goodmen @ L-MT Shared Task: A Comparative Study of Neural Models for English-Hindi Legal Machine Translation

Deeraj S K, Karthik Suryanarayanan, Yash Ingle, Pruthwik Mishra
Sardar Vallabhbhai National Institute of Technology, Surat
{u23ai050, u23cs013, u23ai062}@coed.svnit.ac.in
{pruthwikmishra}@aid.svnit.ac.in

Abstract

In a massively multilingual country like India, providing legal judgments in understandable native languages is essential for equitable justice to all. The Legal Machine Translation (L-MT) shared task focuses on translating legal content from English to Hindi which is the most spoken language in India. We present a comprehensive evaluation of neural machine translation models for English-Hindi legal document translation, developed as part of the L-MT shared task. We investigate four multilingual and Indic focused translation systems. Our approach emphasizes domain specific fine-tuning on legal corpus while preserving statutory structure, legal citations, and jurisdictional terminology. We fine-tune two legal focused translation models, InLegalTrans and IndicTrans2 on the English-Hindi legal parallel corpus provided by the organizers where the use of any external data is constrained. The fine-tuned InLegalTrans model achieves the highest BLEU score of 0.48. Comparative analysis reveals that domain adaptation through fine-tuning on legal corpora significantly enhances translation quality for specialized legal texts. Human evaluation confirms superior coherence and judicial tone preservation in InLegalTrans outputs. Our best performing model is ranked 3^{rd} on the test data.

1 Introduction

Legal translation is one of the most challenging domains in natural language processing, requiring not only linguistic accuracy, but also preservation of legal semantics, statutory structure, and jurisdictional terminology. In multilingual legal systems such as India's, where legal proceedings and documentation occur across multiple languages, accurate translation between English and Indian languages is essential for ensuring access to justice and legal transparency. The linguistic and cultural gap between English legal texts and their Hindi translations demands specialized translation systems that can handle domain specific terminology, complex sentence structures, and formal register. Recent studies in legal NLP highlight concrete failures of general-purpose systems on legal discourse: domain-specific pretraining or fine-tuning consistently improves performance on legal tasks such as judgment classification, statutory retrieval, and terminology preservation (Chalkidis et al., 2020; Zheng et al., 2021; Chu and Wang, 2018). For Indic language pairs, large parallel resources such as Samanantar have enabled improved base models for Indian languages (Ramesh et al., 2021), yet English–Indic legal translation remains underresourced compared to English–European pairs. This gap is further amplified by code-switching and transliteration phenomena in Indian legal texts, which complicate tokenization and lexical alignment (see e.g. (Mujadia et al., 2024)).

Empirical evidence from prior MT research indicates that domain adaptation, either via continued pre-training on in-domain corpora or targeted fine-tuning—yields substantial gains in adequacy and terminology fidelity compared to out-of-domain baselines (Chu and Wang, 2018; Farajian et al., 2017; Rossi and Chevrot, 2019). Furthermore, recent large multilingual models (e.g., NLLB-200) demonstrate strong cross-lingual transfer but often underperform specialized, domain-adapted models on niche corpora unless further adapted (Costa-jussà et al., 2022; Mahapatra et al., 2025).

Traditional rule-based and statistical machine translation approaches have historically struggled with the specialized vocabulary and syntactic complexity present in legal documents. The advent of neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) has significantly improved translation quality across general domains, yet legal translation remains underexplored, particularly for low-resource language pairs such as English-Hindi in

Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025), pages 127–132

legal contexts.

The L-MT shared task on English-Hindi legal translation provides a standardized evaluation framework for developing translation systems in the legal domain. This addresses the critical need for automated translation tools capable of processing Indian legal judgments, statutes, and legal documents while maintaining semantic fidelity and legal phrasing accuracy. In this work, we present a comprehensive evaluation of four translation systems spanning different architectural paradigms and parameter scales: IndicTrans2 (200M), a specialized encoder-decoder model for Indic languages; InLegalTrans (1B), a domain-adapted model pre-trained on legal corpora; NLLB-200 Distilled (1.3B), a massively multilingual baseline; and Gemini 2.0 Flash API, a large-scale commercial model. Our investigation focuses on the impact of domain-specific fine-tuning, legal-aware preprocessing, and terminology preservation strategies on translation quality.

2 Related Work

Neural machine translation has evolved significantly since the introduction of attention-based sequence-to-sequence architectures (Bahdanau et al., 2015) and self-attention based transformer models (Vaswani et al., 2017). Pre-trained multilingual models such as mBART (Liu et al., 2020), mT5 (Xue et al., 2021), and NLLB-200 (Costa-jussà et al., 2022) have demonstrated impressive zero-shot translation capabilities across hundreds of languages through large-scale multilingual pre-training. These models leverage cross-lingual transfer learning to achieve robust performance even on low-resource language pairs.

For Indic languages, the Samanantar corpus (Ramesh et al., 2021) provided large-scale parallel data for English to Indic and Indic to Indic language pairs that led to the development of IndicTrans. This model introduced the first large-scale model specifically designed for Indian language pairs. IndicTrans2 (Gala et al., 2023) extended this work with improved architectures, larger training corpora, better noise filtering, and better handling of script normalization. Recent parameter-efficient fine-tuning methods including LoRA (Hu et al., 2021) and adapter layers (Houlsby et al., 2019) have enabled domain adaptation with minimal computational overhead retaining pre-trained knowledge.

Although all these NMT systems have been seamlessly integrated into many domains through domain adaptation, legal translation poses unique challenges including specialized terminology, formal register, and complex syntactic structures characteristic of statutory language (Cao, 2007; Šarčević, 2000). Domain-specific models such as Legal-BERT (Chalkidis et al., 2020) and legal domain pre-training approaches (Zheng et al., 2021) have demonstrated the value of legal domain pre-training for natural language understanding of English legal text. However, legal NMT for English-Indic language pairs remains critically underexplored despite the practical importance in multilingual legal systems.

Prior work on legal NMT has focused on domain adaptation through continued pre-training on legal corpora (Chu and Wang, 2018; Farajian et al., 2017) and incorporation of legal terminology glossaries (Rossi and Chevrot, 2019). Our work extends this research by conducting a comparative evaluation of multiple neural architectures for English-Hindi legal translation, demonstrating substantial quality improvements through domain-specific fine-tuning while maintaining strict shared-task constraints that prohibit external data usage. MILPac (Mahapatra et al., 2025) consists of MT benchmarks in the legal domain vetted by law practioners for several English and low resource Indian language pairs. The authors also released InLegalTrans, an multilingual NMT model fine-tuned on IndicTrans2. LMs (Zhu et al., 2024) also have been widely used for machine translation where the machine translation is performed through a decoder only model rather than the traditional encoder-decoder models. The translation capabilities from English to diverse Indian languages (Mujadia et al., 2024) of different LLMs have been studied.

3 Methodology

3.1 Dataset

The L-MT English-Hindi Legal Translation corpus (Singh et al., 2025) consists of parallel sentence pairs extracted from Indian legal documents, including court judgments, statutory provisions, and legal proceedings. Each entry contains an English source sentence and its corresponding Hindi translation in Devanagari script. The corpus exhibits domain-specific characteristics typical of Indian legal discourse, including complex syntactic struc-

tures, specialized terminology, and formal register. It includes 50,000 samples provided for fine-tuning, 5,000 samples provided for validation during the training phase (validation set), and 5,000 samples on which the final BLEU score was calculated (test set).

The dataset preserves legal-specific elements including citation patterns (e.g., "AIR 1997 SC 1234", "Section 125 of the CrPC"), constitutional articles, domain-specific legal terminology such as "writ petition", "respondent", "jurisdiction" and "fundamental rights" that require accurate translation or appropriate transliteration, and numerical consistency with dates, case numbers, section identifiers, and monetary amounts preserved across source and target sentences. For submissions and evaluations, we train on the complete training corpus to maximize model exposure to domain-specific patterns.

Table 1: Corpus statistics for English-Hindi legal translation.

Split	# Samples	Usage
Training	50,000	Fine-tuning
Validation	5,000	Dev evaluation
Test	5,000	Final evaluation

The training corpus exhibits linguistic diversity across legal sub-domains including constitutional law, criminal procedure, civil litigation, contract law, and property law. Average sentence length is approximately 28 tokens for English and 32 tokens for Hindi, reflecting the morphological richness of Devanagari script. The corpus maintains authentic translation challenges including ambiguous legal terminology, code-switching patterns where English legal terms are retained in Hindi translations, and syntactic divergence in terms of grammatical structures. All experiments strictly adhere to shared-task constraints by using only the official provided datasets without external corpora, back-translation, or synthetic augmentation.

3.2 Training Details

The models are trained for 3 epochs with batch size 4 and learning rate 2e-4 on an NVIDIA A100 GPU with 94GB RAM. We follow IndicTrans2's preprocessing guidelines for normalization and tokenization. Script tags in the format Eng_Latn → Hin_Deva are added to each translation pair to ensure script consistency.

Table 2: Training configuration for InLegalTrans-en2Indic-1B

Component	Setting
Base model	InLegalTrans-en2Indic-1B
Quantization	4-bit NF4 (bitsandbytes)
LoRA rank/alpha	$r = 16$, $\alpha = 32$
Target modules	q_proj, k_proj, v_proj, o_proj, fc1, fc2
Optimizer	paged_adamw_32bit
Learning rate	2×10^{-4} (linear decay)
Batch size	4 per device
Epochs	3
Precision	FP16 mixed precision
Max seq. length	512 tokens
Hardware	NVIDIA A100 (94 GB)
Notes	Gradient accumulation used to simulate larger batch size

Particular care is taken to preserve legal symbols, citations, and numbering. Symbols like "§," "Sec.," and "Art." are kept exactly as they are. We eliminate pairs with empty target translations and filter out sentence pairs with significant length mismatches (greater than 3:1) to avoid model confusion. To minimize GPU memory consumption and enable efficient domain adaptation, InLegalTrans is optimized with parameter-efficient adapters using LoRA/PEFT techniques.

3.3 Evaluation Metrics

We evaluate the translation quality using three complementary metrics: BLEU (Papineni et al., 2002) for n-gram precision, ROUGE-L (Lin, 2004) for longest common subsequence to assess sentence-level structural preservation, and chrF++ (Popović, 2017) for character-level or subword accuracy. For inference, we use beam search decoding with beam width of 4 and temperature of 0.7, followed by post-processing for punctuation restoration and formatting corrections.

4 Results and Analysis

Table 3 shows detailed evaluation metrics across all three metrics on the final test set.

4.1 Quantitative Analysis

From the results, we observe that the fine-tuned InLegalTrans model achieves the highest BLEU score of 0.48, representing a 55% improvement over the base model (0.31 BLEU). This model

Table 3: Detailed evaluation metrics on test set. (final leaderboard score)

Model	BLEU	ROUGE-L	chrF++
Base Models			
IndicTrans2 (Gala et al., 2023)	0.30	0.42	0.52
Gemini 2.0 Flash (few-shot)	0.16	0.35	0.43
InLegalTrans (Mahapatra et al., 2025)	0.31	0.44	0.54
NLLB-200 (Costa-jussà et al., 2022)	0.27	0.39	0.49
Fine-tuned Models			
IndicTrans2 (Gala et al., 2023)	0.31	0.43	0.53
InLegalTrans (Mahapatra et al., 2025)	**0.48**	**0.56**	**0.73**

demonstrates superior performance in preserving the inherent characteristics of the legal texts. IndicTrans2 maintains stable performance at approximately 0.31 BLEU regardless of fine-tuning, suggesting strong pre-trained generalization to formal text but limited domain adaptation capacity for legal-specific patterns.

Several key observations emerge from the evaluation:

Domain Adaptation Impact: Fine-tuning on legal corpora yields substantial improvements only for InLegalTrans (+0.17 BLEU), while IndicTrans2 showed minimal gains (+0.01 BLEU). This suggests that InLegalTrans's architecture and legal pre-training approach are more receptive to domain-specific adaptation.

Model Scale vs. Specialization: Despite having fewer parameters (1B vs. 1.3B), the fine-tuned InLegalTrans significantly outperformed the larger NLLB-200 model (0.48 vs. 0.27 BLEU). This demonstrates that domain specialization and targeted fine-tuning can be more effective than raw model capacity for specialized translation tasks.

Commercial Model Performance: Gemini 2.0 Flash achieved the lowest BLEU score (0.16) despite being a large-scale commercial model. Manual inspection reveals that while Gemini produces fluent translations, they frequently deviate from legal phrasing conventions and exhibited semantic inconsistencies in handling statutory language, instead preferring generic terms in generation.

Metric Consistency: Performance rankings remains consistent across all three metrics (BLEU, ROUGE-L, chrF++), with InLegalTrans (FT) leading in all categories. The strong correlation between metrics validates the robustness of our evaluation. The model's strong performance in ROUGE-L (0.56) and chrF++ (0.73) metrics indicates robust sentence-level structural preservation and character-level accuracy.

4.2 Qualitative Analysis

Table 4 presents a representative translation example demonstrating the strengths and weaknesses of different models on legal text.

Table 4: Example translations from different models.

Model	Translation
Source	The petitioner has challenged the constitutional validity of Section 377.
Reference	याचिकाकर्ता ने धारा 377 की संवैधानिक वैधता को चुनौती दी है।
InLegalTrans (FT)	याचिकाकर्ता ने धारा 377 की संवैधानिक वैधता को चुनौती दी।
IndicTrans2 (FT)	याचिकाकर्ता ने अनुच्छेद 377 की संवैधानिक मान्यता को चुनौती दी।
NLLB-200	अर्जीदार ने सेक्शन 377 की संवैधानिक वैधता को चुनौती दी है।
Gemini (few-shot)	याचिकाकर्ता ने धारा 377 की वैधता पर सवाल उठाया है।

The qualitative analysis reveals that InLegalTrans (FT) produces translations nearly identical to the reference, maintaining precise legal terminology. IndicTrans2 substitutes वैधता (validity) with मान्यता (recognition), introducing a subtle but significant semantic shift in legal meaning. NLLB-200 uses inconsistent terminology (अर्जीदार instead of याचिकाकर्ता for petitioner) and transliterates "Section" as सेक्शन rather than using the proper Hindi term धारा. Gemini paraphrases excessively, changing "challenged" to "questioned" (सवाल उठाया), which alters the legal force and precision of the statement.

These examples confirm that domain-specific fine-tuning on legal corpora is essential for achieving high-quality English-Hindi legal translation,

and that specialized models outperform general-purpose systems in preserving legal semantics.

5 Conclusion

In this paper, we present a comprehensive evaluation of neural machine translation models for the translation of English-Hindi legal documents as part of the L-MT shared task. We demonstrate that domain-specific fine-tuning on legal corpora substantially enhances translation quality for specialized legal texts, with the fine-tuned InLegalTrans model achieving the highest BLEU score of 0.48, a 55% improvement over its base performance.

Our comparative analysis of four translation systems spanning different architectural paradigms revealed that domain specialization and targeted fine-tuning can be more effective than raw model capacity, as evidenced by the 1B parameter InLegalTrans outperforming the larger 1.3B parameter NLLB-200 model. The results confirm that preserving legal terminology, statutory structure, and formal phrasing requires dedicated domain adaptation rather than relying solely on general-purpose multilingual models. As a natural extention of this work, we would explore the possibility of developing MT models from English to other Indian languages. Since the legal domain is a critical domain, it requires quality legal benchmarks to evaluate the developed models. We would like to work in this direction as well. We plan to introduce linguistic regularization mechanisms during training to explicitly model legal discourse markers and domain-specific cue phrases. The final fine-tuned model (InLegalTrans-FT) is available here - Hugging Face.

Limitations

While our system demonstrates strong performance on the L-MT dataset, several limitations warrant acknowledgment. The fine-tuning is performed exclusively on the provided legal corpus, which may limit generalization to other legal subdomains or regional legal language variations. The evaluation primarily relies on automatic metrics (BLEU, ROUGE-L, chrF++), which may not fully capture nuanced legal semantic equivalence. Although human evaluation at limited is carried out, it is conducted by non-experts. The system's handling of rare legal terminology and emerging legal concepts requires further extensive human evaluation by legal experts.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Deborah Cao. 2007. *Translating Law*. Multilingual Matters.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319.

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*.

M Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137.

Jay Gala, Pranjal A Chitale, AK Raghavan, Sumanth Doddapaneni, Varun Gumma, Aravinth Bheemaraj, Divyanshu Addanki, Divyanshu Kakwani, Anoop Kunchukuttan Sharma, Pratyush Kumar, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. In *Transactions on Machine Learning Research*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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 Transactions on Asian and Low-Resource Language Information Processing*, 24(8):1–30.

Vandan Mujadia, Ashok Urlana, Yash Bhaskar, Penumalla Aditya Pavani, Kukkapalli Shravya, Parameswari Krishnamurthy, and Dipti Sharma. 2024. Assessing translation capabilities of large language models involving English and Indian languages. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 207–228, Sheffield, UK. European Association for Machine Translation (EAMT).

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.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, and 1 others. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. In *Transactions of the Association for Computational Linguistics*, volume 10, pages 145–162.

Carolina Rossi and Jean-Pierre Chevrot. 2019. Legal translation quality in the european union: A multi-method study. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 165–176.

Susan Šarčević. 2000. *New Approach to Legal Translation*. Kluwer Law International.

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 (WMT 2025)*, pages 823–833. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Haoxi Zheng, Mirella Lapata, and 1 others. 2021. Does pre-training on legal corpora improve legal language understanding? a case study on contract understanding. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–104.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.