Findings of the JUST-NLP 2025 Shared Task on English-to-Hindi Legal Machine Translation

Kshetrimayum Boynao Singh[1], Sandeep Kumar[1], Debtanu Datta[2], Abhinav Joshi[3], Shivani Mishra[3], Shounak Paul[2], Pawan Goyal[2], Sarika Jain[4], Saptarshi Ghosh[2], Ashutosh Modi[3], Asif Ekbal[1]

[1]IIT Patna, India [2]IIT Kharagpur, India [3]IIT Kanpur, India [4]NIT Kurukshetra, India

Correspondence: boynfrancis@gmail.com

Abstract

This paper provides an overview of the Shared Task on Legal Machine Translation (L-MT), organized as part of the JUST-NLP 2025 Workshop at IJCNLP-AACL 2025, aimed at improving the translation of legal texts, a domain where precision, structural faithfulness, and terminology preservation are essential. The training set comprises 50,000 sentences, with 5,000 sentences each for the validation and test sets. The submissions employed strategies such as: domain-adaptive fine-tuning of multilingual models, QLoRA-based parameter-efficient adaptation, curriculum-guided supervised training, reinforcement learning with verifiable MT metrics, and from-scratch Transformer training. The systems are evaluated based on BLEU, METEOR, TER, chrF++, BERTScore, and COMET metrics. We also combine the scores of these metrics to give an average score (AutoRank). The top-performing system is based on a fine-tuned distilled NLLB-200 model and achieved the highest AutoRank score of 72.1. Domain adaptation consistently yielded substantial improvements over baseline models, and precision-focused rewards proved especially effective for the legal MT. The findings also highlight that large multilingual Transformers can deliver accurate and reliable English-to-Hindi legal translations when carefully fine-tuned on legal data, advancing the broader goal of improving access to justice in multilingual settings.

1 Introduction

The JustNLP 2025[1] Shared Task on English-to-Hindi Legal Machine Translation (L-MT)[2] was launched to address a critical need in India's multilingual legal system, where more than 44 million cases remain pending, and a significant portion of the judicial process still operates in English. This linguistic mismatch restricts accessibility for Hindi-speaking citizens, highlighting the need for reliable, domain-specific translation tools. Legal texts, however, are complex, characterized by dense syntax, specialized terminology, and strict semantic precision, making legal MT (Joshi et al., 2024) substantially more challenging than general-domain translation.

To foster progress in this area, the shared task provided a curated parallel corpus of Indian legal documents and a standardized evaluation framework. The participating teams explored a wide range of approaches, including multilingual Transformer fine-tuning, QLoRA-based parameter-efficient methods (Jeon and Strube, 2025), curriculum learning, reinforcement learning with verifiable rewards, and training models from scratch (Singh et al., 2023). This overview summarizes their methodologies and results, offering insight into current capabilities and emerging strategies for high-quality legal translation (Mahapatra et al., 2025). The broader goal of the initiative is to stimulate innovation in Indian-language NLP and identify promising, yet underexplored techniques (Dabre and Kunchukuttan, 2024) that can support scalable, accessible, and accurate legal document processing for Indian languages.

Our primary objective in organizing this collaborative effort is to bring together researchers and developers to explore effective strategies for enhancing the quality of Indian-language machine translation, particularly in the legal domain. A secondary objective was to identify interesting but still underexplored practices, even if they do not directly contribute to achieving state-of-the-art MT performance.

[1]https://exploration-lab.github.io/JUST-NLP/
[2]https://www.codabench.org/competitions/10351/

Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025), pages 12–17

2 Task Dataset

2.1 Dataset for Fine-Tuning and Evaluation

For the development and assessment of the model, we introduce our English-to-Hindi legal parallel corpus, which serves as a training and validation dataset. This dataset consists of 50,000 sentence pairs for training[3], along with 5,000 validation sentence pairs. For testing, we evaluate on the WMT25 Legal benchmark dataset (Singh et al., 2025b), which includes 5,000 additional test sentences. To ensure that the model is trained on accurate and high-quality parallel data, a series of pre-processing and filtration procedures was implemented.

Our training corpus is curated specifically for the legal domain and encompasses a diverse range of judicial document types, including court judgments, legal petitions, bail applications, government regulations, and case proceedings. The average sentence length in the training data is approximately 29 words for English and 31 words for Hindi.

Filtering by Sentence Length. We excluded sentence pairs in which either the source or target sentence contained fewer than 5 words or more than 70 words in the training and validation sets. Sentences that are excessively short often provide minimal supervision signals, whereas overly long sentences introduce structural complexity that may compromise training stability. Therefore, this length-based filtering helps preserve a more uniform and trainable distribution of training samples.

Dataset Split	Language	Sentences	Words
Train	English	50,000	1,463,911
	Hindi	50,000	1,577,884
Validation	English	5,000	145,063
	Hindi	5,000	156,850
Test	English	5,000	134,016
	Hindi	5,000	133,471

Table 1: Statistics of the English and Hindi Legal Translation Dataset

Word Count Ratio Filtering. To further enhance the alignment quality, we filtered sentence pairings based on the ratio of word counts between the source and target sentences. Only pairings in which the difference in word count fell within the range of $[-10, +10]$ were preserved. This phase helps remove misaligned pairs where the target sentence is significantly longer or shorter than the source, thereby reducing noise and promoting improved structural correspondence between the languages.

3 Evaluation Metrics

We use six widely used automatic evaluation metrics to fully assess system performance in the English-to-Hindi Legal MT shared task. Each metric examines a different aspect of translation quality, so that systems are evaluated not only on their surface similarity, but also on their accuracy in terms of meaning and context, which are crucial for legal translation.

- **BLEU (Papineni et al., 2002)**: Measures n-gram overlap between the system output and reference translation, focusing on lexical accuracy.
- **METEOR (Banerjee and Lavie, 2005)**: Evaluates translations using exact, stem, and synonym matches, offering a more flexible lexical comparison than BLEU.
- **TER (Snover et al., 2006)**: Computes the number of edits required to transform the MT output into the reference; lower scores indicate better structural alignment.
- **chrF++ (Popović, 2017)**: A character and word-level F-score metric that captures fine-grained morphological and orthographic similarities.
- **BERTScore (Zhang et al., 2020)**: Uses contextual embeddings to measure semantic similarity between the reference and the generated translation.
- **COMET (Rei et al., 2020)**: A neural metric trained on human judgments that captures adequacy and fluency using encoder–decoder representations.

Together, these metrics provide complementary perspectives, lexical, structural, semantic, and contextual, offering a more complete understanding of translation quality.

3.1 AutoRank for Unified Scoring

Each metric provides useful insights, but relying on only one can overemphasize certain translation

[3] https://huggingface.co/datasets/helloboyn/IJCNLP-JustNLP-LMT

characteristics, such as n-gram precision or semantic similarity. To avoid this kind of bias, we utilize AutoRank (Kocmi et al., 2025), a single scoring system that combines all six evaluation metrics into a single balanced score.

$$\text{AutoRank} = \frac{1}{6}\sum_{i=1}^{6} M_{i,\text{norm}} \quad (1)$$

AutoRank score[4] ensures that no single metric is more important than the others by placing all of them on the same scale. Metrics like BLEU, METEOR, chrF++, BERTScore, and COMET are scaled directly, whereas TER (where lower is better) is inverted to maintain consistency. The final AutoRank value is then found by averaging the normalized values on a scale ranging from 0 to 100.

This unified scoring method ensures that systems are evaluated as a whole, taking into account lexical accuracy, semantic meaning, structural fidelity, and contextual alignment, rather than relying on a single metric. Because of this, AutoRank provides the legal MT shared task with a fair, comprehensive, and dependable ranking system.

4 Summary of Participant Systems

A total of 24 teams initially registered for the shared task. Of these, 7 teams submitted system outputs, and 5 teams provided accompanying system description papers. The participating teams and their submitted systems are summarized as follows.

4.1 Team-SVNIT: Domain-Adaptive Fine-Tuning of Multilingual Models for English-Hindi Legal Machine Translation

Team-SVNIT (Dhakad et al., 2025) achieved 1st place in the shared task. Legal translation between English and Hindi is particularly challenging due to domain-specific terminology and long, syntactically complex sentence structures. To address this, the team fine-tunes and evaluates several multilingual pre-trained translation models, including `facebook/nllb-200-distilled-1.3B`, on the 50,000 English-Hindi legal sentence pairs provided by the organizers. Their training pipeline incorporates careful pre-processing, a 512-token context window, and optimized decoding strategies to improve robustness and translation fidelity. The final system attained an AutoRank score of 72.10, securing the top position on the leaderboard. Across evaluation metrics, the model achieved: BLEU 51.61, METEOR 75.80, TER 37.09, CHRF++ 73.29, BERTScore 92.61, and COMET 76.36. These results underscore the effectiveness of domain-adaptive fine-tuning for specialized legal MT tasks. The authors have publicly released their implementation for further research and reproducibility[5].

4.2 FourCorners: Cold Starts and Hard Cases - A Two-Stage SFT–RLVR Approach for Legal Machine Translation

FourCorners (Akarajaradwong and Chaksangchaichot, 2025) introduced one of the most innovative methodologies in the shared task. Their system for the JUST-NLP 2025 English–Hindi Legal Machine Translation challenge is built around a two-stage, data-centric training pipeline designed to handle both easy and difficult examples effectively. In the first stage, the team annotates the training corpus by translation difficulty, creating subsets labeled as 'easy', 'medium', and 'hard'. Supervised fine-tuning (SFT) is then applied to the easy-to-medium partition to establish a strong cold start model. In the second stage, the authors employ Reinforcement Learning with Verifiable Rewards (RLVR) exclusively on the hard subset. The reward signal, derived from standard machine translation metrics such as BLEU, ROUGE-L, and chrF++, enables the model to optimize directly for precision-oriented translation quality. This two-stage SFT-RLVR framework yields substantial improvements over strong SFT baselines, validating the effectiveness of combining difficulty-aware data curation with metric-driven reinforcement learning for high-fidelity legal translation. The system ranked 2nd overall in the official leaderboard. The authors have released their code and model weights[6].

4.3 goodmen: A Comparative Study of Neural Models for English-Hindi Legal Machine Translation

The *goodmen* (K et al., 2025) team ranked 3rd overall in the shared task. In a multilingual country like India, ensuring that legal judgments are accessible in native languages is essential for equitable

[4] https://huggingface.co/datasets/helloboyn/IJCNLP-JustNLP-LMT/blob/main/JustNLP25_L-MT_Result_Final.pdf

[5] https://github.com/Rupeshddhakad06/JUST-NLP-LMT

[6] https://github.com/ppaolong/FourCorners-JustNLP-MT-Shared-Task

justice. The Legal Machine Translation (L-MT) shared task aims to address this need by focusing on English-Hindi legal translation. The authors present a systematic comparative evaluation of neural MT models for this domain, examining four multilingual and Indic-focused systems developed under the shared-task constraints.

Their approach emphasizes domain-specific fine-tuning while preserving statutory structure, legal citations, and jurisdictional terminology. In particular, they fine-tune two legal-oriented models: `InLegalTrans` and `IndicTrans2` on the 50k English-Hindi legal parallel corpus provided by the organizers, with no external data permitted. The fine-tuned `InLegalTrans` model achieves the strongest performance, reaching a BLEU score of 48.56, substantially outperforming its base version. The comparative study reveals that targeted domain adaptation substantially enhances translation quality for specialized legal texts. Human evaluation further confirms that the fine-tuned `InLegalTrans` outputs better preserve legal coherence and judicial tone. The team's best-performing model is ranked 3rd on the test set. The authors have released their model[7].

4.4 JUNLP: From Scratch to Fine-Tuned: A Comparative Study of Transformer Training Strategies for Legal Machine Translation

The JUNLP (Barman et al., 2025) team ranked 4th overall in the shared task. In multilingual countries like India, access to legal information is often constrained by linguistic barriers, as much of the judicial and administrative discourse remains in English. Legal Machine Translation (L-MT) offers a scalable approach to accessible legal communication by enabling the consistent and accurate translation of legal texts. This work, developed for the JUST-NLP 2025 shared task, investigates English–Hindi legal translation using Transformer-based approaches. The authors evaluate two complementary strategies: fine-tuning a pre-trained OPUS-MT model for domain adaptation, and training a Transformer model from scratch using only the provided legal corpus. These systems are assessed across a broad suite of MT metrics, including SacreBLEU, chrF++, TER, ROUGE, BERTScore, METEOR, and COMET. Their fine-tuned OPUS-MT model achieves a SacreBLEU score of 46.03, substantially outperforming both the baseline and from-scratch models. The results demonstrate the clear advantage of domain adaptation for legal MT, showing that pretrained models fine-tuned on in-domain data significantly surpass models trained from scratch. The study highlights the potential of L–MT systems to improve legal accessibility and transparency in multilingual settings[8].

4.5 JUST-MEI: Adapting IndicTrans2 for Legal Domain MT via QLoRA Fine-Tuning

The JUST-MEI (Singh et al., 2025a) system ranked 5th overall in the shared task. Legal Machine Translation presents unique difficulties due to domain-specific terminology, long and formally structured statutes, and the high precision required for legal communication. As part of the JUST-NLP 2025 Shared Task on English-Hindi legal translation, the authors adapt the pretrained `ai4bharat/indictrans2-en-indic-1B` model using QLoRA, a parameter-efficient fine-tuning strategy designed for low-resource computational settings. Using only the domain-specific parallel corpus provided by the organizers, the fine-tuned model achieves substantial improvements over the baseline IndicTrans2 system, especially in handling specialized legal vocabulary and complex syntactic constructions. In automatic evaluation, the system obtains a BLEU score of 46.67 and a chrF++ score of 70.03. Human evaluation further reflects strong performance, with adequacy and fluency scores of 4.085 and 4.006, respectively. The approach achieves a final AutoRank score of 67.98, demonstrating the effectiveness of QLoRA-driven domain adaptation for legal MT[9].

5 Results and Findings

Among the five participating systems, JUST-NLP 2025 Shared Task on English-to-Hindi Legal Machine Translation revealed that domain-adaptive fine-tuning remains the most decisive factor for achieving high-quality legal translations. The top-ranked 'Team-SVNIT' demonstrated that large multilingual models, such as `facebook/nllb-200-distilled-1.3B`, when carefully fine-tuned with long-context training, rigorous preprocessing, and stable optimization, de-

[7] https://huggingface.co/drjk16/InLegalTrans-Finetuned-JUSTNLP2025

[8] https://github.com/atanumandal0491/Legal-Translation

[9] https://drive.google.com/drive/folders/1USkOkqvV3HxFILAPsFpkxWYJ1FjNFRUl

Rank	Team Name	BLEU↑	METEOR↑	TER↓	chrF++↑	BERTScore↑	COMET↑	AutoRank↑
1	Team-SVNIT	51.61	75.80	37.09	73.29	92.61	76.36	72.10
2	FourCorners	50.19	69.54	42.32	73.67	92.70	75.74	69.92
3	goodmen	48.56	67.15	41.63	73.07	92.38	75.16	69.12
4	JUNLP	46.03	71.84	42.08	70.59	91.19	73.72	68.55
5	JUST-MEI	46.67	72.86	44.63	70.03	90.86	72.12	67.98
6	Lawgorithms	46.27	71.80	43.06	68.32	91.03	72.14	67.75
7	Tokenizers	34.08	61.78	55.25	56.75	87.39	65.20	58.32

Table 2: Evaluation results for the English-to-Hindi Legal MT Shared Task.

liver state-of-the-art performance across lexical and semantic metrics. 'FourCorners' contributed the most innovative approach, showing that curriculum-based learning combined with reinforcement learning using verifiable MT metrics can outperform conventional supervised fine-tuning, especially in precision-sensitive domains like law. The 'goodmen' team highlighted the value of domain-specific architectures, with their fine-tuned `InLegalTrans` model excelling in preserving statutory structure and legal terminology. 'JUNLP' results further underscored the superiority of pretrained models, as their fine-tuned OPUS-MT system significantly outperformed a Transformer trained from scratch. Finally, 'JUST-MEI' demonstrated that parameter-efficient adaptation via QLoRA can achieve competitive performance even under limited computational resources. Together, these findings affirm that legal MT requires not only linguistic fluency but rigorous domain adaptation, precision-oriented optimization, and careful engineering, marking a substantial advancement in the development of reliable English-to-Hindi legal translation systems.

6 Conclusion

The shared task demonstrates that high-quality English-to-Hindi legal translation is well within reach when modern Transformer models are carefully tailored to the legal domain. We found that simply using large, pre-trained models is not enough; fine-tuning them on legal data is essential for capturing the precision and structure that legal texts require. Approaches that prioritize accuracy over paraphrasing, especially those optimized with BLEU and chrF++, consistently produced more reliable outputs. Innovative strategies, such as combining supervised training with reinforcement learning, further boosted performance, and even lightweight methods like QLoRA proved effective for teams working with limited compute resources. Overall, these results demonstrate that with the right combination of domain adaptation, careful engineering, and precision-focused training, today's MT systems can produce translations that are accurate, trustworthy, and suitable for real-world legal applications.

Limitations

The shared task is limited only to the English-to-Hindi direction and does not consider the reverse Hindi-to-English direction. Although we follow the AutoRank framework with normalized scores for fair system comparison, the inclusion of TER as an inverted (negative) metric may not always correlate consistently with other positive evaluation metrics, which could potentially influence the final ranking. Finally, the shared task evaluation is conducted only at the sentence level; however, a document-level evaluation would be essential for capturing broader contextual dependencies and discourse-level translation quality, particularly in the legal domain.

Acknowledgement

The authors express gratitude to the COIL-D Project under Bhashini, funded by MeitY, Govt. of India, for the support and resources that facilitated the successful execution of this research.

References

Pawitsapak Akarajaradwong and Chompakorn Chaksangchaichot. 2025. Cold Starts and Hard Cases: A Two-Stage SFT-RLVR Approach for Legal Machine Translation (Just-NLP L-MT shared task). In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Pro-*

ceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Amit Barman, Atanu Mandal, and Sudip Kumar Naskar. 2025. From Scratch to Fine-Tuned: A Comparative Study of Transformer Training Strategies for Legal Machine Translation. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Raj Dabre and Anoop Kunchukuttan. 2024. Findings of WMT 2024's MultiIndic22MT shared task for machine translation of 22 Indian languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 669–676, Miami, Florida, USA. Association for Computational Linguistics.

Rupesh Dhakad, Naveen Kumar, and Shrikant Malviya. 2025. Team-SVNIT at JUST-NLP 2025: Domain-Adaptive Fine-Tuning of Multilingual Models for English–Hindi Legal Machine Translation. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Sungho Jeon and Michael Strube. 2025. Entity tracking in small language models: An attention-based study of parameter-efficient fine-tuning. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 42–53, Suzhou, China. Association for Computational Linguistics.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. IL-TUR: Benchmark for Indian legal text understanding and reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11460–11499, Bangkok, Thailand. Association for Computational Linguistics.

Deeraj S K, Karthik Suryanarayanan, Yash Ingle, and Pruthwik Mishra. 2025. goodmen @ L-MT Shared Task: A Comparative Study of Neural Models for English-Hindi Legal Machine Translation. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Sayan Mahapatra, Debtanu Datta, Shubham Soni, Adrijit Goswami, and Saptarshi Ghosh. 2025. Milpac: A novel benchmark for evaluating translation of legal text to indian languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(8).

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Akoijam Jenil Singh, Loitongbam Sanayai Meetei, and Yumnam Surajkanta. 2025a. Adapting IndicTrans2 for Legal Domain MT via QLoRA Fine-Tuning at JUST-NLP 2025. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025b. Evaluation of LLM for English to Hindi legal domain machine translation systems. In *Proceedings of the Tenth Conference on Machine Translation*, pages 823–833, Suzhou, China. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich 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, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.