When Gujarati Meets English: Toward Robust Translation of Code-Mixed Low Resourced Indian Language

Mukund Agarwalla[1]*, **Himanshu Kumar**[1]*, **Nishat Afshan Ansari**[1]
[1] Indian Institute of Information Technology Nagpur
{ mukundagarwalla2002@gmail.com, himanshukumariiitn@gmail.com, nishat.ansari@iiitn.ac.in }

Abstract

Code-mixing, the practice of blending multiple languages within a single utterance, is a widespread linguistic phenomenon in multilingual societies such as India. While substantial progress has been made in machine translation for Hinglish (Hindi–English), other low-resource code-mixed variants like Gujlish (Gujarati–English) remain largely unexplored. Developing effective translation systems for such languages is challenging due to the scarcity of high-quality parallel corpora. To bridge this gap, we present the first large-scale, general-purpose Gujlish–English parallel corpus comprising approximately 30k sentence pairs. The dataset was curated from the BPCC corpus (AI4Bharat) and translated using GPT-4o, followed by human validation. We fine-tune the multilingual NLLB-200 model on this corpus to establish the first baselines for Gujlish→English translation. Evaluated on the XNLI and IN22 benchmarks, our model significantly outperforms Google Translate, achieving 1.5–2× improvements in BLEU and ChrF++ scores, and shifting COMET scores from near-zero to strongly positive. The code and dataset available at: https://github.com/mukund302002/Gujlish-English-Translation .

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 general-purpose 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 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 with state-of-the-art

*Equal contribution

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

tools such as Google Translate and Devnagri[2], demonstrating substantial improvements across BLEU, ChrF++, and COMET metrics.

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.

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), 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 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 benchmark is a small standardized dataset 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 striking, 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 four 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 metrics. Each component is explained in detail 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, general-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

[2] https://devnagri.com/

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

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. An illustrative example of English-Gujlish pair is shown in Figure 2 in Appendix A.2.

3.2 Model: NLLB-200 Fine-Tuning

For translation modeling, we employ NLLB-200, a massively multilingual neural machine translation system supporting 200 languages. 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 NLLB-200 on the Gujlish–English parallel corpus using a sequence-to-sequence learning setup with label smoothing and early stopping to prevent overfitting. 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. This adaptation proved critical, as zero-shot translation experiments with the base NLLB-200 model yielded poor lexical alignment and weak semantic consistency.

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 Srivastava and Singh (2021b). A randomly selected subset of 3,000 sentence pairs was reviewed by six expert annotators,[4] focusing on evaluating the Gujlish side of the corpus. Each sentence was rated along three dimensions: *Degree of Code-Mixing (DCM)*, *Readability (RA)*, and *Human-likeness (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.

To the best of our knowledge, no existing benchmarks specifically target Gujlish translation. Therefore, we adapted two multilingual evaluation datasets—XNLI and IN22—for this purpose. To construct their Gujlish counterparts, we randomly sampled 3,000 English sentences from XNLI and 800 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. Prior 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 compare our fine-tuned model against two widely used translation systems:

1. **Google Translate API**, a strong general-purpose multilingual baseline.

2. **Devnagri API**, a commercial translation tool specialized for Indian languages.

We did not identify any other publicly available systems capable of handling Gujlish inputs effectively. Including both commercial and research baselines ensures a comprehensive and fair performance comparison.

3.4 Statistical Metrics and Human Evaluation

For quantitative evaluation, we report three widely adopted metrics: BLEU, ChrF++ (Huzaifah et al., 2024) , and COMET. These metrics collectively measure n-gram overlap, character-level similarity, and semantic adequacy.

To complement automatic metrics, we conducted human evaluations to assess translation fluency and adequacy. We randomly selected 1,000 sentences from XNLI and 500 from IN22, and asked three human annotators (different from dataset evaluators) to compare outputs from our fine-tuned NLLB model, Google Translate, and Devnagri. Annotators followed detailed instructions on evaluating linguistic naturalness and meaning preservation, as described in Appendix A.3, to ensure consistency in their judgments.

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

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

[4] All annotators possess native or near-native proficiency in both Gujarati and English.

readable. These results confirm the quality and authenticity of the dataset for downstream translation modeling.

We evaluate our fine-tuned NLLB-200 model against two competitive baselines, Google Translate and Devnagri API, on the XNLI and IN22 datasets. Evaluation is conducted using three standard automatic metrics: BLEU, ChrF++, and COMET, along with human preference studies.

Google Translate, though fluent and widely trained, performs inconsistently on low-resource Indian and code-mixed text. In contrast, the Devnagri API handles Indian languages better through translation memory and human feedback but still struggles to generalize to informal, code-mixed contexts. Some 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	16.10	47.76	-0.54
	Devnagri	27.62	56.86	0.10
	NLLB (Ours)	**41.38**	**67.05**	**0.44**
IN22	Google Translate	34.87	65.21	0.09
	Devnagri	44.59	70.54	0.46
	NLLB (Ours)	**62.44**	**81.06**	**0.79**

Table 1: Comparison of fine-tuned NLLB model with Google Translate and Devnagri API 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.

Our fine-tuned model achieves consistent and substantial improvements across all metrics and datasets. 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. On XNLI, our fine-tuned NLLB yields a +25.28 BLEU and +19.29 ChrF++ gain over Google Translate, while achieving a positive COMET score (0.44) compared to Google's negative value (-0.54). On IN22, performance further improves, with BLEU reaching 62.44, nearly double that of Google Translate (34.87), and COMET rising to 0.79, reflecting superior semantic alignment (Table 1). These quantitative gains are corroborated by human evaluation results (Table 2). 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.

Interestingly, Devnagri performs better than Google Translate on both benchmarks, highlighting the benefits of domain adaptation for Indian languages. However, even Devnagri lags far behind fine-tuned NLLB, underscoring the effectiveness of fine-tuning multilingual models specifically on code-mixed data.

The generally higher scores on IN22 compared to XNLI can be attributed to the relative simplicity and contextual coherence of IN22 sentences, as illustrated by sample outputs from our model (see Appendix A.4). 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.

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 both Google Translate and Devnagri API 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 and Future Work

Despite the promising outcomes, our study has certain limitations. Minor inconsistencies might persist in the dataset, particularly in the orthography of Romanized Gujarati tokens generated by GPT-4o. Moreover, the NLLB-200 model, while effective, is computationally intensive, posing challenges for deployment in low-resource or real-time applications.

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 resourced 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.

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.

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.

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.

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.

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.

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

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

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.

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 1) 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 2), 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 3) 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 4) 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 5) and Devnagri (Figure 6). 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.

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

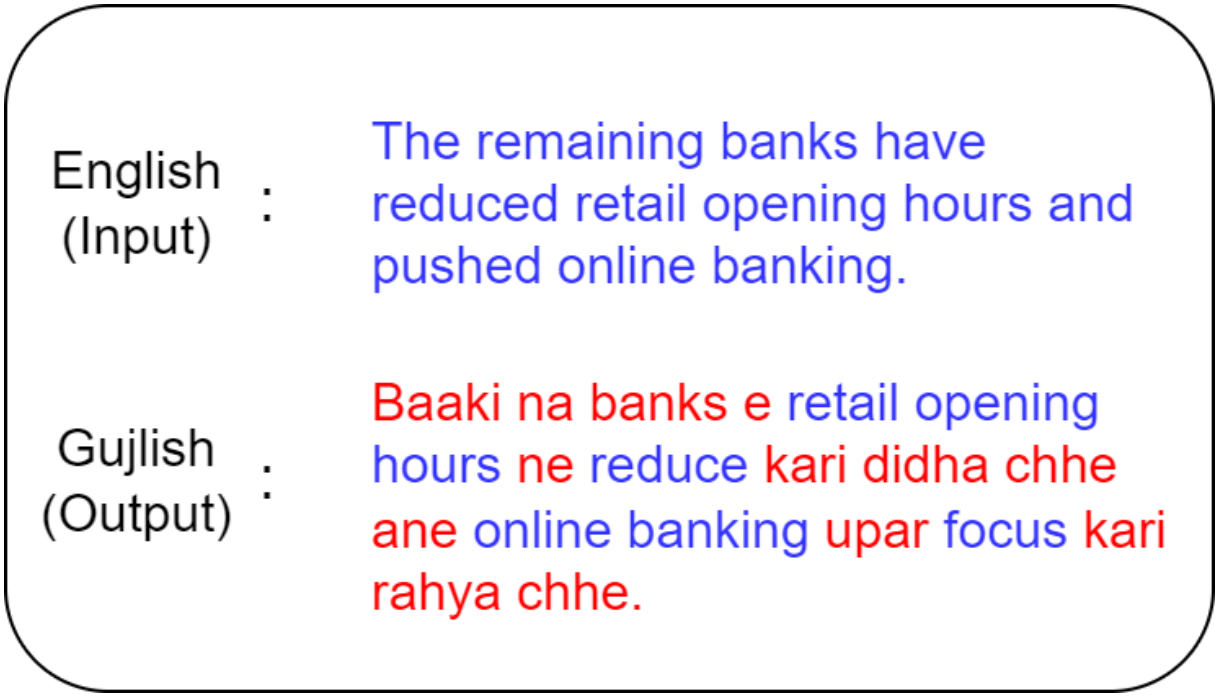

Figure 2: Example of English and it's Gujlish output from GPT-4o model. The words in blue are pure English and words in red are Romanized Gujarati.

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 3: 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 4: 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 5: 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 6: Devnagri outputs on XNLI and IN22, highlighting literal word-for-word translation without proper code-mixing.