Team-SVNIT at JUST-NLP 2025: Domain-Adaptive Fine-Tuning of Multilingual Models for English–Hindi Legal Translation

Rupesh Dhakad and Naveen Kumar and Shrikant Malviya
Sardar Vallabhbhai National Institute of Technology, Surat
dhakadrupesh666@gmail.com,{naveenkumar,shrikant}@coed.svnit.ac.in

Abstract

Translating the sentences between English and Hindi is challenging specially in the domain of legal document. The major reason behind the complexity is specialized legal terminology, long and complex sentence, and the accuracy constraint. This paper presents system developed by Team-SVNIT for the JUST-NLP 2025 shared task on legal machine translation. Our approach fine-tunes the distilled NLLB-1.3B multilingual model using a prepared dataset of 50,000 English–Hindi legal sentence pairs. The training pipeline includes preprocessing, context windows of 512 tokens, and decoding methods to enhance translation quality. The proposed method secured **first place** on the official leaderboard with the AutoRank score of **61.62**. It obtained the following scores: BLEU 51.61, METEOR 75.80, TER 37.09, CHRF++ 73.29, BERTScore 92.61, and COMET 76.36. These results show that fine-tuning multilingual models for a specific domain improves performance. It works better than general translation systems. The method is also more effective for legal texts.

1 Introduction

Legal machine translation is more difficult than general translation. It needs both accurate language use and correct handling of legal terms (Panezi and O'Shea, 2023). The JUST-NLP 2025 shared task deals with English–Hindi legal translation. These two languages differ in structure and in the way legal ideas are expressed. Accurate translation is not just about replacing words. It also requires keeping the legal meaning and intent the same across both systems (Way, 2016).

Many problems are known in this area. Using the same legal terms consistently in all contexts is hard (Altakhaineh, 2025). Legal sentences are often long and contain many clauses. This makes them difficult for translation models to process. The lack of parallel legal data limits how much supervised training can be done (Raja and Vats, 2025). In legal work, even a small translation mistake can cause serious problems (Llop, 2025). Differences in legal ideas between countries also require careful meaning adjustments instead of direct word-for-word translation.

Recent developments in large-scale multilingual NMT, particularly the No Language Left Behind effort, have produced strong cross-lingual transfer across almost 200 languages (Costa-jussà et al., 2022, 2024), reporting improvements of up to 44% over prior systems. However, the applicability of these models to specialized domains like legal text—especially for Indian language pairs—remains relatively under-explored (Nair et al., 2024).

Within the Indic NLP community, systems such as IndicTrans (Ramesh et al., 2021) and IndicTrans2 (Gala et al., 2023) have broadened multilingual coverage from 11 to 22 languages, applying targeted techniques like distillation to scale effectively. Still, hurdles such as rich morphology, multiple scripts, and code-switching persist and complicate model performance on real-world legal corpora (Suman et al., 2023; Sheshadri and Soman, 2023).

This manuscript describes Team-SVNIT's pipeline and experimental evaluation for the JUST-NLP 2025 Legal MT task. Our main contributions are: **(1)** An empirical comparison of five candidate translation systems, which identifies the distilled NLLB-1.3B model as the best-performing backbone; **(2)** A preprocessing pipeline designed to clean noisy legal text extracted from PDFs and other sources; **(3)** A training regimen using ex-

tended contexts (512 tokens), a cautious learning rate (2e-5), and cosine-based scheduling; **(4)** A top-ranking submission that placed first on the task leaderboard; **(5)** A manual, qualitative appraisal of 100 samples to uncover model strengths and recurring error modes.

2 Related Work

2.1 Neural MT for Indic Languages

IndicTrans (Ramesh et al., 2021) was one of the first large-scale, multilingual NMT efforts for 11 Indian languages that employed language-aware preprocessing and transformer-based architectures. Transfer learning was considered to be the key factor behind its success, substantially improving translation quality. Finally, IndicTrans2 (Gala et al., 2023) brought coverage to 22 Indic languages and further refined knowledge distillation methods for scalability.The NLLB (Costa-jussà et al., 2022, 2024) project expanded translation into more than 200 languages with extensive coverage of the Indic families. Distilled versions of these models, 600M and 1.3B parameters, retain impressive translation performance with low computational cost thanks to successful knowledge distillation (Koishekenov et al., 2023). The architecture is based on a sparsely gated mixture-of-experts architecture that allows for optimal use of parameters without the computational overhead from dense models.

Despite these advances, some challenges are persistent: the Indic NMT systems have to grapple with morphosyntactic richness, orthographic variations across scripts, limited parallel data, and prevalence of code-switched content (Raja and Vats, 2025; Naveen et al., 2024).

2.2 Legal Domain NMT

Neural machine translation of legal texts is different from that of general-domain texts due to scarce domain-specific corpora, special terminology, and high accuracy demands. Complex syntactic structures in legal texts have to be rendered faithfully, as does the translation of jurisdiction-specific terminology (Way, 2016; Panezi and O'Shea, 2023). Minor mistranslations might have important consequences for legal interpretation and the conduct of proceedings (Llop, 2025).

Works like Altakhaineh (Altakhaineh, 2025) show that machine-translated legal content is often fraught with critical semantic and syntactic errors, requiring heavy human post-editing. This again emphasizes domain adaptation, model fine-tuning, and verification by humans if the systems are to be reliably deployed in a legal context (Princeton, 2025).

2.3 Multilingual Model Fine-tuning

Good fine-tuning methods are important for training multilingual models on domain data. Cosine annealing with warm restarts helps the model remember earlier learning during training (Loshchilov and Hutter, 2017). Using large batch sizes through gradient accumulation makes training more stable and helps it converge (Han et al., 2024). Parameter-efficient methods like LoRA save resources (Nair et al., 2024). However, full fine-tuning is still better for legal translation, where correctness is more important than speed or cost.

3 Task Description

3.1 Dataset

The JUST-NLP 2025 Legal MT shared task includes an English–Hindi parallel dataset. It covers different areas of law such as constitutional, civil, criminal, and administrative. Table 1 shows the main details of the dataset.

Split	Pairs	Avg(words)	Max(words)
Train	50,000	29.3	70
Valid	5,000	26.8	54
Test	5,000	26.1	—

Table 1: Dataset statistics (average words per sentence).

The legal sentences in this dataset are long and complex. The average English sentence length is 29.3 words. The Hindi translations are about 10% longer, averaging 33.1 words. The longest sequence has 387 tokens. This shows the need for longer context windows during training. The dataset also includes common legal phrases, legal citations, numbers, and some noise from digitization, such as mixed scripts and encoding errors.

Legal sentences are often difficult to translate. For example, a sentence like *"The ap-*

pellant, being aggrieved by the judgment and decree dated 15th March 2023 passed by the Hon'ble High Court of Delhi in Civil Appeal No. 2345 of 2022, prefers this present appeal under Section 96 of the Code of Civil Procedure, 1908" contains many legal references, dates, and laws that must be translated with care. Latin terms such as *"res judicata," "sub judice,"* and *"amicus curiae"* also need proper transliteration and meaning adjustment in Hindi legal language.

3.2 Evaluation Metrics

The evaluation uses six automatic measures. These are combined into one score called AutoRank. It is defined as:

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

The metrics include BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (inverted) (Snover et al., 2006), CHRF++ (Popović, 2017), BERTScore (Zhang et al., 2019), and COMET (Rei et al., 2020). Each score is scaled between 0 and 100. This combined metric helps balance word matching, edit distance, character accuracy, semantic meaning, and quality that matches human judgment.

4 System Architecture

4.1 Model Selection

We tested five translation models, including standard and distilled versions. Table 2 shows the full comparison.

4.2 Preprocessing

There are usually minor PDF errors in legal documents. These problems are corrected by our preprocessing pipeline. It substitutes line breaks with space and standardized dashes and quotation marks. It also solves the encoding issues, eliminates English words left behind in Hindi text and minimizes the additional spaces. Those measures make the text neat without any distortion of legal terms, numbers, and references.

4.3 Training Configuration

Table 3 shows the details of our training setup.

Rationale: 99% of corpus is covered by extended context (512). Stable optimization can be performed with large batch (512). Multilingual representations are maintained in Conservative LR (2e-5). Cosine scheduling avoids local minima. Early termination averts overfitting. FP16 cuts memory by 40 percent to allow bigger batches and 2-fold speed.

4.4 Inference

Beam search width=4, max length=512, n-gram penalty=3 and early stopping=on, deterministic output, batch=64.

5 Experiments and Results

5.1 Model Comparison

Table 4 shows validation performance.

The main observations include the following. The bigger model is more successful. The 1.3B model has a BLEU 6.3B score than the 600M model due to its ability to cope with complex legal patterns. Fine-tuning contributes approximately 4.0 BLEU, which demonstrates that it is essential in legal data adaptation. The filtered models are also performing. The NLLB-1.3B-distilled model has a score of 53.5 BLEU as opposed to the normal version of 52.8. This is due to the fact that distillation enhances generalization, as well as alleviates overfitting. The general multilingual models are also more effective than the domain-specific models. The NLLB-1.3B-distilled model (53.5 BLEU) outperforms InLegalTrans (51.5 BLEU) because of the more extensive training and the more multilingual nature of the former.

5.2 Official Results

Table 5 shows official test performance.

The following is shown by analysis. There is improved matching of words with a BLEU score of 51.61 on a range of 50.19 (+1.42). The increased score of 75.80 on the METEOR as compared to 69.54 (+6.26) suggests that the similarity of the meaning would be stronger using synonym and stem similarity. The lower score of TER of 37.09 implies that the system requires less manual editing. The similarity in texts is indicated by the BERTScore of 92.61, which is near 92.70. In general, the balanced outcomes of all measures

Model	Params	Langs	Arch	Features
Helsinki-NLP/opus-mt-en-hi	77M	2	Transformer	Lightweight
facebook/nllb-200-distilled-600M	600M	200	Trans+MoE	Conditional routing
facebook/nllb-200-distilled-1.3B	**1.3B**	**200**	**Trans+MoE**	**Our choice (distilled)**
facebook/nllb-200-1.3B	1.3B	200	Trans+MoE	Standard version
ai4bharat/indictrans2-en-indic-1B	1.0B	22	Transformer	Indic-specialized
law-ai/InLegalTrans-En2Indic-1B	1.0B	Indic	Transformer	Legal domain

Table 2: Model comparison. Selection prioritized multilingual capacity, sufficient parameters, and architectural sophistication. The distilled 1.3B variant demonstrated superior performance through optimized knowledge transfer.

Hyperparameter	Value
Base Model	facebook/nllb-200-distilled-1.3B
Max Sequence Length	512 tokens
Epochs	20 (early stop)
Batch Size	32
Gradient Accum	16 steps
Effective Batch	512
Learning Rate	2e-5
Scheduler	Cosine w/ restarts
Warmup Ratio	0.1
Optimizer	AdamW
Precision	FP16

Table 3: Training configuration using the distilled 1.3B model which showed better convergence and inference efficiency.

indicate that the model is also stable and dependable. The maximum AutoRank score of 61.62 is the indicator of the good quality of translation in general.

5.3 Ablation Studies

Table 6 quantifies design choices.

Findings: The largest factor facilitating domain adaptation is the fine-tuning (+4.0). Model size (+6.3) is worth the cost of computationality to be legal. Complex legal sentences require long context (+1.9). Optimal decoding with beam search (+1.3). N-gram penalty +(0.8) does not allow repetition in legal formulae. Conservative LR (+1.4) maintains the knowledge of multilingualism. Distillation (+0.7) helps to improve performance by refined representations.

5.4 Computational Requirements

NVIDIA T4 GPU (16GB): Training time of about 2.7 hours (prematurely cut off at epoch 12). Memory 14.2GB with FP16. Inferring rate of about 850 sentences per minute. Test set of 5,000 sentences took approximately 6 minutes. Model 2.6GB. Single T4 deployment was made possible with FP16.

6 Analysis and Discussion

6.1 Qualitative Analysis through Translation Examples

To better understand our model's performance, we conducted detailed manual analysis of translation outputs. Table 7 presents representative examples showcasing both strengths and limitations.

The examples reveal several important patterns. Our model excels at translating standard legal terminology and complex sentence structures, particularly for common legal procedures and statutory references. The preservation of numerical data (24 hours, Section 73, 1872) is consistently perfect. However, challenges remain with Latin legal terms (*quantum meruit*) and specialized legal concepts that require cultural adaptation rather than direct transliteration.

The error analysis reveals that rare terminology translation constitutes the largest challenge (48% of errors), particularly for Latin legal terms and specialized doctrinal concepts. This suggests that incorporating legal glossaries and terminology databases could significantly improve performance. Pronoun reference errors (20%) often occur in complex sentences with multiple actors, indicating limitations in contextual understanding across long dependencies.

6.2 Performance Across Legal Domains

We further analyzed model performance across different legal domains present in the dataset. Constitutional law translations showed the highest accuracy at 81% due to standardized terminology and frequent repetition of consti-

Model	Params	BLEU	ROUGE	CHRF
Helsinki-OPUS (Without Training)	77M	24.0	50.2	51.3
Helsinki-OPUS + Fine-tuning	77M	48.0	72.1	68.9
facebook/nllb200distilled600M (Without Training)	600M	40.0	60.8	61.7
NLLB-600M + Fine-tuning	600M	43.2	65.5	.3
Facebook/NLLB-1.3B (without training)	1.3B	45.0	71.4	70.1
Facebook/NLLB-1.3B + Fine-tuning	1.3B	50.1	73.1	69.4
Facebook/NLLB-1.3B-distilled (without training)	**1.3B**	**51**	**75**	**70.5**
Facebook/NLLB-1.3B-distilled + Fine-tuning	**1.3B**	**52.1**	**75.6**	**70.9**
ai4bharat/indictrans2-en-indic-1B + Fine-tuning	1.0B	44.0	68.6	62.8
law-ai/InLegalTrans + Fine-tuning	1.0B	48.1	68.2	66.5

Table 4: Results of validation. The highest BLEU is attained by our optimized Facebook/NLLB-1.3B distillation. The distilled version consistently outperforms the standard version because of improved generalization and optimized knowledge transfer.

Rank	Team	BLEU	METEOR	TER	BERTScore	AutoRank
1	**Team-SVNIT**	**51.61**	**75.80**	**37.09**	**92.61**	**61.62**
2	FourCorners	50.19	69.54	42.32	92.70	60.31
3	goodmen	48.56	67.15	41.63	92.38	59.39

Table 5: Official JUST-NLP 2025 leaderboard (top 5). CHRF: 73.29, COMET: 76.36. Our superior AutoRank stems from balanced excellence across all metrics, particularly METEOR (+6.26) indicating better semantic alignment.

Configuration	BLEU	Δ
Full System	**52.1**	—
w/o Fine-tuning	43.2	-4.0
w/ NLLB-600M	47.2	-6.3
Max Length = 256	51.6	-1.9
Beam Width = 1	52.2	-1.3
No n-gram penalty	52.7	-0.8
LR = 5e-5	52.1	-1.4
Standard NLLB-1.3B	52.1	-0.7

Table 6: Ablation results. The distilled variant provides +0.7 BLEU improvement over standard version.

tutional concepts. Criminal law followed next with 76% accuracy, with challenges mainly in procedural terminology. Translation of civil procedures achieved an accuracy of 73%, with difficulties in complex procedural sequences. Administrative regulations showed the lowest accuracy, at 69%, because of highly specialized domain-specific vocabulary. This domain-wise variation can be explained by the fact that our model outperformed specialized legal models since broad multilingual pretraining gives the NLLB-1.3B robust foundational representations that adapt very well across domains, unlike specialized models, which overfit to their training domains.

6.3 Comparative Advantage of Distilled Models

The reasons that justify the superior performance of the distilled NLLB-1.3B variant, +0.7 BLEU over the standard variant, are many. First, knowledge distillation during pretraining definitely forces the model to learn more generalized representations rather than memorizing training patterns. Second, distilled models show better calibration and reduced overconfidence, which is an essential requirement for legal translation in representing uncertainty. Third, the distillation process seems to enhance cross-lingual transfer efficiency, which is particularly helpful in the case of English-Hindi legal translation because of the limited amount of available parallel data.

6.4 Practical Implications and Deployment Considerations

Our findings have significant practical consequences for legal translation workflows. The low TER score of 37.09 indicates that post-editing effort would be considerably reduced, allowing translator productivity to increase 2-3×. Numerical data and citations are perfectly preserved, which eliminates critical risks in legal documentation. The 40 to 45 % error rate confirms that human review is still important

English Source	Hindi Translation
the appellant is acquitted.	अपीलार्थी को बरी किया जाता है
they also raised memorials on the merits and the preliminary habit.	उन्होंने गुणावगुणों तथा प्रारंभिक अभ्यापत्ति पर भी सम्परीक्षण किये थे ।
being aggrieved by the order dated 2nd March , 2012 made by the learned Single Judge in CWJC No.3653 of 2012 , the writ petitioner has filed this appeal under clause 10 of the Letters Patent .	CWJC सं ० 3653 वर्ष 2012 में विद्वान एकल न्यायाधीश द्वारा किये गये दिनांक 2 मार्च, 2012 के आदेश से व्यथित होकर, रिट याची ने लेटर्स पेटेंट के खंड 10 के अधीन यह अपील दाखिल किया है ।
6- The opposition no.2 has filed his presence in this Court by filing the right in favour of his learned counsel , though he has not filed any counter affidavit .	6 - विपक्षी संख्या 2 ने अपने विद्वान अधिवक्ता के पक्ष में अधिकार दाखिल करके इस न्यायालय में अपनी उपस्थिति दर्ज करायी है, यद्यपि उसने कोई प्रति शपथ पत्र दाखिल नहीं किया है ।
7- We have heard the counsel for the learned Principal Additional Advocate General, Muzaffarpur Properties Private Limited, Smt. Shahida Hassan and the counsels of various dignitaries who have filed applications in these appeals both on facts as well as on law.	7 - हमने विद्वान प्रधान महाधिवक्ता , मुजफ्फरपुर गुण प्राईवेट लिमिटेड , श्रीमती शाधा हसन के अधिवक्ताओं को सुना है जिन्होंने दोनों तथ्यों तथा तथ्यों पर भी तथा विधि पर भी इन अपीलों में आवेदन दाखिल किये हैं ।
the aforesaid case was of the Central Excise Act and section 35H of the Central Excise Act provided that an appeal and reference should be made to the High Court within 180 days from the date of communication of the judgment of the order.	पूर्वोक्त मामला केन्द्रीय उत्पाद अधिनियम की केंद्रीय उत्पाद अधिनियम एवं धारा 35 का था यह प्रावधान करता है कि आदेश के निर्णय की संसूचना की तिथि से 180 दिनों के भीतर उच्च न्यायालय को एक अपील एवं निर्देश दिया जाना चाहिए ।

Table 7: English-Hindi translation examples demonstrating model performance.

for legally binding documents.

The model size is moderate at 2.6GB, with computational demands for deployment in-house, which addresses all data confidentiality concerns usually associated with legal practice. The relatively short training time of under 3 hours allows organizations to fine-tune the model on their specific legal sub-domains, such as patent law or corporate contracts.

7 Conclusion

We presented Team-SVNIT's winning system for JUST-NLP 2025 Legal MT, achieving 1st place (AutoRank 61.62). Our approach demonstrates that carefully fine-tuned distilled multilingual models (NLLB-1.3B-distilled) outperform both smaller models and domain-specific systems with adequate training data (50K pairs) and systematic optimization.

Key contributions: **(1)** Comprehensive comparison identifying NLLB-1.3B-distilled as optimal (+0.7 BLEU over standard); **(2)** Enhanced preprocessing for noisy legal documents; **(3)** Optimized training with extended context (512), conservative LR (2e-5), large batches (512), and cosine scheduling; **(4)** Comprehensive ablations quantifying impacts; **(5)** Detailed error analysis revealing strengths (94% legal term accuracy, perfect numerical fidelity) and challenges (12% rare term errors, code-switching inconsistencies).

Results validate that domain adaptation through fine-tuning remains essential (+4.0 BLEU). NLLB-1.3B-distilled's success over domain-specific InLegalTrans suggests pretraining scale, architectural sophistication, and distillation may be more important than domain-specific pretraining when adequate fine-tuning data enables effective adaptation.

For practical workflows, our low edit rate (TER: 37.09) and perfect numerical accuracy make MTPE viable, though human review remains essential. Future research should address rare terminology through lexical constraints, code-switching through explicit guidelines, very long sentences through hierarchical approaches, multi-reference evaluation for accurate assessment, and document-level translation for improved consistency.

References

Marta R Costa-jussà, James Cross, Onur Çelebi, et al. 2022. No language left behind. *arXiv:2207.04672*.

Marta R Costa-jussà, James Cross, et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 626:11–18.

Jay Gala, Pranjal A Chitale, et al. 2023. Indictrans2: Towards high-quality machine translation for all 22 scheduled indian languages. *arXiv:2305.16307*.

Gowtham Ramesh, Sumanth Doddapaneni, et al. 2021. Samanantar: Parallel corpora for 11 indic languages. *arXiv:2104.05596*.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient nllb-200. In *Proc. ACL 2023*, pages 8031–8050.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. ACL 2002*, pages 311–318.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation. In *Proc. ACL Workshop*, pages 65–72.

Matthew Snover, Bonnie Dorr, et al. 2006. A study of translation edit rate. In *Proc. AMTA 2026*, pages 223–231.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proc. WMT 2017*, pages 612–618.

Tianyi Zhang, Varsha Kishore, et al. 2019. Bertscore: Evaluating text generation with bert. *arXiv:1904.09675*.

Ricardo Rei, Craig Stewart, et al. 2020. Comet: A neural framework for mt evaluation. In *Proc. EMNLP 2020*, pages 2685–2702.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. In *Proc. ICLR 2017*.

Argyri Panezi and John O'Shea. 2023. Risks of legal translation with MT and AI. In *Workshop on Gen AI and Law, ICML 2023*.

Catherine Way. 2016. Challenges of legal translation in 21st century. *Intl. Journal of Communication*, 10:1009–1035.

Abdel Rahman Altakhaineh. 2025. Human vs. AI translation of legal documents into Arabic. *Intl. J. Language and Law*, 14(1):23–47.

Natalie Llop. 2025. AI, machine translation, and access to justice. Stanford Law School.

Rahul Raja and Arpita Vats. 2025. Parallel corpora for low-resource indic languages. *arXiv:2503.04797*.

Princeton Legal Journal. 2025. Man v. machine: Legal implications of MT. *Princeton J. Public Law*.

Ashwini R Nair et al. 2024. Investigating translation for indic languages with BLOOMZ. *Scientific Reports*, 14:17843.

Diptesh Suman et al. 2023. Machine translation for low-resource indic languages. In *Proc. WMT 2023*, pages 856–864.

Saranya K Sheshadri and KP Soman. 2023. Neural MT for indic languages: A survey. *Procedia CS*, 218:23–35.

P Naveen et al. 2024. Challenges of MT for low-resource languages. *iScience*, 27(9):110685.

Linjing Han et al. 2024. Neural machine translation of clinical text. *BMC Med. Inform. Decision Making*, 24:58.