# Adapting IndicTrans2 for Legal Domain MT via QLoRA Fine-Tuning at JUST-NLP 2025

**Akoijam Jenil Singh**[1], **Loitongbam Sanayai Meetei**[2], and **Yumnam Surajkanta**[1]

[1]NIT Manipur, Manipur, India
[2]SOA University, Odisha, India
{akoijamjenilsingh, loisanayai, ysurajkanta}@gmail.com

## Abstract

Machine Translation (MT) in the legal domain presents substantial challenges due to its complex terminology, lengthy statutes, and rigid syntactic structures. The JUST-NLP 2025 Shared Task on Legal Machine Translation [1] was organized to advance research on domain-specific MT systems for legal texts. In this work, we propose a fine-tuned version of the pretrained large language model (LLM) ai4bharat/indictrans2-en-indic-1B [2], a transformer-based English-to-Indic translation model. Fine-tuning was performed using the parallel corpus provided by the JUST-NLP 2025 Shared Task organizers.Our adapted model demonstrates notable improvements over the baseline system, particularly in handling domain-specific legal terminology and complex syntactic constructions. In automatic evaluation, our system obtained BLEU = 46.67 and chrF = 70.03.In human evaluation, it achieved adequacy = 4.085 and fluency = 4.006. Our approach achieved an AutoRank score of 58.79, highlighting the effectiveness of domain adaptation through fine-tuning for legal machine translation. [3]

## 1 Introduction

India is a linguistically diverse country, with 22 officially recognized languages listed under the Eighth Schedule of the Constitution as of 2004. Despite this multilingual landscape, English serves as the official language of the judiciary throughout the country. In certain states such as Rajasthan, Madhya Pradesh, Uttar Pradesh, and Bihar, the use of Hindi is also permitted in High Court proceedings (PBI, 2025), highlighting the need for high-quality legal translation systems between English and Hindi.

However, legal translation is uniquely complex due to the presence of domain-specific terminology, lengthy statutes, and highly formalized language structures. General-purpose machine translation systems are not designed to handle such intricacies. Even minor translation errors in legal contexts can result in significant misunderstandings, making precision and domain awareness critical requirements.

The JUST-NLP 2025 Shared Task aims to advance machine translation in the legal domain, focusing on the English–Hindi language pair. In this paper, we present a domain-adapted legal machine translation system built upon the pretrained indictrans2-en-indic-1B model (Gala et al., 2023). The pretrained was originally developed for general-purpose translation across the 22 languages listed in the Eighth Schedule of the Indian Constitution. We fine-tune the model on the legal parallel corpus provided by the JUST-NLP 2025 Shared Task. Fine-tuning on this domain-specific corpus enhances the system's robustness in translating legal texts from English to Hindi, ensuring better preservation of legal terminology and contextual accuracy.

As part of the JUST-NLP 2025 Shared Task on Legal Machine Translation, our system demonstrated strong performance, achieving an AutoRank score of 58.79. This result provides empirical evidence that domain adaptation substantially enhances translation quality in the legal domain.

To support reproducibility and facilitate further research, we release the fine-tuned weights of our model, built on top of indictrans2-en-indic-1B. The model weights are publicly available at our repository [4].

[1] JUST-NLP 2025 Shared Task.
[2] Hugging-Face ai4bharat/indicTrans2-en-indic-1B.
[3] The final result announced by JUST-NLP 2025 Shared Task organizers
[4] Repository of the Model Weight



*Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025)*, pages 142–147
December 24, 2025 ©2025 Association for Computational Linguistics

## 2 Related Work

Machine Translation (MT) is a core task in Natural Language Processing (NLP), aiming to automatically translate text across languages. In Indian language and legal translation, Haque et al. (2019) applied Phrase-Based SMT for English–Hindi, and Das et al. (2025) extended SMT to fifteen Indic languages. Evaluation of English–Hindi systems by Shetty (2025) found Google Translate and IndicTrans2 to achieve the highest automatic scores. More recently, Singh et al. (2025) assessed thirty-seven LLMs for English-to-Hindi legal translation, identifying Gemini-2.5-Pro, ONLINE-B, and Claude-4 as top performers. The MultiIndic22MT 2024 shared-task (Singh et al., 2024) focused on English–Manipuri translation using Transformer-based NMT with OpenNMT, comparing sequence-to-sequence models and Byte Pair Encoding (BPE) tokenization.

Overall, research shows a shift from rule-based and phrase-based methods to neural and transformer architectures. Nevertheless, accurate legal translation between English and Hindi remains challenging due to limited domain-specific corpora, complex terminology, and contextual ambiguities. The next section addresses these challenges using transformer-based architectures combined with domain adaptation techniques for legal machine translation.

## 3 Dataset

The JUST-NLP 2025 Shared Task focuses on translating legal texts from English (source) to Hindi (target). The organizers provided three Excel files: `English-hindi-train.xlsx` (tra, 2025), `English-hindi-valid.xlsx` (val, 2025), and `WMT25-TS_eng-hin-test.xlsx` (tes, 2025). The training file contains 50,000 English–Hindi parallel sentence pairs from the legal domain, while the validation and test files each contain 5,000 English-only sentences for evaluation and testing, respectively.

To facilitate model training and hyperparameter tuning, we further split the training data into 48,000 sentence pairs for training and 2,000 for internal validation. The official test set (tes, 2025) is used for final evaluation of our system using automatic metrics such as BLEU, chrF, and METEOR. Table 1 summarizes the dataset structure.

| **Dataset** | **Size (pairs)** |
|---|---|
| `Train (full)` | 50,000 |
| `Train (used)` | 48,000 |
| `Validation (used)` | 2,000 |
| **Dataset** | **Source Only** |
| `Validation (official)` | 5,000 |
| `Test (official)` | 5,000 |

Table 1: JUST-NLP 2025 dataset split statistics for English–Hindi legal text translation.

## 4 Methodology

We propose an English-to-Hindi machine translation system tailored for the legal domain. Our approach builds upon the pretrained IndicBART model indictrans2-en-indic-1B which we fine-tune using a domain-specific parallel corpus. This fine-tuning process allows the model to more effectively learn and translate legal terminology and contextual nuances, resulting in translations that are both accurate and contextually appropriate for legal texts. Following the fine-tuning, a post-processing step is applied to remove any unwanted characters produced during translation. A visual overview of this process is provided in Figure 1.

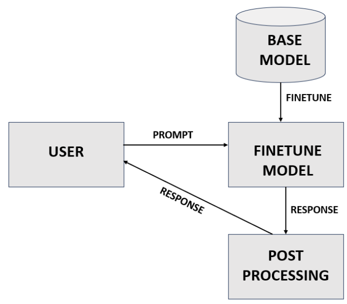


Figure 1: Workflow of the English→Hindi legal machine translation system, including user prompt, fine-tuning the base model, and post processing.

### 4.1 Data Preprocessing

We prepare our dataset by loading and tokenizing all the training, validation and test corpora using the sequence-to-sequence tokenizer provided by the base model indictrans2-en-indic-1B. This ensures consistency with the input format expected by IndicBART and preserves syntactic and semantic structures necessary for high-quality translation.

### 4.2 Parameter-Efficient Fine-Tuning via QLoRA

Due to computational constraints, we adopt QLoRA (Quantized Low-Rank Adapter) (Dettmers et al., 2023) for efficient fine-tuning. QLoRA combines 4-bit quantization of the pretrained model with Low-Rank Adaptation (LoRA), which introduces trainable adapter layers into specific transformer components while keeping the base model weights frozen. This method significantly reduces GPU memory usage and training cost, enabling fine-tuning of large language models (LLMs) without substantial degradation in performance.

| Parameter | Setting |
|---|---|
| 4 bit quantization | True |
| Device map | auto |
| LoRA rank (r) | 16 |
| LoRA alpha | 16 |
| LoRA dropout | 0.05 |
| Task type | Seq2Seq LM |

Table 2: Key Hyperparameters for QLoRA-based Fine-Tuning.

### 4.3 Training Strategy

We fine-tuned the model indictrans2-en-indic-1B using 4-bit QLoRA (Quantized Low-Rank Adaptation) from the Hugging Face library (Hug, 2025) for parameter-efficient training.

| Training Args | Values |
|---|---|
| Optimizer | AdamW |
| Learning Rate | 2e-4 |
| Scheduler | Cosine Scheduler |
| Weight Decay | 0.01 |
| GPU | NVIDIA T4 |
| Batch Size | 16 |
| Mixed Precision | fp16 |
| Checkpoint | Every 1000 steps |
| Tokenizer | Seq2Seq Tokenizer |
| Dynamic Padding | DataCollatorForSeq2Seq |

Table 3: Training parameters for finetuning ai4bharat/indictrans2-en-indic-1B.

We train the model using an early stopping mechanism with a patience value of 5. The three best-performing checkpoints are selected based on validation loss. These checkpoints are then ensembled to form the final model, aggregating outputs to improve robustness and translation quality.

We apply a post-processing step to clean the model outputs. Specifically, we remove extraneous characters, such as punctuation marks, which are occasionally generated at the end of translated sentences. This step helps improve the fluency and readability of the final output and ensures conformity with the target language conventions.

### 4.4 Inference

During inference, the fine-tuned model is loaded, and source sentences are tokenized accordingly. Target sequences are generated using beam search with a beam width of 5 and a maximum length of 512 tokens. We evaluated multiple beam widths and observed that this setting yields the best translation performance.

## 5 Experiments and Results

We fine-tuned the pretrained indictrans2-en-indic-1B model on the English–Hindi legal parallel corpus to adapt it to the legal domain.

### 5.1 Evaluation Metrics

We conducted a human evaluation focusing on adequacy and fluency. In addition, the translations produced by our model were evaluated by the shared task organizers using automatic metrics. The evaluation procedures are described below.

#### 5.1.1 Human Evaluation Metrics

Human evaluation remains the most reliable approach for assessing translation quality, as it captures linguistic and semantic nuances that automatic metrics may overlook. We conducted human evaluation along two qualitative dimensions: adequacy and fluency. These metrics provide complementary insights into translation performance and are described below.

**Adequacy Evaluation :** Adequacy (Snover et al., 2009) measures the extent to which the translated text preserves the meaning of the source sentence, regardless of its grammatical quality.

**Fluency Evaluation :** Fluency (Snover et al., 2009) assesses the grammatical correctness and naturalness of the translation in the target language, independent of the source content.

The scoring criteria of Adequacy and Fluency Evaluation is given in table 4

| Score | Adequacy | Fluency |
|---|---|---|
| 1 | Does not retain any of the information from the source sentence. | Unintelligible due to grammatical errors. |
| 2 | Conveys only a minimal amount of information. | Contains grammatical errors that impede comprehension. |
| 3 | Retains a moderate amount of information. | Includes some mistakes or phrasing that feels unnatural. |
| 4 | Retains almost all relevant information. | Conforms to accepted grammatical norms. |
| 5 | Accurately reflects all information in the source. | Flawless, natural, and stylistically appropriate. |

Table 4: Human evaluation criteria for fluency and adequacy. Reprodced from (Meetei et al., 2024)

#### 5.1.2 Automatic Evaluation Metrics

Automatic evaluation is widely adopted in machine translation research for its scalability, reproducibility, and efficiency. The automatic metrics employed in this work are described below.

**BLEU :** The Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2002) measures the n-gram precision of a candidate translation with respect to reference translations, penalizing short translations with a brevity penalty.

**ChrF :** The Character F-score (ChrF) (Popović, 2015) calculates F-scores over character n-grams rather than word n-grams, which makes it more suitable for morphologically rich languages.

**METEOR :** METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005) aligns hypothesis and reference sentences based on exact, stem, synonym and paraphrase matches. A higher METEOR score reflects better adequacy and fluency.

**TER :** Translation Edit Rate (TER) (Snover et al., 2006) measures the number of edits required to transform the system output into the reference translation. Lower TER values indicate higher translation quality, as fewer edits are needed to match the human reference.

**BERTScore :** BERTScore (Zhang et al., 2019)leverages contextual embeddings from pre-trained language models to compute semantic similarity between hypothesis and reference. Higher scores indicate a stronger semantic alignment.

**COMET :** COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020) is a neural evaluation metric trained to predict human judgments of translation quality. Higher COMET scores indicate closer agreement with human assessments of adequacy and fluency.

### 5.2 Results

The performance of our fine-tuned English→Hindi legal MT system is summarized through human evaluation and official leaderboard results from the JUST-NLP 2025 Shared Task.

Human evaluation was conducted by bilingual experts fluent in English and Hindi. Adequacy and fluency scores are reported in Table 5. The results indicate strong preservation of meaning and natural readability in Hindi translations.

| Model | Adequacy | Fluency |
|---|---|---|
| Finetuned Model | 4.085 | 4.006 |

Table 5: Human evaluation of the English→Hindi legal MT system. Scores range from 1 (poor) to 5 (excellent).

On the official leaderboard, our system achieved strong n-gram overlap, morphological robustness, and semantic preservation: BLEU = 46.67, METEOR = 72.86, TER = 44.63, chrF++ = 70.03, BERTScore = 90.86, and COMET = 72.12. The AutoRank score, computed by the organizers as a weighted combination of these metrics, is 58.79, indicating high-quality translations. The AutoRank calculation is given in Equation 1.

**Leaderboard Results.** Table 6 presents the top 7 participants for English→Hindi legal translation. Metrics include BLEU, METEOR, TER, chrF++, BERTScore, COMET, and AutoRank. Our system, **JUST-MEI**, ranked 5th, demonstrating competitive performance across all metrics.

Overall, both automatic and human evaluations confirm that our QLoRA fine-tuned IndicTrans2 model reliably translates English legal texts into Hindi, maintaining high lexical, semantic, and stylistic accuracy while effectively preserving legal terminology.

$$\text{AutoRank} = \frac{1}{6}\Big(BLEU_{\text{norm}} + METEOR_{\text{norm}} + (1 - TER_{\text{norm}}) + CHRF^{++}_{\text{norm}} + BERTScore_{\text{norm}} + COMET_{\text{norm}}\Big) \tag{1}$$

| Rank | Team | BLEU↑ | METEOR↑ | TER↓ | chrF++↑ | BERTScore↑ | COMET↑ | AutoRank↑ |
|---|---|---|---|---|---|---|---|---|
| 1 | Team-SVNIT | 51.61 | 75.80 | 37.09 | 73.29 | 92.61 | 76.36 | 61.62 |
| 2 | FourCorners | 50.19 | 69.54 | 42.32 | 73.67 | 92.70 | 75.74 | 60.31 |
| 3 | goodmen | 48.56 | 67.15 | 41.63 | 73.07 | 92.38 | 75.16 | 59.39 |
| 4 | JUNLP | 46.03 | 71.84 | 42.08 | 70.59 | 91.19 | 73.72 | 58.90 |
| **5** | **JUST-MEI** | **46.67** | **72.86** | **44.63** | **70.03** | **90.86** | **72.12** | **58.79** |
| 6 | Lawgorithms | 46.27 | 71.80 | 43.06 | 68.32 | 91.03 | 72.14 | 58.26 |
| 7 | Tokenizers | 34.08 | 61.78 | 55.25 | 56.75 | 87.93 | 65.20 | 50.87 |

Table 6: Top 7 participants in the JUST-NLP 2025 Shared Task for English→Hindi legal translation. Automatic metrics reflect both formal correctness and semantic accuracy. Our system (rank 5) is highlighted in bold.

| English (Source) | Hindi (Finetuned Model) | Legal Term Correctness |
|---|---|---|
| plaintiff No.1 was dead. | वादी संख्या 1 की मृत्यु हो चुकी थी। | correct |
| hence, this appeal. | अतएव, यह अपील | correct |
| writ petition is dismissed. | रिट याचिका खारिज की जाती है। | correct |
| they were employees employed under the defendant-appellants. | वे प्रतिवादीगण – अपीलार्थीगण के अधीनियोजित कर्मचारीगण थे। | correct |
| other allegations were denied by the defendant. | प्रतिवादी द्वारा अन्य अभिकथनों से इनकार किया गया था। | correct |
| accordingly, the title appeal was dismissed. | तद्नुसार, अभिधान अपील खारिज कर दी गयी थी। | Legal term correct; minor lexical mismatch |
| PW-36 is the plaintiff himself. | अ जनतादल – 36 स्वयं वादी है। | Partially correct; witness designation mistranslated |

Table 7: Sample English→Hindi legal translations showing preservation of legal terminology. Each entry is evaluated for correctness of domain-specific terms.

### 5.3 Preservation of Legal Terminology

We evaluated whether the translations correctly preserve the legal terminology. Most legal terms were accurately rendered in Hindi, reflecting the model's ability to capture domain-specific terminology. However, a small portion of terms were mistranslated or rendered in a non-standard form, indicating that while the system is largely effective in maintaining legal terminology, occasional inconsistencies remain. Table 7 shows the sample output.

## 6 Conclusion and Future Work

We presented a domain-adapted English-to-Hindi legal machine translation system built on the pretrained indictrans2-en-indic-1B model and finetuned with QLoRA on the JUST-NLP 2025 legal corpus. Our approach effectively captures domain-specific terminology and contextual nuances, yielding substantial improvements over a general-purpose baseline across multiple automatic metrics (BLEU, METEOR, TER, chrF++, BERTScore, COMET) and human evaluation dimensions (adequacy and fluency). The results demonstrate that the proposed system produces accurate and natural translations, highlighting the effectiveness of domain adaptation and the importance of combining automatic and human evaluations for comprehensive evaluation in specialized translation settings such as the legal domain.

Although our study is limited to a single model variant and limited computational resources, future work can investigate larger architectures, multilingual legal translation, and advanced domain adaptation techniques to further enhance performance. In general, our results highlight the importance of targeting domain adaptation for producing accurate and reliable legal machine translation systems in the Indian context.

## Limitation

Although our fine-tuned model demonstrates strong performance, several limitations remain. First, we only explored a single variant of Indic-

Trans2; other architectures and larger models were not evaluated. Additionally, our experiments were constrained by hardware limitations, including limited GPU resources and batch sizes. To accommodate these constraints during fine-tuning, we employed 4-bit quantization of the base model.

## Acknowledgments

We thank the organizers of the JUST NLP 2025 Shared Task for providing the dataset and the competition platform. We also acknowledge the support and valuable feedback from our colleagues and team.

## References

2025. Huggin face peft docs. https://huggingface.co/docs/peft/index. Accessed: 2025-10-25.

2025. testing dataset. https://huggingface.co/datasets/helloboyn/WMT25-TS/tree/main. Accessed: 2025-10-25.

2025. training dataset. https://huggingface.co/datasets/helloboyn/IJCNLP-JustNLP-LMT/blob/main/english-hindi-train.xlsx. Accessed: 2025-10-25.

2025. Using regional language in court. https://www.pib.gov.in/Pressreleaseshare.aspx?PRID=2042983. Accessed: 2025-10-24.

2025. validation dataset. https://huggingface.co/datasets/helloboyn/IJCNLP-JustNLP-LMT/resolve/main/english-hindi-valid.xlsx. Accessed: 2025-10-25.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Sudhansu Bala Das, Divyajyoti Panda, Tapas Kumar Mishra, and Bidyut Kr Patra. 2025. Statistical machine translation for indic languages. *Natural Language Processing*, 31(2):328–345.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

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

Rejwanul Haque, Md Hasanuzzaman, and Andy Way. 2019. Investigating terminology translation in statistical and neural machine translation: a case study on english-to-hindi and hindi-to-english. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 437–446.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2024. An empirical study of a novel multimodal dataset for low-resource machine translation. *Knowledge and Information Systems*, 66(11):7031–7055.

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. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Ahan Prasannakumar Shetty. 2025. Evaluating machine translation models for english-hindi language pairs: A comparative analysis. *arXiv preprint arXiv:2505.19604*.

Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025. Evaluation of llm for english to hindi legal domain machine translation systems. In *Proceedings of the Tenth Conference on Machine Translation, China. Association for Computational Linguistics*.

Ningthoujam Justwant Singh, Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Sanjita Phijam, and Thoudam Doren Singh. 2024. Wmt24 system description for the multiindic22mt shared task on manipuri language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 797–803.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter? exploring different human judgments with a tunable mt metric. In *Proceedings of the fourth workshop on statistical machine translation*, pages 259–268.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.