Assessing and Improving Punctuation Robustness in English-Marathi Machine Translation

Kaustubh Shivshankar Shejole, Sourabh Deoghare and Pushpak Bhattacharyya
Computation for Indian Language Technology (CFILT)
Department of Computer Science and Engineering
Indian Institute of Technology Bombay, Mumbai, India
{kaustubhshejole, sourabhdeoghare, pb}@cse.iitb.ac.in

Abstract

Neural Machine Translation (NMT) systems rely heavily on explicit punctuation cues to resolve semantic ambiguities in a source sentence. Inputting user-generated sentences, which are likely to contain missing or incorrect punctuation, results in fluent but semantically disastrous translations. This work attempts to highlight and address the problem of punctuation robustness of NMT systems through an English-to-Marathi translation. First, we introduce ***Virām***, a human-curated diagnostic benchmark of 54 punctuation-ambiguous English-Marathi sentence pairs to stress-test existing NMT systems. Second, we evaluate two simple remediation strategies: cascade-based *restore-then-translate* and *direct fine-tuning*. Our experimental results and analysis demonstrate that both strategies yield substantial NMT performance improvements. Furthermore, we find that current Large Language Models (LLMs) exhibit relatively poorer robustness in translating such sentences than these task-specific strategies, thus necessitating further research in this area. The code and dataset are available at `https://github.com/KaustubhShejole/Viram_Marathi`.

1 Introduction

Punctuation is an essential component of written language, playing a critical role in resolving both structural and semantic ambiguity. By signaling how textual elements should be grouped and interpreted, punctuation enables readers to accurately infer the intended meaning of a sentence. Broadly, punctuation serves two complementary functions. First, it marks boundaries between segments of a larger statement and encodes grammatical relationships among those segments. Second, it provides rhetorical cues by indicating emphasis, tone, or nuance associated with particular words or phrases (Kirkman, 2006).

The importance of punctuation can be illustrated through classic examples. For instance, the omission of a comma in the phrase *"Let's eat, Grandma."* transforms an innocent dinner invitation into a cannibalistic implication. Such cases demonstrate how ambiguity naturally arises when punctuation is absent or misused. Similarly, in the sentence "This is known as 'exact' recovery.", quotation marks signal specific emphasis on the term exact, guiding the reader's interpretation. In general, punctuation errors that affect grammatical structure are more consequential than those that affect rhetorical emphasis as the former can fundamentally alter semantic interpretation (Kirkman, 2006; Carey, 1980).

The advent of Transformer (Vaswani et al., 2017) has led to rapid improvements in NMT quality over the last few years. Consequently, the applicability of encoder-decoder and Large Language Model (LLM)-based systems has expanded significantly, now encompassing diverse domains and low-resource languages (Kocmi et al., 2025; Pakray et al., 2025). In this paper, we focus on Marathi[1], an Indo-Aryan language primarily spoken by over 80 million people in the complex linguistic landscape of India, yet considered a low- to mid-resource language (Dabre et al., 2024; Lahoti et al., 2022; Gaikwad et al., 2021).

Figure 1 illustrates an example in which a missing comma in an instruction written on a fire extinguisher could lead to a disaster, highlighting the punctuation sensitivity of NMT systems. Hence, we consider it important to analyze the punctuation sensitivity of current models and to develop techniques to improve their robustness to punctuation, along with an examination of the associated trade-offs. In addition, we emphasize the need to create resources for evaluating punctuation robust-

[1] `https://en.wikipedia.org/wiki/Marathi_language`

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 151–167

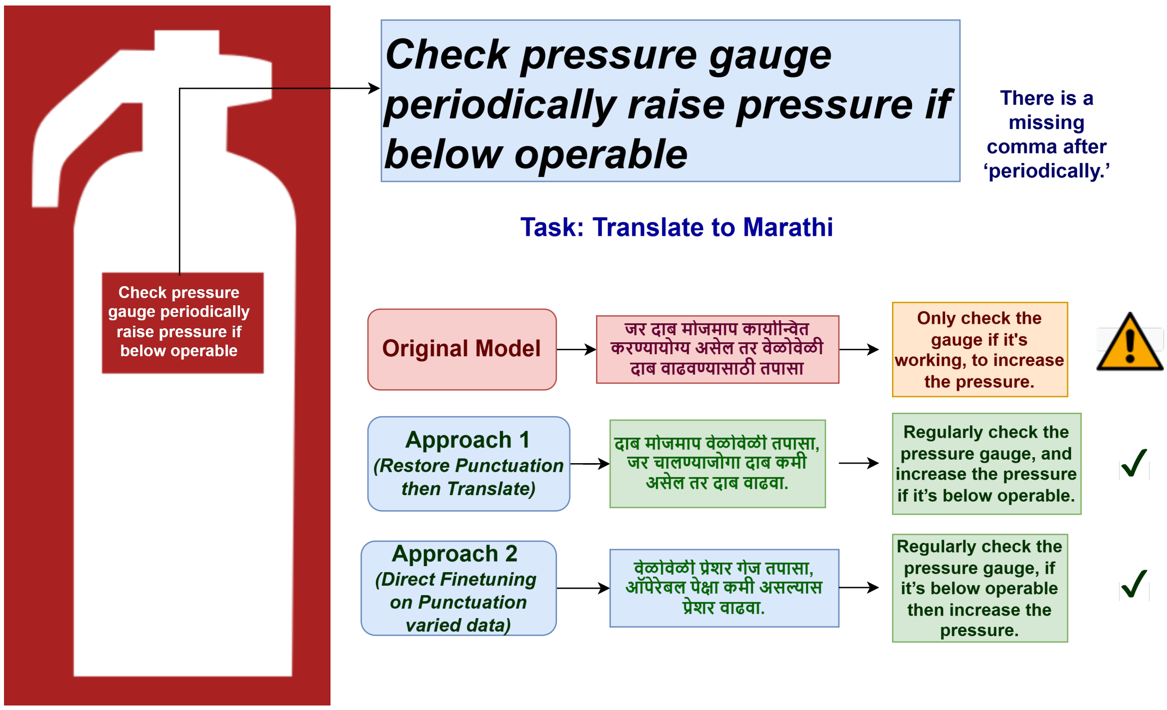

Figure 1: A missing comma can lead to a disaster in English–Marathi machine translation.

ness and to explore strategies for improving translation reliability under punctuation variability.

We first describe the data collection process for assessing the punctuation sensitivity of current Indic models, which was carried out via two native speakers of Marathi and a book by Kirkman (2006), leading to ***Virām***[2], the first English (Written)–English (Meant)–Marathi (Meant) benchmark, where 'Meant' refers to the disambiguated semantic representation. We then apply two approaches for improving the punctuation robustness of current models, and carry out quantitative and qualitative comparison. We also attempt to evaluate the translation quality via prompting of LLMs. All LLMs we considered exhibit lower performance, indicating the need of punctuation-robust approaches to be developed further. Finally, we analyze the performance of our models on standard benchmarks and observe that on the cost of punctuation robustness we might lose slightly on evaluation metrics. Our contributions are as follows:

1. The first study of punctuation robustness in English-Marathi machine translation.

2. A novel diagnostic benchmark called ***Virām*** for English–Marathi punctuation sensitivity analysis. It consists of 54 manually curated, punctuation-ambiguous instances of the form English (Written) – English (Meant) – Marathi (Meant).

3. An analysis of improving punctuation robustness using two complementary approaches: (i) *punctuation restoration in English* then *translate to Marathi*, and (ii) *Direct translation to Marathi*. This dual formulation enables systematic comparison of restoration paradigms. This analysis will help in proliferating further approaches for improving punctuation robustness.

4. A detailed qualitative analysis of model outputs, highlighting strengths, limitations, and error patterns, and identifying directions for future research on punctuation robustness in machine translation.

2 Related Work

Punctuation has long been studied in linguistics for its role in disambiguation, grammatical structure, and rhetorical emphasis (Kirkman, 2006; Carey, 1980; Lukeman, 2011; Trask, 2019). These works establish how punctuation errors can introduce semantic ambiguity, motivating its importance in downstream language technologies.

[2] *Virām chinhe* (विराम चिन्हे) is a Marathi word for punctuation, i.e., signs for marking boundaries by stopping.

English (Written)	English (Meant)	Marathi (Meant)	Punctuation
As the machine develops the forms we use to record data from past projects will be amended.	As the machine develops, the forms we use to record data from past projects will be amended.	जसजशी यंत्रणा विकसित होईल, तसतसे मागील प्रकल्पांतील डेटा रेकॉर्ड करण्यासाठी आम्ही वापरत असलेले फॉर्म्स सुधारित केले जातील.	Comma
What we see, we believe what we hear, we register	What we see, we believe; what we hear, we register	जे पाहतो, त्यावर विश्वास ठेवतो; जे ऐकतो, त्याची नोंद घेतो.	Semi Colon

Table 1: Examples of punctuation ambiguity with English sentences and their Marathi translations in the Virām Benchmark

In Natural Language Processing (NLP), punctuation restoration has been explored primarily as a preprocessing task for text and speech. Early neural approaches modeled the problem using recurrent architectures, including LSTM-based models (Tilk and Alumäe, 2015) and bidirectional RNNs with attention (Tilk and Alumäe, 2016), particularly for spoken language transcripts. Subsequent work extended punctuation restoration to multilingual and transformer-based settings, including large pretrained models for automatic punctuation and capitalization (Nagy et al., 2021; Păiş and Tufiş, 2022). More recently, systems such as Punktuator (Chordia, 2021) and Cadence (Pulipaka et al., 2025) have demonstrated robust multilingual and cross-domain punctuation restoration for both text and speech.

Within machine translation, prior studies have acknowledged the role of punctuation in preserving meaning across languages. For example, Mogahed (2012) examined punctuation effects in English–Arabic MT, highlighting its impact on translation quality. However, explicit modeling of punctuation robustness in MT pipelines remains limited.

Recent research on Indic languages has focused on improving translation quality and evaluation, with models like IndicTrans2 (Gala et al., 2023) supporting translation across all 22 scheduled Indian languages, alongside work on MT metric meta-evaluation (Dixit et al., 2023) and zero-shot evaluation in low-resource settings (Singh et al., 2024). However, English-to-Marathi translation remains highly sensitive to punctuation cues: standard models such as IndicTrans2 often misinterpret syntactic and semantic relations when punctuation is altered or removed. This highlights a critical gap in current MT systems for Marathi. To address it, we develop punctuation-robust MT models tailored for English–Marathi translation, aiming to improve reliability under punctuation variability.

In contrast to prior work, our study lies at the intersection of punctuation restoration and English–Marathi machine translation. We explicitly examine punctuation sensitivity in MT models and analyze the improvement using punctuation-robust modeling approaches, addressing a gap in both Indic MT and punctuation restoration literature.

3 Creating the *Virām* Benchmark

Kirkman (2006) analyze punctuation in the English language, examining how ambiguity can arise from the omission of punctuation marks. For instance, in the sentence, "As the machine develops the forms we use to record data from past projects will be amended," readers must insert a comma after *develops* to derive the intended meaning. This example illustrates the human ability to extract meaning from syntactically ambiguous sentences. Given that Kirkman (2006) is a well-established resource, we manually curated English sentences from this work and, with the assistance of two native Marathi speakers, translated the English (Meant) sentences into Marathi. The resulting diagnostic benchmark comprises 54 punctuation-ambiguous instances, structured as English (Written) – English (Meant) – Marathi (Meant). While the benchmark size is relatively modest, it is commensurate with the significant challenges inherent in data acquisition and curation within this specific domain. Despite this, the rigor applied to its curation ensures that it serves as a high-quality, representative sample for diagnostic evaluation. Table 1 presents selected examples from *Virām*, illustrating the nature of punctuation ambiguities and their corresponding translations. Details regarding the annotation process are provided in Appendix A.1.

4 Methodology

We explore two primary paradigms for achieving punctuation robustness in English-to-Marathi translation.

4.1 Approach 1: *Restore Punctuation then Translate*

In this decouple-and-conquer approach, punctuation is first restored in the English source text before translation, reducing the task to punctuation restoration. We adopt two modeling paradigms for punctuation restoration. In the **token classification** approach, `bert-large-uncased` (Devlin et al., 2019) and `microsoft-mpnet-base` (Song et al., 2020) are used to treat punctuation prediction as a sequence labeling task. In the **text-to-text generation** approach, we fine-tune `google-t5-base` (Raffel et al., 2020) to generate punctuated text from unpunctuated input and also evaluate AI4Bharat's `Cadence` model (Pulipaka et al., 2025) without fine-tuning it.

4.2 Approach 2: *Direct Translation*

This approach aims to improve MT robustness to noisy input. We fine-tune the `IndicTrans2` model[3] on four variants of our internal dataset[4]. We construct four variants using the original data with punctuation (**With Punct**) as a baseline, removing all source punctuation (**Without Punct**), combining both original and punctuation-removed data (**Combined 2x**), and alternately retaining or removing punctuation on a per-sentence basis (**Combined x**). Please note that 'x' refers to the size of the internal fine-tuning dataset. Details for data handling are provided for both the approaches in Appendix B.1 and B.2 respectively. Details regarding fine-tuning and hyperparameter selection are provided in Appendix F.

5 Prompting LLMs

We attempt to evaluate the translation quality of punctuation-ambiguous sentences from English to Marathi using zero-shot and few-shot prompting across three LLMs. The models considered are Sarvam-2b-v0.5[5], a 2-billion-parameter model; Gemma-2-9b[6] (Team, 2024), a 9-billion-parameter model; and LLaMA-3.1-8b[7] (Grattafiori et al., 2024), an 8-billion-parameter model. All three models have been exposed to Indian languages during pre-training. Notably, Sarvam-2b-v0.5 has been trained exclusively on English and Indian languages, including Marathi, using a corpus of approximately one trillion tokens per language. We adopt the same methodology described in Section 4 for each prompting strategy:

1. **Zero-shot prompting**
   (a) Restore punctuation, then translate (see Appendix E.2 for details about the prompt).
   (b) Direct translation (see Appendix E.3 for details about the prompt).
2. **Three-shot prompting**
   (a) Restore punctuation, then translate (see Appendix E.4 for details about the prompt).
   (b) Direct translation (see Appendix E.5 for details about the prompt).

For outputs using correctly punctuated inputs (original sentence-meant), we employed the direct translation strategy in Appendix E.1. For three-shot prompting, we selected three examples from the Virām benchmark, each illustrating a distinct punctuation error involving commas, semicolons, and colons. During evaluation, these examples were excluded from the test set to ensure a fair and unbiased assessment.

6 Results and Analysis

In this section, we present a comprehensive evaluation of the proposed approaches. We first provide a quantitative comparison of all methods (§6.1). We then analyze the impact of the two proposed improvement strategies on the original model's performance across standard benchmarks (§6.2). Next, we examine the effectiveness of different prompting strategies for large language models (§6.3). Finally, we complement the quantitative results with a qualitative analysis to better understand the strengths and limitations of the models (§6.4).

[3] `https://huggingface.co/ai4bharat/indictrans2-en-indic-dist-200M`

[4] It is an in-house corpus created by professional human translators as part of another project. The internal dataset details are provided at `https://github.com/KaustubhShejole/Viram_Marathi`

[5] `https://huggingface.co/sarvamai/sarvam-2b-v0.5`

[6] `https://huggingface.co/google/gemma-2-9b`

[7] `https://huggingface.co/meta-llama/Llama-3.1-8B`

Type	Model Name	BLEU	BLEURT-20	COMET	chrF++	chrF2++	LabSE	MuRIL
Baseline	IndicTrans2 en indic 200M (Original Model)	21.72	0.7916	0.7391	59.45	55.38	0.9126	0.7619
Upper Performance Boundary	IndicTrans2 en indic 200M + Input as 'sentence meant'	**26.20**	**0.8082**	**0.7606**	**61.15**	**57.41**	**0.9313**	**0.7915**
Approach 1	Fine-tuned bert-large-uncased + Original Model	23.84	0.7955	0.7595	60.02	56.11	0.9199	0.7806
	Fine-tuned microsoft-mpnet + Original Model	**25.12**	**0.7996**	**0.7597**	60.56	56.79	0.9210	0.7813
	Fine-tuned t5-base + Original Model	24.74	0.7977	0.7586	**60.68**	**56.92**	**0.9230**	**0.7838**
	AI4Bharat's cadence + Original Model	23.44	0.7980	0.7516	60.49	56.69	0.9210	0.7809
Approach 2	Finetuned (w/ punct) (x)	21.21	**0.7830**	0.7426	58.90	54.72	**0.9145**	0.7685
	Finetuned (w/o punct) (x)	24.66	0.7774	0.7417	60.30	56.56	0.9122	**0.7830**
	Finetuned (with and w/o punct) (2x)	24.27	0.7785	**0.7443**	**60.61**	**56.83**	0.9120	0.7794
	Finetuned (alternate with and w/o punct) (x)	24.28	0.7745	0.7433	60.21	56.52	0.9047	0.7761
LLM	GPT-5-mini (Zero-Shot + Direct Translation)	18.69	0.7786	0.7420	52.50	48.82	0.9096	0.7394
	DeepSeek-V3.2 (Zero-Shot + Direct Translation)	**23.41**	**0.7858**	**0.7590**	**58.48**	**54.82**	**0.9197**	**0.7765**

Table 2: Quantitative Analysis on the Virām Benchmark

Benchmark	Model Name	BLEU	BLEURT-20	COMET	chrF++	chrF2++	LabSE	MuRIL
IN22 (CONV)	IndicTrans2 en indic 200 M (Original)	18.95	0.8209	0.8117	51.05	47.74	0.9051	0.7451
	Fine-tuned t5-base + Original Model	17.82	0.8121	0.8077	50.24	47.17	0.8982	0.7404
	Finetuned (w/ punct) (x)	16.08	0.7963	0.7920	49.77	46.65	0.8869	0.7300
	Finetuned (w/o punct) (x)	16.67	0.8034	0.7995	49.68	46.72	0.8927	0.7354
	Finetuned (with and w/o punct) (2x)	17.93	0.8106	0.8065	50.45	47.33	0.8962	0.7408
	Finetuned (alternate with and w/o punct) (x)	17.87	0.8122	0.8082	50.34	47.26	0.8983	0.7411
IN22 (GEN)	IndicTrans2 en indic 200 M (Original)	21.01	0.7920	0.7539	54.22	49.99	0.9153	0.7389
	Fine-tuned t5-base + Original Model	16.86	0.7819	0.7439	53.19	48.84	0.9129	0.7309
	Finetuned (w/ punct) (x)	16.98	0.7627	0.7405	53.23	48.86	0.9098	0.7290
	Finetuned (w/o punct) (x)	17.07	0.7745	0.7422	53.09	48.77	0.9116	0.7307
	Finetuned (with and w/o punct) (2x)	17.11	0.7809	0.7440	53.50	49.12	0.9131	0.7315
	Finetuned (alternate with and w/o punct) (x)	17.06	0.7822	0.7450	53.37	48.99	0.9132	0.7312
FLORES-22	IndicTrans2 en indic 200 M (Original)	18.69	0.7894	0.7616	54.48	50.19	0.9226	0.7425
	Fine-tuned t5-base + Original Model	19.34	0.7818	0.7531	55.02	50.82	0.9213	0.7413
	Finetuned (w/ punct) (x)	19.20	0.7786	0.7513	54.78	50.57	0.9198	0.7400
	Finetuned (w/o punct) (x)	19.27	0.7795	0.7528	54.73	50.57	0.9201	0.7413
	Finetuned (with and w/o punct) (2x)	19.37	0.7810	0.7521	55.11	50.90	0.9204	0.7418
	Finetuned (alternate with and w/o punct) (x)	19.46	0.7825	0.7533	55.11	50.90	0.9218	0.7421

Table 3: Performance Comparison across Benchmark Datasets

6.1 Quantitative Performance

Table 2 reports quantitative results on the Virām benchmark across lexical, semantic, and embedding-based metrics. Details about metrics are provided in Appendix D. The original model with 'sent-written' input serves as the baseline, while providing oracle sentence boundaries establishes an upper bound, yielding substantial gains in BLEU, chrF, and embedding similarity. This gap highlights the impact of correct sentence boundary recovery on translation quality. Pipeline-based punctuation restoration (Approach 1) consistently outperforms the baseline, with `t5-base` and `mpnet` restorers approaching the oracle upper bound, indicating that higher-quality punctuation directly improves translation. Direct fine-tuning (Approach 2) on unpunctuated data yields clear gains in BLEU and chrF, while as expected, training only on punctuated data offers limited improvement. Mixed training improves robustness, particularly in COMET and chrF, but still falls short of the strongest pipeline-based results.

Among LLMs, DeepSeek-V3.2 outperforms GPT 5-mini across all metrics, achieving competitive semantic similarity scores in a zero-shot setting. However, both LLMs remain below the strongest pipeline and oracle-segmentation configurations. Overall, accurate sentence boundary recovery is critical for translation quality on Virām, with pipeline-based restoration most effective when segmentation quality is high, while fine-tuning improves robustness to punctuation variability.

Prompting Strategy	Approach	Model Name	BLEU	BLEURT-20	COMET	chrF++	chrF2++	LabSE	MuRIL
Zero-Shot	Restore then Translate	Llama 3.1 8b	5.16	0.6548	0.6026	37.13	33.18	0.8030	0.6093
		Gemma 2 9b	11.69	0.7260	0.6783	45.89	41.98	0.8783	0.6952
		Sarvam 2b v0.5	9.97	0.6961	0.6736	42.93	38.86	0.8228	0.6468
	Direct Translation	Llama 3.1 8b	6.37	0.6660	0.6192	39.89	35.76	0.8230	0.6314
		Gemma 2 9b	8.94	0.7248	0.6880	45.40	41.33	0.8559	0.6712
		Sarvam 2b v0.5	5.67	0.6300	0.6392	37.47	33.67	0.8102	0.6149
3-Shot Prompting	Restore then Translate	Llama 3.1 8b	4.56	0.6641	0.6113	37.80	33.68	0.8194	0.6120
		Gemma 2 9b	14.84	0.7407	0.6889	49.27	45.50	0.8926	0.7056
		Sarvam 2b v0.5	8.55	0.7352	0.6982	44.48	40.02	0.8509	0.6710
	Direct Translation	Llama 3.1 8b	7.98	0.6776	0.6272	43.21	38.91	0.8216	0.6192
		Gemma 2 9b	10.99	0.7388	0.6938	48.98	44.88	0.8883	0.7013
		Sarvam 2b v0.5	11.01	0.7571	0.7224	49.27	44.74	0.8803	0.7019
Direct Translation using Sentence Meant (Original)	Direct Translation	Llama 3.1 8b	7.87	0.6785	0.6330	39.70	35.84	0.8404	0.6535
		Gemma 2 9b	13.75	0.7256	0.6941	45.85	42.47	0.8829	0.6896
		Sarvam 2b v0.5	10.36	0.6866	0.6725	41.30	37.63	0.8422	0.6550

Table 4: Quantitative Analysis of LLMs via various prompting strategies on the Virām Benchmark

6.2 Performance Analysis on Standard Benchmarks

Table 3 reports automatic evaluation results on IN22 (CONV)[8], IN22 (GEN)[9], and FLORES-22[10]. We compare the original model with pipeline-based punctuation restoration (Approach 1) and direct fine-tuning variants on punctuated, unpunctuated, and mixed data (Approach 2).

On IN22 (CONV), the original model achieves the highest BLEU, while fine-tuned variants show modest drops. Models trained on both punctuated and unpunctuated data outperform single-input variants, with the alternate mixed strategy narrowing the gap with the original on BLEURT-20, COMET, LabSE, and MuRIL. On IN22 (GEN), the original model again outperforms fine-tuned variants. Pipeline-based punctuation restoration harms BLEU and semantic scores, indicating error propagation. Mixed fine-tuning outperforms single-condition models but remains below the original. On FLORES-22, all models perform similarly, with some fine-tuned variants slightly exceeding the original in BLEU and chrF without reducing semantic scores. This suggests that gains in punctuation robustness may come at the cost of slight reductions in certain evaluation metrics.

6.3 Analysis of Prompting Strategies in LLMs

The results in Table 4 show that, among zero-shot prompting strategies, Gemma 2 9B consistently outperforms the other evaluated LLMs across most metrics. When considering all prompting strategies, LLaMA 3.1 8B exhibits comparatively lower performance than Gemma 2 and Sarvam 2B, highlighting the impact of model architecture and pretraining scale on multilingual translation quality. In zero-shot settings, Sarvam performs better under Approach 1, whereas the other models achieve higher scores with Approach 2. Under 3-shot prompting, both LLaMA and Sarvam benefit more from Approach 2, while Gemma continues to achieve superior results with Approach 1.

When these LLM results are compared to the quantitative baselines reported in Table 2, it becomes apparent that sub-10B parameter models generally underperform relative to closed-source models such as DeepSeek-V3.2 and GPT 5-mini, which benefit from more specialized capabilities. For instance, DeepSeek-V3.2 achieves a BLEU score of 23.41 and a BLEURT-20 score of 0.7858, whereas GPT 5-mini in a zero-shot direct translation setting attains BLEU 18.69 and BLEURT-20 0.7786. In contrast, Gemma 2 9b, the best-performing model under zero-shot and three-shot prompting conditions, reaches a BLEU score of

[8] https://huggingface.co/datasets/ai4bharat/IN22-Conv
[9] https://huggingface.co/datasets/ai4bharat/IN22-Gen
[10] https://indictrans2-public.objectstore.e2enetworks.net/flores-22_dev.zip

Sentence (English Input)	Model	Marathi Output	Word-by-Word [English Gloss]	Inferred Meaning	Status
1a. Police say suspect confessed lawyer denies claim	Original	संशयिताच्या वकिलाने दावा फेटाळल्याचे पोलिसांचे म्हणणे	संशयिताच्या (Suspect's) वकिलाने (by lawyer) दावा (claim) फेटाळल्याचे (denied) पोलिसांचे (Police's) म्हणणे (saying)	Police say the suspect's lawyer denied the claim. (Missing the confession part of suspect).	Incorrect
	Approach 2 (with and w/o punct) (2x)	पोलिसांचे म्हणणे आहे की संशयिताने कबूल केलेला दावा वकील नाकारतो.	पोलिसांचे (Police's) म्हणणे (say) आहे (is) की (that) संशयिताने (by suspect) कबूल (confessed) केलेला (done) दावा (claim) वकील (lawyer) नाकारतो (denies).	The lawyer denies the claim that the suspect confessed.	Correct
	Approach 1 (t5-base)	पोलिसांचे म्हणणे आहे की संशयिताने कबूल केलेला वकील दावा नाकारतो.	पोलिसांचे (Police's) म्हणणे (say) आहे (is) की (that) संशयिताने (by suspect) कबूल (confessed) केलेला (done) वकील (lawyer) दावा (claim) नाकारतो (denies).	Police say that the claim is denied by the lawyer that the suspect confessed.	Correct
1b. Police say suspect confessed, lawyer denies claim.	All Models	पोलिसांनी सांगितले की संशयिताने कबुली दिली, वकील दावा नाकारतो.	पोलिसांनी (Police) सांगितले (said) की (that) संशयिताने (suspect) कबुली (confession) दिली (gave), वकील (lawyer) दावा (claim) नाकारतो (denies).	Two separate reports: one confession, one denial.	Correct
2a. Minister says reform failed opposition celebrates	Original	सुधारणांमध्ये अपयशी ठरलेल्या विरोधी पक्षांचा जल्लोषः मंत्री	सुधारणांमध्ये (In reforms) अपयशी (failed) ठरलेल्या (proven) विरोधी (opposition) पक्षांचा (parties') जल्लोष (celebration): मंत्री (Minister)	The Minister notes the celebration of the opposition failed in reforms.	Incorrect
	Approach 2 (with and w/o punct) (2x)	मंत्री म्हणतात, सुधारणा अयशस्वी झाल्याबद्दल विरोधकांनी जल्लोष केला.	मंत्री (Minister) म्हणतात (says), सुधारणा (reform) अयशस्वी (unsuccessful) झाल्याबद्दल (about becoming) विरोधकांनी (by opposition) जल्लोष (celebration) केला (did).	The Minister says the opposition celebrated the failure of reforms.	Correct
	Approach 1 (t5-base)	मंत्री म्हणतात की सुधारणा अयशस्वी झाल्या, विरोधक जल्लोष करतात.	मंत्री (Minister) म्हणतात (says) की (that) सुधारणा (reforms) अयशस्वी (unsuccessful) झाल्या (became), विरोधक (opposition) जल्लोष (celebration) करतात (do).	Minister says reforms failed and the opposition celebrates.	Correct
2b. Minister says reform failed, opposition celebrates.	All models	मंत्री म्हणतात की सुधारणा अयशस्वी झाल्या, विरोधक जल्लोष करतात.	(Same as Approach 1 (t5-base) above)	(Same as Approach 1 (t5-base) above)	Correct
3a. Check pressure gauge periodically raise pressure if below operable	Original	जर दाब मोजमाप कार्यान्वित करण्यायोग्य असेल तर वेळोवेळी दाब वाढवण्यासाठी तपासा	जर (If) दाब (pressure) मोजमाप (gauge) कार्यान्वित (operable) असेल (is) तर (then) वेळोवेळी (periodically) दाब (pressure) वाढवण्यासाठी (to increase) तपासा (check).	Only check the gauge if it's working, to increase pressure.	Incorrect
	Approach 2 (with and w/o punct) (2x)	प्रेशर गेज वेळोवेळी तपासा, जर ऑपेरेबल पेक्षा कमी असेल तर प्रेशर वाढवा.	प्रेशर गेज (Pressure gauge) वेळोवेळी (periodically) तपासा (check), जर (if) ऑपेरेबल (operable) पेक्षा (than) कमी (less) असेल (is) तर (then) प्रेशर (pressure) वाढवा (increase).	Regular checks; increase pressure only if it's too low.	Correct
	Approach 1 (t5-base)	दाब मोजमाप वेळोवेळी तपासा, जर चालण्याजोगा दाब कमी असेल तर दाब वाढवा.	दाब (Pressure) मोजमाप (gauge) वेळोवेळी (periodically) तपासा (check), जर (if) चालण्याजोगा (operable) दाब (pressure) कमी (low) असेल (is) तर (then) दाब (pressure) वाढवा (increase).	Check the gauge; if pressure is not operable, increase it.	Correct
3b. Check pressure gauge periodically, raise pressure if below operable.	Original	दाब मोजमाप वेळोवेळी तपासा, जर चालण्याजोगा दाब कमी असेल तर दाब वाढवा.	(Same as Approach 1 (t5-base) above)	(Same as above)	Correct
	Approach 2 (with and w/o punct) (2x)	वेळोवेळी प्रेशर गेज तपासा, ऑपेरेबल पेक्षा कमी असल्यास प्रेशर वाढवा.	वेळोवेळी (Periodically) प्रेशर गेज (gauge) तपासा (check), ऑपेरेबल (operable) पेक्षा (than) कमी (less) असल्यास (if being) प्रेशर (pressure) वाढवा (increase).	(Same as above)	Correct
	Approach 1 (t5-base)	दाब मोजमाप वेळोवेळी तपासा, जर चालण्याजोगा दाब कमी असेल तर दाब वाढवा.	(Same as Approach 1 (t5-base) above)	(Same as above)	Correct
4a. What we see we believe what we hear we register	Original	आपण जे पाहतो त्यावर आपण विश्वास ठेवतो की आपण जे ऐकतो त्यावर आपण नोंदणी करतो.	आपण (We) जे (what) पाहतो (see) त्यावर (on that) आपण (we) विश्वास (believe) ठेवतो (keep) की (OR) आपण (we) जे (what) ऐकतो (hear) त्यावर (on that) आपण (we) नोंदणी (register) करतो (do).	A choice: Do we believe what we see OR register what we hear?	Incorrect
	Approach 2 (with and w/o punct) (2x)	आपण जे पाहतो त्यावर आपण विश्वास ठेवतो, आपण जे ऐकतो त्यावर आपण नोंदणी करतो.	आपण (We) जे (what) पाहतो (see) त्यावर (on that) आपण (we) विश्वास (believe) ठेवतो (keep), आपण (we) जे (what) ऐकतो (hear) त्यावर (on that) आपण (we) नोंदणी (register) करतो (do).	We believe what we see, and we register what we hear.	Correct
	Approach 1 (t5-base)	आपण जे पाहतो, विश्वास ठेवतो, जे ऐकतो, ते नोंदवतो.	आपण (We) जे (what) पाहतो (see), विश्वास (believe) ठेवतो (keep), जे (what) ऐकतो (hear), ते (that) नोंदवतो (register).	Seeing leads to belief, hearing leads to registration.	Correct
4b. What we see we believe; what we hear we register.	All Models	आपण जे पाहतो त्यावर आपण विश्वास ठेवतो; जे ऐकतो त्यावर आपण नोंदणी करतो.	आपण (We) जे (what) पाहतो (see) त्यावर (on that) आपण (we) विश्वास (believe) ठेवतो (keep); जे (what) ऐकतो (hear) त्यावर (on that) आपण (we) नोंदणी (register) करतो (do).	Parallel statements of two human actions.	Correct

Table 5: Qualitative comparison of translation outputs of the original models and fine-tuned models on two approaches.

14.84 and BLEURT-20 of 0.7407. These results further suggest that targeted fine-tuning or augmented input strategies continue to offer substantially higher translation quality, particularly on metrics that are sensitive to semantic adequacy, such as COMET and LabSE.

6.4 Qualitative Analysis

Tables 5 presents a qualitative comparison of translations produced by the original model, Approach 1 (punctuation restoration using T5-base followed by translation), and Approach 2 (Combined 2x: direct fine-tuning on punctuated and unpunctuated data). The examples evaluate the models' ability to resolve syntactic ambiguity, clause boundaries, and semantic scope in punctuation-sparse inputs. The original model consistently struggles with unpunctuated sentences, particularly in headline-style constructions (e.g., 1a, 2a) and instructional text (3a). In news headlines containing multiple reporting verbs, the model frequently misidentifies clause attachment and argument scope, either omitting one of the reported events (1a) or incorrectly attributing actions to the wrong entity (2a). In procedural sentences, it often embeds conditional phrases incorrectly, conflating the condition with the action itself rather than expressing a sequence of operations (3a). Similarly, in parallel or contrastive constructions (4a), the absence of punctuation leads the model to misinterpret coordination as disjunction, resulting in unintended semantic alternation. These errors indicate a strong dependence on explicit punctuation cues for recovering sentence structure.

Approach 1 substantially improves translation quality by restoring punctuation prior to translation. This enables more accurate recovery of clause boundaries, coordination, and reporting structures, yielding correct interpretations in most ambiguous cases (e.g., 1a, 2a, 4a). However, as a pipeline approach, it remains sensitive to errors introduced during punctuation restoration, which can occasionally propagate into the final translation and lead to less natural or slightly misaligned syntactic realizations. Approach 2 consistently produces the most accurate and stable translations across all examples. The fine-tuned model correctly resolves implicit coordination, reporting structures, and conditional logic even in the absence of punctuation, as demonstrated in both news headlines and procedural instructions. Performance remains robust across punctuated and unpunctuated variants, suggesting that the model learns to infer latent sentence structure directly from contextual and syntactic cues rather than relying on surface punctuation.

For punctuated inputs (b variants), all models produce correct translations, confirming that most observed errors in the original model arise from difficulties in handling missing punctuation rather than lexical or morphological limitations. These results demonstrate that fine-tuning may help in combating the punctuation-sensitivity of the original model for English-Marathi machine translation. While automatic metrics show moderate gains, qualitative evaluation reveals substantial improvements in semantic fidelity that are not captured by standard scores.

7 Conclusion and Future Directions

In this study, we focused on assessing and improving punctuation robustness of English–to-Marathi NMT systems. We manually constructed ***Virām***, a diagnostic benchmark that contains punctuation-ambiguous instances. We evaluated two primary approaches: a pipeline-based *restore-then-translated* and *direct fine-tuning* on punctuated and unpunctuated data. Our quantitative and qualitative analyses reveal that both approaches significantly improve punctuation robustness compared to the baseline model. Through qualitative analysis, we identified specific failure modes where NMT models fail to capture the intended meaning in the absence of punctuation. We also evaluated LLMs via zero-shot and few-shot prompting, finding that few-shot prompting improves performance. However, these models lag behind task-specific approaches in preserving meaning for punctuation-ambiguous text, highlighting the need for further research in this area.

We plan to extend this work to other Indic languages to assess whether similar qualitative patterns emerge across language families. Future work should focus on better assessment metrics that check meaning preservation and nuances similar to human judgment, and on exploring hybrid model architectures capable of handling punctuation ambiguity natively, without relying on multi-stage pipelines like multi-task learning approaches. This work opens various research directions for punctuation-robust machine translation.

Limitations

While our study provides valuable insights into punctuation robustness, several inherent limitations bound its scope. The Virām benchmark consists of only 54 manually curated instances; although this size is sufficient for diagnostic evaluation of specific semantic ambiguities, it is not intended as a large-scale test set, with the focus deliberately placed on quality and linguistic complexity rather than volume. Our analysis is restricted to the English–Marathi language pair, and while Marathi represents a morphologically rich, low-resource Indic language, the punctuation-induced errors we observed may differ in nature and frequency for other language families or syntactic structures. Finally, as noted in our qualitative analysis, standard automated metrics such as BLEU and chrF are often insensitive to the subtle semantic shifts introduced by punctuation. While we supplemented these with manual inspection, the scalability of such qualitative evaluation is inherently limited due to the need for expert linguistic annotators.

Ethical Considerations

In alignment with the ACL Ethics Policy, we provide the following disclosures regarding our data, annotation process, and potential societal impact. The English source sentences for the Virām benchmark were manually curated from a well-established linguistic resource (Kirkman, 2006), and the fine-tuning of models in Approach 2 utilized an internal in-house corpus created by professional translators, which we plan to release publicly upon project completion to support further research. Translations for the benchmark were performed by two native Marathi speakers with advanced academic backgrounds (Master's and PhD) in Computer Science. Annotators were fairly compensated for their specialized expertise, and all translations were developed through collaborative discussions to ensure semantic accuracy and cultural relevance. We recognize that machine translation is increasingly used in India for critical applications such as digital governance and agricultural assistance, where punctuation errors can lead to significant semantic shifts and the potential dissemination of incorrect information. While our work aims to improve model robustness and mitigate such risks, we caution that no MT system is entirely error-free, and users in sensitive domains should verify automated translations with human experts. To ensure transparency and reproducibility, we have detailed our experimental setups and prompting strategies in the Appendix and are committed to releasing the Virām diagnostic benchmark publicly to encourage more robust evaluations in Indic language technologies.

Acknowledgments

We thank the anonymous reviewers and the meta-reviewer of LoResMT 2026 for their constructive comments and suggestions, which substantially improved this work. The first author gratefully acknowledges Shalaka Thorat for insightful discussions and significant assistance in developing the evaluation pipelines. We also thank seniors at CFILT, Satyam Kumar and Dhara Gorasiya, for their helpful feedback on the manuscript. We are grateful to the senior linguists at the CFILT Lab, Dr. Nilesh Joshi and Dr. Irawati Kulkarni, for their valuable remarks. Finally, the first author thanks his friends for their continued support and motivation.

References

Gordon Vero Carey. 1980. *Punctuation*. 14. CUP Archive.

Varnith Chordia. 2021. Punktuator: A multilingual punctuation restoration system for spoken and written text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 312–320.

Raj Dabre, Mary Dabre, and Teresa Pereira. 2024. Machine translation of Marathi dialects: A case study of kadodi. In *Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024)*, pages 36–44, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and 1 others. 2023. Indicmt eval: A dataset to meta-evaluate machine translation metrics for indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891.

Saurabh Sampatrao Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan. 2021. Cross-lingual offensive language identification for low resource languages: The case of marathi. In *Proceedings of the international conference on recent advances in natural language processing (RANLP 2021)*, pages 437–443.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, and 1 others. 2021. MuRIL: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

John Kirkman. 2006. *Punctuation matters: Advice on punctuation for scientific and technical writing*. Routledge.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Pawan Lahoti, Namita Mittal, and Girdhari Singh. 2022. A survey on nlp resources, tools, and techniques for marathi language processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–34.

Noah Lukeman. 2011. *The art of punctuation*. OUP Oxford.

Mogahed M Mogahed. 2012. Punctuation marks make a difference in translation: Practical examples. *Online Submission*.

Attila Nagy, Bence Bial, and Judit Ács. 2021. Automatic punctuation restoration with bert models. *arXiv preprint arXiv:2101.07343*.

Vasile Păiş and Dan Tufiş. 2022. Capitalization and punctuation restoration: a survey. *Artificial Intelligence Review*, 55(3):1681–1722.

Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of WMT 2025 shared task on low-resource Indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 532–553, Suzhou, China. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618.

Sidharth Pulipaka, Sparsh Jain, Ashwin Sankar, and Raj Dabre. 2025. Mark my words: A robust multilingual model for punctuation in text and speech transcripts. *arXiv preprint arXiv:2506.03793*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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 (EMNLP)*, pages 2685–2702.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Thibault Sellam, Amy Pu, Hyung Won Chung, and 1 others. 2020b. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927.

Anushka Singh, Ananya Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M Khapra. 2024. How good is zero-shot mt evaluation for low resource indian languages? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in* *neural information processing systems*, 33:16857–16867.

Gemma Team. 2024. Gemma.

Ottokar Tilk and Tanel Alumäe. 2015. Lstm for punctuation restoration in speech transcripts. In *Interspeech*, pages 683–687.

Ottokar Tilk and Tanel Alumäe. 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech*, volume 3, page 9.

Robert Lawrence Trask. 2019. *The Penguin guide to punctuation*. Penguin UK.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A More details about the *Virām* Benchmark

A.1 Annotation Procedure

For the creation of the Virām benchmark translations, we hired two annotators, one pursuing a Master's degree and the other a PhD, both in the Computer Science and Engineering department. Both annotators are native speakers of Marathi. All sentences were discussed and translated collaboratively to ensure high-quality and consistent translations. The annotators received appropriate honorarium for their work.

A.2 Data Statistics

The human-validated English–Marathi test set contains a total of 54 instances with various punctuation marks. Commas are the most frequent, appearing 38 times, followed by colons and hyphens, each occurring 3 times. Parentheses and quotation marks appear twice each, while em dashes, question marks, semi-colons, and slashes are less frequent, with one or two occurrences. This distribution reflects the diversity of punctuation in the dataset, which may affect the complexity of translation and evaluation.

B Dataset construction for training models

B.1 Data Handling for Approach 1

For training the punctuation restoration models, we used the English data from the IWSLT 2017 MT challenge[11]. We considered only the English portion of the dataset, where the source sentences were stripped of punctuation and the target sentences retained the original punctuation. This setup enables the models to learn to predict and restore punctuation in English sentences. Figure 2 illustrates the data handling process for Approach 1.

B.2 Data Handling for Approach 2

For direct fine-tuning of machine translation models, we created four variants of the dataset to evaluate the effect of punctuation on translation quality:

- **Original data**: Used as a baseline without expecting any punctuation robustness.
- **Data without punctuation**: All punctuation marks were removed from the source sentences to give models the ability to predict punctuation.
- **Data with and without punctuation (alternate)**: Punctuation is alternately removed and retained, keeping the dataset size equal to the original.
- **Data with and without punctuation (doubled)**: Each sentence is included twice, once with punctuation and once without, effectively doubling the dataset size.

Figure 3 shows the data handling process for Approach 2.

C Statistics of the Datasets Used

Table 6 summarizes the statistics of the `english_punctuation_restoration` dataset. The training split contains 206,112 instances, while the validation and test splits include 888 and 8,079 instances, respectively.

Table 7 shows the dataset statistics for the internal `eng_mar_finetuning_data`. The training set consists of 189,740 instances, and both the validation and test sets contain 23,717 instances each. These datasets provide the necessary coverage for training and evaluating models for punctuation restoration and English-to-Marathi fine-tuning tasks.

[11] https://huggingface.co/datasets/IWSLT/iwslt2017

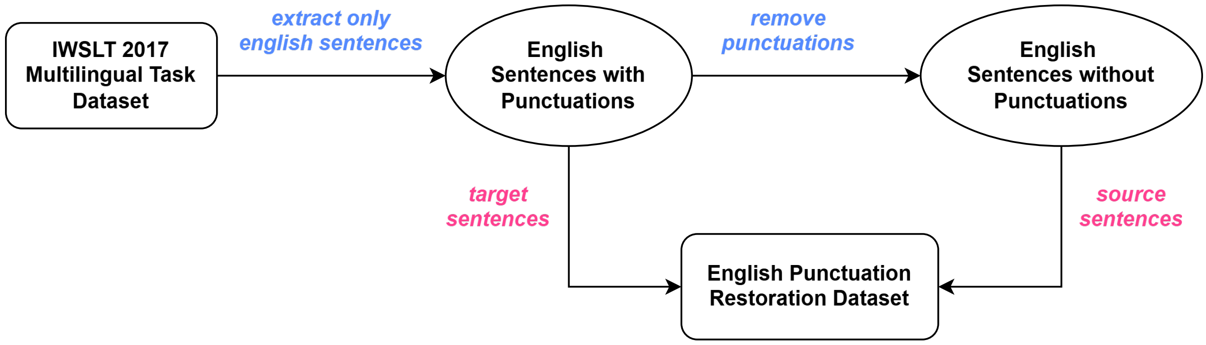

Figure 2: Data handling for the punctuation restoration task: Approach 1.

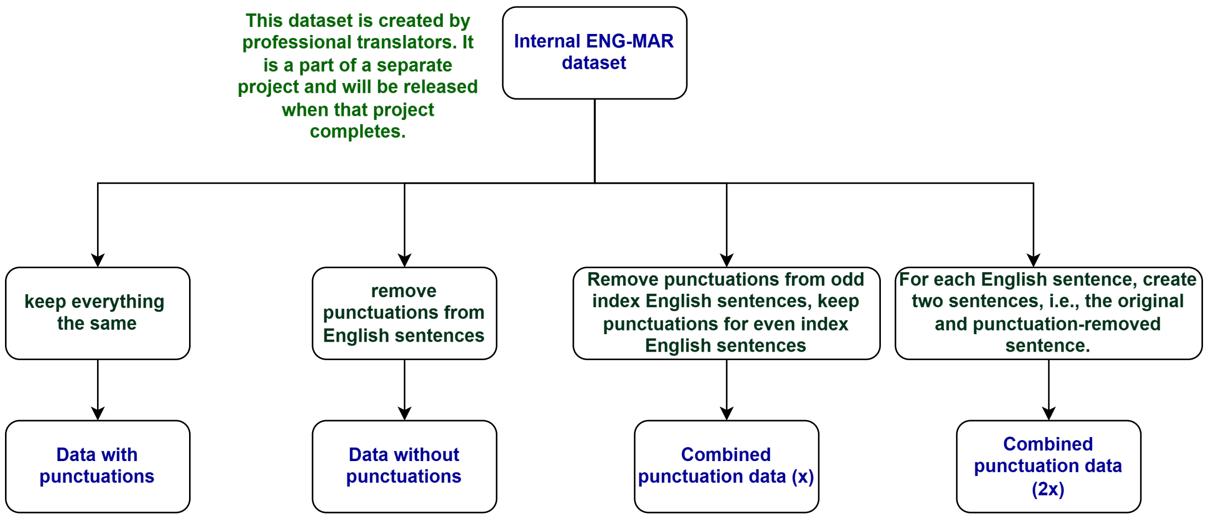

Figure 3: Data handling for the direct fine-tuning task: Approach 2.

Split	Number of Instances
Train	206,112
Validation	888
Test	8,079

Table 6: Dataset statistics for `english_punctuation_restoration`.

Split	Number of Instances
Train	189740
Validation	23717
Test	23717

Table 7: Dataset statistics for internal eng_mar_finetuning_data

D Evaluation Metrics Used

Recent advancements in Natural Language Processing (NLP), particularly in Machine Translation (MT) and cross-lingual transfer, have been driven by robust evaluation metrics and high-quality multilingual representations. This section briefly describes the evaluation metrics used in our study.

1. **BLEU (Bilingual Evaluation Understudy)** (Papineni et al., 2002) remains one of the most widely used automatic evaluation metrics for machine translation. It computes the geometric mean of modified n-gram precision between a candidate translation and one or more reference translations, combined with a brevity penalty to discourage overly short outputs.

2. **chrF++ and chrF2++** (Popović, 2017) are character n-gram–based F-score metrics that improve upon BLEU by capturing subword-level similarities, making them particularly effective for morphologically rich languages. While *chrF++* incorporates both character and word n-grams, *chrF2++* sets the β parameter to 2 (i.e., an F_2-score), placing greater emphasis on recall than precision.

3. **BLEURT-20** (Sellam et al., 2020a,b) represents a shift toward learned, neural evaluation metrics. Built on a BERT-based architecture, BLEURT is pre-trained on millions of

Table 8: Implementation details and repositories for the evaluation metrics and models.

Metric	Library / Implementation	Link / Repository
BLEU	Hugging Face `evaluate` (SacreBLEU)	`https://huggingface.co/spaces/evaluate-metric/sacrebleu`
chrF / chrF++	Hugging Face `evaluate` (SacreBLEU)	`https://huggingface.co/spaces/evaluate-metric/chrf`
COMET	Hugging Face `evaluate` (Unbabel/COMET)	`https://github.com/Unbabel/COMET`
BLEURT-20	Google Research BLEURT	`https://github.com/google-research/bleurt`
BERTScore	Hugging Face `evaluate` (MuRIL)	`https://huggingface.co/spaces/evaluate-metric/bertscore`
LaBSE	`sentence-transformers`	`https://huggingface.co/sentence-transformers/LaBSE`
MuRIL	`google/muril-base-cased`	`https://huggingface.co/google/muril-base-cased`

synthetic examples and fine-tuned using human judgment data. The "-20" checkpoint corresponds to the refined version released for the WMT 2020 Metrics shared task and exhibits strong correlation with human evaluation scores.

4. **COMET (Cross-lingual Optimized Metric for Evaluation of Translation)** (Rei et al., 2020) is a neural evaluation framework that leverages multilingual encoders such as XLM-RoBERTa. Unlike surface-level metrics such as BLEU, COMET jointly models the source sentence, hypothesis, and reference translation to directly predict translation quality.

5. **LaBSE (Language-agnostic BERT Sentence Embedding)** (Feng et al., 2022) is a dual-encoder model designed to produce language-agnostic sentence representations across 109 languages. It is trained using masked language modeling and translation ranking objectives, making it particularly effective for bitext mining and cross-lingual similarity tasks.

6. **MuRIL (Multilingual Representations for Indian Languages)** (Khanuja et al., 2021) is a BERT-based model tailored for the Indian linguistic landscape. Trained on 17 Indian languages and English, it incorporates both monolingual and translated/transliterated data, significantly outperforming general-purpose multilingual models (e.g., mBERT) on South Asian language tasks.

Implementation details and repositories for the evaluation metrics and models is provided in Table 8. The evaluation code used in this work follows the Indic MT Eval framework of Dixit et al. (2023).

E Prompting Details

E.1 Original Prompt: Direct Translation without Examples

The original prompt style instructs the model to directly translate an English sentence into Marathi without any example demonstrations or intermediate punctuation restoration. This approach tests the model's ability to perform translation with minimal guidance (see Figure 4). We used this prompt to directly input the correctly punctuated sentences to the model.

E.2 Zero-shot Prompt: Restore Punctuation then Translate

The zero-shot prompting strategy instructs the model to first restore punctuation in the input sentence and subsequently translate the punctuated sentence from English to Marathi. The prompt explicitly guides the model to perform punctuation restoration as an intermediate step before translation (see Figure 5).

E.3 Zero-shot Prompt: Direct Translation

The zero-shot direct translation prompt directly instructs the model to translate punctuation-ambiguous English sentences into Marathi without any intermediate punctuation restoration step (see Figure 6).

E.4 Three-shot Prompt: Restore Punctuation then Translate

The three-shot prompting strategy incorporates example demonstrations. Each prompt includes three input–output examples illustrating punctuation restoration followed by translation, after

Prompt for Original Translation

Prompt:
You are an expert linguist and translator specializing in English-to-Marathi machine translation. Translate the given English sentence into Marathi.

Input English: sentence

Make sure that the translation is in Devanagari Script.
Please provide the response in the following format:
Marathi Translation (Devanagari Script):

Figure 4: Prompt used for original translation.

Prompt for Zero-Shot Reasoning with Approach 1 (Restore then Translate)

Prompt:
You are an expert linguist and translator specializing in English-to-Marathi translation. You specialize in "Punctuation Restoration," resolving ambiguities caused by missing punctuation in English.

Steps for Analysis:
1. Analyze the English sentence for ambiguity.
2. Identify missing punctuation.
3. Generate the punctuated "English (Meant)" sentence.
4. Translate it into Marathi, ensuring the meaning is preserved.

Input English: sentence

Please provide the response in the following format:
Step 1 (Restoration): [The English (Meant) sentence]
Step 2 (Translation): [The Marathi translation (Devanagari Script)]
Reasoning: [Briefly explain your punctuation choices]

Figure 5: Prompt used for zero-shot reasoning with Approach 1 (Restore then Translate)

which the model applies the same process to a new punctuation-ambiguous sentence (see Figure 7).

E.5 Three-shot Prompt: Direct Translation

The three-shot direct translation prompting strategy provides three input–output examples illustrating direct translation of punctuation-ambiguous English sentences into Marathi, without any intermediate punctuation restoration. The model is then asked to translate a new sentence using the same approach (see Figure 8).

F Model Fine-tuning and Hyperparameter Tuning Details

For machine translation experiments, we fine-tuned all models on a server equipped with four NVIDIA A100 GPUs. We conducted a comprehensive hyperparameter search, experimenting with learning rates in [1e-3, 3e-3, 5e-3, 1e-4, 3e-4, 5e-4, 1e-5, 3e-5, 5e-5], varying the number of training epochs [2, 5, 8, 10], and testing different batch sizes [8, 16, 32]. This systematic exploration allowed us to identify the most effective hyperparameter configurations for

Prompt for Zero-Shot Translation with Approach 2 (direct translation)

Prompt:
You are an expert linguist and translator specializing in English-to-Marathi machine translation. Translate the given English sentence into Marathi, identifying the most logical intended meaning behind missing punctuation.

Input English: sentence

Make sure that the translation is in Devanagari Script.
Please provide the response in the following format:
Marathi Translation (Devanagari Script):

Figure 6: Prompt used for zero-shot translation withwith Approach 2 (direct translation).

each model. The final models were selected based on their performance on the validation sets.

Prompt for Few-Shot inference with Approach 1 (Restore then Translate)

Prompt:
You are an expert linguist and translator specializing in English-to-Marathi translation. Use punctuation restoration to resolve ambiguity.

Definitions:
1. English (Written): Unpunctuated input.
2. English (Meant): Restored punctuation version.
3. Marathi (Translation): Translation matching "English (Meant)".

Steps:
1. Analyze ambiguity.
2. Restore punctuation.
3. Translate to Marathi.

Some examples are as follows:

1. **Input English:** These are the components required motor brushes, bearings, and wiring.

   English Meant: These are the components required: motor brushes, bearings, and wiring.

   Marathi Translation: आवश्यक असलेले घटक खालीलप्रमाणे आहेत: मोटार ब्रशेस, बेअरिंग्ज आणि वायरिंग.

2. **Input English:** As the machine develops the forms we use to record data from past projects will be amended.

   English Meant: As the machine develops, the forms we use to record data from past projects will be amended.

   Marathi Translation: जसजशी यंत्रणा विकसित होईल, तसतसे मागील प्रकल्पांतील डेटा रेकॉर्ड करण्यासाठी आम्ही वापरत असलेले फॉर्म्स सुधारित केले जातील.

3. **Input English:** What we see, we believe what we hear, we register

   English Meant: What we see, we believe; what we hear, we register.

   Marathi Translation: जे पाहतो, त्यावर विश्वास ठेवतो; जे ऐकतो, त्याची नोंद घेतो.

Input English: sentence

Please provide the response in the following format:
Step 1 (Restoration): [The English (Meant) sentence]
Step 2 (Translation): [The Marathi translation (Devanagari Script)]
Reasoning: [Briefly explain your punctuation choice]

Figure 7: Prompt used for few-shot inference with Approach 1 (Restore then Translate)

Prompt for Few-Shot Translation

Prompt:
You are an expert linguist and translator specializing in English-to-Marathi machine translation. Translate the English sentence into Marathi, resolving ambiguity caused by missing punctuation.

Some examples are as follows:

1. **Input English:** These are the components required motor brushes, bearings, and wiring.

   English Meant: These are the components required: motor brushes, bearings, and wiring.

   Marathi Translation: आवश्यक असलेले घटक खालीलप्रमाणे आहेत: मोटार ब्रशेस, बेअरिंग्ज आणि वायरिंग.

2. **Input English:** As the machine develops the forms we use to record data from past projects will be amended.

   English Meant: As the machine develops, the forms we use to record data from past projects will be amended.

   Marathi Translation: जसजशी यंत्रणा विकसित होईल, तसतसे मागील प्रकल्पांतील डेटा रेकॉर्ड करण्यासाठी आम्ही वापरत असलेले फॉर्म्स सुधारित केले जातील.

3. **Input English:** What we see, we believe what we hear, we register

   English Meant: What we see, we believe; what we hear, we register.

   Marathi Translation: जे पाहतो, त्यावर विश्वास ठेवतो; जे ऐकतो, त्याची नोंद घेतो.

Input English: sentence

Make sure that the translation is in Devanagari Script.
Please provide the response in the following format:
Marathi Translation (Devanagari Script):

Figure 8: Prompt used for few-shot translation with Approach 2 (direct translation)