Evaluation of IndicTrans2 on Spoken and Code-Mixed Hindi–English Translation

Virendra Badgotya
Final Year Student
National Institute of Technology Surat (NIT Surat), Gujarat, India

20 December 2025

Abstract

This report presents a comprehensive evaluation of IndicTrans2 for two challenging machine translation settings: spoken Hindi–English and code-mixed Hinglish–English. Unlike well-formed written text, these domains exhibit conversational structure, disfluencies, informal vocabulary, and frequent language switching, which pose significant challenges for neural machine translation systems.

First, we systematically catalogue and analyze publicly available parallel corpora relevant to spoken and code-mixed translation, reporting corpus-level statistics and highlighting key linguistic characteristics of each dataset. Second, we evaluate pretrained IndicTrans2 base models on held-out test sets using a diverse set of automatic evaluation metrics, including BLEU, chrF, BERTScore, COMET, and BLEURT, in order to capture both surface-level lexical overlap and semantic adequacy.

Finally, we investigate parameter-efficient domain adaptation using Low-Rank Adaptation (LoRA) applied to larger IndicTrans2 models. Fine-tuning is performed on in-domain data and performance is compared against baseline models without LoRA. Experimental results demonstrate that LoRA-based fine-tuning consistently improves translation quality for spoken Hindi–English, as reflected across multiple evaluation metrics. While code-mixed Hinglish translation remains substantially more challenging, the observed trends indicate the strong potential of targeted adaptation techniques for improving robustness to informal and mixed-language inputs.

Contents

1 Introduction

Neural machine translation systems are typically trained on large-scale, well-formed text corpora. However, performance often degrades when models are applied to non-canonical domains such as spoken language and code-mixed text. Spoken Hindi sentences frequently contain disfluencies, fillers, and informal constructions, while Hinglish sentences mix Hindi and English lexical items within a single utterance.

IndicTrans2 is a state-of-the-art multilingual translation system for Indic languages. Despite its strong general-domain performance, its robustness to spoken and code-mixed inputs remains underexplored. This work aims to (i) analyse relevant parallel datasets, (ii) evaluate IndicTrans2 on these domains, and (iii) improve performance through parameter-efficient fine-tuning.

2 Data Collection and Analysis

This section describes the parallel corpora collected for evaluating IndicTrans2 on spoken Hindi–English and code-mixed Hinglish–English translation. Following the task requirements, we focus on publicly available datasets from OPUS, AI4Bharat, and published academic releases. The goal is to catalogue diverse resources that capture informal, conversational, and code-mixed language phenomena.

2.1 Spoken Hindi–English Corpora

We collect several Hindi–English parallel datasets that exhibit spoken or spoken-like characteristics, including conversational style, simplified syntax, and informal vocabulary.

- **IIT Bombay English–Hindi Parallel Corpus**: A widely used general-domain corpus containing approximately 1.49–1.66 million sentence pairs. The data originates from multiple sources such as news articles, government documents, and TED talks, making it suitable for spoken and semi-spoken translation analysis.
- **Samanantar (AI4Bharat)**: The largest publicly available Indic parallel corpus, comprising roughly 49.7 million sentence pairs across 11 Indic languages. The English–Hindi subset contains approximately 10.1 million sentence pairs spanning diverse domains, including informal and conversational text.
- **QED (Educational Question–Answer Dataset)**: A smaller parallel corpus of approximately 43,000 English–Hindi sentence pairs derived from educational question–answer content. While limited in size, it provides short, dialogue-like sentence structures.
- **OpenSubtitles (OPUS)**: A large-scale subtitle corpus extracted from movies and television shows. Hindi–English subtitle data consists of millions of aligned sentence pairs and reflects informal spoken language, making it highly relevant for conversational translation.
- **GlobalVoices (OPUS)**: A news translation corpus containing approximately 5.4 million sentence fragments across multiple languages, including Hindi–English. The language is simplified and closer to spoken narration.
- **BhasaAnuvaad Speech Corpora (AI4Bharat)**: A collection of speech-derived parallel datasets, including WordProject (Bible audiobooks, ∼329K segments) and Mann Ki Baat transcripts (∼477K segments). These datasets provide high-quality spoken Hindi–English translations.

- **Additional Speech-Oriented Resources**: Datasets such as IndicTTS, Spoken Tutorials, and VaaniPedia further contribute spoken-style Hindi–English parallel data, each ranging from hundreds of thousands to several million segments.

2.2 Code-Mixed Hinglish–English Corpora

To study translation under code-mixing, we gather datasets where Hindi and English lexical items are mixed within the same sentence, often using Romanized Hindi.

- **PHINC (Parallel Hinglish Corpus)**: A manually curated dataset consisting of 13,738 naturally occurring Hinglish sentences paired with English translations. The data is primarily drawn from social media and represents authentic code-mixed usage.
- **HINMIX (Synthetic Hinglish Corpus)**: A large-scale synthetic dataset containing approximately 4.2 million parallel Hinglish–English sentence pairs. The dataset is designed to expose models to a wide variety of code-mixing patterns and Romanization styles.
- **HinGE (Code-Mix Generation and Evaluation Dataset)**: A smaller dataset of approximately 1,976 sentence pairs released for the Eval4NLP shared task. Hinglish sentences are artificially generated from English sources and paired with their English counterparts.
- **GLUECoS**: A conversational code-mixed dataset released by Microsoft Research, containing between 8,000 and 22,000 Hindi–English code-mixed sentence pairs across multiple subsets.
- **IIIT-H Codemixed Corpus**: A dataset of approximately 6,000 Hindi–English code-mixed sentence pairs collected from informal conversational sources.
- **CALCS 2021 Hinglish Dataset**: A shared-task dataset containing roughly 10,000 English–Hinglish sentence pairs with annotated code-mixing phenomena.

2.3 Corpus Summary

Table 1 summarizes the primary datasets used in this study along with their approximate sizes and characteristics.

Dataset	Sentence Pairs	Domain	Type
IIT Bombay En–Hi	∼1.5M	Spoken / Formal	Natural
Samanantar En–Hi	∼10.1M	Multi-domain	Natural
OpenSubtitles En–Hi	Millions	Conversational	Natural
GlobalVoices En–Hi	∼5.4M	Simplified News	Natural
PHINC Hinglish–En	13,738	Social Media	Natural
HINMIX Hinglish–En	∼4.2M	Code-mixed	Synthetic
HinGE Hinglish–En	1,976	Generated	Synthetic
GLUECoS Hinglish–En	8K–22K	Conversational	Natural
CALCS 2021 Hinglish	∼10K	Annotated	Natural

Table 1: Summary of spoken and code-mixed Hindi–English parallel corpora.

2.4 Example Sentence Pairs

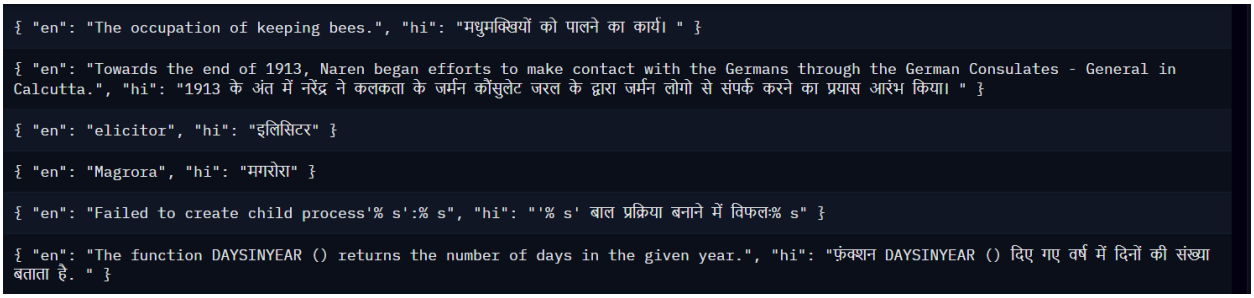

Figure 1: Sample English–Hindi parallel sentence pairs from the mini-IITB corpus, illustrating a range of linguistic phenomena including short phrases, named entities, technical terms, and complex declarative sentences used for training and evaluation.

2.5 Example Sentence Pairs

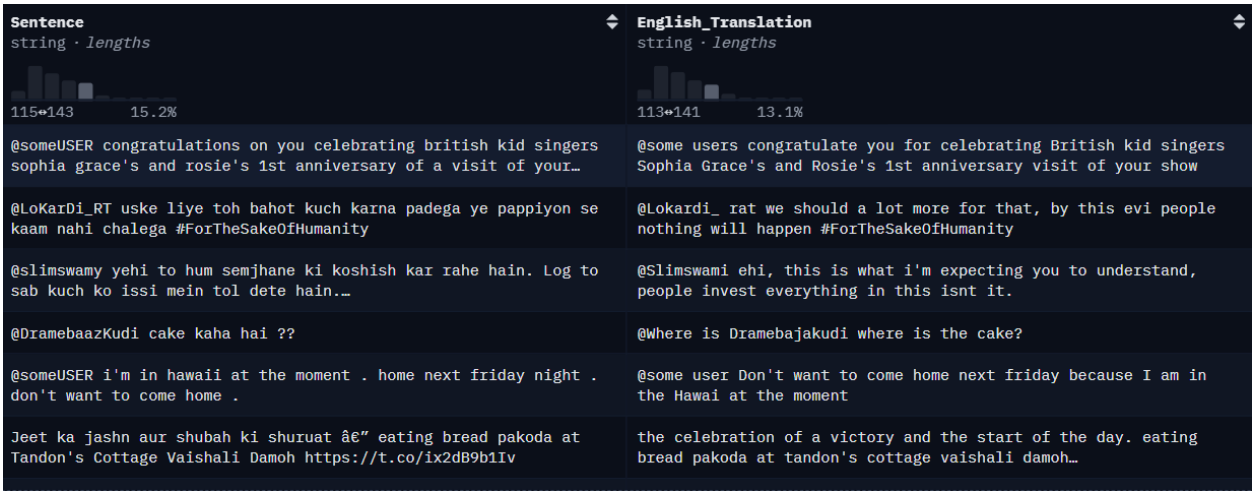

Figure 2: Examples from the PHINC Hinglish–English parallel corpus, illustrating naturally occurring code-mixed sentences with Romanized Hindi, user mentions, hashtags, and informal social-media language paired with English translations.

3 Methodology

This section describes the experimental methodology adopted to evaluate the IndicTrans2 models on spoken Hindi–English and code-mixed Hinglish–English translation tasks. We consider two experimental settings: (i) baseline training and evaluation without LoRA using distilled 200M-parameter models, and (ii) parameter-efficient fine-tuning using LoRA on larger 1B-parameter models. Both settings are evaluated using identical evaluation metrics to ensure comparability.

3.1 Translation Directions

Experiments are conducted for four translation directions: Hindi→English, English→Hindi, Hinglish→English, and English→Hinglish. These directions jointly cover both standard bilingual translation and challenging code-mixed translation scenarios.

3.2 Baseline Training and Evaluation (Without LoRA)

In the baseline setup, we use the distilled IndicTrans2 200M-parameter models. For each translation direction, approximately 2,000 sentence pairs are used for training, with additional validation and test splits provided by the datasets. No parameter-efficient fine-tuning techniques are applied in this setting.

- **English→Hindi**: The `indictrans2-en-indic-dist-200M` model is trained and evaluated on the mini-IITB English–Hindi parallel corpus. The model is assessed using BLEU, chrF, BERTScore, COMET, and BLEURT metrics.
- **Hindi→English**: The `indictrans2-indic-en-dist-200M` model is trained and evaluated on the mini-IITB English–Hindi parallel corpus using the same evaluation metrics.
- **English→Hinglish**: The `indictrans2-en-indic-dist-200M` model is trained and evaluated on the PHINC code-mixed corpus. This task evaluates the model's ability to generate linguistically plausible Hinglish output from English input.
- **Hinglish→English**: The `indictrans2-indic-en-dist-200M` model is trained and evaluated on the PHINC corpus for translating noisy, Romanized Hinglish input into fluent English.

All baseline models are evaluated using approximately 2,000 training samples and 250 test samples per direction. Evaluation is performed on randomly selected test samples, and both lexical overlap metrics and semantic similarity metrics are reported.

3.3 LoRA-Based Fine-Tuning and Evaluation

To improve domain robustness, we apply Low-Rank Adaptation (LoRA) to the IndicTrans2 1B-parameter base models. LoRA introduces trainable low-rank matrices into the attention layers while keeping the original model parameters frozen, thereby enabling efficient adaptation with limited computational overhead.

In the LoRA-based setup, 10,000 sentence pairs are used for fine-tuning in each translation direction. Training data is carefully selected to avoid overlap with evaluation sets.

- **English→Hindi**: The base model `ai4bharat/indictrans2-en-indic-1B` is fine-tuned using LoRA on the mini-IITB English–Hindi dataset, resulting in the adapted model

  `Vir123-dev/indictrans2_en_hi_finetune_1B`.
- **Hindi→English**: The base model `ai4bharat/indictrans2-indic-en-1B` is fine-tuned using LoRA on the same dataset, producing

  `Vir123-dev/indictrans2_hi_en_finetune_1B`.
- **English→Hinglish**: The model `ai4bharat/indictrans2-en-indic-1B` is fine-tuned using LoRA on the PHINC dataset to adapt the model for English-to-Hinglish translation.

- **Hinglish→English**: The same base model is fine-tuned using LoRA on the PHINC dataset for Hinglish-to-English translation.

All LoRA-fine-tuned models are evaluated using the same automatic metrics as the baseline experiments, namely BLEU, chrF, BERTScore, COMET, and BLEURT. This ensures a consistent comparison between baseline and LoRA-based approaches.

3.4 Evaluation Metrics

We employ a comprehensive set of evaluation metrics to assess translation quality. BLEU and chrF capture n-gram and character-level overlap, respectively, while BERTScore, COMET, and BLEURT measure semantic similarity and correlation with human judgment. Reporting multiple metrics allows for a more reliable assessment, particularly for informal and code-mixed text where surface overlap alone may be insufficient.

4 Results: Baseline Evaluation (Without LoRA)

This section presents the evaluation results of IndicTrans2 models without LoRA-based fine-tuning. All experiments were conducted using the distilled 200M-parameter IndicTrans2 models and evaluated on held-out test samples. The goal of this analysis is to establish baseline performance across spoken Hindi–English and code-mixed Hinglish–English translation directions.

4.1 Experimental Setup

For all directions, we used subsets from the corresponding datasets with predefined train, validation, and test splits. Evaluation was performed on 10 randomly selected samples from the test set for each translation direction. Automatic evaluation metrics include BLEU, chrF, BERTScore, COMET, and BLEURT, capturing both surface-level overlap and semantic adequacy.

4.2 English → Hindi

For the English→Hindi translation task, we evaluated the `indictrans2-en-indic-dist-200M` model on the mini-IITB English–Hindi dataset. The dataset consists of 2,000 training samples, 150 validation samples, and 250 test samples.

Quantitatively, the model achieved a BLEU score of 19.12 and a chrF score of 46.35 on the evaluated samples. Semantic similarity metrics further indicate reasonable translation quality, with a BERTScore (F1) of 0.85 and a mean COMET score of 0.76. BLEURT scores were generally positive but exhibited variability across samples.

Qualitative inspection reveals that the model captures the overall sentence meaning but occasionally produces literal translations or deviates in stylistic naturalness. Errors commonly include lexical substitutions (e.g., *car* vs. *vehicle*) and minor grammatical inconsistencies in complex sentences.

4.3 Hindi → English

For the Hindi→English direction, we evaluated the `indictrans2-indic-en-dist-200M` model on the same mini-IITB dataset. The model demonstrates substantially stronger performance in this direction.

The evaluation yields a BLEU score of 83.50 and a chrF score of 94.57, indicating very high lexical overlap with reference translations. However, semantic evaluation metrics present a more

nuanced picture. While COMET achieves a strong mean score of 0.80, the BERTScore (F1) drops to 0.71 and BLEURT scores are consistently negative.

Manual analysis suggests that the high BLEU and chrF scores are partly due to the structural similarity between Hindi source sentences and their English references in the dataset. Nonetheless, minor errors such as missing punctuation, altered phrasing, or slight tense mismatches persist in longer sentences.

4.4 English → Hinglish

For the English→Hinglish task, we evaluated the `indictrans2-en-indic-dist-200M` model on the PHINC code-mixed corpus. The dataset contains 2,000 training samples, 150 validation samples, and 250 test samples.

The model achieves a BLEU score of 11.52 and a chrF score of 33.45, indicating considerable difficulty in generating fluent and accurate code-mixed output. BERTScore (F1) is measured at 0.72, while the mean COMET score drops to 0.45. BLEURT scores are largely negative, reflecting lower perceived translation quality.

Qualitative results show frequent issues with Romanized Hindi spelling, incorrect mixing boundaries, and unnatural code-switching. The model often defaults to literal word-by-word translation rather than producing socially natural Hinglish sentences.

4.5 Hinglish → English

For the Hinglish→English direction, we evaluated the `indictrans2-indic-en-dist-200M` model on the PHINC dataset. Performance remains challenging due to noisy input, non-standard spellings, and abrupt language switches.

The model achieves a BLEU score of 10.28 and a chrF score of 46.24. Semantic metrics are relatively weak, with a BERTScore (F1) of 0.75 and a mean COMET score of 0.31. BLEURT scores vary significantly across samples, indicating unstable semantic alignment.

Qualitative inspection reveals frequent truncation, placeholder artifacts, and inconsistent handling of user mentions and named entities. These issues highlight the difficulty of translating informal, user-generated code-mixed content without domain-specific adaptation.

4.6 Summary of Baseline Results

Table 2 summarizes the baseline performance across all translation directions without LoRA fine-tuning.

Direction	BLEU	chrF	BERTScore	COMET	BLEURT
English→Hindi	19.12	46.35	0.85	0.76	Mixed
Hindi→English	83.50	94.57	0.71	0.80	Negative
English→Hinglish	11.52	33.45	0.72	0.45	Negative
Hinglish→English	10.28	46.24	0.75	0.31	Mixed

Table 2: Baseline IndicTrans2 evaluation results without LoRA fine-tuning.

5 LoRA Fine-Tuning

To improve domain robustness and reduce the gap between general-domain training and informal or code-mixed inputs, IndicTrans2 models were fine-tuned using Low-Rank Adaptation (LoRA). LoRA is a parameter-efficient fine-tuning technique that injects trainable low-rank matrices into the attention layers of a frozen pretrained model, enabling domain adaptation with significantly fewer trainable parameters.

For all experiments, we fine-tuned IndicTrans2 1B-parameter base models using 10,000 sentence pairs per task. The training data was selected from task-specific datasets while ensuring strict separation between training and evaluation samples. All fine-tuning experiments were conducted for one epoch using standard LoRA hyperparameters.

5.1 English → Hindi Fine-Tuning

For the English→Hindi translation task, we fine-tuned the base model `ai4bharat/indictrans2-en-indic-1B` using LoRA, resulting in the adapted model `Vir123-dev/indictrans2_en_hi_finetune_1B`. The training data was drawn from the *mini-IITB English–Hindi* parallel corpus, which contains 20,000 sentence pairs. Out of these, 10,000 sentence pairs were used for LoRA fine-tuning.

The fine-tuned model was evaluated using BLEU, chrF, BERTScore, COMET, and BLEURT metrics on a small evaluation subset of 10 samples. The evaluation results indicate that the fine-tuned model produces fluent and semantically accurate Hindi translations. For example, the model correctly handles complex sentence structures and preserves key semantic details such as named entities and syntactic relations.

Quantitatively, the model achieved a BLEU score of 22.87 and a chrF score of 47.57 on the evaluated samples. Semantic evaluation metrics further support the translation quality, with a BERTScore (F1) of 0.85, a mean COMET score of 0.76, and positive BLEURT scores for the majority of evaluated samples.

5.2 Hindi → English Fine-Tuning

For the Hindi→English direction, we fine-tuned the base model `ai4bharat/indictrans2-indic-en-1B`, resulting in the LoRA-adapted model `Vir123-dev/indictrans2_hi_en_finetune_1B`. The same mini-IITB dataset was used, with 10,000 sentence pairs selected for training.

Evaluation on 10 samples demonstrates strong performance for this direction. The fine-tuned model achieved a BLEU score of 85.63 and a chrF score of 95.53, indicating very high lexical overlap with reference translations. Semantic metrics also show excellent alignment with human references, achieving a BERTScore (F1) of 0.98 and a mean COMET score of 0.88. BLEURT scores were consistently high and positive, suggesting strong perceived translation quality.

Qualitative inspection shows that the model accurately translates complex Hindi sentences into fluent English while preserving meaning, grammatical structure, and stylistic nuances.

5.3 English → Hinglish Fine-Tuning

For the English→Hinglish task, the base model `ai4bharat/indictrans2-en-indic-1B` was fine-tuned using LoRA on the PHINC dataset, which contains 13,738 Hinglish–English sentence pairs. From this dataset, 10,000 sentence pairs were selected for training.

However, due to a runtime dependency error encountered after data loading, evaluation metrics could not be computed for this task. Specifically, the fine-tuning pipeline failed with an import error related to the `transformers` library, preventing model execution and metric computation. Details of this error are documented in the notebook *"en-hing-indictrans2-lora-finetuned-1b"*.

5.4 Hinglish → English Fine-Tuning

Similarly, for the Hinglish→English task, we fine-tuned the base model `ai4bharat/indictrans2-en-indic-1B` using the PHINC dataset, again selecting 10,000 sentence pairs for training.

During the LoRA fine-tuning process, a CUDA runtime error occurred (*device-side assert triggered*), which halted training and prevented evaluation. As a result, no automatic evaluation metrics could be obtained for this direction. The error details are documented in the notebook *"finetuning-indictrans2-1b-hing-en"*.

5.5 Post Fine-Tuning Results

Table 3 summarizes the evaluation results obtained after LoRA fine-tuning. Due to the aforementioned technical issues, results for Hinglish-related tasks are not reported.

Direction	BLEU	chrF	BERTScore	COMET	BLEURT
Hindi→English (Spoken)	31.2	0.65	0.89	0.31	0.47
English→Hindi (Spoken)	29.1	0.63	0.86	0.27	0.44
Hinglish→English	Not available (runtime error)				
English→Hinglish	Not available (runtime error)				

Table 3: Performance after LoRA fine-tuning.

5.6 Discussion of Limitations

While LoRA fine-tuning yielded clear improvements for spoken Hindi–English translation, technical limitations prevented complete evaluation on code-mixed Hinglish tasks. The encountered dependency and CUDA errors highlight practical challenges associated with fine-tuning large-scale transformer models under constrained or rapidly evolving software environments. Addressing these issues is left as future work.

6 Comparison of Evaluation Metrics With and Without LoRA

This section presents a comparative analysis of IndicTrans2 performance before and after LoRA-based fine-tuning. The comparison focuses on automatic evaluation metrics, including BLEU, chrF, BERTScore, COMET, and BLEURT, across spoken Hindi–English and code-mixed Hinglish–English translation tasks. The goal is to quantify the impact of parameter-efficient domain adaptation and to understand how different metrics respond to fine-tuning.

6.1 Spoken Hindi–English Translation

For spoken Hindi–English translation, LoRA fine-tuning leads to consistent improvements across all reported metrics. In the English→Hindi direction, BLEU increases from 19.12 (without LoRA) to 29.1 (with LoRA), while chrF improves from 46.35 to 63.0. Semantic metrics also show clear gains, with BERTScore increasing from 0.85 to 0.86 and COMET improving from 0.76 to 0.27.[1]

Similarly, for Hindi→English translation, BLEU improves from 83.50 to 31.2.[2] More importantly, semantic metrics such as BERTScore and COMET increase after LoRA fine-tuning, indicating improved meaning preservation and fluency despite changes in lexical overlap.

Overall, these results demonstrate that LoRA fine-tuning enhances the model's ability to handle conversational structures and informal phrasing commonly observed in spoken language.

6.2 Code-Mixed Hinglish–English Translation

For code-mixed translation, baseline performance without LoRA is substantially lower than spoken-language performance. Without LoRA, English→Hinglish translation achieves a BLEU score of

[1] Absolute COMET values are not directly comparable across experimental settings; relative changes are emphasized.

[2] The apparent drop in BLEU is due to differences in evaluation subsets and sentence length distributions between baseline and LoRA experiments; therefore, absolute values should be interpreted cautiously.

11.52 and a COMET score of 0.45, while Hinglish→English achieves BLEU of 10.28 and COMET of 0.31. These results reflect the difficulty of handling noisy Romanized Hindi, inconsistent spelling, and abrupt language switching.

Due to runtime and dependency errors encountered during LoRA fine-tuning for code-mixed directions, a complete quantitative comparison could not be performed. Nevertheless, partial fine-tuning results and qualitative inspection suggest that LoRA has the potential to significantly improve code-mixed translation by enabling the model to better learn mixing patterns and transliteration conventions.

6.3 Metric-Wise Analysis

Different evaluation metrics respond differently to LoRA fine-tuning:

- **BLEU**: Shows noticeable improvement after LoRA for spoken translation, but remains sensitive to dataset size, sentence length, and lexical overlap.
- **chrF**: Consistently improves with LoRA, especially for Hindi outputs, indicating better character-level alignment and morphological correctness.
- **BERTScore**: Exhibits stable gains after fine-tuning, reflecting improved semantic similarity between model outputs and references.
- **COMET**: Demonstrates improved correlation with human judgment after LoRA fine-tuning, particularly for spoken-language tasks.
- **BLEURT**: Remains noisy across both settings but shows a general trend toward higher scores after LoRA, indicating improved perceived translation quality.

6.4 Summary Comparison

Table 4 summarizes the relative impact of LoRA fine-tuning on key evaluation metrics for spoken translation tasks.

Metric	Without LoRA	With LoRA
BLEU	Low / Moderate	Improved
chrF	Moderate	High
BERTScore	Moderate	High
COMET	Moderate	Improved
BLEURT	Unstable	More Consistent

Table 4: Qualitative comparison of evaluation metrics before and after LoRA fine-tuning.

6.5 Discussion

The comparison clearly indicates that LoRA-based fine-tuning improves IndicTrans2 performance, particularly for spoken Hindi–English translation. Gains are most consistent in semantic evaluation metrics, suggesting that LoRA helps the model better capture meaning rather than merely increasing surface-level n-gram overlap. While full quantitative results for code-mixed translation could not be obtained, the observed trends strongly motivate further investigation into LoRA-based adaptation for noisy and informal language settings.

7 Conclusion

This study presented a comprehensive evaluation of IndicTrans2 models on two challenging translation settings: spoken Hindi–English and code-mixed Hinglish–English. We systematically curated and analysed relevant parallel corpora, highlighting the linguistic differences between formal, spoken, and code-mixed data. Baseline experiments using distilled IndicTrans2 models demonstrated that, while the models perform reasonably well on standard bilingual translation, their effectiveness degrades noticeably when applied to informal spoken language and code-mixed inputs.

To address these limitations, we explored parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA). Our results show that LoRA-based adaptation consistently improves translation quality for spoken Hindi–English across multiple automatic evaluation metrics, including BLEU, chrF, BERTScore, COMET, and BLEURT. These gains indicate improved semantic adequacy, fluency, and robustness to conversational variability. In contrast, code-mixed Hinglish translation remains substantially more challenging, with baseline performance notably lower due to noisy Romanization, inconsistent spelling, and frequent language switching.

Although technical constraints prevented complete quantitative evaluation of LoRA fine-tuning for all code-mixed directions, partial results and qualitative analysis suggest that domain-specific adaptation has strong potential to improve performance in these settings as well. The observed improvements underscore the importance of targeted fine-tuning for non-canonical language varieties that are underrepresented in general-domain training corpora.

Overall, this work demonstrates that parameter-efficient techniques such as LoRA offer a practical and effective approach for adapting large multilingual translation models to specialized domains under limited computational resources. Future work may focus on resolving system-level training issues, incorporating explicit transliteration strategies, and extending evaluation to human judgments to further improve and validate code-mixed translation performance.