# When Gujarati Meets English: Bridging the Code-Mixed Translation Gap for Gujarati-English

**Anonymous ACL submission**

## Abstract

Code-mixing — the blending of multiple languages within a single utterance — is widespread in multilingual societies such as India. While substantial progress has been made for Hinglish (Hindi–English), other low-resource code-mixed variants like Gujlish (Gujarati–English) remain largely unexplored, primarily due to the scarcity of parallel corpora. We address this gap with three contributions. First, we construct the first general-domain Gujlish–English parallel corpus of 30,000 sentence pairs, generated from the BPCC corpus via GPT-4o and validated by six bilingual annotators (Krippendorff's $\alpha$ = 0.79–0.83). Second, we additionally release a gold-standard set of 500 human-annotated Gujlish–English pairs reflecting naturally occurring code-mixing, addressing the synthetic-data evaluation gap. Third, we structurally validate that our synthetic corpus matches naturally occurring code-mixing patterns: CMI and SPF distributions of our generated data lie within the range established by human-authored Hinglish (HINMIX) and Gujlish references, with differences below 0.6 SPF points. We fine-tune NLLB-200 on this corpus, establishing strong Gujlish→English baselines. Across three test sets (XNLI, IN22, and our Gujlish-real), our model outperforms six baselines — including Google Translate, Devnagri, IndicTrans2, mBART, mT5, and base NLLB-200 — by +16 to +27 BLEU points. Human evaluators preferred our translations in 68.64% (XNLI) and 63.14% (IN22) of cases, with substantial inter-annotator agreement (Fleiss' $\kappa$ = 0.62–0.66).

## 1 Introduction

Code-mixing and code-switching are natural linguistic phenomena in multilingual societies, where speakers fluidly combine lexical items and syntactic structures from multiple languages during everyday communication (Thara and Poornachandran, 2018). In India, this practice is particularly prevalent across digital platforms such as WhatsApp, YouTube comments, and social media posts, where English is frequently interwoven with regional languages. One such variant is Gujlish, a hybrid form of Gujarati and English written in Roman script. Unlike standard Gujarati written in its native script, Gujlish emerges in informal online communication, making it a low-resource and underexplored variety in natural language processing (NLP).

While multilingual neural machine translation (NMT) models have achieved remarkable success across high-resource language pairs (Wang et al., 2024), their performance drops significantly in code-mixed and low-resource settings (Kartik et al., 2024; Huzaifah et al., 2024; Vavre et al., 2022). Widely used commercial tools like Google Translate are often unable to handle Gujlish adequately, primarily due to the lack of curated training data and the orthographic variations that arise when speakers write Gujarati words phonetically in the Roman alphabet. This shortcoming limits the accessibility of translation technologies in domains such as education, healthcare, and digital inclusion, where Gujlish serves as a practical medium of communication.

In this work, we aim to address this gap by making two primary contributions:

1. **Dataset Development:** We construct the first large-scale parallel corpus for Gujlish–English translation, covering diverse domains. This resource provides a foundational benchmark for developing and evaluating translation systems on Gujlish, which has thus far lacked any standardized dataset.

2. **Model Fine-Tuning and Evaluation:** We fine-tune the NLLB-200 multilingual translation model (Team et al., 2022) on the curated Gujlish–English corpus, establishing

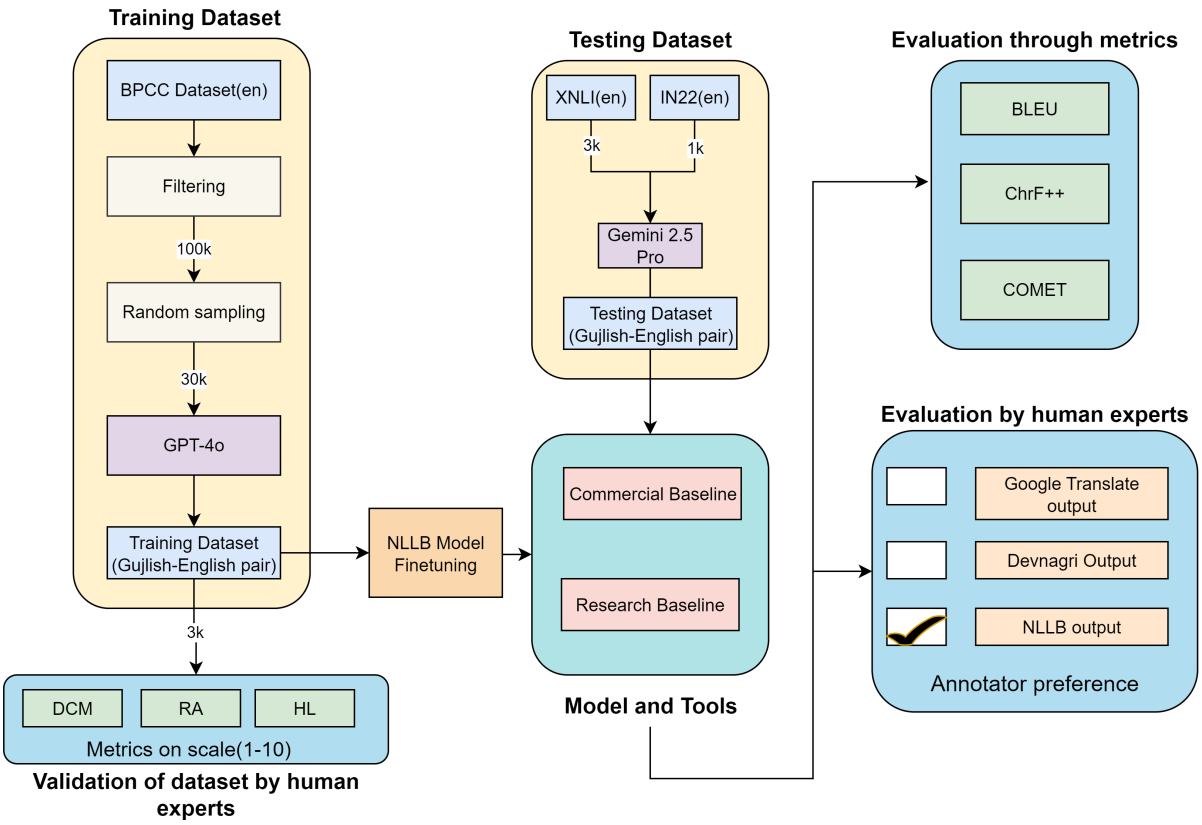


Figure 1: Overall system architecture illustrating dataset construction, model fine-tuning, evaluation pipeline, and human validation framework for Gujlish-English translation.

strong baselines for Gujlish→English translation. We evaluate the model on the XNLI (Conneau et al., 2018) and IN22[1] datasets and compare its performance against six baselines including Google Translate, Devnagri[2], IndicTrans2, mBART, mT5, and base NLLB-200 demonstrating substantial improvements across BLEU (Papineni et al., 2002), ChrF++ (Huzaifah et al., 2024), and COMET (Rei et al., 2020) scores.

By combining large-scale dataset construction with systematic model evaluation, our work provides both a practical resource and a methodological framework for extending machine translation research to other underrepresented code-mixed languages.

The remainder of this paper is organized as follows. Section 2 reviews related work on code-mixed translation and synthetic data generation. Section 3 details our dataset construction, fine-tuning setup, and structural validation methodology. Section 4 presents quantitative and qualitative results, followed by error analysis. Section 5 concludes with limitations and future directions.

## 2 Literature Review

Research on code-mixed language processing has accelerated in recent years, reflecting the increasing prevalence of multilingual communication in digital platforms. A significant portion of work in this domain has focused on Hinglish (Hindi–English code-mixing) (Srivastava and Singh, 2021a; Gahoi et al., 2022; Khan et al., 2022; Gupta et al., 2023), where challenges such as language identification (Ansari et al., 2021), Named Entity Recognition (NER) (Singh et al., 2018), and sentiment analysis (Singh and Lefever, 2020) have been documented. To address these, studies have employed synthetic corpus generation (Srivastava and Singh, 2020), transliteration-based preprocessing, and domain-specific fine-tuning strategies.

Multilingual pretrained models like mBART (Liu et al., 2020) and mT5 (Xue et al., 2021) have shown promise for code-mixed inputs when fine-tuned (Gautam et al., 2021) due to their ability to generalize across multiple languages via subword tokenization. However, they have been shown to

[1] https://huggingface.co/datasets/ai4bharat/IN22-Gen

[2] https://devnagri.com/

struggle in the presence of Romanized orthography and frequent language switching. Other works have explored transfer learning (Tatariya et al., 2023) from high-resource to low-resource languages.

Parallel corpus construction remains a critical bottleneck for code-mixed research. Existing efforts have relied on crowd-sourced annotations, social media conversations, and synthetic augmentation (Pei et al., 2025) techniques to generate parallel data. For example, datasets for Hinglish have been curated from online platforms and extended through lexical substitution or transliteration to enable robust evaluation benchmarks (Makhija et al., 2020). These resources have been instrumental in advancing machine translation, classification, and sentiment analysis for Hinglish. Recently, however, there has been growing interest in leveraging large language models (LLMs) to generate synthetic code-mixed data. (Yong et al., 2023; Zeng, 2024; de Gibert et al., 2025) demonstrates that carefully prompted LLMs can produce natural and diverse code-mixed sentences that augment scarce human-annotated corpora. (de Gibert et al., 2025) further shows that LLM-generated synthetic parallel data can significantly improve low-resource MT systems when combined with human verification, outperforming earlier augmentation methods.

Despite such progress, low-resource code-mixed variants such as Gujlish (Gujarati–English) remain largely absent from the literature. The only available resource is a small standardized dataset of specific domain consisting of roughly 6,733 examples, created with human linguistic effort and limited to a single domain—restaurant conversations (Banerjee et al., 2018). Beyond this, general-purpose systems like Google Translate perform poorly, as they fail to capture the orthographic irregularities and lexical nuances of Romanized Gujarati or similar code-mixed languages. This gap is particularly significant, given that millions of speakers use such languages in daily communication but still lack tailored NLP resources.

## 3 Methodology

In this section, we describe the complete workflow for developing the Gujlish–English machine translation system. The methodology comprises five major components: (i) dataset creation, (ii) fine-tuning of a multilingual baseline model, (iii) evaluation setup, and (iv) performance assessment through statistical and human evaluation measures. (v) structural validation of synthetic code-mixing.

**System Overview:** Figure 1 illustrates the overall pipeline for our Gujlish–English machine translation system. The architecture comprises three stages. First, English sentences from the Bharat Parallel Corpus Collection (BPCC) corpus are filtered, sampled, and translated into Gujlish via GPT-4o to construct the training corpus, with linguistic quality validated by human annotators along Degre of Code-Mixing (DCM), RA (Readability), and HL (Human-likeness) dimensions. Second, evaluation benchmarks (XNLI and IN22) are adapted into Gujlish using Gemini 2.5 Pro a different model from training to avoid source model bias and the fine-tuned NLLB-200 is benchmarked against six baselines. Finally, translation quality is assessed through both automatic metrics (BLEU, ChrF++, COMET) and a human preference study. Each component is described in detail in the subsections below.

### 3.1 Dataset Creation

We begin with the BPCC dataset released by AI4Bharat, which provides high-quality parallel corpora across several Indian languages. This dataset was selected due to its diverse, multi-domain coverage, ensuring that the resulting Gujlish corpus represents a variety of contexts, beyond news or conversational text, suitable for robust translation modeling.

To maintain linguistic diversity, sentences were filtered by length: 188[3] datapoints in the 0–5 word range, 50k between 5–10 words, 25k between 10–15 words, and another 25k between 15–20 words. From this pool of roughly 100k English sentences, we uniformly sampled 30k sentences across length categories. These were translated into Gujlish using the GPT-4o model, guided by a few-shot prompting strategy that included exemplar translations to encourage natural code-mixing and preserve contextual meaning. The specific prompting setup used for generating our Gujlish training dataset is illustrated in Appendix A.1. To complement the synthetic corpus, we also constructed a gold-standard dataset Gujlish (real), of 500 Gujlish–English sentence pairs using professional bilingual annotators proficient in Gujarati and English, aged between 25–35 years. Two annotators were presented with randomly sampled English sentences from BPCC and asked to write

[3] only 188 were available in the 0-5 words range

the Gujlish version that appeared as the most natural spontaneous rendering upon reading the sentence. Since no standardized Romanization scheme exists for Gujarati, annotators were allowed to transliterate Gujarati words according to their own phonological intuition, resulting in natural orthographic variation that reflects real-world Gujlish usage. An illustrative example of English-Gujlish pair is shown in Figure 5 in Appendix A.2.

### 3.2 Model: NLLB-200 Fine-Tuning

For translation modeling, we employ NLLB-200 with eng-Latn tag, a massively multilingual neural machine translation system supporting 200 languages, for the task of Gujlish→English translation. NLLB's encoder–decoder transformer architecture is particularly suited for low-resource scenarios due to its cross-lingual transfer capabilities and subword-level generalization.

We fine-tuned the NLLB-200 on the Gujlish–English parallel corpus using a sequence-to-sequence learning setup. Since NLLB-200 does not include a dedicated language tag for Gujlish, we used eng-Latn as the source language tag, leveraging the shared Latin script between Romanized Gujarati and English. Training was performed using the AdamW optimizer with a learning rate of 5e-5, weight decay of 0.01, and a batch size of 8 for 5 epochs, with mixed-precision (fp16) enabled. The model was evaluated after each epoch on a held-out 20% validation split. Full hyperparameter details are provided in Appendix A.6. Fine-tuning enables the model to adapt to the Romanized Gujarati tokens and the syntactic irregularities characteristic of Gujlish, which general-purpose multilingual models typically fail to capture.

### 3.3 Evaluation Method

To assess the linguistic quality and authenticity of our constructed Gujlish–English dataset, we conducted a human evaluation following the protocol of (Kartik et al., 2024). A randomly selected subset of 3,000 sentence pairs was reviewed by six bilingual annotators proficient in both Gujarati and English in a blind evaluation setting (annotators were not informed whether the sentences were human generated or GPT-4o generated), focusing on evaluating the Gujlish side of the corpus. Each sentence was rated along three dimensions: DCM, RA and HL, on a scale of 1–10. These ratings provide an intrinsic measure of the dataset's linguistic naturalness and representativeness of real-world code-mixed usage (Srivastava and Singh, 2021b).

To validate annotation reliability, we computed Krippendorff's $\alpha$ (Krippendorff, 2013) with ordinal distance weighting across all six annotators, yielding $\alpha$ = 0.81 (DCM), 0.79 (RA), and 0.83 (HL), indicating substantial inter-annotator agreement (Artstein and Poesio, 2008).

To the best of our knowledge, no existing benchmarks specifically target Gujlish translation. Therefore, we adapted two multilingual evaluation datasets, i.e XNLI and IN22 for this purpose. To construct their Gujlish counterparts, we randomly sampled 3,000 English sentences from XNLI and 1,000 from IN22, and translated them into Gujlish using the Gemini 2.5 Pro model. Since our training corpus was generated using GPT-4o, employing a different model for evaluation data preparation helps mitigate potential source-model bias.

Previous work (Gupta et al., 2024) has also shown that Gemini-based models effectively produce natural and contextually coherent code-mixed text, ensuring high-quality and unbiased evaluation data. Hence, in this work, XNLI and IN22 refer to these code-mixed Gujlish–English variants created for evaluation.

We evaluate our fine-tuned model against six baselines spanning commercial and research systems. Among commercial tools, we include Google Translate, a strong general-purpose multilingual system, and Devnagri API, a translation service specifically tailored for Indian languages. For research baselines, we evaluate IndicTrans2 (Gala et al., 2023), a state-of-the-art multilingual model for Indian language translation; mBART and mT5, two widely adopted sequence-to-sequence pretrained models; and the base NLLB-200 (Team et al., 2022) without fine-tuning, which serves as a direct ablation isolating the contribution of our curated corpus. This combination of commercial and research baselines provides a comprehensive and fair evaluation across diverse system types.

### 3.4 Evaluation Measures and Human Evaluation

For quantitative evaluation, we report three widely adopted evaluation measures: BLEU , ChrF++ , and COMET. These measures collectively capture n-gram overlap, character-level similarity, and semantic adequacy.

To complement automatic metrics, we conducted human evaluations to assess translation fluency and adequacy. From the full test sets, we randomly

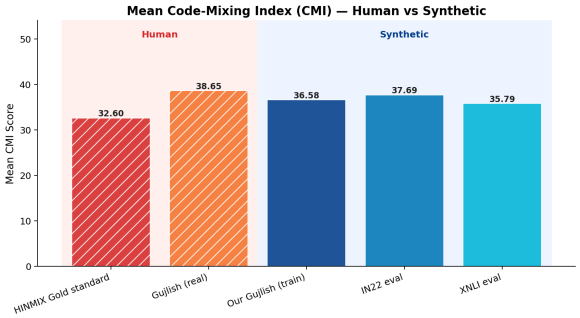


Figure 2: Mean Code-Mixing Index (CMI) across human and synthetic corpora. Synthetic data closely mirrors the CMI range of real code-mixed datasets, confirming linguistic authenticity of our corpus.

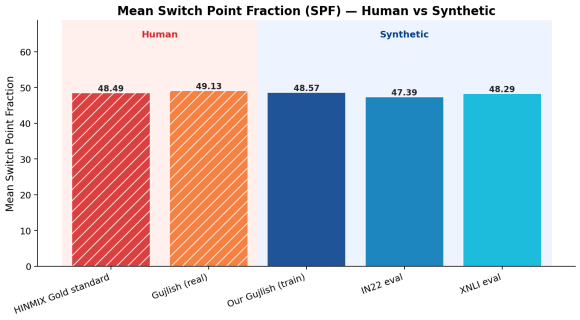


Figure 3: Mean Switch Point Fraction (SPF) across human and synthetic corpora. Hatched bars denote human-authored data; solid bars denote LLM-generated data.

selected a subset of 1,000 (XNLI) and 500 (IN22) sentences for human preference evaluation, and asked three bilingual annotators proficient in both Gujarati and English (different from dataset evaluators) to compare outputs from our fine-tuned NLLB model, Google Translate, and Devnagri whose results are shown in Table 2. Annotators followed detailed instructions on evaluating linguistic naturalness and meaning preservation, as described in Appendix A.3, to ensure consistency in their judgments.

To assess annotation reliability, we computed Fleiss' $\kappa$ (Fleiss, 1971) across the three annotators, yielding $\kappa = 0.62$ (XNLI) and $\kappa = 0.66$ (IN22), indicating moderate to substantial agreement (Landis and Koch, 1977). Additionally, 71% of XNLI and 74% of IN22 sentences received unanimous 3/3 agreement, with the remainder resolved by 2/3 majority, confirming that preference margins are not driven by a single annotator.

Together, these evaluations provide both surface-level and semantic insights into model performance, ensuring a robust understanding of translation quality and generalization.

### 3.5 Structural Validation of Synthetic Code-Mixing via CMI and SPF

A central concern in LLM-generated code-mixed corpora is whether the resulting text exhibits the structural properties of *naturally occurring* code-mixing, or whether it produces degenerate patterns (e.g., excessive language dominance or unnaturally uniform switching). To address this, we conduct a quantitative linguistic analysis of our synthetic datasets using two established metrics: the Code-Mixing Index (CMI) (Das and Gambäck, 2014) and the Switch Point Fraction (SPF) (Gupta et al., 2020). We acknowledge that CMI and SPF have known limitations (Srivastava and Singh, 2021a), particularly their surface-level nature. We use them as structural sanity checks rather than absolute quality indicators, complemented by human evaluation and downstream MT performance for comprehensive validation.

**Metrics:** CMI measures the degree of inter-language mixing within a sentence, computed as the fraction of tokens belonging to the non-matrix language normalized by sentence length. SPF measures switching frequency specifically, the proportion of token boundaries at which a language switch occurs providing a complementary view of code-mixing density. Together, these metrics characterize both the *extent* and the *rhythm* of code-mixing in a corpus.

**Datasets:** We compute CMI and SPF over five corpora spanning both human-authored and synthetically generated data (Figure 2 and 3). For human-authored reference, we use: (i) the HINMIX gold-standard corpus (Kartik et al., 2024), a benchmark dataset of 2,500 real Hinglish sentence pairs annotated by native speakers, and (ii) our own **Gujlish (real)** dataset of 500 sentence pairs produced by professional bilingual annotators (described in Section 3.1). For synthetic data, we evaluate: (iii) our GPT-4o generated training corpus (*Our Gujlish train*), (iv) the Gemini 2.5 Pro generated IN22 evaluation set, and (v) the Gemini 2.5 Pro generated XNLI evaluation set.

**Results and Interpretation:** As shown in Figures 2 and 3, the synthetic corpora exhibit CMI and SPF values that are closely aligned with both human-authored reference corpora. Specifically, the SPF of our training corpus (48.57) is nearly identical to that of the real Gujlish corpus (49.13) and the HINMIX gold standard (48.49), with a dif-

ference of less than 0.6 points. Similarly, the CMI of our training corpus (36.58) closely tracks the real Gujlish corpus (38.65), and all three synthetic sets fall within the range established by the two human corpora.

Crucially, this alignment holds across *both* a different language pair (Hinglish vs. Gujlish) and *both* generation models (GPT-4o and Gemini 2.5 Pro), indicating that the structural properties of natural code-mixing are reliably reproduced by LLM-based generation under our prompting strategy. This provides strong evidence that our synthetic corpora are not merely fluent paraphrases with token substitution, but exhibit the same inter-sentential switching patterns and language balance as real code-mixed speech.

## 4 Results and Discussion

For the dataset quality evaluation, the average scores across all annotators were: DCM = 8.66, RA = 8.98, and HL = 9.01, indicating that the generated Gujlish text is both natural and highly readable. These results confirm the quality and authenticity of the dataset for downstream translation modeling.

We evaluate our fine-tuned NLLB-200 model against six baselines — Google Translate, Devnagri API, IndicTrans2, mBART, mT5, and the base NLLB-200 model on the XNLI and IN22 datasets. Evaluation is conducted using three standard automatic measures: BLEU, ChrF++, and COMET (specifically the `wmt22-comet-da` checkpoint (Rei et al., 2022), a reference-based direct assessment model trained on human judgments).

Google Translate, though fluent and widely trained, performs inconsistently on low-resource Indian and code-mixed text. Devnagri API handles Indian languages better through translation memory and human feedback but still struggles to generalize to informal, code-mixed contexts. IndicTrans2 and mBART, despite being strong multilingual models for standard Indian languages, fail to handle the Romanized orthography of Gujlish effectively. mT5 performs worst among all baselines, likely due to its tokenization being poorly suited for informal code-mixed text without task-specific adaptation. Sample outputs from Google Translate and Devnagri are shown in Appendix A.5, highlighting their tendency to produce monolingual or literal translations and their struggle with code-mixed contexts

### 4.1 Performance Insights

| Dataset | Model | BLEU ↑ | ChrF++ ↑ | COMET ↑ |
|---|---|---|---|---|
| XNLI | Google Translate | 28.74 | 54.64 | 0.70 |
| | Devnagri | 27.62 | 56.86 | 0.10 |
| | IndicTrans2 | 10.06 | 38.26 | 0.51 |
| | mBART | 9.73 | 36.07 | 0.49 |
| | NLLB (Base) | 11.2 | 38.79 | 0.54 |
| | mT5 | 0.63 | 16.6 | 0.37 |
| | **NLLB (Ours)** | **41.52** | **64.09** | **0.79** |
| IN22 | Google Translate | 38.39 | 64.78 | 0.74 |
| | Devnagri | 44.59 | 70.54 | 0.46 |
| | IndicTrans2 | 19.38 | 51.98 | 0.55 |
| | mBART | 19.39 | 52.01 | 0.54 |
| | NLLB (Base) | 25.16 | 54.13 | 0.65 |
| | mT5 | 1.21 | 22.66 | 0.34 |
| | **NLLB (Ours)** | **62.2** | **79.46** | **0.88** |

Table 1: Comparison of various models on XNLI and IN22 datasets. Arrows (↑) indicate higher is better.

| Dataset | NLLB | Google Translate | Devnagri |
|---|---|---|---|
| XNLI | **68.64%** | 7.34% | 24.02% |
| IN22 | **63.14%** | 10.45% | 26.41% |

Table 2: Human preference study results showing the percentage of sentences for which each system's translation was preferred by annotators.

| Dataset | Fleiss' $\kappa$ | 3/3 Unanimous | 2/3 Split |
|---|---|---|---|
| XNLI | 0.62 | 71% | 29% |
| IN22 | 0.66 | 74% | 26% |

Table 3: Inter-annotator agreement (Fleiss' $\kappa$) and vote distribution for the human preference study.

| Dataset | Model | BLEU ↑ | ChrF++ ↑ | COMET ↑ |
|---|---|---|---|---|
| Gujlish(real) | Google Translate | 42.34 | 65.99 | 0.77 |
| | Devnagri | 45.31 | 69.09 | 0.82 |
| | IndicTrans2 | 9.05 | 44.59 | 0.51 |
| | mBART | 9.18 | 42.68 | 0.51 |
| | NLLB (Base) | 13.39 | 45.68 | 0.58 |
| | mT5 | 0.58 | 20.02 | 0.35 |
| | **NLLB (Ours)** | **61.47** | **79.12** | **0.88** |

Table 4: Comparison of various models on our Gujlish(real) dataset. Arrows (↑) indicate higher is better.

Our fine-tuned NLLB-200 model achieves consistent and substantial improvements across all metrics and both evaluation datasets, outperforming every baseline by a significant margin (Table 1). Sample outputs from our fine-tuned NLLB model on XNLI and IN22 are shown in Appendix A.4, illustrating how the model captures lexical borrowing and intra-sentential code-mixing effectively.

Among commercial tools, Google Translate achieves moderate performance (BLEU: 28.74, ChrF++: 54.64 on XNLI), while Devnagri performs comparably (BLEU: 27.62, ChrF++: 56.86).

Our model surpasses Google Translate by +12.78 BLEU and +9.45 ChrF++ on XNLI, and by +23.81 BLEU and +14.68 ChrF++ on IN22 (Table 1). Among research baselines, IndicTrans2 and mBART achieve BLEU scores of only around 10 and 19 on XNLI and IN22 respectively, confirming that standard multilingual models do not transfer well to the Romanized, code-mixed register of Gujlish. Among all base models, NLLB-200 achieves the strongest baseline performance on both XNLI and IN22 (Table 1), confirming it as the most suitable choice for fine-tuning on our Gujlish–English corpus. A direct comparison between NLLB (Base) and NLLB (Ours) most cleanly isolates the contribution of our curated corpus the untuned model achieves only 11.2 BLEU on XNLI and 25.16 on IN22, whereas fine-tuning raises this to 41.52 and 62.2 respectively, a 3.7× and 2.5× improvement (Table 1). We did not include direct LLM zero-shot/few-shot baselines as their use of the same model class (GPT-4o) that generated our training data would conflate model capability with training-test overlap. Also comparing fine-tuned smaller models (NLLB-200, 600M parameters) against LLMs with 100+× parameters also introduces unfair compute comparison.

To focus human evaluation where discrimination is most meaningful, we restricted the preference study to the three highest-performing systems in Table 1: our fine-tuned NLLB-200, Google Translate, and Devnagri. The remaining baselines (IndicTrans2, mBART, mT5, and base NLLB-200) were excluded as their substantially lower automatic scores made meaningful human discrimination unlikely. These quantitative gains are further corroborated by human evaluation results (Table 2). Inter-annotator agreement for the preference study is reported in Table 3. Annotators preferred NLLB outputs in 68.64% of XNLI cases and 63.14% of IN22 cases, demonstrating that improvements in automatic metrics translate directly into perceived translation quality.

Table 4 presents results on our gold-standard Gujlish(real) dataset of 500 human-annotated sentence pairs. Our fine-tuned NLLB model achieves the highest scores across all three metrics (BLEU: 61.47, ChrF++: 79.12, COMET: 0.88),consistent with its performance on XNLI and IN22. Notably, commercial systems perform relatively stronger on this dataset compared to the synthetic benchmarks Devnagri achieves a BLEU of 45.31 and COMET of 0.82, outperforming Google Translate (BLEU: 42.34, COMET: 0.77). This can be attributed to the real-world nature of our Gujlish(real) sentences, which exhibit more naturalistic code-mixing patterns that partially align with the translation memory of commercial systems. Despite this, our fine-tuned model outperforms all baselines by a substantial margin (+16.16 BLEU over Devnagri), confirming that targeted fine-tuning on curated code-mixed data generalizes effectively to real annotator-produced inputs.

Generally higher scores on IN22 compared to XNLI, observed consistently across all models and metrics including Google Translate, Devnagri, IndicTrans2, mBART, mT5, NLLB (Base), and our fine-tuned NLLB can be attributed to the relative simplicity and contextual coherence of IN22 sentences (which tend to be longer and more contextually structured by nature of the IN22 benchmark), as illustrated by sample outputs from our model (see Appendix A.4). Notably, on IN22, commercial systems such as Google Translate and Devnagri also achieve comparatively stronger performance with Devnagri even surpassing Google Translate in BLEU (44.59 vs. 38.39) (Table 1) suggesting that IN22 sentences are more tractable for general-purpose translation systems. This stronger competition on IN22 is reflected in the human preference study (Table 2), where annotator preference for our fine-tuned NLLB model drops slightly from 68.64% (XNLI) to 63.14% (IN22), as the relatively higher quality of baseline outputs makes discrimination harder for human evaluators. To better contextualize the improvements, we also computed baseline BLEU scores between the raw Gujlish inputs and their corresponding English reference sentences: 10.45 for XNLI and 19.73 for IN22. These values represent the natural lexical overlap between the two languages, even without model translation. Any subsequent increase in BLEU achieved by the models therefore reflects genuine translation capability rather than inherent word similarity between Gujlish and English.

### 4.2 Error Analysis

To better understand the limitations of our fine-tuned NLLB-200 model and characterize the residual translation errors, we conducted a qualitative error analysis on 150 randomly sampled outputs (50 each from XNLI, IN22, and Gujlish-real) where the model's translation diverged from the reference. Three bilingual annotators (proficient in both Gujarati and English, distinct from those involved in

dataset and preference evaluation) independently categorized each error into one of four predefined categories, with disagreements resolved through majority voting. Inter-annotator agreement on error categorization was substantial (Fleiss' $\kappa = 0.64$).

**Error Categories and Distribution:** We identified four recurring error types: (i) Orthographic Misinterpretation (33%), where non-standard Romanization of Gujarati tokens (e.g., chhe vs. che, thase vs. thashe) led to incorrect lexical mapping; (ii) Lexical Substitution Errors (26%), where the model selected a semantically related but contextually incorrect English equivalent for a Gujarati token (e.g., translating kaam as "job" instead of "work" in domain-specific contexts); (iii) Named Entity Distortion (21%), where proper nouns embedded within Gujarati syntactic frames were either dropped or partially translated; (iv) Function Word Omission (20%), particularly affecting Gujarati postpositions (ne, thi, ma, e) which carry critical case information but lack direct English counterparts.

## 5 Conclusion

In this work, we addressed the challenge of machine translation for low-resource, code-mixed Indian languages by focusing on Gujlish. We curated a 30k Gujlish-English sentence pairs and fine-tuned the NLLB-200 multilingual model to establish the first Gujlish→English baseline dataset and model.

Comprehensive evaluations across XNLI and IN22 benchmarks demonstrate that our fine-tuned model consistently outperforms all baselines including Google Translate, Devnagri API, IndicTrans2, mBART, mT5, and the base NLLB-200 — in terms of BLEU, ChrF++, and COMET scores. Human preference studies further validate that the improvements are perceptually meaningful, with annotators overwhelmingly favoring our fine-tuned model's outputs. These findings confirm that targeted fine-tuning on well-curated code-mixed data can yield significant improvements in translation quality, even for underrepresented language varieties.

## Limitations

Despite the promising outcomes, our study has certain limitations. As the dataset was constructed using GPT-4o, the orthographic representation of Romanized Gujarati tokens may not fully capture all regional spelling variations, which future work incorporating native speaker contributions could address. Moreover, the NLLB-200 model, while effective, is computationally intensive, posing challenges for deployment in low-resource or real-time applications. We note that LLMs used in our pipeline (GPT-4o, Gemini 2.5 Pro) may have been exposed to BPCC corpus during pre-training. We did not perform formal contamination analysis but mitigate this concern through (a) BLEU comparison against raw input-reference overlap (showing genuine translation capability), and (b) evaluation on Gujlish-real, which uses human-authored Gujlish independent of LLM generation.

## Future Work

For future work, we plan to explore meta-learning frameworks to enable efficient cross-task adaptation and improved generalization from limited data. Further, we aim to expand the Gujlish corpus to include domain-specific and conversational contexts and to extend the proposed methodology to other low-resource code-mixed Indian languages. We believe that these efforts will advance the scalability, inclusivity, and real-world applicability of code-mixed low resource NLP systems.

## References


Mohd Zeeshan Ansari, M M Sufyan Beg, Tanvir Ahmad, Mohd Jazib Khan, and Ghazali Wasim. 2021. Language identification of hindi-english tweets using code-mixed BERT. *Preprint*, arXiv:2107.01202.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M. Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3766–3780, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*,

pages 378–387, Goa, India. NLP Association of India.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource MT via synthetic data generation with LLMs. *Preprint*, arXiv:2505.14423.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Akshat Gahoi, Jayant Duneja, Anshul Padhi, Shivam Mangale, Saransh Rajput, Tanvi Kamble, Dipti Sharma, and Vasudev Varma. 2022. Gui at MixMT 2022 : English-Hinglish : An MT approach for translation of code mixed data. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1126–1130, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. CoMeT: Towards code-mixed translation using parallel monolingual sentences. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 47–55, Online. Association for Computational Linguistics.

Ayushman Gupta, Akhil Bhogal, and Kripabandhu Ghosh. 2024. Code-mixer ya nahi: Novel approaches to measuring multilingual LLMs' code-mixing capabilities. *Preprint*, arXiv:2410.11079.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280, Online. Association for Computational Linguistics.

Rahul Gupta, Vivek Srivastava, and Mayank Singh. 2023. MUTANT: A multi-sentential code-mixed Hinglish dataset. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 744–753, Dubrovnik, Croatia. Association for Computational Linguistics.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating code-switching translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394, Torino, Italia. ELRA and ICCL.

Kartik Kartik, Sanjana Soni, Anoop Kunchukuttan, Tanmoy Chakraborty, and Md. Shad Akhtar. 2024. Synthetic data generation and joint learning for robust code-mixed translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15480–15492, Torino, Italia. ELRA and ICCL.

Abdul Khan, Hrishikesh Kanade, Girish Budhrani, Preet Jhanglani, and Jia Xu. 2022. SIT at MixMT 2022: Fluent translation built on giant pre-trained models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1136–1144, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*, 3rd edition. Sage Publications, Thousand Oaks, CA.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

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

Piyush Makhija, Ankit Kumar, and Anuj Gupta. 2020. hinglishnorm – a corpus of hindi-english code mixed sentences for text normalization. *Preprint*, arXiv:2010.08974.

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.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). 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*, pages 2685–2702.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. Language identification and named entity recognition in Hinglish code mixed tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 52–58, Melbourne, Australia. Association for Computational Linguistics.

Pranaydeep Singh and Els Lefever. 2020. Sentiment analysis for Hinglish code-mixed tweets by means of cross-lingual word embeddings. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 45–51, Marseille, France. European Language Resources Association.

Vivek Srivastava and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2021a. Challenges and limitations with the metrics measuring the complexity of code-mixed text. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 6–14, Online. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2021b. HinGE: A dataset for generation and evaluation of code-mixed Hinglish text. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 200–208, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kushal Tatariya, Heather Lent, and Miryam de Lhoneux. 2023. Transfer learning for code-mixed data: Do pretraining languages matter? In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 365–378, Toronto, Canada. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

S Thara and Prabaharan Poornachandran. 2018. Code-mixing: A brief survey. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2382–2388.

Aditya Vavre, Abhirut Gupta, and Sunita Sarawagi. 2022. Adapting multilingual models for code-mixed translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 7133–7141.

Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024. Findings of the WMT 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700, Miami, Florida, USA. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Linda Zeng. 2024. Leveraging large language models for code-mixed data augmentation in sentiment analysis. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 85–101, Miami, Florida, USA. Association for Computational Linguistics.

# A Appendix

## A.1 Prompt for Translating English to Gujlish using GPT-4o

To generate our Gujlish training dataset, we used GPT-4o with carefully designed translation instructions emphasizing natural code-mixing and context preservation. The prompt encouraged maintaining English keywords (especially named entities and domain terms) while blending Gujarati syntax and morphology to ensure grammatical fluency and cultural relevance.

(Figure 4) illustrates our few-shot prompting setup, where one translation example was included to guide GPT-4o toward producing more natural and coherent Gujlish outputs. Empirically, we observed that including this exemplar significantly improved the fluency and balance of code-mixing in the generated sentences.

## A.2 GPT-4o Generated example

An example English sentence and its corresponding Gujlish translation generated by the GPT-4o model are shown in (Figure 5), where blue tokens represent English words and red tokens denote Gujarati insertions, illustrating the natural code-mixing achieved through our prompting approach.

## A.3 Instruction Template for Human Annotators

To evaluate translation quality, human annotators were instructed to compare system-generated English translations of Gujlish inputs. Each Gujlish sentence was accompanied by three English outputs (A, B, and C) produced by our model, Google Translate, and Devnagri. Annotators were asked to read all translations carefully and select the one that best preserved the original meaning while maintaining natural English phrasing and fluency.

(Figure 6) shows the evaluation prompt emphasizing two key aspects: (1) semantic adequacy — ensuring that the translation accurately conveyed the meaning of the Gujlish input, and (2) linguistic naturalness — preferring outputs that sounded fluent and idiomatic in English. Each annotator was required to select only one option (A, B, or C) for each input sentence.

## A.4 Sample Outputs of Fine-tuned NLLB on XNLI and IN22

(Figure 7) shows sample outputs from our fine-tuned model on the XNLI and IN22 test sets. The examples demonstrate that our model captures both lexical borrowing and intra-sentential code-mixing effectively, while maintaining semantic fidelity to the original English sentence.

## A.5 Google Translate and Devnagri Outputs on XNLI and IN22

To benchmark translation quality, we also generated Gujlish translations from Google Translate (Figure 8) and Devnagri (Figure 9). Both systems predominantly produce monolingual Gujarati translations, lacking code-mixed structure and exhibiting literal word-for-word mapping. This demonstrates the need for a specialized Gujlish translation system.

## A.6 Training and Inference Hyperparameters

Table 5 summarizes the hyperparameters used for fine-tuning and inference of our NLLB-200 model.

| Hyperparameter | Value |
|---|---|
| Base model | `nllb-200-distilled-600M` |
| Source language tag (training) | `eng_Latn` |
| Target language tag (training) | `eng_Latn` |
| Source language tag (inference) | `eng_Latn` |
| Target language tag (inference) | `eng_Latn` |
| Optimizer | AdamW |
| Learning rate | $5 \times 10^{-5}$ |
| Weight decay | 0.01 |
| Train batch size | 8 |
| Eval batch size | 8 |
| Epochs | 5 |
| Max input length(tokens) | 64 |
| Max output length(tokens) | 64(train), 256(inference) |
| Decoding strategy | Beam search (num_beams = 5) |
| Precision | fp16 |
| Train / val split | 80/20 |
| seed | 42 |
| Eval strategy | Per epoch |

Table 5: Hyperparameters for fine-tuning and inference of the NLLB-200 model.

You are a translator. Convert English to code-mixed English-Gujarati. Follow natural code-switching patterns.

Examples:
- "What will the temperature be tomorrow?" → "Kaal no temperature shu hase?"
- "Is the rain going to die down today?" → "Aaj no rain thoda oochu thase ke nahi?"

Rules: Proper grammar, only English/Gujarati words, romanized output.
"""

Figure 4: Few-shot prompting for GPT-4o to generate Gujlish translations, showing the inclusion of one exemplar sentence to guide model output.

| | | |
|---|---|---|
| English (Input) | : | The remaining banks have reduced retail opening hours and pushed online banking. |
| Gujlish (Output) | : | Baaki na banks e retail opening hours ne reduce kari didha chhe ane online banking upar focus kari rahya chhe. |

Figure 5: Example of English and it's Gujlish output from GPT-4o model. The words in blue are pure English and words in red are Romanized Gujarati.

**Annotation Instructions**

**Task:** You are given a Gujlish (Gujarati–English mixed) input sentence and three English translations (A, B, C) generated by different systems. Read all three outputs carefully and choose the one that best preserves the meaning of the Gujlish sentence and is most fluent in English.

**Instruction:**

- Select only one translation (A, B, or C) per input.
- Focus solely on meaning preservation and naturalness of English phrasing.

---

**Gujlish :** Live-attenuated vaccines safe chhe ane ek majboot ane effective immune response stimulate kare chhe je lamba samay sudhi chale.

☐ A Live-attenuated vaccines are safe and stimulate a stronger and more effective immune response that lasts for a long time.

☐ B Live-attenuated vaccines are safe, effective and can stimulate an effective immune response.

☐ C Live-attenuated vaccines are clean and stimulating a strong and effective immune response that lasts longer.

Figure 6: Prompt provided to human annotators for evaluating translation outputs based on semantic adequacy and linguistic naturalness.

| Dataset | Gujlish | English |
|---|---|---|
| XNLI | Mane lagyu ke aa ek khub cute movie hati ane mane enjoy kari. | I thought this was a very cute movie and I enjoyed it. |
| | Postal Service na costing systems mujab, 1996 ma basic mail ni per-piece cost 26 hati. | According to the Postal Service's costing systems, the basic cost per-piece of mail in 1996 was 26. |
| | Location ane restaurant ni popularity ne lidhe, tane ek table reserve karvu j pade. | Due to the location and popularity of the restaurant, you should always reserve a table. |
| IN22 | A badhaj addictive chhe pan brain upar adverse effect pan kare chhe. | All this is addictive but it also has an adverse effect on the brain. |
| | Tethi Gotama ne bijo ek religious teacher shodhvo mushkil nathi je pahla karta vadhu competent hoy. | So Gotama had no difficulty in finding another religious teacher who was more competent than her predecessor. |
| | E writers ni social responsibility thai gayi ke loko ne nyay ane slavery thi freedom na vichar ma jagruti laava mate lakhvu. | Due to the location and popularity of the restaurant, you should always reserve a table. |

Figure 7: Sample outputs from our fine-tuned model on XNLI and IN22 datasets illustrating correct code-mixing and semantic fidelity.

| Dataset | Gujlish | English |
|---|---|---|
| XNLI | Mane lagyu ke aa ek khub cute movie hati ane mane enjoy kari. | I'm so happy to see you enjoying this cute movie. |
| | Postal Service na costing systems mujab, 1996 ma basic mail ni per-piece cost 26 hati. | Postal Service na costing systems mujab, 1996 ma basic mail ni per-piece cost 26 hati. |
| | Location ane restaurant ni popularity ne lidhe, tane ek table reserve karvu j pade. | The location of the restaurant is very popular, so it is difficult to reserve a table. |
| IN22 | A badhaj addictive chhe pan brain upar adverse effect pan kare chhe. | A badhaj addictive chhe pan brain upar adverse effect pan kare chhe. |
| | Tethi Gotama ne bijo ek religious teacher shodhvo mushkil nathi je pahla karta vadhu competent hoy. | Gotama, do not be afraid of a religious teacher, it is difficult to find a competent one. |
| | E writers ni social responsibility thai gayi ke loko ne nyay ane slavery thi freedom na vichar ma jagruti laava mate lakhvu. | Due to the location and popularity of the restaurant, you should always reserve a table. |

Figure 8: Google Translate outputs on XNLI and IN22, showing predominantly monolingual Gujarati translations lacking code-mixed structure.

| Dataset | Gujlish | English |
|---|---|---|
| XNLI | Mane lagyu ke aa ek khub cute movie hati ane mane enjoy kari. | I thought it was a very cute movie and I enjoyed it. |
| | Postal Service na costing systems mujab, 1996 ma basic mail ni per-piece cost 26 hati. | According to the Postal Service's costing systems, the per-piece cost of basic mail was 26 in 1996. |
| | Location ane restaurant ni popularity ne lidhe, tane ek table reserve karvu j pade. | Because of the location and the popularity of the restaurant, you have to reserve a table. |
| IN22 | A badhaj addictive chhe pan brain upar adverse effect pan kare chhe. | A is all additive but also has adversarial effects on the brain. |
| | Tethi Gotama ne bijo ek religious teacher shodhvo mushkil nathi je pahla karta vadhu competent hoy. | So, it is not difficult to find another consistent teacher in Gotham who is more competitive than the first one. |
| | E writers ni social responsibility thai gayi ke loko ne nyay ane slavery thi freedom na vichar ma jagruti laava mate lakhvu. | It became the social responsibility of the writers to write to the people to create awareness about the idea of justice and freedom from slavishness. |

Figure 9: Devnagri outputs on XNLI and IN22, highlighting literal word-for-word translation without proper code-mixing.