies and ensemble decoding. Bold text highlights major errors such as hallucination, repetition and semantic distortion in individual pivot models. English shown for readability only. Task remains Kashmiri to Manipuri.

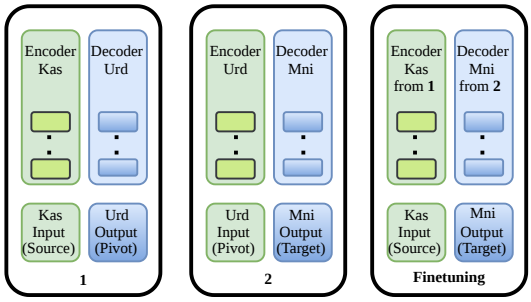


Figure 4: Plain transfer learning (TL) strategy via Urdu pivot: pre-training Kashmiri-Urdu and Urdu-Manipuri models followed by Kashmiri-Manipuri finetuning.

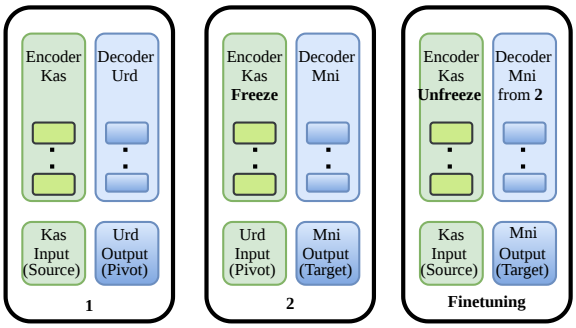


Figure 5: Stepwise transfer learning (SW) framework via Urdu pivot: combining Kashmiri-Urdu pre-training, frozen-encoder Urdu-Manipuri training, and final finetuning on Kashmiri-Manipuri corpus.

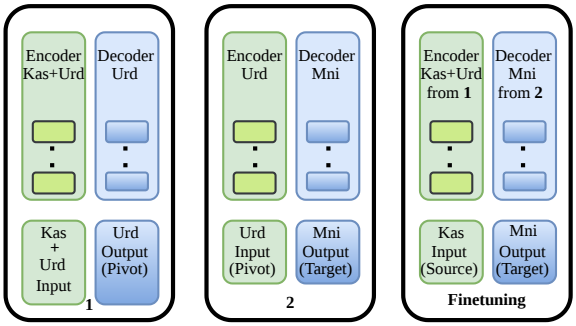


Figure 6: A shared Kashmiri-Urdu cross-lingual encoder (XL) is pre-trained on translation and denoising autoencoding objectives to learn aligned Kashmiri-Urdu representations. This is followed by Urdu-Manipuri pre-training. The combined Kashmiri+Urdu encoder and Manipuri decoder are finetuned on Kashmiri-Manipuri corpus.

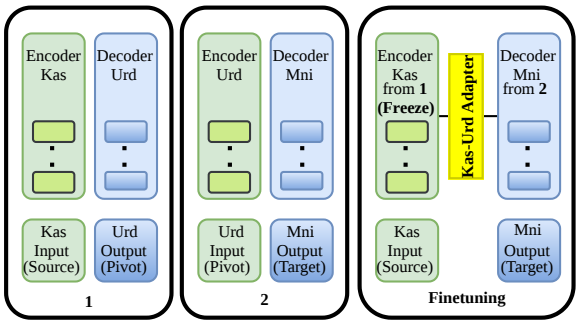


Figure 7: A source to pivot adapter (PA) maps source encodings to the pivot representation space: A Kashmiri-Urdu adapter maps the Kashmiri encodings to the Urdu semantic space with which the Manipuri decoder is pre-trained. Finally only the decoder is finetuned on the direct Kashmiri-Manipuri corpus.

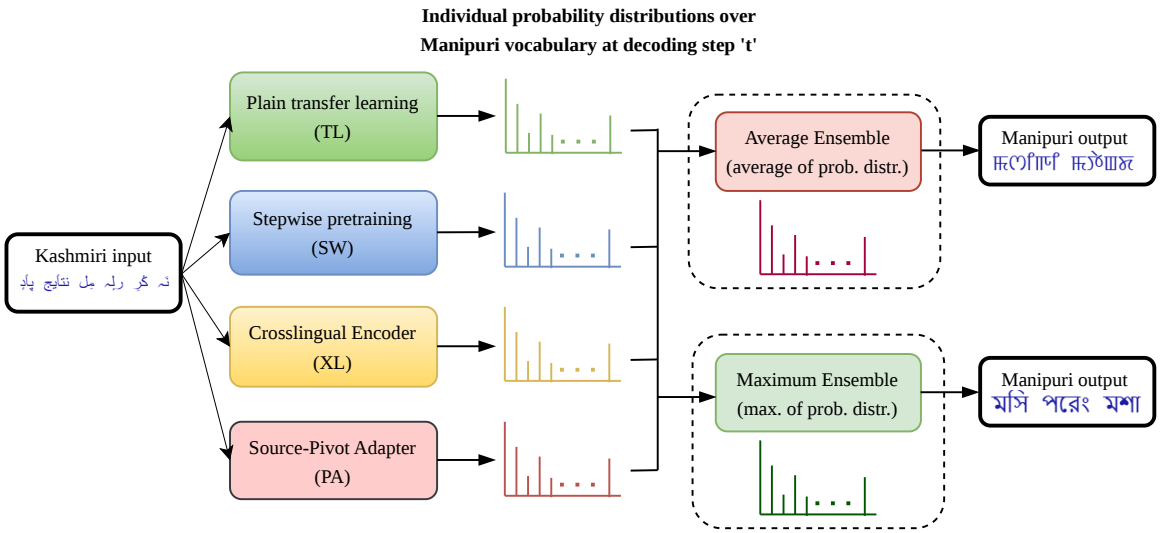


Figure 8: At each decoding step, the ensemble aggregates the probability distributions (prob. distr.) predicted by the individual pivot-based models, and the resulting distribution is used to select the next token candidates during beam search decoding.