STAR-IL: A Dataset for Style-Aware Machine Translation of Product Reviews in Indian Languages

Ketaki Mangesh Shetye, Dipti Misra Sharma, Parameswari Krishnamurthy
International Institute of Information Technology, Hyderabad
ketaki.shetye@research.iiit.ac.in,{dipti,param.krishna}@iiit.ac.in

Abstract

Product reviews on e-commerce platforms are a critical form of user-generated content that influence consumer decisions. However, these reviews are predominantly in English, creating a significant accessibility barrier for users who are not fluent in English. When translating into major Indian languages using the current models, the outputs often fail to capture domain-specific features and colloquial style, resulting in stylistically unnatural texts. To address this gap, we introduce **STAR-IL**, a human-annotated, multilingual, parallel corpus for style-aware translation of product reviews. We evaluate the performance of several state-of-the-art models on our dataset for the task of product review translation. Our experiments show that models fine-tuned on STAR-IL achieve significant average performance gain of **5.77** points in BLEU and **3.78** points in COMET, when compared to their baselines, across all languages. Our dataset provides a valuable benchmark for future research in style-aware product review translation. The STAR-IL dataset is publicly available at https://github.com/ltrc/STAR-IL-Corpus.

Keywords: Language Resource Development, Machine Translation, Evaluation, Multilinguality

1. Introduction

The rise of online platforms has led to a large amount of User-Generated Content (UGC). Product reviews on e-commerce sites are a vital type of UGC influencing buying decisions, most of which are in English. This creates major accessibility barriers for non-English speakers, especially in multilingual countries like India, where only 10.6% of the population speak English (2011 Census[1]). To improve accessibility, the e-commerce sites provide automatic translations of product reviews into Indian languages. However, these translations are frequently inaccurate. They often fail to preserve the style and tone of the original review, which distorts the intended message (Saadany and Orasan, 2020). This issue arises because product reviews are inherently challenging to translate, as they contain various attributes like spelling errors, colloquialisms, sentiment, and abbreviations that are difficult to preserve.

The development of reliable translation systems for product reviews is also hindered by the lack of high-quality, domain-specific parallel data for widely spoken Indian languages[2] , including Hindi and Bengali. Past research has focused on areas such as sentiment analysis (Yadav et al., 2021), handling noisy text, and creating resources like the English-Hindi parallel corpus of translated product reviews (Gupta et al., 2021). However, the critical problem of preserving the review's original style remains largely unsolved.

As illustrated in Table 1, even advanced translation systems produce unnatural translations for the product reviews. For example, in the first review, the phrase "amazing specifications" is translated into a formal Hindi phrase " अद्भुत विशेषताओं " *(adbhuta viśēṣatāōṁ)* by Llama-3.1-8B (Grattafiori et al., 2024), " अद्भुत विशिष्टताओं " *(adbhuta viśiṣṭatāōṁ)* by Google Translate[3] and " अद्भुत विनिर्देशों " *(adbhuta vinirdēśōṁ)* by Bhashaverse (Mujadia and Sharma, 2025) rather than a colloquial, human-translated phrase like " इतनी सारी खूबियों " *(itanī sārī khūbiyōṁ)*. Therefore, to address the stylistic gap in product review translation, we introduce a new dataset and evaluate its effectiveness. The main contributions of our work are:

1. We introduce **STAR-IL**, a novel, human-annotated parallel corpus designed for style-aware translation, containing over 55,000 product reviews translated from English into eight major Indian Languages (Hindi, Marathi, Bengali, Gujarati, Urdu, Kannada, Tamil, and Telugu) across the '*fashion*' and '*electronics*' e-commerce domains.

2. We demonstrate the effectiveness of our dataset by evaluating and fine-tuning the current state-of-the-art (SOTA) models, achieving average gains of **5.77** points in BLEU and **3.78** points in COMET over baselines, across all languages.

3. We establish a new, human-validated benchmark to guide future research in this domain.

[1] https://www.census2011.co.in/
[2] https://www.ethnologue.com/insights/ethnologue200/
[3] https://translate.google.co.in/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8780–8793
11–16 May 2026.

Product Review (Source)	STAR-IL (Human-translated)	Llama-3.1-8B	Bhashaverse	Google Translate
value for money with these amazing specifications	इतनी सारी खूबियों के साथ पैसा वसूल *(itanī sārī khūbiyōṁ kē sātha paisā vasūla)*	मूल्य के लिए मूल्य इन अद्भुत विशेषताओं के साथ *(mūlya kē liē mūlya ina adbhuta viśēṣatāōṁ kē sātha)*	इन अद्भुत विनिर्देशों के साथ पैसे के लिए मूल्य *(ina adbhuta vinirdēśōṁ kē sātha paisē kē liē mūlya)*	इन अद्भुत विशिष्टताओं के साथ पैसे का पूरा मूल्य *(ina adbhuta viśiṣṭatāōṁ kē sātha paisē kā pūrā mūlya)*
crazy phone it is really ossum	வித்தியாசமான போன் ரொம்ப நல்லா இருக்கு *vidhdhiyājhamāṉa bhōṉ roṁbha nallā irughghu)*	அசத்தலான பொன்னிலக்கி இது சரியானது *(ajhadhdhalāṉa bhoṉṉilaghghi idhu jhariyāṉadhu)*	பைத்தியக்காரத்தனமான தொலைபேசி இது உண்மையில் ஓசம் *(bhaidhdhiyaghghāradhdhaṉamāṉa dholaibhējhi idhu uṇmaiyil ōjhaṁ)*	பைத்தியக்கார போன், இது உண்மையிலேயே முட்டாள்தனம்தான். *(bhaidhdhiyaghghāra bhōṉ, idhu uṇmaiyilēyē muḍhḍhāḷdhaṉaṁdhāṉ.)*

Table 1: Qualitative comparison of the translation from STAR-IL against model-generated Hindi and Tamil translations for the English product reviews. The ISO 15919 transliterations are provided in parenthesis.

2. Related Works

Prior work in translation of product reviews has largely focused on preserving the sentiment of the original review. This has been achieved through methods such as fine-tuning models for sentiment polarity (Saadany and Orasan, 2020) or constructing specialized sentiment analysis models for languages like Hindi (Yadav et al., 2021). However, these studies have placed less emphasis on preserving the colloquial style and tone inherent in the reviews.

A relevant study by Gupta et al. (2021) provides a English-Hindi parallel corpus of 22,595 manually verified machine-translated product reviews. However, our comparative automatic and human evaluation (see Section 5.1) reveals that these translations lack the source review's colloquial style and tone. Due to this stylistic issue and the dataset's limited scope to only Hindi and the *'electronics'* domain, we use their English source text but create our own human-annotated, style-aware translations across multiple Indian languages.

When evaluating UGC translations, studies show that while LLMs can be used as reference-free quality estimators, they often suffer from reliability issues (Qian et al., 2024). This finding motivates us to use a combination of standard automatic metrics and human evaluation to thoroughly assess our style-aware translations.

3. Dataset Construction

The **STAR-IL** dataset is a multilingual, parallel corpus for product review translation. It covers English and eight major Indian languages from two language families: **Indo-Aryan** (Hindi (*Hin*), Marathi (*Mar*), Bengali (*Ben*), Gujarati (*Guj*), Urdu (*Urd*)) and **Dravidian** (Kannada (*Kan*), Tamil (*Tam*), Telugu (*Tel*)).

We collect a total of 7,700 English product reviews from *'electronics'* and *'fashion'* e-commerce domains. Approximately, 71% of the reviews are from the *'electronics'* domain, obtained from the corpus by Gupta et al. (2021) (IIT Patna dataset; originally from *Flipkart*). The remaining 29% are from the *'fashion'* domain, obtained from Arnob and Khan (2024) (originally from *Myntra*). This dual-domain design ensures that STAR-IL is more generalizable and captures a diverse range of review styles common to popular retail categories.

As shown in Table 2, the data is split as 90% training and 10% benchmark data, with both splits maintaining the 71% *'electronics'* and 29% *'fashion'* domain distribution. To preserve the authenticity and context of the review, we do not perform any additional post-processing on the source text. Common features of product reviews, such as emojis, informal markers, and spelling errors, are deliberately retained to reflect real-world data.

Language Family	Indo-Aryan					Dravidian		
Language	**Ben**	**Guj**	**Hin**	**Mar**	**Urd**	**Kan**	**Tam**	**Tel**
Training	6091	6990	5986	5989	7794	5988	5488	5991
Benchmark	700	700	699	700	700	699	699	700

Table 2: Distribution of training and benchmark samples per Indian Language in the STAR-IL dataset.

3.1. Data Annotation

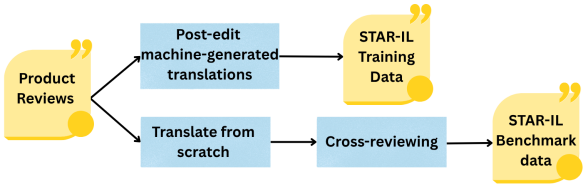

Figure 1: Two-stage human annotation workflow for STAR-IL: Post-editing machine translations for training data, and translation from scratch with cross-reviewing for benchmark data.

As discussed in Section 1 and illustrated in Table 1, modern translation systems fail to produce colloquial style-aware translations for product reviews. Hence, they cannot be used for dataset construction independently. This highlights the need for manual annotation and verification to ensure better quality of the translations.

As described in Figure 1, we implement a rigorous annotation process with an in-house team of experienced translators (3 to 4 native-speakers per language pair), using distinct methodologies to annotate the translations for the training and benchmark data in STAR-IL.

To construct the parallel training data cost-effectively, we adopt a post-editing technique. Rather than translating reviews from scratch, annotators are provided with initial machine-generated translations from either Google Translate or Bhashaverse. Then, they correct and refine these outputs to meet our annotation guidelines[4], and ensure consistency. The guidelines are based on three key principles:

- **Prioritizing Colloquial Style**: Use a natural, vernacular style that matches the source review, allowing relaxed grammar and code-mixing.
- **Retaining Domain-Specific Terms**: Keep English technical terms and brand names in their original form (or transliterated).
- **Ensuring Idiomatic Equivalence**: Translate idioms and sarcastic comments to their closest target-language equivalent to preserve the original intent.

[4]https://github.com/ltrc/STAR-IL-Corpus

The benchmark data is created adhering to the same guidelines, but it is translated entirely from scratch by the annotators. This approach ensures a high-quality test dataset, free from any potential biases introduced by machine-generated outputs. To ensure consistency, each sample in the benchmark data undergoes two rounds of independent cross-reviewing. After annotating the training and benchmark sets, we filtered out unsuitable translations to ensure a high-quality final corpus.

3.2. Dataset Analysis

Table 3 details the STAR-IL dataset statistics for both *'electronics'* and *'fashion'* domain. A key observation is that the average number of tokens per sample, measured using the *IndicTrans2-en-indic-1B tokenizer*[5], is consistently higher in the *'fashion'* domain than *'electronics'*. This shows that *'fashion'* domain reviews are more descriptive, which is also illustrated through few examples from our dataset in Table 4. Further, the Type-Token Ratio (TTR%) is consistently low across all languages which is mainly because of the colloquial style of the reviews and the dataset's focus on just two e-commerce domains, where specific terms are frequently repeated.

4. Experiments

As discussed in Section 4.1, we independently validate the STAR-IL dataset, and establish baselines using several multilingual models. We then finetune the best-performing baseline models, on our training data and measure the resulting performance gains (see Section 4.2). We use the language-wise training and benchmark splits of STAR-IL as reported in Table 2 for all our experiments.

4.1. Baselines

To establish baseline performance on the STAR-IL benchmark, we evaluate five multilingual models. All the models are selected such that they are pre-trained on few or all of the Indian languages in our dataset. The models include one encoder-decoder

[5]https://huggingface.co/ai4bharat/indictrans2-en-indic-1B

	Electronics								Fashion							
Language Family	**Indo-Aryan**					**Dravidian**			**Indo-Aryan**					**Dravidian**		
Language	**Ben**	**Guj**	**Hin**	**Mar**	**Urd**	**Kan**	**Tam**	**Tel**	**Ben**	**Guj**	**Hin**	**Mar**	**Urd**	**Kan**	**Tam**	**Tel**
Samples	4391	4491	4486	4489	4471	4488	4490	4491	2400	3199	2199	2200	3963	2199	1697	2200
Sample Length	47.86	56.55	15.93	14.23	16.52	55.13	67.15	61.97	259.79	310.23	80.98	72.65	87.90	307.15	334.91	355.34
TTR%	0.26	0.32	6.89	9.45	5.72	0.44	0.32	0.25	0.10	0.12	3.96	0.09	1.82	0.37	0.07	0.10

Table 3: Statistics for the STAR-IL dataset across the *'electronics'* and *'fashion'* domains, showing number of samples, average sample length (tokens), and Type-Token Ratio (TTR%) for the eight Indian languages.

Domain	**Product Review (Source)**
Electronics	camera not as expected compared to other mobiles in this range .
Fashion	Best clothing app ever ,with numerous collections of variant products. The quality of the products are 10/10 ,no compromise with the materials, it's a trust worthy store .Not like other online shopping apps ,where we order a product of a particular size & color and then receive a different one with tampered piece. Always on time delivery and Myntra maintains it's standard. I love Myntra

Table 4: Examples of product reviews from the *'electronics'* and *'fashion'* domains in the STAR-IL dataset.

model, **IndicTrans2** (Gala et al., 2023), and several open-weight LLMs. The LLMs are **Llama-3.1-8B** and **Llama-3.2-3B** (Grattafiori et al., 2024), which include Hindi in their pre-training; **Sarvam-Translate** (Sarvam AI Team, 2024) a Gemma3-4B-IT model (Team et al., 2025) fine-tuned on 23 Indian languages; and **Qwen3-8B** (Yang et al., 2025) , which covers all languages in the STAR-IL dataset.

We tailor our baseline experiments to each model's architecture. For IndicTrans2, we perform direct translation of the English product review into Indian Languages. On the other hand, we assess the performance of the LLMs, using different prompting techniques. In the zero-shot setting we use a detailed prompt which contains the e-commerce domain (*'electronics'* or *'fashion'*) and explicitly instructs the model to preserve the review's original style (see Prompt Template).

We evaluate the two best-performing LLMs (selected on the basis of higher COMET score) from our zero-shot experiments using dynamic few-shot prompting (C et al., 2024). This approach assesses the model's performance when relevant, in-context examples are provided in the prompt. For each source review, we dynamically retrieve semantically similar examples from the training data using the text embeddings from *nomic-embed-text-v1.5* (Nussbaum et al., 2025). These examples are then incorporated directly into the prompt after the instructions (see Prompt Template) to generate the translation.

Prompt Template

Your task is to translate a `{domain_cat}` product review from `{source_lang}` to `{target_lang}` for an e-commerce website. You are a professional translator who understands casual language and customer sentiment.
Key Instructions:

1. **Preserve Sentiment:** The translation must accurately reflect the original review's sentiment (e.g., happy, frustrated, disappointed).
2. **Maintain Tone:** Keep the tone informal and natural, just as a real customer would write. Translate slang and colloquialisms appropriately.
3. **Do Not Translate Entities:** Keep brand names, product model numbers, and technical specifications in their original form.

Translate the following review, providing ONLY the translated text.
Text:
`{source}`
Translation:

4.2. Model Finetuning

To evaluate the impact of our training data, we finetune IndicTrans2 and the two best-performing baseline LLMs (Llama-3.1-8B and Sarvam-

Hyperparameter	Value
Batch Size	1
Gradient Accumulation Steps	8
Learning Rate	2×10^{-5}
16-bit Floating Point Precision	True
LoRA Rank	16
LoRA Scaling Factor	32
LoRA Dropout	0.05

Table 5: Hyperparameter settings for fine-tuning Llama-3.1-8B and Sarvam-Translate. IndicTrans2 share the same settings, except LoRA configuration.

Translate) (see Table 8) on the human-annotated STAR-IL training data.

For the parametrically smaller IndicTrans2 model, we perform full fine-tuning for 5 epochs as it is computationally viable. The LLMs are fine-tuned for 3 epochs using Low-Rank Adaptation (LoRA) (Hu et al., 2021) with the detailed prompt (see Prompt Template). This technique efficiently adapts them to the task while preserving their pre-trained knowledge. All models are finetuned on a single NVIDIA L40S GPU with 45GB VRAM. The detailed hyperparameter settings are reported in Table 5.

5. Results and Analysis

In this section, first, we present a comparative automatic and human evaluation of the STAR-IL benchmark against the existing IIT Patna dataset (Gupta et al., 2021). For automatic evaluation we employ both lexical-level and semantic-level metrics. For lexical analysis, we use BLEU (Papineni et al., 2002) to measure n-gram precision, CHrF (Popović, 2015) to assess character n-gram overlap, and TER (Translation Edit Rate) (Snover et al., 2006) to calculate the edit distance between the predicted and reference translation. For semantic-level evaluation, we use COMET (Rei et al., 2022), which scores translation quality against both the source and reference, and BERTScore (F1) (Zhang et al., 2020), which measures semantic similarity using token embeddings.

5.1. Comparative Analysis

To evaluate our STAR-IL benchmark, we conduct both quantitative and qualitative comparisons against the existing IIT Patna dataset (Gupta et al., 2021). The IIT Patna dataset contains English-to-Hindi translated product reviews and is also the source for the English reviews in the *electronics* domain of STAR-IL.

5.1.1. Automatic Evaluation

Model	Dataset	Metrics				
		BLEU ↑	CHrF ↑	TER ↓	COMET ↑	BS ↑
IndicTrans2	IIT Patna	16.66	37.22	69.30	77.53	92.47
	STAR-IL	**25.57**	**49.92**	**57.69**	**85.87**	**94.64**
Llama 3.1	IIT Patna	**21.36**	**48.51**	**75.11**	**80.93**	**93.36**
	STAR-IL	14.12	36.54	85.05	77.71	92.28
Sarvam	IIT Patna	**34.16**	**59.99**	**57.64**	**84.81**	**94.68**
	STAR-IL	22.06	42.84	66.82	80.89	92.99

Table 6: Zero-shot performance comparison between the IIT Patna dataset and STAR-IL dataset, using IndicTrans2, Llama 3.1, and Sarvam, where BS denotes BERTScore (F1). The best score for each metric is in **bold**.

We conduct zero-shot inference for English-to-Hindi product review translation, using the three best baseline models (see Section 5.2), namely, IndicTrans2, Llama-3.1-8B (Llama 3.1), and Sarvam-Translate (Sarvam). As reported in Table 6, IndicTrans2, an encoder-decoder model specifically trained for translation, aligns better with our colloquial STAR-IL dataset, suggesting that its focused training helps to capture the stylistic nuances of a product review better. Conversely, general-purpose LLMs like Llama 3.1 and Sarvam (finetuned Gemma3-4B-IT model), which lack domain-specific training, fail to preserve the original review's style. Hence, their less-colloquial outputs align more closely to the translations in the IIT Patna dataset than to those in STAR-IL.

5.1.2. Human Evaluation

Dataset	Accuracy (Crit. 1)	Fluency (Crit. 2)	Style (Crit. 3)	Tone (Crit. 4)
IIT Patna	7.00	5.67	8.33	6.67
STAR-IL	**36.00**	**36.00**	**41.67**	**37.67**
Equal	57.00	58.33	50.00	55.67

Table 7: The values represent the percentage of times annotators preferred translations from the IIT Patna dataset or the STAR-IL dataset, or rated them as 'Equal', averaged across all criteria, with STAR-IL's preferred percentages in **bold**.

Recognizing that automatic metrics struggle to capture stylistic nuances (Agrawal et al., 2024), we conduct a comparative human evaluation. We randomly sample 100 English product reviews and compare their Hindi translations in STAR-IL benchmark against those in the IIT Patna dataset. Three annotators, who are native speakers of both English and Hindi, are presented with the source review alongside the two Hindi translations. The order is randomized and the dataset source is hid-

den to prevent bias. For each pair, annotators select the superior translation or rate them as 'Equal' based on four criteria (detailed in Section 12.2):

- **Accuracy (Criterion 1)**: Faithfulness to the original meaning and information.
- **Fluency (Criterion 2)**: Grammatical correctness and smooth readability.
- **Style (Criterion 3)**: Colloquial and natural to a native speaker.
- **Tone (Criterion 4)**: Preservation of the degree of politeness or tone.

We assess the inter-annotator agreement using Fleiss' Kappa (Fleiss, 1971). It achieves an average score of 0.74 (standard deviation = 0.02) across all criteria, which indicates substantial agreement among annotators. As reported in Table 7, for all the four criteria, the preference of rating the paired translations as 'Equal' is more than 50%, indicating that our STAR-IL benchmark maintains the quality comparable to the IIT Patna dataset, especially for fundamental aspects like 'Accuracy' and 'Fluency' of a translation. However, in cases where annotators express a clear preference, STAR-IL translations consistently outperform the translations from the IIT Patna dataset across all criteria. This is particularly observed in criteria of Style (Criterion 3) and Tone (Criterion 4), with preference rates of 41.67% and 37.67% respectively. These results show that the STAR-IL benchmark effectively captures the colloquial and stylistic nuances crucial for product review translation, as compared to the existing IIT Patna dataset.

5.2. Baseline Performance

As discussed in Section 4.1, we perform automatic evaluation of our baseline models, namely, IndicTrans2, Llama-3.1-8B (Llama 3.1), Llama-3.2-3B (Llama 3.2), Sarvam-Translate (Sarvam), and Qwen3-8B (Qwen3) on our STAR-IL benchmark.

As reported in Table 8, for the zero-shot experiment, our results indicate that Sarvam demonstrates strong instruction-following capabilities for translation, with Llama 3.1 as the second-best performing LLM (selected based on higher COMET score). Conversely, Qwen3 and Llama 3.2 achieve relatively low scores across all metrics, which shows their limited ability to translate product reviews from English to the target Indian languages in the STAR-IL benchmark. For Qwen3, this failure is explicitly linked to its architecture. As a Chain-of-Thought (CoT) instruct model, it structurally overrides the explicit 'translation-only' instructions in the prompt, resulting in severe penalties across exact-match metrics. Notably, IndicTrans2 performs competitively with Sarvam, achieving similar scores across all metrics (see Table 9).

Lang.	Metrics				
	BLEU ↑	CHrF ↑	TER ↓	COMET ↑	BS ↑
Ben					
Qwen3	0.43	8.85	1152.72	39.44	80.45
Llama 3.2	8.43	38.68	85.53	78.84	90.28
Llama 3.1	12.54	44.04	77.51	84.33	92.62
Sarvam	**26.06**	**58.29**	**57.95**	**88.74**	**93.94**
Guj					
Qwen3	0.53	7.46	939.00	39.85	80.99
Llama 3.2	4.98	25.32	97.84	71.36	89.21
Llama 3.1	8.19	31.05	88.64	78.26	90.62
Sarvam	**29.19**	**53.16**	**57.01**	**89.15**	**93.79**
Hin					
Qwen3	0.63	8.02	907.30	35.44	81.17
Llama 3.2	12.09	36.19	79.68	74.41	91.32
Llama 3.1	15.87	40.32	73.47	78.69	92.56
Sarvam	**22.31**	**45.91**	**62.70**	**81.52**	**93.38**
Mar					
Qwen3	0.40	8.41	1169.36	32.74	80.69
Llama 3.2	3.49	25.86	127.84	55.28	88.40
Llama 3.1	9.89	37.72	91.05	67.68	91.19
Sarvam	**23.43**	**51.39**	**63.81**	**78.61**	**93.46**
Urd					
Qwen3	0.68	8.08	1112.18	36.11	80.25
Llama 3.2	7.83	33.17	98.05	68.84	89.27
Llama 3.1	14.56	41.03	80.51	77.44	90.99
Sarvam	**31.48**	**56.34**	**54.82**	**85.78**	**93.82**
Kan					
Qwen3	0.26	8.70	1094.71	40.44	80.86
Llama 3.2	1.93	23.57	97.97	61.45	87.76
Llama 3.1	2.88	26.92	94.54	70.93	90.03
Sarvam	**9.33**	**41.24**	**82.29**	**84.50**	**92.22**
Tam					
Qwen3	0.08	8.17	1354.77	41.68	79.52
Llama 3.2	0.64	24.70	106.01	63.86	87.10
Llama 3.1	**4.64**	30.33	**89.18**	79.05	**91.51**
Sarvam	3.55	**36.77**	100.57	**85.74**	90.78
Tel					
Qwen3	0.24	7.73	1055.68	39.45	80.63
Llama 3.2	3.16	27.38	93.18	72.39	89.58
Llama 3.1	4.64	30.33	89.19	79.05	91.51
Sarvam	**11.40**	**44.82**	**75.53**	**86.60**	**93.22**

Table 8: Zero-shot baseline performance of LLMs on the STAR-IL benchmark. Scores are reported across the eight Indian languages, where BS denotes BERTScore (F1). The best score for each metric is in **bold** and the second-best is underlined.

Our analysis reveals two key performance trends across all models. First, Indo-Aryan languages consistently outperform Dravidian languages. This is likely due to better representation of Indo-Aryan languages in the model's pretraining data. Second, as reported in Table 12, reviews from the *'fashion'* domain consistently achieve lower TER scores than those from *'elec-*

Lang.	Metrics				
	BLEU ↑	CHrF ↑	TER ↓	COMET ↑	BS ↑
Ben	24.27	54.88	62.38	86.14	93.60
Guj	19.80	43.39	67.29	86.51	92.68
Hin	16.84	38.30	69.87	76.60	92.41
Mar	20.73	47.46	66.00	74.70	92.37
Urd	26.43	50.35	59.90	82.53	93.20
Kan	10.18	42.33	80.90	83.16	92.10
Tam	2.08	29.68	99.46	81.93	89.89
Tel	10.98	40.99	78.22	82.63	92.33

Table 9: Baseline performance of IndicTrans2 on the STAR-IL benchmark, across eight Indian languages, where BS denotes BERTScore (F1).

Model	Metrics				
	BLEU ↑	CHrF ↑	TER ↓	COMET ↑	BS ↑
Ben					
Sarvam	8.35	32.27	99.24	75.98	90.24
Llama 3.1	**14.59**	**47.05**	**73.35**	**85.70**	**92.90**
Guj					
Sarvam	10.46	30.82	95.15	80.07	90.64
Llama 3.1	**14.39**	**39.13**	**76.67**	**84.77**	**92.31**
Hin					
Sarvam	7.68	26.92	93.47	71.59	90.75
Llama 3.1	**19.22**	**43.35**	**69.67**	**80.24**	**93.00**
Mar					
Sarvam	1.37	12.32	122.13	54.16	87.70
Llama 3.1	**13.81**	**42.35**	**81.87**	**70.54**	**91.65**
Urd					
Sarvam	13.99	35.79	93.04	78.00	90.04
Llama 3.1	**19.37**	**44.61**	**72.63**	**78.75**	**92.06**
Kan					
Sarvam	4.16	27.95	102.68	72.80	89.71
Llama 3.1	**7.74**	**34.79**	**86.51**	**77.45**	**91.07**
Tam					
Sarvam	2.08	27.42	113.31	74.36	87.85
Llama 3.1	**2.19**	**30.47**	**98.17**	**81.16**	**89.88**
Tel					
Sarvam	3.98	27.58	104.15	73.77	89.76
Llama 3.1	**8.75**	**34.14**	**82.82**	**78.35**	**91.63**

Table 10: Few-shot performance of Llama 3.1 and Sarvam across the eight Indian languages, where BS denotes BERTScore (F1). The best score for each metric is in **bold**.

tronics' across all languages. This suggests that models are more prone to failure when translating *'electronics'* reviews, likely due to the high presence of technical and domain-specific terminology. In contrast, translations of *'fashion'* reviews exhibit superior lexical overlap with the ground truth as they are descriptive in nature and rely on generic vocabulary, resulting in a higher probability of successful word matching, as observed in the examples in Table 4. Despite these quantitative differences, a key qualitative finding is that all baseline models fail to produce translations that are colloquial and convey the true intent of the source review.

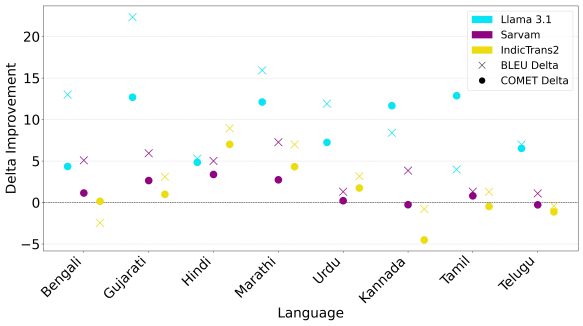

Figure 2: Performance improvement (Delta) in BLEU (x) and COMET (o) scores for three models across eight Indian languages after finetuning.

Next, we analyze in-context learning by applying few-shot prompting on the two best-performing LLMs, Llama 3.1 and Sarvam. As shown in Table 10, providing Llama 3.1 with two dynamically retrieved examples yields significant improvements over its zero-shot baseline, with average gains of 3.83 points in BLEU and 3.54 points in COMET across all languages. Being a general-purpose model, its pattern-recognition abilities allow it to adapt to the task better. Conversely, when Sarvam is provided with one dynamically retrieved example, its performance drops by an average of 13.08 points in BLEU and 12.49 points in COMET compared to its zero-shot baseline. Aligning with recent findings (Chitale et al., 2024; Ponce and Etchegoyhen, 2025), our results suggest that as Sarvam is a smaller, multilingual, instruction-tuned model, its training is too rigid to learn from in-context example. Instead, it performs better with instructions alone, as it is fine-tuned specifically to follow instructions. Given this trade-off, fine-tuning is clearly required to enhance the model's performance in this domain.

5.3. Performance After Finetuning

Following the methodology in Section 4.2, we finetune IndicTrans2 and the two best-performing LLMs Llama-3.1-8B (Llama 3.1) and Sarvam-Translate (Sarvam), selected based on higher COMET score among the zero-shot baselines. As shown in Figure 2, fine-tuning yields significant performance gains over baselines. The absolute score difference (delta) is predominantly positive in BLEU and COMET for the three models, across all languages. As reported in Table 11, Llama 3.1 improves substantially, with an average increase of 10 points in BLEU and 0.1 points in COMET over its baseline, across all languages. Similarly,

Lang.	Model	Electronics					Fashion				
		BLEU↑	CHrF↑	TER↓	COMET↑	BS↑	BLEU↑	CHrF↑	TER↓	COMET↑	BS↑
Ben	IndicTrans2	22.40	55.87	61.60	88.97	94.32	21.22	50.73	65.39	83.61	93.30
	Llama 3.1	24.09	55.31	62.03	89.23	94.40	26.99	59.96	58.17	88.11	94.35
	Sarvam	**27.73**	**58.07**	**58.83**	**90.04**	**94.68**	**34.55**	**65.49**	**49.23**	**89.72**	**95.13**
Guj	IndicTrans2	26.08	52.10	59.96	89.94	93.96	19.72	43.69	68.69	85.06	93.44
	Llama 3.1	27.46	51.95	61.53	91.30	94.20	33.62	58.80	51.75	90.59	95.04
	Sarvam	**30.16**	**55.03**	**55.53**	**91.98**	**94.51**	**40.09**	**64.06**	**46.09**	**91.60**	**95.72**
Hin	IndicTrans2	25.57	49.92	57.69	85.87	94.64	25.98	52.72	58.62	81.34	93.90
	Llama 3.1	18.38	48.48	87.65	85.54	94.62	23.91	51.48	60.75	81.52	93.73
	Sarvam	**27.22**	**50.77**	**55.81**	**86.78**	**94.87**	**27.42**	**53.78**	**57.09**	**83.01**	**94.17**
Mar	IndicTrans2	26.77	56.42	60.02	82.24	94.03	28.68	59.56	57.32	75.79	94.26
	Llama 3.1	25.98	54.82	59.79	82.64	94.42	25.64	57.00	60.58	76.93	94.44
	Sarvam	**29.88**	**57.12**	**56.09**	**83.19**	**94.49**	**31.51**	**60.96**	**54.56**	**79.50**	**95.04**
Urd	IndicTrans2	29.49	52.91	55.91	85.61	93.55	29.71	56.09	56.43	82.94	93.60
	Llama 3.1	25.16	50.63	69.47	86.30	93.51	27.77	55.12	59.29	83.05	93.58
	Sarvam	**32.27**	**54.93**	**52.36**	**86.38**	**93.85**	**33.27**	**59.41**	**52.64**	**85.60**	**94.35**
Kan	IndicTrans2	14.53	45.78	77.15	85.13	92.84	4.27	30.32	88.66	72.18	89.72
	Llama 3.1	15.30	46.29	74.92	86.25	93.12	7.26	38.42	85.28	78.93	90.42
	Sarvam	**17.51**	**48.47**	**73.02**	**86.97**	**93.13**	**8.82**	**41.12**	**81.52**	**81.49**	**91.02**
Tam	IndicTrans2	**3.76**	**32.44**	**94.29**	83.84	90.70	2.97	31.13	**90.94**	79.07	89.22
	Llama 3.1	3.15	32.09	98.36	84.48	90.80	**6.52**	40.57	97.57	85.70	90.04
	Sarvam	3.74	32.05	94.57	**85.27**	**90.88**	5.99	**41.32**	94.39	**87.81**	**90.41**
Tel	IndicTrans2	**14.99**	**45.41**	71.76	86.04	93.28	6.16	32.43	83.94	76.96	90.88
	Llama 3.1	12.96	43.84	72.10	86.23	93.40	10.28	45.14	76.95	84.90	92.16
	Sarvam	14.53	44.99	**71.66**	**86.82**	**93.58**	**10.45**	**45.63**	**75.72**	**85.82**	**92.41**

Table 11: Domainwise performance of finetuned models on the STAR-IL benchmark after finetuning IndicTrans2, Llama 3.1 and Sarvam. Scores are reported across eight Indian languages (Lang.), where BS denotes BERTScore (F1). The best score for each metric is in **bold** and the second-best is underlined.

Sarvam shows considerable gains of 3.61 points in BLEU and 0.02 points in COMET, while IndicTrans2 shows average increase of 3.14 points in BLEU and 0.02 points in COMET. This improvement is more prominent in the *'electronics'* domain, indicating that all the models successfully learn the relevant domain-specific terminology.

However, IndicTrans2's performance decreases in the *'fashion'* domain for some languages, with a 2.22 points drop in BLEU for Bengali and a 0.02 points drop in COMET for Kannada (Figure 2). This performance drop is a likely consequence of the inherent challenges of full-model finetuning, which is more sensitive to the smaller dataset size of the *'fashion'* domain compared to the *'electronics'*.

Another key observation is the limited performance gain for Tamil across all three models and metrics, particularly with only 1.85 points gain in BLEU. This is mainly due to Tamil's diglossic nature (Steever, 2019). The models which are pretrained on formal, written Tamil register, struggle to adapt to the colloquial style in our dataset, as it provides limited data to overcome the strong pretraining bias. For instance, the English review *"phone display is very nice."* is translated by finetuned IndicTrans2 into a slightly formal review as "போன் டிஸ்பிலே மிகவும் அருமையாக உள்ளது." *(Pōṉ ṭispiḷē mikavum arumaiyāka uḷḷatu)* instead of the preferred vernacular translation in our STAR-IL dataset as "போன் டிஸ்பிலே ரொம்ப நல்ல இருக்கு" *(Pōṉ ṭispiḷē roṁpa nalla irukku)*. So, while the model correctly handles domain-specific terms like "phone display" it is using formal words like "mikavum" (meaning: more) and "arumaiyāka" (meaning: nice) instead of their spoken equivalents "roṁpa" and "nalla irukku" respectively.

Overall, the results show that STAR-IL is highly effective for task adaptation, significantly improving both lexical and semantic metrics.

6. Conclusion

In this work, we introduce **STAR-IL**, a novel, human-annotated parallel corpus for style-aware product review translation, covering eight English–Indian Language pairs across the *'electronics'* and *'fashion'* e-commerce domains. Our evaluation confirms that current SOTA models struggle with the task of translating product reviews, and a comparative study against the existing dataset by Gupta et al. (2021) validates that our benchmark preserves colloquial style and tone. Fine-tuning on STAR-IL yields significant performance gains over baselines, with average increase of **5.77** points in BLEU and **3.78** points in COMET, across all languages. This demonstrates STAR-IL's effectiveness in adapting models for this task. Future work will focus on expanding the size and domain coverage of STAR-IL, to build more robust English–Indian language product review translation models. We believe STAR-IL provides a strong foundation for future research in UGC translation.

7. Limitations

The STAR-IL dataset is primarily limited by small number of samples restricted to just two e-commerce domains. This is a direct consequence of the substantial cost required for large-scale, high-quality human translation by professional annotators. Synthetic data augmentation through back-translation is not feasible due to the scarcity of monolingual corpora for the selected Indian languages in the domain of product reviews. Despite its smaller size, we demonstrate the significant performance gains achievable by models when trained using high-quality data to adapt to the task of style-aware translation. Future work will focus on different methods to expand the data in terms of size and domain-coverage.

8. Ethical Considerations

We have taken care to address the ethical considerations of this work. The source product reviews for the STAR-IL dataset are obtained from existing public corpora, and our usage respects their original licensing agreements. All the translation and annotation tasks were performed by professional translators who were formally contracted and fairly compensated for their work.

9. Acknowledgments

This work is supported by HIMANGY (HIndustani Machini ANuvaad TechnoloGY), a consortium-based initiative under BHASHINI, funded by the Ministry of Electronics and Information Technology (MeitY), Government of India. We gratefully acknowledge their support in funding and providing compensation to the annotators involved in creating the STAR-IL dataset.

10. Bibliographical References

Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can automatic metrics assess high-quality translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502, Miami, Florida, USA. Association for Computational Linguistics.

Meethu Mohan C, Sneha Shaji Punnan, and Jeena Kleenankandy. 2024. Improving few-shot prompting using cluster-based sample retrieval for medical NER in clinical text. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 37–44, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. An empirical study of in-context learning in LLMs for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406, Bangkok, Thailand. Association for Computational Linguistics.

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

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle

Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Jun-

jie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Kamal Gupta, Soumya Chennabasavaraj, Nikesh Garera, and Asif Ekbal. 2021. Product review translation using phrase replacement and attention guided noise augmentation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 243–255, Virtual. Association for Machine Translation in the Americas.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Vandan Mujadia and Dipti Misra Sharma. 2025. Bhashaverse : Translation ecosystem for indian subcontinent languages.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2025. Nomic embed: Training a reproducible long context text embedder.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

David Ponce and Thierry Etchegoyhen. 2025. In-context learning vs. instruction tuning: The case of small and multilingual language models.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Shenbin Qian, Constantin Orasan, Diptesh Kanojia, and Félix Do Carmo. 2024. Are large language models state-of-the-art quality estimators for machine translation of user-generated content? In *Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024)*, pages 45–55, Miami, Florida, USA. Association for Computational Linguistics.

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

Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hadeel Saadany and Constantin Orasan. 2020. Is it great or terrible? preserving sentiment in neural machine translation of Arabic reviews. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 24–37, Barcelona, Spain (Online). Association for Computational Linguistics.

Sarvam AI Team. 2024. SARVAM - TRANSLATE.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Sanford B. Steever, editor. 2019. *The Dravidian Languages*, 2nd edition. Routledge.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Vandana Yadav, Parul Verma, and Vinodini Katiyar. 2021. E-commerce product reviews using aspect based hindi sentiment analysis. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–8.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang,

Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

11. Language Resource References

Arnob and Noor Mairukh Khan. 2024. *ShoppingAppReviews Dataset*. Mendeley Data.

12. Appendix

12.1. Domain-wise Baseline Performance

As discussed in Section 4.1, we evaluate the domain-specific performance of the baseline models on the *electronics* and *fashion* domains from the STAR-IL benchmark. The evaluated models include IndicTrans2, Llama-3.1-8B (Llama 3.1), Llama-3.2-3B (Llama 3.2), Sarvam-Translate (Sarvam), and Qwen3-8B (Qwen).

The detailed results across all lexical and semantic evaluation metrics are presented in Table 12.

12.2. Comparative Evaluation Guidelines

We present the guidelines provided to the annotators for the human evaluation phase discussed in Section 5.1 below:

Task

For each of the 100 translations from English to Hindi, you will be presented with three pieces of text:

- **Source Text:** The original product review.
- **Translation A:** The first translated version.
- **Translation B:** The second translated version.

Your task is to compare Translation A and Translation B and decide which one is better.

How to evaluate?

After reviewing the source and both translations, please select **one** of the following three options:

- **A" : Translation A is Better:** Choose this if Translation A is clearly superior to B.
- **B" : Translation B is Better:** Choose this if Translation B is clearly superior to A.
- **"Equal" : They are Equal in Quality:** Choose this if both translations are of roughly the same quality (whether good or bad).

Evaluation Criteria

What Makes a Translation "Better"?
When comparing the two translations, please base your decision on the following three criteria, with special emphasis on Style & Tone.

Criterion 1: Accuracy (Which is more faithful to the meaning?)

The more accurate translation is the one that is truer to the source.

- Does the translation preserve all the key information, facts, opinions from the original review?

Criterion 2: Fluency (Which sounds more natural?)

- Which one is more grammatically correct and reads more smoothly in the target language with respect to the source.

Criterion 3: Style

- The more colloquial translation sounds like something a native speaker would actually write. Does the translation use a natural equivalent, or does it produce a literal and awkward result?
- In which translation are the slang, idioms, and casual phrases if any, handled better?

Criterion 4: Tone

- Depending on the degree of politeness (tone) of the source review, does the translation match same tone?

Lang.	Model	Electronics					Fashion				
		BLEU↑	CHrF↑	TER↓	COMET↑	BS↑	BLEU↑	CHrF↑	TER↓	COMET↑	BS↑
Ben	IndicTrans2	**22.47**	**52.37**	**65.43**	85.98	**93.37**	**29.27**	**61.80**	**53.94**	86.57	94.24
	Qwen	0.23	5.56	3247.09	39.69	79.77	0.88	13.00	455.10	39.16	81.89
	Llama 3.2	7.57	35.18	99.19	78.45	89.67	8.41	39.64	81.46	76.61	90.51
	Llama 3.1	11.82	42.48	84.99	83.48	92.90	12.52	44.30	75.11	82.60	91.67
	Sarvam	19.03	50.28	67.53	**86.81**	93.15	27.98	60.59	54.75	**89.86**	**94.55**
Guj	IndicTrans2	18.13	40.87	69.45	86.50	92.44	24.42	50.36	61.35	86.53	93.36
	Qwen	0.31	4.92	2891.91	39.80	80.06	1.02	10.42	360.38	39.55	82.65
	Llama 3.2	4.70	24.54	122.97	72.50	89.31	4.41	25.48	89.93	65.63	88.12
	Llama 3.1	7.60	30.06	105.69	79.04	90.63	7.63	31.25	83.81	74.50	89.91
	Sarvam	**23.39**	**45.37**	**63.98**	**88.26**	**93.29**	**30.85**	**55.62**	**54.83**	**89.44**	**94.43**
Hin	IndicTrans2	16.23	37.01	70.88	77.34	92.49	18.52	41.85	67.07	74.57	92.20
	Qwen	0.29	4.68	2770.41	35.71	80.28	1.41	12.36	347.13	34.85	82.63
	Llama 3.2	10.00	32.10	96.50	73.45	90.88	12.23	37.45	74.60	72.86	91.13
	Llama 3.1	14.12	36.54	85.05	77.71	92.28	16.07	41.41	70.63	77.29	92.06
	Sarvam	**22.06**	**42.84**	**66.82**	**80.89**	**92.99**	**22.33**	**46.77**	**62.03**	**80.10**	**93.13**
Kan	IndicTrans2	**11.35**	**43.87**	79.57	84.92	92.73	6.97	38.10	84.55	78.30	90.37
	Qwen	0.20	6.61	3024.87	40.63	80.42	0.36	10.79	436.43	39.92	81.74
	Llama 3.2	3.34	25.63	104.80	64.44	88.36	1.37	22.62	95.95	54.23	86.46
	Llama 3.1	4.73	32.10	98.91	73.95	90.87	2.06	24.84	93.19	62.78	87.89
	Sarvam	11.29	43.25	**79.06**	**86.02**	**92.80**	**8.70**	**40.60**	**83.34**	**81.77**	**90.87**
Mar	IndicTrans2	**20.62**	46.36	**66.90**	76.18	92.18	21.06	50.50	63.53	70.60	92.91
	Qwen	0.23	4.88	3453.38	34.23	79.73	0.76	12.85	454.40	30.00	82.47
	Llama 3.2	4.61	24.82	134.70	59.31	88.90	3.05	26.13	125.81	46.42	86.58
	Llama 3.1	8.18	33.67	105.38	68.14	91.05	10.45	38.98	86.19	64.62	91.00
	Sarvam	20.61	**46.86**	69.50	**77.65**	**92.68**	**23.95**	**52.67**	**62.57**	**77.66**	**94.06**
Tam	IndicTrans2	2.38	28.16	97.74	82.86	90.40	1.26	33.87	104.22	79.36	88.49
	Qwen	0.03	4.68	3578.78	40.58	78.82	0.18	12.65	547.44	44.34	81.04
	Llama 3.2	0.56	19.98	121.99	65.20	87.45	0.65	26.44	99.16	61.32	86.38
	Llama 3.1	0.75	24.14	111.73	74.03	89.25	0.82	27.92	**97.97**	69.79	87.34
	Sarvam	**5.59**	**37.01**	**92.09**	**86.57**	**91.59**	**2.91**	**37.02**	103.56	**85.08**	**89.41**
Tel	IndicTrans2	11.80	40.58	78.02	82.74	92.62	8.73	42.22	78.77	82.35	91.54
	Qwen	0.17	5.78	3078.85	39.75	80.05	0.36	9.66	413.78	39.10	81.85
	Llama 3.2	5.23	28.04	100.60	72.70	89.28	2.34	26.96	91.01	68.37	88.77
	Llama 3.1	8.10	34.33	91.51	79.52	91.96	3.37	28.78	88.92	73.27	89.68
	Sarvam	**13.07**	**43.23**	**74.40**	**85.21**	**93.26**	**10.83**	**45.16**	**76.10**	**86.37**	**92.42**
Urd	IndicTrans2	26.38	49.23	60.11	82.94	93.18	26.62	54.05	59.17	81.19	93.27
	Qwen	0.35	4.62	3030.70	36.60	79.55	1.61	13.95	394.53	35.97	82.17
	Llama 3.2	6.68	30.82	115.38	71.26	89.34	8.23	33.94	91.77	60.15	88.65
	Llama 3.1	12.39	36.96	96.13	77.05	90.88	15.32	42.41	74.82	74.53	91.28
	Sarvam	**31.65**	**53.37**	**54.18**	**85.15**	**93.69**	**31.35**	**57.37**	**55.05**	**84.78**	**93.99**

Table 12: Domainwise baseline performance of IndicTrans2, Qwen, Llama 3.2, Llama 3.1 and Sarvam on the STAR-IL benchmark. Scores are reported across eight Indian languages (Lang.), where BS denotes BERTScore (F1). The best score for each metric is in **bold** and the second-best is underlined.