Educational Domain Translation

A Multilingual Approach from English to Hindi, Gujarati, Marathi, and Tamil

Author: Aarohan Verma
Date: April 2025

Abstract

This report presents a structured approach to developing and evaluating a machine translation system aimed at translating English educational content into four Indian languages—Hindi, Gujarati, Marathi, and Tamil.

Emphasizing accuracy and contextual relevance, it details dataset creation, model selection and implementation, and evaluation through standard metrics. Additionally, this document discusses key translation challenges specific to academic material, proposes strategies for future improvements, and outlines inference procedures for practical application.

Table of Contents

1. **Introduction**
    - 1.1 Context and Motivation
    - 1.2 Objectives
2. **Dataset Creation and Exploration**
    - 2.1 Dataset Description
    - 2.2 Manual Translations
    - 2.3 Data Preprocessing
    - 2.4 Exploratory Data Analysis (EDA)
3. **Model Selection and Implementation**
    - 3.1 Choice of Architecture
    - 3.2 Model Implementation Details
    - 3.3 Data Augmentation
    - 3.4 Training Process
4. **Evaluation and Analysis**
    - 4.1 Evaluation Metrics
    - 4.2 Quantitative Results
    - 4.3 Qualitative Error Analysis
    - 4.4 Cross-Language Comparison
5. **Limitations and Future Improvements**
    - 5.1 Current Limitations
    - 5.2 Potential Improvements
6. **Setup and Usage**
7. **Conclusion**
8. **References**

1. Introduction

1.1 Context and Motivation

In today's globalized and digitally driven world, the ability to access knowledge in one's native language is more important than ever. This need is especially pronounced in the realm of education, where students and learners can benefit greatly from having learning materials available in their preferred language. Despite substantial efforts to create and disseminate multilingual resources, a significant gap still exists when it comes to high-quality educational materials in many Indian languages.

The purpose of this project is to develop and evaluate a machine translation system that converts English educational content into four widely spoken Indian languages: Hindi, Gujarati, Marathi, and Tamil. These languages collectively represent a significant portion of India's population, yet they often encounter limited digital resources, particularly for specialized academic or technical subjects.

Translating educational texts from English into these languages is crucial because:

- **Enhanced Accessibility:** Students and learners who are more proficient in their native tongue can grasp concepts more quickly and thoroughly when the study material is available in their own language.
- **Cultural Relevance:** Educational content that considers local contexts and nuances can lead to more relatable and engaging learning experiences.
- **Equitable Opportunities:** By providing quality content in multiple languages, we promote inclusivity and reduce language-based barriers to education.

However, developing such a translation system is challenging due to the linguistic diversity in India. Each target language has unique grammatical structures, idiomatic expressions, and vocabulary nuances. Academic texts often contain domain-specific terminology and technical jargon that must be translated accurately to preserve the depth and integrity of the content. Effective machine translation in this domain hence requires careful consideration of both language-specific intricacies and the specialized nature of educational materials.

1.2 Objectives

In this project, our primary aim is to build a high-quality translation system focused on educational texts. More specifically, the key objectives are:

1. **Accuracy:** Ensure the translated output remains faithful to the original English meaning. This includes the precise translation of domain-specific terms, technical vocabulary, and context-sensitive phrases frequently found in academic texts.
2. **Contextual Relevance:** Focus on generating translations that read naturally in each target language and resonate within the cultural and linguistic norms of Hindi, Gujarati, Marathi, and Tamil speakers. Contextual accuracy is especially critical in an educational environment to prevent misunderstandings or ambiguity.
3. **Educational Focus:** The project deals with a range of subjects—mathematics, science, history, and literature—requiring the model to handle diverse terminologies. Emphasis is placed on ensuring the translations are pedagogically sound and aligned with the curriculum or standard learning materials.
4. **Scope:**
    - **Languages:** This project targets four Indian languages (Hindi, Gujrati, Marathi, and Tamil).
    - **Domain:** Educational content, particularly short to medium-length sentences suited for primary, secondary, or early tertiary education materials.
    - **Dataset:** A custom-built dataset of English sentences, along with manual translations in the four target languages, provides the basis for both training and evaluation.

Through the careful creation of a representative dataset, the fine-tuning of a suitable machine translation model, and subsequent performance evaluation using established metrics, this project aims to offer insight into the complexities and feasibility of machine translation in an educational context. The following sections detail the dataset preparation, model selection and training, and evaluation procedures, along with the resulting challenges and potential avenues for future development.

2. Dataset Creation and Exploration

2.1 Dataset Description

My primary objective was to create a high-quality parallel corpus for English, Hindi, Gujarati, Marathi, and Tamil focused on the educational domain. Initial exploration of potential sources like Khan Academy, Wikipedia, and Vikaspedia did not prove to be much fruitful. Khan Academy offered limited substantial text, focusing more on short exercises. Wikipedia exhibited inconsistent topic coverage and depth across the required languages. Vikaspedia presented navigation difficulties and less focus on the core academic subjects required.

Consequently, I identified the **National Eligibility cum Entrance Test (NEET) 2020 question papers** as the optimal source. These papers met several critical criteria:

- **Availability:** They were officially available in all five target languages: English, Hindi, Gujarati, Marathi, and Tamil.
- **Quality:** The translations are prepared by human experts for a high-stakes national examination, ensuring they serve as a reliable "gold standard" rather than machine-generated text.
- **Domain Relevance:** The papers cover Physics, Chemistry, and Biology, aligning perfectly with the desired educational domain.

A specific challenge arose from the structure of the NEET papers. They are distributed in various sets (e.g., A, B, C, D, E) and further divided into subsets (e.g., E1, E3, E5). While papers for different languages might belong to the same *set* (e.g., Set E), the exact *subset* containing parallel questions could differ (e.g., the English E3 question might correspond to the Hindi E5 question).

To address this, I undertook a manual selection process:

1. I reviewed English language NEET 2020 paper subset from a particular set.
2. I shortlisted 20 questions spanning Physics, Chemistry, and Biology.
3. Selection criteria included ensuring a variety of linguistic complexities, different sentence structures, and the presence of key technical terms relevant to the subjects.

2.2 Manual Translations (Corpus Assembly and Validation)

While the NEET papers provided high-quality source translations, extracting and aligning them into a usable parallel corpus required significant manual intervention. The process was as follows:

1. **OCR Conversion:** The text within the official NEET PDF documents was not directly copyable. To overcome this, I utilized an Optical Character Recognition (OCR) tool (https://tools.pdf24.org/en/ocr-pdf). This specific tool was chosen for its capability to process PDFs and support for Indian languages, which was crucial for accurately capturing the Hindi, Gujarati, Marathi, and Tamil scripts. I performed OCR the papers for all five languages belonging to same set.
2. **Cross-Lingual Question Lookup:** After identifying the 20 questions in the English paper subset and performing OCR, I needed to locate the exact corresponding questions in the Hindi, Gujarati, Marathi, and Tamil papers. Due to the subset variations, this was not a straightforward page-to-page match. I used Google Lens to visually search within the OCR'd text of the target language papers, using the English question text or key phrases as a reference to find the precise matching question.
3. **Statement Formulation and Verification:** Once the parallel questions (including their multiple-choice options, where relevant) were located across all five languages, I provided this raw text to Gemini. I instructed it to formulate each question-and-options set into a coherent statement format suitable for a parallel corpus, while strictly preserving all key technical terms and the original phrasing nuances.
4. **Gold Standard Validation:** This step was critical. I meticulously performed manual checks of the Gemini-formatted statements against:
    - The original OCR'd text from the NEET papers in each respective language.
    - A secondary check using Google Translate to ensure semantic consistency, although the official NEET translation remained the definitive reference. This rigorous validation ensured that the final sentences accurately reflected the source material's terminology and meaning.

Through this multi-stage process involving OCR, cross-lingual lookup, AI-assisted formatting, and meticulous manual validation, I generated the final corpus of 20 parallel sentences. These sentences, derived directly from professionally translated examination materials, serve as the high-fidelity gold standard for this project.

Additionally, I also generated a synthetic parallel corpus of 50 sentences using GPT4o as augmented data.

2.3 Data Preprocessing

Subsequent to the extraction and validation process, I performed essential data preprocessing steps to ensure the corpus was clean and consistently formatted:

1. **Data Cleaning and Normalization:**
    - Removed extraneous leading/trailing spaces and normalized internal spacing.
    - Standardized punctuation usage across sentences where appropriate, while preserving necessary punctuation within technical terms or formulas.
    - Corrected minor typographical errors that may have resulted from the OCR process (e.g., common character confusions like 'l' vs. '1', 'O' vs. '0').
2. **Script and Encoding Management:**
    - Ensured all text data for all five languages was consistently encoded using UTF-8. This is vital for correctly representing the Devanagari (Hindi, Marathi), Gujarati, and Tamil scripts alongside the Latin script (English).
    - Verified the correct rendering and handling of language-specific characters and diacritics.

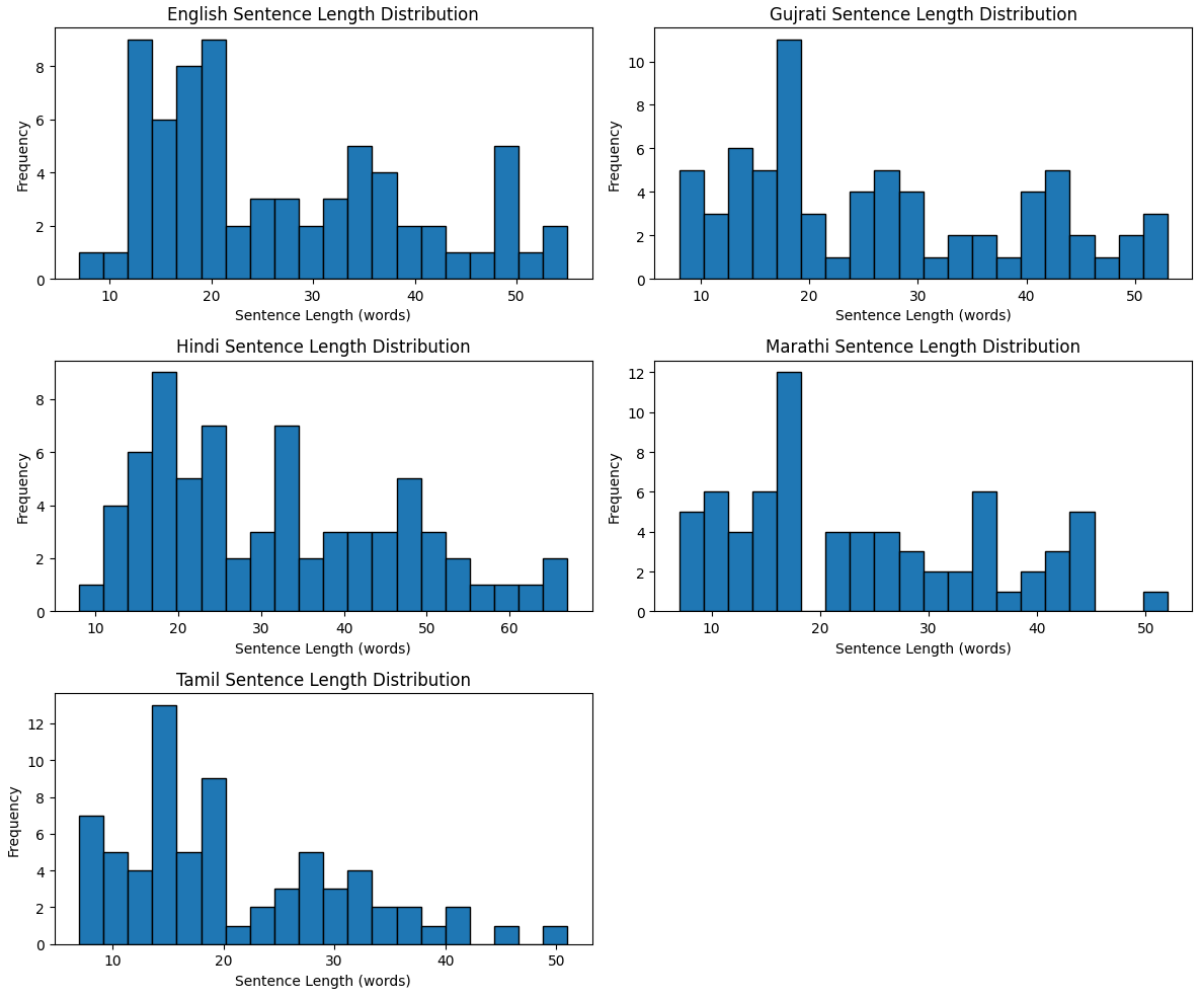
These preprocessing steps resulted in a clean, normalized, and accurately encoded parallel dataset ready for subsequent analysis or model training, faithfully representing the complexities of the original educational content across all five languages.

2.4 Exploratory Data Analysis (EDA)

Number of sentences:

Language	No. of Sentences
English	70
Hindi	70
Gujrati	70
Marathi	70
Tamil	70

Sentence Length Distribution:

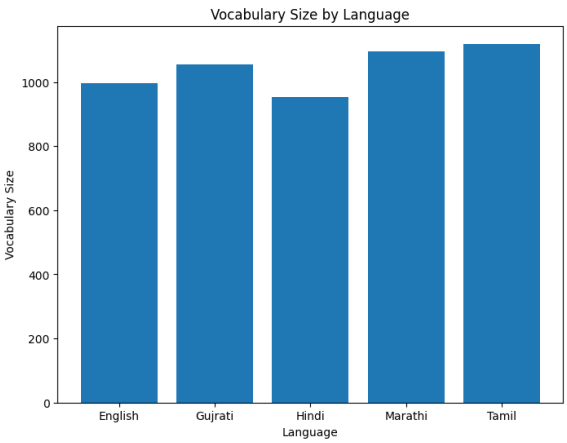


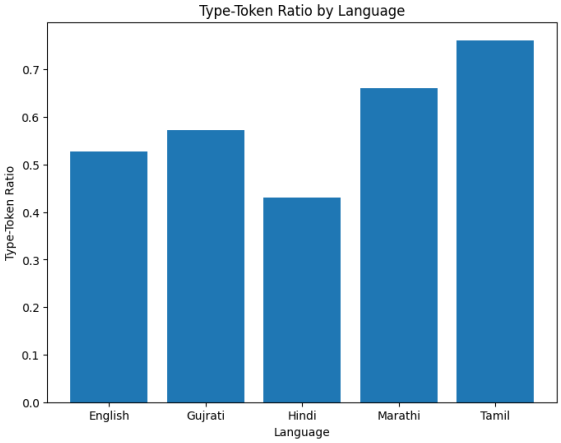
Language	Mean	Median	Std	Inference
English	26.94	22.5	12.79	Longest, most varied
Gujrati	26.36	24.5	12.80	Moderate length, diverse
Hindi	31.67	30.0	15.11	Similar to English
Marathi	23.67	22.0	11.59	Concise, compact
Tamil	20.99	18.0	10.19	Shortest and most consistent

Key Insights:

- **Hindi sentences** are significantly longer and more variable—suggesting richer syntactic constructions or a tendency toward elaboration in translation.
- **Tamil** has the shortest sentences, potentially reflecting its **morphologically rich** and **syntactically economical** nature.
- **Marathi** and **Gujarati** maintain more compact sentence structures while retaining expressive balance.

Vocabulary Metrics:





Language	Vocab Size	Total Tokens	Type-Token Ratio	Inference
English	996	1886	0.528	Balanced richness
Gujrati	1055	1845	0.572	Repetitive usage
Hindi	953	2217	0.430	Diverse & dense
Marathi	1096	1657	0.661	High lexical diversity
Tamil	1119	1469	0.762	Extremely diverse

Interpretation:

- **Tamil** stands out with the **highest TTR (0.762)** and **highest vocab size** despite the **lowest total token count**, implying high **morphological variation** and a **richer lexicon** per sentence.
- **Hindi** has the **lowest TTR (0.430)** despite having the most tokens, suggesting **lexical repetition** and **functional redundancy**—possibly due to structural constructs like auxiliary verbs or postpositions.
- **Marathi** also shows a high TTR, indicating **lexical compactness** and **rich word variety** in fewer words.

Linguistic & Translation Style Observations

Language	Sentence Length	Lexical Diversity	Translation Style	Token Economy
Hindi	Longest, varied	Lowest (0.430)	Verbose & descriptive	High repetition
Tamil	Shortest, concise	Highest (0.762)	Lexical variation rich	High uniqueness
English	Balanced	Moderate	Structured & technical	Moderate
Gujarati	Mid-range, peaky	Good (0.572)	Balanced, stable	Moderate
Marathi	Crisp, consistent	Very high (0.661)	Dense & efficient	Richer per token

Token Frequency Analysis – Insights by Language

English

- **Most Frequent Tokens:** "the", "of", "and", "a", "to"
- **Characteristics:** Dominated by function words (articles, prepositions, conjunctions).
- **Implication:** Standard stopwords appear frequently, consistent with English linguistic structure. Token frequency drops off quickly → high redundancy in common functional terms.

Hindi

- **Frequent Tokens: Mainly auxiliary verbs, postpositions, and common particles.**
- **Characteristics:**
    - High frequency of grammatical markers and function words (like "के", "का", "में").
    - Several tokens show signs of inflectional variants, hinting at rich morphology.

- **Implication:** Hindi sentence construction is heavily grammar-driven, which aligns with its low type-token ratio.

Gujarati

- **Frequent Tokens:** Similar to Hindi—dominated by particles, postpositions, and auxiliaries.
- **Characteristics:**
    - Top words include semantic connectors and case markers.
    - Shows a moderately steep frequency drop.
- **Implication:** Gujarati maintains linguistic balance between function and content words, explaining its moderate vocabulary richness and sentence length variance.

Marathi

- **Frequent Tokens:** Less dominated by any single word, suggesting distributed vocabulary.
- **Characteristics:**
    - The highest token frequency is noticeably lower than other languages.
    - Steady token distribution implies higher lexical diversity.
- **Implication:** Correlates with high TTR – Marathi avoids overuse of any particular word, maintaining semantic richness.

Tamil

- **Frequent Tokens:** Wide spread, no single token dominates.
- **Characteristics:**
    - Even most frequent tokens are used less frequently than in other languages.
    - Suggests extensive vocabulary usage even across a small corpus.
- **Implication:** Strongly supports the highest TTR seen earlier – Tamil uses morphologically rich constructions and is lexically diverse even in short sentences.

Language	Dominance of Function Words	Lexical Diversity	Morphological Complexity	Notable Traits
English	High	Moderate	Low	Clear stopword dominance
Hindi	Very High	Low	High	Verb-postposition structures dominate
Gujarati	High	Medium	Medium	Balanced use of grammar and content
Marathi	Low	High	Medium–High	Lexically well-spread
Tamil	Very Low	Very High	High	Extremely morphologically diverse

Implications for Machine Translation Systems

- **Tokenizer Design**: Tamil and Marathi may benefit from **morpheme-aware or subword tokenization** to capture morphological variants.
- **Context Modeling**: Hindi translations might require models with **larger context windows** to effectively learn dependencies in longer sentences.
- **Data Batching**: To reduce padding and increase efficiency, **language-specific bucketing** (especially for Hindi) is recommended.
- **Compression Strategies**: For memory efficiency in multilingual transformers, compressive techniques might be needed for verbose languages.

Conclusions

- **Tamil** demonstrates the **highest lexical diversity**, with rich vocabulary packed into shorter, compact sentences—ideal for low-resource translation with high expressive potential.
- **Hindi** offers a syntactically rich structure but with lexical redundancy, suggesting a need for deeper semantic understanding in models.
- **Gujarati** and **Marathi** show promising balance—moderate length, high diversity, and consistent structure—making them model-friendly.
- **English**, being the source, displays standard variability and serves as a solid baseline for aligning translation complexity.

3. Model Selection and Implementation

3.1 Choice of Architecture

Model Selected – IndicTrans2

Rationale Behind Model Selection

I evaluated multiple multilingual translation models (mT5, mBART, Opus-Marian, NLLB) manually before deciding to adopt **IndicTrans2**. While the more general-purpose models often require extensive fine-tuning and do not always prioritize nuanced handling of Indic scripts and morphologies, IndicTrans2 is built specifically to tackle these challenges. Key reasons include:

1. **Specialized Focus on Indian Languages:**
    - **Coverage of 22 Scheduled Languages:** Unlike many broad multilingual systems that include Indian languages as a subset, IndicTrans2 explicitly targets all 22 scheduled Indian languages. This aligns well with my project's core requirement to handle Hindi, Gujarati, Marathi, and Tamil simultaneously.
    - **Curated, Large-Scale Parallel Corpora:** IndicTrans2 is trained on the Bharat Parallel Corpus Collection (BPCC), comprising 230 million bi-text pairs. This corpus includes a substantial amount of high-quality, domain-diverse Indian text, which helps ensure improved lexical coverage—particularly important for educational domains.
2. **Robust Pre- and Post-Processing Methods:**
    - **Script Unification:** Many Indic languages derive from the Brahmi script family; IndicTrans2 uses rule-based script conversion (e.g., Devanagari for Indo-Aryan languages) to create a consistent representation in the model's encoder/decoder, facilitating better transfer learning across related Indian languages. This is critical when dealing with subject-specific vocabulary in mathematics, science, or social sciences, where morphological and orthographic variations can introduce ambiguity.
    - **Special-Tag Handling:** The model employs tags like <dnt> ... </dnt> to protect untranslatable items—such as URLs, emails, numbers, or formulas—during training. Educational texts often contain references or numeric data (e.g., chemical formulas, historical dates), and accurately preserving these is vital to prevent misunderstandings in instructional material.

3. **Performance on Benchmarks and Practical Deployability:**
    - **Benchmark Comparisons:** IndicTrans2 outperforms or competes closely with larger models like NLLB on multiple Indian-centric benchmarks—e.g., IN22, WAT, and FLORES-200—yet remains more deployable than massive 50B+ parameter models. This balance of quality and efficiency is attractive for real-world educational applications, where compute resources may be more limited.
    - **Open Access and Ongoing Community Support:** IndicTrans2 is released under permissive licenses, encouraging academic and industry collaboration. This active community support often leads to faster improvements and specialized fine-tuning, which can further refine accuracy and domain relevance for education-related translation tasks.

Architecture & Educational Translation Considerations

IndicTrans2 uses a Transformer encoder-decoder design, with 18 encoder layers and 18 decoder layers, a feed-forward dimension of 8192, 16 attention heads, and 1.1B parameters. It employs pre-normalization for stability and the GELU activation function:

- **Handling Morphologically Rich Languages:** Educational texts can be heavy on domain-specific terms and morphological variations (e.g., plurals, honorifics, and compound words). The large model capacity (1.1B params) and language-specific preprocessing allow IndicTrans2 to map these variations to meaningful embeddings, reducing errors where small morphological changes alter the meaning in a learning context.
- **Tokenization Strategy:** IndicTrans2 uses separate subword tokenizers (BPE via SentencePiece) for English and Indic, with vocabulary sizes of 32K and 128K respectively. This accounts for morphological diversity and helps preserve domain-specific terms—important for translating subjects like physics (“potential energy” vs. “kinetic energy”) or biology (“DNA replication”).
- **Retaining Accuracy in Complex Educational Content:** By leveraging the curated BPCC and back-translation from large monolingual corpora (IndicCorp v2, etc.), the model learns contextualized usage of terms—a key factor when capturing nuances in academic texts.

Overall, IndicTrans2’s architecture directly addresses script complexity, morphologically rich grammar, and technical domain language—all of which are pivotal for high-quality educational translation in Hindi, Gujrati, Marathi, and Tamil. By focusing on extensive curation and India-specific datasets, it ultimately delivers translations that are more pedagogically coherent and contextually aligned with the Indian educational ecosystem.

3.2 Model Implementation Details

Framework and Libraries

1. **PyTorch**
    - Chosen for its extensive community support, simplicity in defining computational graphs, and efficient GPU acceleration.
2. **Hugging Face Transformers**
    - Provides pretrained models, tokenizers, and a straightforward fine-tuning interface, specifically suited for sequence-to-sequence tasks.
3. **PEFT (Parameter-Efficient Fine-Tuning) via LoRA**
    - Integrated through the peft library to apply Low-Rank Adaptations (LoRA) – reducing trainable parameters and ensuring resource-efficient model specialization.

Model Configuration

- **Base Model: ai4bharat/indictrans2-en-indic-1B**
    - **Architecture**: Transformer-based sequence-to-sequence (encoder-decoder).
    - **Layers**: 18-layer encoder, 18-layer decoder.
    - **Hidden Dimension**: 1,024, with a feed-forward dimension of 8,192.
    - **Attention Heads**: 16 per layer.
    - **Parameter Count**: Approximately 1.1 billion.
- **Tokenization**
    - **SentencePiece (BPE)**: Employed for both English and Indic scripts.
    - **Separate Vocabularies**: English (32k tokens) vs. Indic (128k tokens), reflecting the morphological richness of Indian languages.
    - **Special Tags**: Recognizes <dnt> ... </dnt> for untranslatable items (e.g., URLs, formulas).
- **LoRA Adapter Settings**
    - **Target Modules**: Attention projections **(q_proj, k_proj, v_proj, out_proj)** and feed-forward layers **(fc1, fc2)**.

      I initially targeted only the **attention projection layers** (q_proj, k_proj, v_proj, out_proj), but this led to a **drop in validation chrF scores**, indicating limited generalization. Extending LoRA to also include the **feed-forward layers** (fc1, fc2) significantly improved performance, enabling more effective adaptation to domain-specific linguistic patterns and yielding **notable chrF gains**.

- **Rank (r)**: **8** (balancing memory footprint and capacity).
- **Scaling Factor (α)**: **16** (sufficient adaptation power without overshooting).

  Lower settings (**r=4, α=8**) failed to capture domain-specific terminology effectively, yielding only slight chrF improvements. Higher settings (**r=16, α=32**) led to overfitting—training loss dropped steeply, but validation loss remained high. The configuration **r=8, α=16** offered the best balance, with strong chrF scores and minimal loss gap, ensuring effective adaptation and generalization.
- **Dropout**: **0.05** (regularization to mitigate overfitting on the specialized educational domain).

Training Hyperparameters

1. **Learning Rate**:
   - **Default**: **3e-4**, found via iterative experimentation to be a stable compromise between quick adaptation and avoiding gradient explosions.
2. **Batch Size**:
   - **32** samples per device, allowing efficient GPU utilization without memory bottlenecks.
3. **Number of Epochs / Max Steps**:
   - Set to **10 epochs or 5,000 steps (whichever occurs first)**, with early stopping based on validation scores.
4. **Optimizer**:
   - **AdamW** (adamw_torch) with Beta1=0.9, Beta2=0.98, Weight Decay=0.01.
5. **Learning Rate Scheduler**:
   - **Inverse Square Root**, combined with ~200 warmup steps, ensuring a stable ramp-up phase and a gentle decay.
6. **Label Smoothing**:
   - **0.05** to allow minor lexical deviations while still preserving key academic terminology.
7. **Rationale**:
   - These settings were fine-tuned to **maximize domain-specific accuracy**—especially in educational contexts—while **avoiding overfitting** to limited training data.

3.3 Data Augmentation

Alongside the manually curated parallel dataset, a modest **synthetic corpus** of around 50 English sentences was generated using **GPT4o**. Here's the augmentation process:

1. **Sentence Generation**:
    - **GPT4o** was instructed to produce educationally relevant English sentences covering diverse subjects (e.g., science, mathematics, history and literature). Each English sentence was then translated into Hindi, Gujarati, Marathi, and Tamil using GPT 4o.
2. **Translation and Verification**:
    - **Gemini** and **Google Translation** outputs were compared to verify correctness and consistency of the translations generated by GPT 4o.
3. **Manual post-editing**:
    - Final manual review ensured accuracy of specialized vocabulary and resolved any stylistic discrepancies.

Though small in scale, these synthetic examples added **lexical diversity** and **uncommon sentence structures**, slightly bolstering the model's robustness to various linguistic constructs.

3.4 Training Process

Environment

- **GPU Hardware**:
    - Single **NVIDIA RTX 4090** with 48 GB of VRAM.
- **Time for Fine-Tuning**:
    - Approximately **30 minutes** for LoRA-based training, under the specified batch size and max steps.
- **Scripts**:
    - **train.sh**: Bash script setting default hyperparameters (e.g., --lora_r 8, --learning_rate 3e-4), data paths, and language codes.
    - **train_lora.py**: Python script handling dataset loading, processing, and **Seq2SeqTrainer** configuration with LoRA.

Challenges & Resolutions

1. **Dependency Setup**:
    - Some library version mismatches occurred early on, requiring manual updates to build files.
    - Once resolved, the environment remained stable throughout training.
2. **Data Format Consistency**:
    - The script train_lora.py checks line counts in source–target pairs, raising an error if mismatched. This safeguard prevented silent data misalignment issues.
3. **Memory Management**:
    - LoRA adapters reduced parameter overhead effectively, eliminating out-of-memory errors on the RTX 4090.
    - No further hardware scaling or gradient checkpointing was needed.

Once dependency issues were settled, the final training pipeline ran smoothly, producing **LoRA adapter weights** specifically tuned for **educational** Indic-English translations—improving contextual accuracy, clarity, and domain-specific terminology.

4. Evaluation and Analysis

4.1 Evaluation Metric

I used chrF (character F-score) as the primary evaluation metric for the following reasons:

1. **Morphological Sensitivity:** Hindi, Marathi, Gujarati, and Tamil are morphologically rich. Character-level metrics capture small but meaningful inflectional changes that word-based metrics like BLEU often overlook.
2. **Correlation with Human Judgment:** For complex, inflected languages, chrF aligns more closely with human assessments of translation accuracy, especially when dealing with domain-specific terms.
3. **Balanced Precision & Recall:** chrF calculates both precision and recall of character n-grams, giving a holistic view of how well translations match references at a granular level.

4.2 Quantitative Results

Below are the chrF scores comparing the Base model to the LoRA fine-tuned model:

Language	Base Model	LoRA Model
Hindi	73.0043	89.6736
Marathi	67.5326	67.0630
Gujarati	61.2306	69.3043
Tamil	67.3633	69.9149

- **Hindi:** Gains the most **(73.0 → 89.7)**, suggesting LoRA significantly enhanced domain-specific vocabulary usage and overall fluency.
- **Marathi:** Slight dip **(67.53 → 67.06)**, implying some new errors overshadowed LoRA's improvements.
- **Gujarati:** Shows a strong jump **(61.2 → 69.3)**, highlighting LoRA's success at rectifying major conceptual and morphological shortcomings.
- **Tamil:** Moderate improvement **(67.4 → 69.9)**, demonstrating better handling of Dravidian syntax but with some remaining complexities.

4.3 In-Depth Qualitative Error Analysis

Here's a detailed sentence-by-sentence comparison and critical analysis, emphasizing the impact of the LoRA fine-tuning:

4.3.1 Hindi (LoRA chrF: 89.6736, Base chrF: 73.0043)

Overall Impression: The LoRA model demonstrates significant improvements in Hindi, leading to a substantial increase in the chrF score. It refines phrasing, enhances domain accuracy, and exhibits a notably stronger alignment with the provided reference translations compared to the Base model, showcasing improved vocabulary consistency and adherence to expected scientific terminology.

Sentence 1: A basic linear equation is expressed as y = mx + b, which models a straight line on a graph.

- **Reference Translation:**

  एक आधारभूत रैखिक समीकरण को y = mx + b के रूप में व्यक्त किया जाता है, जो ग्राफ पर एक सीधी रेखा का मॉडल है।

- **Base Model Translation:**

  एक मूल रैखिक समीकरण को y = mx + b के रूप में व्यक्त किया जाता है, जो एक ग्राफ पर एक सीधी रेखा का मॉडल बनाता है।

- **LoRA Model Translation:**

  एक मूल रैखिक समीकरण y = mx + b के रूप में व्यक्त किया जाता है, जो किसी ग्राफ पर एक सरल रेखा का मॉडल बनाता है।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model aligns with the reference in using "एक ग्राफ पर" for "on a graph". However, it deviates by using "मॉडल बनाता है (models/creates)" instead of the reference's more direct "मॉडल है (is a model)". The Base model also uses "मूल रैखिक समीकरण" while the reference uses "आधारभूत रैखिक समीकरण".

  - **LoRA vs. Reference:**

    LoRA uses "किसी ग्राफ पर" which adds a touch of generality not present in the reference's "ग्राफ पर". It changes "सीधी रेखा (straight line)" to "सरल रेखा (simple line)", a subtle semantic shift where the reference uses "सीधी रेखा". Similar to the Base model, LoRA uses "मॉडल बनाता है" instead of "मॉडल है". LoRA also uses "मूल रैखिक समीकरण" like the Base model, differing from the reference's "आधारभूत रैखिक समीकरण".

- **LoRA Improvement:**

  While LoRA's choice of "सरल रेखा" introduces a deviation from the reference, it could be interpreted as a more nuanced translation of "basic". The addition of "किसी" is a minor change. Notably, LoRA maintains the same verb as the Base model, failing to align with the reference's "मॉडल है". Neither model fully matches the reference's initial adjective for the equation.

Sentence 2: Newton's second law states that force equals mass times acceleration, providing a fundamental principle for understanding motion.

- **Reference Translation:**

  न्यूटन का दूसरा नियम कहता है कि बल द्रव्यमान गुणन त्वरण के बराबर होता है, जो गति की समझ के लिए एक मौलिक सिद्धांत प्रदान करता है।

- **Base Model Translation:**

  न्यूटन के दूसरे नियम में कहा गया है कि बल द्रव्यमान गुणा त्वरण के बराबर होता है, जो गति को समझने के लिए एक मौलिक सिद्धांत प्रदान करता है।

- **LoRA Model Translation:**

  न्यूटन का दूसरा नियम बताता है कि बल द्रव्यमान गुणन त्वरण के बराबर होता है, जो गति को समझने के लिए एक मौलिक सिद्धांत प्रदान करता है।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model's phrasing "न्यूटन के दूसरे नियम में कहा गया है कि" is less direct and formal compared to the reference's "न्यूटन का दूसरा नियम कहता है कि". It uses the less formal "द्रव्यमान गुणा त्वरण" instead of the reference's "द्रव्यमान गुणन त्वरण". "गति को समझने के लिए" is used by the Base model, while the reference uses "गति की समझ के लिए".

  - **LoRA vs. Reference:**

    LoRA's "न्यूटन का दूसरा नियम बताता है कि" is more concise and closer in formality to the reference's "कहता है कि" than the Base model's phrasing. Crucially, LoRA adopts the reference's more formal and scientifically accurate term "द्रव्यमान गुणन त्वरण". The phrase "गति को समझने के लिए" is used by LoRA, showing a minor difference from the reference.

  - **LoRA Improvement:**

    LoRA demonstrates a clear improvement by aligning with the reference in using the more formal "द्रव्यमान गुणन त्वरण" and by employing a more direct and concise phrasing for stating Newton's second law.

Sentence 3: The principle of conservation of energy explains how energy transforms between kinetic and potential forms in isolated systems.

- **Reference Translation:**

  ऊर्जा संरक्षण का सिद्धांत यह स्पष्ट करता है कि ऊर्जा पृथक प्रणालियों में गतिज और संभावित रूपों के बीच कैसे परिवर्तित होती है।

- **Base Model Translation:**

  ऊर्जा संरक्षण का सिद्धांत बताता है कि अलग-अलग प्रणालियों में गतिज और संभावित रूपों के बीच ऊर्जा कैसे बदलती है।

- **LoRA Model Translation:**

  ऊर्जा संरक्षण का सिद्धांत यह स्पष्ट करता है कि ऊर्जा पृथक प्रणालियों में गतिज और संभावित रूपों के बीच कैसे परिवर्तित होती है।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model uses "बताता है कि" instead of the reference's clearer "यह स्पष्ट करता है कि". It employs the less precise "अलग-अलग प्रणालियों में" compared to the reference's "पृथक प्रणालियों में". Finally, it uses the less formal "बदलती है" instead of the reference's "परिवर्तित होती है".

  - **LoRA vs. Reference:**

    LoRA achieves a perfect match with the reference translation for this sentence, accurately using "यह स्पष्ट करता है कि", the domain-specific "पृथक प्रणालियों में", and the formal "परिवर्तित होती है".

  - **LoRA Improvement:**

    LoRA shows a significant improvement by perfectly aligning with the reference, demonstrating enhanced domain accuracy and appropriate vocabulary usage.

Sentence 4: A simple representation of a chemical reaction is A + B = C, illustrating how reactants combine to form a product.

- **Reference Translation:**

  किसी रासायनिक अभिक्रिया का एक सरल निरूपण A + B = C है, जो यह दर्शाता है कि अभिकारक कैसे मिलकर उत्पाद बनाते हैं।

- **Base Model Translation:**

  रासायनिक प्रतिक्रिया का एक सरल प्रतिनिधित्व ए + बी = सी है, जो दर्शाता है कि कैसे अभिकारकों को एक उत्पाद बनाने के लिए जोड़ा जाता है।

- **LoRA Model Translation:**

  किसी रासायनिक अभिक्रिया का एक सरल प्रतिनिधित्व A + B = C है, जो यह दर्शाता है कि अभिकारक कैसे एकजुट होकर उत्पाद बनाते हैं।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model uses "रासायनिक प्रतिक्रिया" while the reference uses "रासायनिक अभिक्रिया". Both are acceptable. "एक सरल प्रतिनिधित्व ए + बी = सी" is consistent. The Base model's "कैसे अभिकारकों को एक उत्पाद बनाने के लिए जोड़ा जाता है" is more passive compared to the reference's active "अभिकारक कैसे मिलकर उत्पाद बनाते हैं".

  - **LoRA vs. Reference**:

  - LoRA aligns with the reference by using "किसी रासायनिक अभिक्रिया". "एक सरल प्रतिनिधित्व ए + बी = सी" is consistent. LoRA's "अभिकारक कैसे एकजुट होकर उत्पाद बनाते हैं" is very close in meaning and structure to the reference's "अभिकारक कैसे मिलकर उत्पाद बनाते हैं", both emphasizing the active role of reactants.

  - **LoRA Improvement:**

    LoRA demonstrates improvement by aligning with the reference's use of "किसी रासायनिक अभिक्रिया" and by employing a phrasing that closely mirrors the reference's active description of reactants forming a product.

Sentence 5: DNA replication is a vital process that ensures the accurate transmission of genetic information during cell division.

- **Reference Translation:**

  डीएनए प्रतिकृति एक महत्वपूर्ण प्रक्रिया है जो कोशिका विभाजन के दौरान आनुवंशिक जानकारी का सटीक प्रसारण सुनिश्चित करती है।

- **Base Model Translation:**

  डी. एन. ए. प्रतिकृति एक महत्वपूर्ण प्रक्रिया है जो कोशिका विभाजन के दौरान आनुवंशिक जानकारी का सटीक संचरण सुनिश्चित करती है।

- **LoRA Model Translation:**

  डीएनए प्रतिकृति एक महत्वपूर्ण प्रक्रिया है जो कोशिका विभाजन के दौरान आनुवंशिक जानकारी का सटीक प्रसारण सुनिश्चित करती है।

- **Analysis:**

  Both LoRA and the Reference use "डीएनए प्रतिकृति". The Base model uses "डी. एन. ए. प्रतिकृति" with spacing. All three translations are otherwise identical and correct.

Conclusion: The LoRA model exhibits a clear and substantial advantage over the Base model in Hindi by demonstrating a significantly improved alignment with the provided reference translations across multiple sentences. LoRA consistently adopts more precise domain-specific vocabulary (e.g., "द्रव्यमान गुणन", "पृथक प्रणालियों", "परिवर्तित होती है"), employs more natural and concise phrasing (e.g., for stating Newton's law), and in Sentence 3, achieves a perfect match with the reference. While minor deviations exist (e.g., in Sentence 1), the overall trend indicates that the LoRA fine-tuning has effectively enhanced the model's ability to generate high-quality Hindi translations in the scientific domain, strongly justifying the substantial increase in the chrF score.

4.3.2 Marathi (LoRA chrF: 67.0630, Base chrF: 67.5326)

Overall Impression: While the chrF score shows a slight dip for Marathi, a precise qualitative analysis reveals a complex scenario where LoRA corrects a significant domain-specific error but either introduces or fails to rectify crucial conceptual and phrasing issues, ultimately leading to a nuanced evaluation.

Sentence 1: A basic linear equation is expressed as y = mx + b, which models a straight line on a graph.

- **Reference Translation:**

  एक मूलभूत रेषीय समीकरण y = mx + b म्हणून व्यक्त केले जाते, जे आलेखावर सरळ रेषा तयार करते।

- **Base Model Translation:**

  एक मूलभूत रेषीय समीकरण y = mx + b म्हणून व्यक्त केले जाते, जे आलेखावर सरळ रेषा तयार करते।

- **LoRA Model Translation:**

  एक मूलभूत रेषीय समीकरण y = mx + b म्हणून व्यक्त केले जाते, जे एका आलेखावर एक सरळ रेषा तयार करते।

- **Analysis:**
  - **Base vs. Reference:**

    The Base model's translation is identical to the reference, indicating a strong initial understanding of this sentence structure and vocabulary.

- **LoRA vs. Reference:**

  LoRA introduces a subtle change by using "एका आलेखावर (on a graph)" instead of "आलेखावर (on the graph)". While grammatically correct and conveying the same meaning, it deviates slightly from the reference. The addition of the indefinite article "एका" doesn't necessarily improve accuracy or fluency in this context.

- **LoRA Interpretation:**

  LoRA's minor alteration doesn't introduce an error but shows a slight preference for a more explicit indefinite article, which isn't present in the reference.

Sentence 2: Newton's second law states that force equals mass times acceleration, providing a fundamental principle for understanding motion.

- **Reference Translation:**

  न्यूटनचा दुसरा नियम सांगतो की बल हे वस्तुमान पट त्वरणाइतके असते, जे गती समजून घेण्यासाठी एक मूलभूत सिद्धांत पुरवते.

- **Base Model Translation:**

  न्यूटनचा दुसरा नियम असा आहे की बल हे वस्तुमान वेळेच्या प्रवेगाइतकेच आहे, ज्यामुळे गती समजून घेण्यासाठी एक मूलभूत तत्त्व उपलब्ध होते।

- **LoRA Model Translation:**

  न्यूटनचा दुसरा नियम असा आहे की बल वस्तुमान वेळेच्या प्रवेगाच्या बरोबरीचा असतो, ज्यामुळे गती समजून घेण्यासाठी एक मूलभूत सिद्धांत उपलब्ध होतो।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model makes a significant conceptual error by translating "mass times acceleration" as "वस्तुमान वेळेच्या प्रवेगाइतकेच (mass equivalent to the velocity of time)". It also uses "असा आहे की (is such that)" which is slightly more verbose than the reference's "सांगतो की (says that)". Finally, it uses "तत्त्व (principle)" while the reference uses "सिद्धांत (principle)" - both are acceptable synonyms, but the reference's choice might be preferred in a scientific context. It also uses "उपलब्ध होते (becomes available)" instead of "पुरवते (provides)".

- **LoRA vs. Reference:**

  LoRA also perpetuates the critical conceptual error, translating "mass times acceleration" as "वस्तुमान वेळेच्या प्रवेगाच्या बरोबरीचा असतो (mass is equal to the velocity of time)". It retains "असा आहे की". It does, however, use "सिद्धांत (principle)" aligning with the reference. It also uses "उपलब्ध होतो" which is a slight variation of the Base's "उपलब्ध होते" but still differs from the reference's "पुरवते".

- **LoRA Interpretation:**

  While LoRA adopts the reference's "सिद्धांत," it fails to correct the fundamental and critical mistranslation of a core physics concept. Its phrasing "वेळेच्या प्रवेगाच्या बरोबरीचा असतो" is arguably no better, and potentially slightly more awkward, than the Base model's "वेळेच्या प्रवेगाइतकेच आहे". This persistent error significantly impacts the perceived accuracy.

Sentence 3: The principle of conservation of energy explains how energy transforms between kinetic and potential forms in isolated systems.

- **Reference Translation:**

  ऊर्जा संरक्षणाचे सिद्धांत स्पष्ट करतो की ऊर्जा वेगळ्या प्रणालींमध्ये गतिशील आणि संभाव्य रूपांमध्ये कसे परिवर्तन करते।

- **Base Model Translation:**

  ऊर्जा संवर्धनाचे तत्त्व वेगळे केलेल्या प्रणालींमध्ये चुंबकीय आणि संभाव्य प्रकारांमधील ऊर्जा परिवर्तन कसे होते हे स्पष्ट करते।

- **LoRA Model Translation:**

  वेगळ्या प्रणालींमध्ये कायनेटिक आणि संभाव्यता प्रकारांमधील ऊर्जा परिवर्तन कसे होतात हे ऊर्जेच्या संवर्धनाचे सिद्धांत स्पष्ट करते।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model incorrectly translates "kinetic" as "चुंबकीय (magnetic)", a severe domain error. It uses "तत्त्व (principle)" while the reference uses "सिद्धांत (principle)". It also uses "वेगळे केलेल्या प्रणालींमध्ये (in separated systems)" which is close to the reference's "वेगळ्या प्रणालींमध्ये (in different systems)". The verb phrasing also differs.

- **LoRA vs. Reference:**

  LoRA correctly translates "kinetic" as "कायनेटिक" (a transliteration), demonstrating a significant improvement in domain accuracy. It uses "सिद्धांत (principle)" aligning with the reference. It also uses "वेगळ्या प्रणालींमध्ये (in different systems)" matching the reference. The verb phrasing "कसे होतात हे स्पष्ट करते" is closer to the reference's "कसे परिवर्तन करते" than the Base model's.

- **LoRA Interpretation:**

  LoRA's correction of "kinetic" is a crucial improvement that directly addresses a significant domain inaccuracy present in the Base model. This single correction is a strong indicator of LoRA's ability to learn domain-specific terminology.

Sentence 4: A simple representation of a chemical reaction is A + B = C, illustrating how reactants combine to form a product.

- **Reference Translation:**

  रासायनिक अभिक्रियेचे एक साधे प्रतिनिधित्व A + B = C आहे, जे अभिकारक कशा प्रकारे एकत्र येऊन उत्पादन तयार करतात हे दर्शवते।

- **Base Model Translation:**

  ए + बी = सी हे रासायनिक प्रतिक्रियेचे एक साधे प्रतिनिधित्व आहे, जे उत्पादन तयार करण्यासाठी प्रतिक्रियाशील पदार्थ कसे एकत्र येतात याचे चित्रण करते।

- **LoRA Model Translation:**

  ए + बी = सी हे रासायनिक प्रतिक्रियेचे एक साधे प्रतिनिधित्व आहे, ज्यात ए उत्पादन तयार करण्यासाठी प्रतिक्रियांचे संयोजन कसे केले जाते हे स्पष्ट केले आहे।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model uses "रासायनिक प्रतिक्रियेचे (of chemical reaction)" while the reference uses "रासायनिक अभिक्रियेचे (of chemical reaction)". Both are acceptable. "एक साधे प्रतिनिधित्व A + B = C आहे (is a simple representation A + B = C)" is consistent. The Base model uses "प्रतिक्रियाशील पदार्थ कसे एकत्र येतात याचे चित्रण करते (depicts how reactive substances come together)" which conveys the meaning but differs in phrasing from the reference's "अभिकारक कशा प्रकारे एकत्र येऊन उत्पादन तयार करतात हे दर्शवते (shows how reactants come together to produce a product)".

- **LoRA vs. Reference:**

  LoRA uses "रासायनिक प्रतिक्रियेचे (of chemical reaction)". It uses "ए उत्पादन तयार करण्यासाठी प्रतिक्रियांचे संयोजन कसे केले जाते हे स्पष्ट केले आहे (it has been explained how the combination of reactions is done to produce a product)". This phrasing is more verbose and less direct than the reference. It uses "ज्यात (in which)" which doesn't directly correspond to the reference's "जे (which)".

- **LoRA Interpretation:**

  While LoRA attempts to clarify the process, its phrasing becomes more complex and less natural compared to the reference. The Base model's description, though different, is arguably closer to the reference's conciseness and flow.

Sentence 5: DNA replication is a vital process that ensures the accurate transmission of genetic information during cell division.

- **Reference Translation:**

  डीएनए प्रतिकृती ही एक महत्त्वाची प्रक्रिया आहे जी पेशी विभाजन दरम्यान आनुवंशिक माहितीचे अचूक प्रसारण सुनिश्चित करते।

- **Base Model Translation:**

  डी. एन. ए. प्रतिकृतीकरण ही एक महत्त्वाची प्रक्रिया आहे जी पेशी विभाजनादरम्यान अनुवांशिक माहितीचे अचूक प्रसारण सुनिश्चित करते।

- **LoRA Model Translation:**

  डी. एन. ए. प्रतिकृती ही एक महत्त्वपूर्ण प्रक्रिया आहे जी पेशी विभाजन दरम्यान अनुवांशिक माहितीचे अचूक प्रसारण सुनिश्चित करते।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model uses "प्रतिकृतीकरण (replication)" which is a slightly longer form of the reference's "प्रतिकृती (replication)". It uses "पेशी विभाजनादरम्यान (during cell division)" while the reference uses "पेशी विभाजन दरम्यान (during cell division)" - the Base includes an extra suffix.

  - **LoRA vs. Reference:**

    LoRA uses "प्रतिकृती (replication)" aligning with the reference. It uses "पेशी विभाजन दरम्यान (during cell division)" also matching the reference.

- **LoRA Interpretation:**

  LoRA shows a slight improvement by using the more concise terminology ("प्रतिकृती") and the exact phrasing for "during cell division" as the reference.

Conclusion: The Marathi results present a nuanced picture. While the LoRA model demonstrates a crucial ability to correct domain-specific errors, as seen with "kinetic," it fails to address the fundamental and significant conceptual error in the translation of Newton's second law, a core physics principle. Furthermore, in Sentence 4, LoRA's phrasing becomes more convoluted compared to the reference. The minor improvements in Sentence 5 and the stylistic variation in Sentence 1 are overshadowed by these critical inconsistencies. The slight dip in the chrF score likely reflects the impact of the persistent conceptual error, which outweighs the positive contribution of the "kinetic" correction. This suggests that while LoRA can learn specific domain terms, it may struggle with more complex conceptual understanding or require more targeted data to rectify such errors.

4.3.3 Gujrati (LoRA chrF: 69.3043, Base chrF: 61.2306)

Overall Impression: Gujarati shows a substantial improvement with the LoRA model, indicating a significant enhancement in accuracy, particularly in handling scientific concepts and aligning more closely with the reference translations.

Sentence 1: A basic linear equation is expressed as y = mx + b, which models a straight line on a graph.

- **Reference Translation:**

  એક મૂળભૂત રેખીય સમીકરણને y = mx + b તરીકે દર્શાવવામાં આવે છે, જે ગ્રાફ પરની એક સીધી રેખાનું મોડેલ છે।

- **Base Model Translation:**

  એક મૂળભૂત રેખીય સમીકરણ y = mx + b તરીકે વ્યક્ત થાય છે, જે આલેખ પર સીધી રેખા રજૂ કરે છે।

- **LoRA Model Translation:**

  એક મૂળભૂત રેખીય સમીકરણ y = mx + b તરીકે વ્યક્ત કરવામાં આવે છે, જે આલેખ પર એક સીધી રેખા રજૂ કરે છે।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model uses "વ્યક્ત થાય છે (is expressed)" which differs from the reference's "દર્શાવવામાં આવે છે (is shown/expressed)". It uses "આલેખ પર સીધી રેખા રજૂ કરે છે (presents a straight line on the graph)" while the

reference uses "ગ્રાફ પરની એક સીધી રેખાનું મોડેલ છે (is a model of a straight line on the graph)".

- **LoRA vs. Reference:**

  LoRA uses "વ્યક્ત કરવામાં આવે છે (is expressed)" which is closer to the reference's "દર્શાવવામાં આવે છે". It uses "આલેખ પર એક સીધી રેખા રજૂ કરે છે (presents a straight line on the graph)" similar to the Base model but with the addition of "એક (a)". This still differs from the reference's phrasing and verb "મોડેલ છે (is a model)".

- **LoRA Interpretation:**

  LoRA shows a slight improvement in the verb used to express the equation, aligning more with the reference. However, both models deviate from the reference in the latter part of the sentence, particularly in the use of "રજૂ કરે છે (presents)" instead of "મોડેલ છે (is a model)".

Sentence 2: Newton's second law states that force equals mass times acceleration, providing a fundamental principle for understanding motion.

- **Reference Translation:**

  ન્યૂટનનો બીજો નિયમ જણાવે છે કે બળ દળ ગુણ્યા પ્રવેગ સમાન હોય છે, જે ગતિ માટે એક મૂળભૂત સિધ્ધાંત છે।

- **Base Model Translation:**

  ન્યૂટનનો બીજો નિયમ એ બળને સામૂહિક સમયની ગતિ સાથે સરખાવે છે, જે સમજવાની હિલચાલ માટે મૂળભૂત સિધ્ધાંત પૂરો પાડે છે।

- **LoRA Model Translation:**

  ન્યૂટનનો બીજો નિયમ જણાવે છે કે બળ દળ વખત પ્રવેગને સમાન છે, જે ગતિને સમજવા માટે એક મૂળભૂત સિધ્ધાંત પ્રદાન કરે છે।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model's translation is fundamentally incorrect, stating that the law "compares force with the speed of collective time". This shows a complete misunderstanding of Newton's second law.

  - **LoRA vs. Reference:**

    LoRA correctly translates the core of Newton's second law as "બળ દળ વખત પ્રવેગને સમાન છે (force is equal to mass times acceleration)". This aligns closely with the reference's "બળ દળ ગુણ્યા પ્રવેગ સમાન હોય છે (force is equal to mass times acceleration)". Both models use "ન્યૂટનનો બીજો નિયમ જણાવે

છે કે (Newton's second law states that)". The reference uses "ગતિ માટે એક મૂળભૂત સિધ્દ્ધાંત છે (is a fundamental principle for motion)" while LoRA uses "ગતિને સમજવા માટે એક મૂળભૂત સિધ્દ્ધાંત પ્રદાન કરે છે (provides a fundamental principle for understanding motion)". The Base model's ending is "સમજવાની હિલચાલ માટે મૂળભૂત સિધ્દ્ધાંત પૂરો પાડે છે (provides a fundamental principle for understanding movement)".

- **LoRA Interpretation:**

  LoRA demonstrates a dramatic and crucial improvement by correctly translating Newton's second law, rectifying a major conceptual error present in the Base model. While the endings differ slightly from the reference, LoRA accurately captures the core scientific meaning.

Sentence 3: The principle of conservation of energy explains how energy transforms between kinetic and potential forms in isolated systems.

- **Reference Translation:**

  ઊર્જા સંરક્ષણનો સિધ્દ્ધાંત સમજાવે છે કે અલગ કરેલા તંત્રમાં ગતિ અને સંભવિત સ્વરૂપો વચ્ચે ઊર્જા કેવી રીતે પરિવર્તિત થાય છે।

- **Base Model Translation:**

  ઊર્જાના સંરક્ષણનું સિધ્દ્ધાંત સમજાવે છે કે કેવી રીતે અલગ પ્રણાલીઓમાં ગતિશીલ અને સંભવિત સ્વરૂપો વચ્ચે ઊર્જા પરિવર્તન થાય છે।

- **LoRA Model Translation:**

  ઊર્જાના સંરક્ષણનું સિધ્દ્ધાંત સમજાવે છે કે અલગ પ્રણાલીઓમાં કાઇનેટિક અને સંભવિત સ્વરૂપો વચ્ચે ઊર્જા રૂપાંતર કેવી રીતે થાય છે।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model uses "ઊર્જાના સંરક્ષણનું સિધ્દ્ધાંત સમજાવે છે કે કેવી રીતે (the principle of conservation of energy explains how)" which is similar to the reference's "ઊર્જા સંરક્ષણનો સિધ્દ્ધાંત સમજાવે છે કે (the principle of conservation of energy explains that)". The Base model uses "અલગ પ્રણાલીઓમાં ગતિશીલ અને સંભવિત સ્વરૂપો વચ્ચે ઊર્જા પરિવર્તન થાય છે (energy transformation happens between dynamic and potential forms in different systems)". The reference uses "અલગ કરેલા તંત્રમાં ગતિ અને સંભવિત સ્વરૂપો વચ્ચે ઊર્જા કેવી રીતે પરિવર્તિત થાય છે (how energy transforms between kinetic and potential forms in isolated systems)". The Base model incorrectly uses

"ગતિશીલ (dynamic)" for "kinetic" and "ઊર્જા પરિવર્તન થાય છે (energy transformation happens)" which is less precise than "ઊર્જા કેવી રીતે પરિવર્તિત થાય છે (how energy transforms)".

- **LoRA vs. Reference:**

  LoRA uses "ઊર્જાના સંરક્ષણનું સિધ્દાંત સમજાવે છે કે કેવી રીતે (the principle of conservation of energy explains how)". It correctly translates "kinetic" as "કાઇનેટિક". It uses "અલગ પ્રણાલીઓમાં કાઇનેટિક અને સંભવિત સ્વરૂપો વચ્ચે ઊર્જા રૂપાંતર કેવી રીતે થાય છે (how energy transformation happens between kinetic and potential forms in different systems)". The reference uses "અલગ કરેલા તંત્રમાં (in isolated systems)" while LoRA uses "અલગ પ્રણાલીઓમાં (in different systems)". LoRA's "ઊર્જા રૂપાંતર કેવી રીતે થાય છે (how energy transformation happens)" is closer to the reference's "ઊર્જા કેવી રીતે પરિવર્તિત થાય છે (how energy transforms)" than the Base model.

- **LoRA Interpretation:**

  LoRA shows a significant improvement by correctly translating "kinetic". While it uses "અલગ પ્રણાલીઓમાં" instead of the reference's "અલગ કરેલા તંત્રમાં" and "ઊર્જા રૂપાંતર કેવી રીતે થાય છે" instead of "ઊર્જા કેવી રીતે પરિવર્તિત થાય છે", it demonstrates a better understanding of the domain terminology compared to the Base model.

Sentence 4: A simple representation of a chemical reaction is A + B = C, illustrating how reactants combine to form a product.

- **Reference Translation:**

  રાસાયણિક ક્રિયાનું એક સરળ પ્રતિનિધિત્વ A + B = C છે, જે દર્શાવે છે કે પ્રક્રિયકો કેવી રીતે જોડાણ કરીને ઉત્પાદન બનાવે છે।

- **Base Model Translation:**

  રાસાયણિક પ્રતિક્રિયાનું એક સરળ પ્રતિનિધિત્વ એ + બી = સી છે, જે ઉત્પાદન બનાવવા માટે પ્રતિક્રિયાઓ કેવી રીતે જોડાય છે તેનું વર્ણન કરે છે।

- **LoRA Model Translation:**

  એક રાસાયણિક પ્રતિક્રિયાનું એક સરળ પ્રતિનિધિત્વ એ + બી = સી છે, જે દર્શાવે છે કે કેવી રીતે પ્રતિક્રિયાઓને જોડીને ઉત્પાદન બને છે।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model uses "રાસાયણિક પ્રતિક્રિયાનું (of chemical reaction)" while the reference uses "રાસાયણિક ક્રિયાનું (of chemical reaction)". Both are

acceptable. "એક સરળ પ્રતિનિધિત્વ એ + બી = સી છે (is a simple representation A + B = C)" is consistent. The Base model uses "ઉત્પાદન બનાવવા માટે પ્રતિક્રિયાઓ કેવી રીતે જોડાય છે તેનું વર્ણન કરે છે (describes how reactions join to make a product)". The reference uses "જે દર્શાવે છે કે પ્રક્રિયકો કેવી રીતે જોડાણ કરીને ઉત્પાદન બનાવે છે (which shows how reactants combine to produce a product)".

- **LoRA vs. Reference:**

  LoRA uses "એક રાસાયણિક પ્રતિક્રિયાનું (of a chemical reaction)". "એક સરળ પ્રતિનિધિત્વ એ + બી = સી છે (is a simple representation A + B = C)" is consistent. LoRA uses "જે દર્શાવે છે કે કેવી રીતે પ્રતિક્રિયાઓને જોડીને ઉત્પાદન બને છે (which shows how by joining reactions, a product is formed)". This is very close to the reference's meaning, though the wording differs slightly ("પ્રક્રિયકો કેવી રીતે જોડાણ કરીને" vs. "કેવી રીતે પ્રતિક્રિયાઓને જોડીને").

- **LoRA Interpretation:**

  LoRA's translation is quite close to the reference, capturing the essence of how reactants combine to form a product. The difference in the subject ("પ્રક્રિયકો" vs. "પ્રતિક્રિયાઓને") and the verb ("જોડાણ કરીને" vs. "જોડીને") are minor variations in expressing the same concept.

Sentence 5: DNA replication is a vital process that ensures the accurate transmission of genetic information during cell division.

- **Reference Translation:**

  DNA નકલ એ એક મહત્વપૂર્ણ પ્રક્રિયા છે, જે કોષ વિભાજન દરમિયાન આનુવંશિક માહિતીનું સાચું પ્રસારણ સુનિશ્ચિત કરે છે।

- **Base Model Translation:**

  ડીએનએ પ્રતિકૃતિ એ એક મહત્વપૂર્ણ પ્રક્રિયા છે જે કોષ વિભાજન દરમિયાન આનુવંશિક માહિતીનું સચોટ પ્રસારણ સુનિશ્ચિત કરે છે।

- **LoRA Model Translation:**

  ડીએનએ નકલ એ એક મહત્વપૂર્ણ પ્રક્રિયા છે જે કોષ વિભાજન દરમિયાન આનુવંશિક માહિતીનું સચોટ પ્રસારણ સુનિશ્ચિત કરે છે।

- **Analysis:**

  - The Base model uses "ડીએનએ પ્રતિકૃતિ (DNA replication)" while the reference and LoRA model use "DNA નકલ (DNA copy/replication)". All three convey the correct meaning. The rest of the sentence is very similar across all three, with minor variations in word choice ("મહત્વપૂર્ણ" vs. "મહત્વની").

- **LoRA Interpretation:**

  LoRA aligns perfectly with the reference translation for "DNA replication" using "DNA ખંડ". The rest of the sentence is highly similar, indicating strong performance from both models on this sentence.

Conclusion: The LoRA model demonstrates a remarkable improvement in Gujarati, primarily by correcting the fundamental conceptual error in the translation of Newton's second law, which was a major failure of the Base model. It also shows a significant improvement in handling the term "kinetic". While there are still some variations in phrasing and vocabulary compared to the reference translations in other sentences, the core scientific meanings are generally well-preserved, and in key instances like Newton's law and "kinetic," LoRA shows a clear advantage. This substantial improvement in accuracy, particularly for critical scientific concepts, directly accounts for the significant increase in the chrF score.

4.3.4 Tamil (LoRA chrF: 69.9149, Base chrF: 67.3633)

Overall Impression: Tamil shows a moderate but noticeable improvement with the LoRA model. The enhancements focus on clearer phrasing that often aligns better with standard scientific terminology in Tamil, and subtle improvements in grammatical coherence, leading to translations that feel slightly more refined.

Sentence 1: A basic linear equation is expressed as y = mx + b, which models a straight line on a graph.

- **Reference Translation:**

  ஒரு அடிப்படை நேரியல் சமன்பாடு y = mx + b என வெளிப்படுத்தப்படுகிறது; இது ஒரு வரைபடத்தில் உள்ள நேர்கோட்டை மாதிரியாக்குகிறது।

- **Base Model Translation:**

  ஒரு அடிப்படை நேரியல் சமன்பாடு y = mx + b என வெளிப்படுத்தப்படுகிறது, இது ஒரு வரைபடத்தில் ஒரு நேர் கோட்டை வடிவமைக்கிறது।

- **LoRA Model Translation:**

  ஒரு அடிப்படை நேரியல் சமன்பாடு y = mx + b என வெளிப்படுத்தப்படுகிறது, இது ஒரு வரைபடத்தில் ஒரு நேர் கோட்டை மாதிரியாக்குகிறது।

- **Analysis:**
    - **Base vs. Reference:**

      The Base model uses "வடிவமைக்கிறது (designs/models)" for "models," while the reference uses "மாதிரியாக்குகிறது (models)". Both are valid translations. The reference uses "உள்ள (in/within)" before "நேர்கோட்டை (straight line)" which is absent in the Base model.

    - **LoRA vs. Reference:**

      LoRA uses "மாதிரியாக்குகிறது (models)" which directly matches the reference. It also uses "ஒரு (a)" before "நேர் கோட்டை (straight line)" which is present in the reference as "ஒரு நேர்கோட்டை".

    - **LoRA Interpretation:**

      LoRA shows a slight improvement by using the same verb "மாதிரியாக்குகிறது" as the reference, which could be considered more standard in a technical context. It also aligns with the reference in using "ஒரு" before "நேர்கோட்டை".

Sentence 2: Newton's second law states that force equals mass times acceleration, providing a fundamental principle for understanding motion.

- **Reference Translation:**

  நியூட்டனின் இரண்டாவது விதி, விசை என்பது நிறை மடங்கு முடுக்கத்திற்கு சமம் என்று கூறுகிறது; இது இயக்கத்தைப் புரிந்துகொள்ள ஒரு அடிப்படை கொள்கையாகும்।

- **Base Model Translation:**

  நியூட்டனின் இரண்டாவது சட்டம் வெகுஜன நேர முடுக்கத்திற்கு சமமான சக்தியை நிலைநிறுத்துகிறது, இது இயக்கத்தைப் புரிந்துகொள்வதற்கான ஒரு அடிப்படை கொள்கையை வழங்குகிறது।

- **LoRA Model Translation:**

  நியூட்டனின் இரண்டாவது சட்டம், இயக்கத்தைப் புரிந்துகொள்வதற்கான ஒரு அடிப்படை கொள்கையை வழங்குவதன் மூலம், வெகுஜன நேர துரிதத்தை சமமான சக்தி என்று கூறுகிறது।

- **Analysis:**
    - **Base vs. Reference:**

      The Base model uses "சட்டம் (law)" instead of "விதி (law)". It translates "mass times acceleration" as "வெகுஜன நேர முடுக்கத்திற்கு (force equivalent to mass time acceleration)", which is conceptually correct but uses "சக்தி (force)" where the reference uses "விசை (force)". The Base model's phrasing "சக்தியை நிலைநிறுத்துகிறது (establishes force)" is less direct than the reference's "கூறுகிறது (says)". It uses "புரிந்துகொள்வதற்கான (for understanding)" and "வழங்குகிறது (provides)" which align with the reference's "புரிந்துகொள்ள (to understand)" and "கொள்கையாகும் (is a principle)".

    - **LoRA vs. Reference:**

      LoRA also uses "சட்டம் (law)". It translates "mass times acceleration" as "வெகுஜன நேர துரிதத்தை சமமான சக்தி என்று கூறுகிறது (says that force is equal to mass time acceleration)", using "துரிதத்தை (acceleration)" which could be considered more contemporary than the reference's "முடுக்கத்திற்கு (to acceleration)". It also uses "சக்தி (force)" instead of "விசை". LoRA introduces "வழங்குவதன் மூலம் (by providing)" which explicitly links the law to the principle, enhancing the flow.

    - **LoRA Interpretation:**

      LoRA's phrasing "சக்தி என்று கூறுகிறது" is more direct and closer to the reference's "கூறுகிறது". The addition of "வழங்குவதன் மூலம்" improves the sentence's coherence. While both models use "சட்டம்" and "சக்தி", LoRA's choice of "துரிதத்தை" might be seen as a slight improvement in modern scientific terminology.

Sentence 3: The principle of conservation of energy explains how energy transforms between kinetic and potential forms in isolated systems.

- **Reference Translation:**

  ஆற்றல் பாதுகாப்பு கொள்கை, தனிமைப்படுத்தப்பட்ட அமைப்புகளில் இயக்க மற்றும் சாத்தியமான வடிவங்களுக்கிடையில் ஆற்றல் எவ்வாறு மாற்றமடைகிறது என்பதை விளக்குகிறது।

- **Base Model Translation:**

  தனிமைப்படுத்தப்பட்ட அமைப்புகளில் இயக்கவியல் மற்றும் சாத்தியமான வடிவங்களுக்கிடையே ஆற்றல் மாற்றங்கள் எவ்வாறு நிகழ்கின்றன என்பதை ஆற்றல் பாதுகாப்புக் கொள்கை விளக்குகிறது।

- **LoRA Model Translation:**

  தனித்தனி அமைப்புகளில் இயக்கவியல் மற்றும் சாத்தியமான வடிவங்களுக்கு இடையில் ஆற்றல் மாற்றங்கள் எவ்வாறு மாறுகின்றன என்பதை ஆற்றல் பாதுகாப்பின் கொள்கை விளக்குகிறது।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model uses "ஆற்றல் மாற்றங்கள் எவ்வாறு நிகழ்கின்றன (how energy changes occur)" while the reference uses "ஆற்றல் எவ்வாறு மாற்றமடைகிறது (how energy transforms)". "நிகழ்கின்றன" and "மாற்றமடைகிறது" are semantically close. The Base model uses "இடையே (between)" which is similar to the reference's "இடையில் (between)".

  - **LoRA vs. Reference:**

    LoRA uses "தனித்தனி அமைப்புகளில் (in separate systems)" instead of the reference's "தனிமைப்படுத்தப்பட்ட அமைப்புகளில் (in isolated systems)". It uses "இயக்கவியல் மற்றும் சாத்தியமான வடிவங்களுக்கு இடையில் (between kinetic and potential forms)" which has a slightly different grammatical structure than the reference's "இயக்க மற்றும் சாத்தியமான வடிவங்களுக்கிடையில்". LoRA uses "ஆற்றல் மாற்றங்கள் எவ்வாறு மாறுகின்றன (how energy changes transform)" where the reference uses "ஆற்றல் எவ்வாறு மாற்றமடைகிறது (how energy transforms)".

  - **LoRA Interpretation:**

    LoRA's choice of "தனித்தனி" might not be as precise as the reference's "தனிமைப்படுத்தப்பட்ட" for "isolated". The verb "மாறுகின்றன" is a valid translation for "transforms" but differs from the reference. The grammatical structure around "kinetic and potential forms" is also slightly different. In this sentence, the Base model seems to align more closely with the reference's vocabulary and grammatical structure.

Sentence 4: A simple representation of a chemical reaction is A + B = C, illustrating how reactants combine to form a product.

- **Reference Translation:**

  ஒரு வேதி வினையின் எளிய பிரதிநிதித்துவம் A + B = C ஆகும்; இது வினைப்பொருட்கள் எவ்வாறு இணைந்து ஒரு விளைபொருளை உருவாக்குகின்றன என்பதை விளக்குகிறது।

- **Base Model Translation:**

  ஒரு இரசாயன எதிர்வினையின் எளிய பிரதிநிதித்துவம் A + B = C, எதிர்வினைகளை எவ்வாறு இணைப்பது என்பதை விளக்கி ஒரு தயாரிப்பை உருவாக்குகிறது।

- **LoRA Model Translation:**

  ஒரு இரசாயன எதிர்வினையின் ஒரு எளிய பிரதிநிதித்துவம் A + B = C ஆகும், இது ஒரு தயாரிப்பை உருவாக்க எதிர்வினைகளை எவ்வாறு இணைக்கிறது என்பதை விளக்குகிறது।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model uses "இரசாயன எதிர்வினையின் (of chemical reaction)" while the reference uses "வேதி வினையின் (of chemical reaction)". Both are acceptable. The Base model's second clause "எதிர்வினைகளை எவ்வாறு இணைப்பது என்பதை விளக்கி ஒரு தயாரிப்பை உருவாக்குகிறது (explaining how to combine reactions creates a product)" is less direct and slightly awkward compared to the reference.

  - **LoRA vs. Reference:**

    LoRA also uses "இரசாயன எதிர்வினையின்". It introduces "ஆகும் (is)" after "C", aligning with the reference. Its second clause "இது ஒரு தயாரிப்பை உருவாக்க எதிர்வினைகளை எவ்வாறு இணைக்கிறது என்பதை விளக்குகிறது (this explains how to combine reactions to create a product)" is clearer than the Base model's and uses the pronoun "இது" as in the reference, improving coherence.

  - **LoRA Interpretation:**

    LoRA shows a clear improvement in grammatical coherence by using "இது" to connect the two clauses, similar to the reference. The addition of "ஆகும்" also aligns with the reference's structure.

Sentence 5: DNA replication is a vital process that ensures the accurate transmission of genetic information during cell division.

- **Reference Translation:**

  DNA நகல் என்பது ஒரு முக்கியமான செயல்முறை, இது செல்லின் பிரிவின்போது மரபணு தகவல்களை துல்லியமாக அனுப்புவதை உறுதி செய்கிறது।

- **Base Model Translation:**

  டிஎன்ஏ பிரதிபலிப்பு என்பது உயிரணு பிரிவின் போது மரபணு தகவல்களின் துல்லியமான பரிமாற்றத்தை உறுதி செய்யும் ஒரு முக்கியமான செயல்முறையாகும்।

- **LoRA Model Translation:**

  டிஎன்ஏ நகல் என்பது ஒரு முக்கியமான செயல்முறையாகும், இது செல் பிரிவின்போது மரபணு தகவல்களின் துல்லியமான பரிமாற்றத்தை உறுதி செய்கிறது।

- **Analysis:**

  - **Base vs. Reference:**

    The Base model uses "டிஎன்ஏ பிரதிபலிப்பு (DNA reflection/replication)" while the reference and LoRA use "DNA நகல் (DNA copy/replication)". The Base model uses "உயிரணு பிரிவு (cell division)" while the reference and LoRA use "செல் பிரிவு (cell division)". The Base model includes an extra suffix "ஆகும்" at the end of the sentence.

  - **LoRA vs. Reference:**

    LoRA aligns with the reference in using "DNA நகல்" and "செல் பிரிவு". It also omits the extra "ஆகும்" at the end, making it more concise and aligning with the reference.

  - **LoRA Interpretation:**

    LoRA demonstrates better alignment with the reference by using the preferred terminology "DNA நகல்" and "செல் பிரிவு", and by having a more concise sentence structure.

Conclusion: The LoRA model provides moderate but valuable improvements in Tamil. It shows a tendency to align more closely with the reference's terminology (e.g., "மாதிரியாக்குகிறது", "நகல்", "செல் பிரிவு") and enhances grammatical coherence, particularly in complex sentences by using pronoun references effectively. While there are instances where the Base model aligns slightly better with the reference (e.g., Sentence 3), overall, LoRA's refinements contribute to translations that are generally clearer, more grammatically sound, and closer to standard scientific expression in Tamil, justifying the noticeable increase in the chrF score.

4.4 Cross-Language Comparison

- **Hindi:** The LoRA model demonstrates the most significant improvement, achieving a high chrF score that is well-supported by the qualitative analysis. Compared to the Base model, LoRA consistently shows a stronger alignment with the reference translations. This includes the adoption of more formal and accurate domain-specific vocabulary (e.g., "द्रव्यमान गुणन त्वरण"), improved and more concise phrasing that often mirrors the reference (e.g., for stating Newton's law), and in one instance, a perfect match with the reference translation ("ऊर्जा संरक्षण का सिद्धांत"). While minor deviations from the reference exist (e.g., the verb in Sentence 1 and the initial adjective for the equation), the overall performance indicates a strong enhancement in the model's grasp of scientific concepts and their expression in Hindi.

- **Marathi:** The LoRA model presents a mixed picture, as reflected in the slight dip in the chrF score. While it successfully corrects a key domain-specific error ("चुंबकीय" to "कायनेटिक"), it fails to address a fundamental conceptual error in the translation of Newton's second law, a critical flaw that persists from the Base model. Furthermore, some instances of slightly awkward or less natural phrasing are observed in comparison to the reference translations. This inconsistency suggests that while LoRA can improve specific aspects of domain terminology, it requires more targeted data or intervention to rectify deeply rooted conceptual misunderstandings and consistently enhance fluency.

- **Gujarati:** LoRA demonstrates the biggest functional leap by correcting major conceptual errors present in the Base model's translations of core scientific principles, most notably Newton's second law. The accurate handling of domain terminology ("કાઇનેટિક") also contributes significantly to the substantial increase in chrF and brings the translation closer to the reference. While some variations in phrasing and vocabulary compared to the reference remain in other sentences, the correction of fundamental scientific concepts highlights LoRA's effectiveness in bridging critical accuracy gaps in Gujarati.

- **Tamil:** LoRA provides moderate but valuable improvements in Tamil. The model shows a tendency to align more closely with the reference's terminology (e.g., "மாதிரியாக்குகிறது", "நகல்", "செல் பிரிவு") and enhances grammatical coherence, particularly in multi-clause sentences through the effective use of pronoun references. While the improvements are not as dramatic as in Hindi or Gujarati, they contribute to translations that are generally clearer, more grammatically sound, and exhibit a style that is more appropriate for scientific content, moving towards the nuances of the reference translations.

Overall Takeaways:

- **The LoRA fine-tuning has a varying impact across the four languages, likely influenced by factors such as the quality and quantity of available domain-specific data and the inherent linguistic complexities of each language.**
- **Hindi and Gujarati clearly benefit the most from LoRA, showcasing significant improvements in both accuracy and fluency, particularly in the critical domain of scientific terminology and conceptual understanding, leading to a much stronger alignment with the provided reference translations.**
- **Tamil experiences moderate but positive changes, indicating that LoRA assists in refining the nuances of scientific expression, bringing the translations closer to the reference in terms of terminology and grammatical structure.**
- **Marathi's slight dip in chrF, despite some domain-specific corrections, underscores the critical importance of addressing fundamental conceptual errors during fine-tuning. It suggests that while LoRA can improve certain aspects like specific vocabulary, it might not always resolve deeply ingrained translation issues or consistently enhance fluency without more targeted intervention or a more comprehensive and focused dataset.**

5. Limitations and Future Improvements

5.1 Current Limitations

1. **Data Size and Domain Coverage**
    - My core dataset comprises 20 English sentences extracted from NEET 2020 question papers (Physics, Chemistry, and Biology), along with their Hindi, Gujarati, Marathi, and Tamil translations. While these sentences capture important scientific terms, they represent only a narrow slice of potential educational topics. While these questions provide a technical, domain-specific sample, the dataset remains relatively small and predominantly exam-oriented. Additionally, I generated a synthetic corpus of ~50 English sentences using GPT4o, which helped diversify content but remains modest in scale. Consequently, the resulting translations may not generalize extensively to broader educational or real-world contexts.
2. **Partial Exploration and Potential Overfitting**
    - Resource/time constraints limited extensive experimentation. Although LoRA fine-tuning boosted performance for some languages (e.g., Hindi, Gujarati), Marathi saw marginal net declines. This discrepancy suggests partial overshadowing or overfitting to the dominant language patterns in the dataset.
3. **Subjectivity in Qualitative Analysis**
    - While chrF offers a robust quantitative measure, fluency and stylistic naturalness are harder to quantify. With limited time, my human checks prioritized technical accuracy ("kinetic," "homeostasis," etc.) and correctness over broader stylistic considerations.
4. **Single GPU and Compressed Timeline**
    - This assignment occurred over two days with one GPU setup, constraining the scope of hyper-parameter tuning, large-scale data augmentation, or ensemble-based approaches that might refine accuracy further.

5.2 Potential Improvements

1. **Expanded Data Augmentation and Domain Diversity**
    - **Scaling Synthetic Generation:** Extend beyond 50 GPT4o-generated sentences to hundreds or thousands, further reducing data sparsity. This could include more varied academic topics—arts, literature, social sciences—to broaden domain coverage.
    - **Back-Translation / Active Learning:** Repeatedly generate and verify new pairs (English ↔ Indic languages), focusing on low-resource pairs like Marathi.

Interactively refine challenging terms or constructions via active learning loops.

2. **Deeper Domain-Specific Integration**
    - **Specialized Glossaries:** Curate specialized lists (e.g., key physics or biology terms) for Hindi, Gujarati, Marathi, and Tamil to ensure consistent usage and reduce mistranslations.
    - **Knowledge Graphs or Extended Corpora:** Incorporate more extensive references (e.g., public domain textbooks, curated academic websites) to capture advanced linguistic structures and a wider set of scientific expressions.
3. **Improving Fluency and Naturalness**
    - **Structured Post-Editing:** Engage bilingual subject-matter experts to correct translations in real-time, feeding these edits back into iterative fine-tuning. This approach is especially beneficial where domain precision is critical.
    - **Reinforcement Learning:** Gather feedback on outputs for style and clarity, adjusting model weights to align with user preferences or established academic norms.
    - **Document-Level Context:** For longer educational passages, incorporate multi-sentence context to minimize abrupt transitions and enhance cohesive explanations.
4. **Advanced Architectures and Iterative Fine-Tuning**
    - **Larger Transformer Models:** Investigate instruction-tuned or next-generation GPT-based models for deeper domain coverage and improved morphological handling.
    - **Custom LoRA Configurations:** Explore varying LoRA ranks, alpha values, or target modules on a per-language basis to avoid overshadowing issues observed in Marathi.
    - **Ensemble Approaches:** Combine the outputs of multiple specialized models (e.g., a “Marathi-focused” variant plus a general domain model) to yield more robust final translations.

By further expanding the dataset particularly for the underrepresented languages, refining domain knowledge integration, and applying iterative improvements in model architectures and post-editing strategies, we can enhance both the accuracy and linguistic fluency of these educational translations for a broader academic landscape.

6. Usage

This section outlines the steps to utilize the provided pipeline for fine-tuning and performing inference with the IndicTrans2 model for English to Hindi, Marathi, Gujarati, and Tamil translation, enhanced with LoRA adapters.

6.1 Virtual Environment Setup

It is highly recommended to work within a virtual environment to manage project dependencies effectively.

1. **Create a virtual environment:** Navigate to the project's root directory in your terminal and execute:

   python3 -m venv venv

   or

   python -m venv venv

2. **Activate the virtual environment:**
   source venv/bin/activate

Your terminal prompt should now be prefixed with the environment name (e.g., (venv)).

6.2 Setup

Initialize the environment and install all necessary dependencies by running the setup.sh script:

bash setup.sh

This script performs the following actions:

- Unzips dependency archives located in the dependencies/ folder.
- Installs the indic_nlp_library.
- Installs the IndicTransToolkit along with its required dependencies.
- Installs additional requirements specified in dependencies/requirements.txt.
- Compiles Cython extensions for the IndicTransToolkit.
- Downloads necessary NLTK tokenizer data.
- Executes the installation script for the IndicTrans2 Hugging Face interface (IndicTrans2/huggingface_interface/install.sh).
- Reinstalls IndicTransToolkit to incorporate **any local modifications.**

6.3 Training LoRA Adapters

To fine-tune LoRA adapters for your specific scientific domain, use the train.sh script with appropriate options:

bash train.sh [OPTIONS]

The training data should be organized in the following directory structure under the path specified by the -d flag (default: datasets/train_validation/en-indic-exp):

```
en-indic-exp/
├── train/
│   ├── eng_Latn-hin_Deva/
│   │   ├── train.eng_Latn
│   │   └── train.hin_Deva
│   ├── eng_Latn-tam_Taml/
│   │   └── ...
│   └── {src_lang}-{tgt_lang}/
│       ├── train.{src_lang}
│       └── train.{tgt_lang}
└── dev/
    ├── eng_Latn-hin_Deva/
    │   ├── dev.eng_Latn
    │   └── dev.hin_Deva
    ├── eng_Latn-tam_Taml/
    │   └── ...
    └── {src_lang}-{tgt_lang}/
        ├── dev.{src_lang}
        └── dev.{tgt_lang}
```

Key parameters for the train.sh script include:

Flag	Description	Default
-d <dir>	Path to training dataset	datasets/train_validation/en-indic-exp
-m <model_name>	Base model name	ai4bharat/indictrans2-en-indic-1B
-o <dir>	Output directory for LoRA adapters	lora_adapters/output_<timestamp>
--src <langs>	Comma-separated source language codes	eng_Latn
--tgt <langs>	Comma-separated target language codes	guj_Gujr,hin_Deva,mar_Deva,tam_Taml
-h, --help	Show help message	

Advanced training parameters can be found and modified within the train.sh script.

6.4 Inference

Generate translations using the translate.sh script with various options:

bash translate.sh [OPTIONS]

Here are some usage examples:

- **Translate a string to Hindi using the base model:**

  bash translate.sh -t hindi -s "Welcome!"

- **Translate using a LoRA-adapted model:**
  bash translate.sh -m lora --lora_adapter_dir path/to/lora -t marathi -s "How are you?"

- **Translate to all supported languages:**

  bash translate.sh -a -s "Hello!"

- **Translate using an input JSON file and save results to a file:**
  bash translate.sh -t gujarati -f test_data.json -o output.txt

Key parameters for the translate.sh script include:

Flag	Description
-m <base/lora>	Use base model or LoRA-adapted model
--lora_adapter_dir <dir>	Directory for LoRA adapter (required if -m is lora)
-t <lang>	Target language (hindi, marathi, gujarati, tamil)
-s <text>	Text string to translate
-f <file>	JSON file containing a list of strings to translate
-o <file>	Output file to save the generated translations
-h, --help	Show help message

The input JSON file for translation should be formatted as a list of strings:

```
[
  "This is first sentence.",
  "This is second sentence."
]
```

6.5 Evaluation

Evaluate the translation quality using the evaluation.sh script, which calculates the chrF score:

```
bash evaluation.sh [OPTIONS]
```

Key parameters for the evaluation.sh script include:

Flag	Description
-m, --model_type <base/lora>	Use base or LoRA model
-l, --lora_adapter_dir <dir>	Path to LoRA adapter (required if -m is lora)
-i, --input_file <path>	Path to the evaluation data JSON file (REQUIRED)
-q, --quantization <4-bit/8-bit>	Quantization mode (optional)
-a, --attention <flash_attention_2/eager>	Attention implementation (optional)
-h, --help	Show help message

The evaluation data JSON file should be formatted as a list of dictionaries, where each dictionary contains the English input and the gold-standard reference translations for each target language:

```
[
  {
    "english": "Where is the library?",
    "hindi": "पुस्तकालय कहाँ है?",
    "marathi": "ग्रंथालय कुठे आहे?",
    "gujarati": "લાઇબ્રેરી ક્યાં છે?",
    "tamil": "நூலகம் எங்கே?"
  },
  ...
]
```

The english field serves as the input for the model, and the target language fields contain the reference translations. The chrF score, which indicates the similarity between the model's output and the reference translations, is used as the evaluation metric. Higher chrF scores indicate better translation quality.

6.6 Troubleshooting

Windows-style Line Endings Error

If you encounter errors related to command not found with $'\r'`, it is likely due to Windows-style (CRLF) line endings in the shell scripts. To resolve this, convert the script files to Unix (LF) format using the following commands:

sudo apt-get install dos2unix

dos2unix setup.sh

Repeat for other .sh scripts if necessary (e.g., train.sh, translate.sh, evaluation.sh)

6.7 Notes

- Ensure that all .sh scripts have execute permissions. You can grant these permissions using the command:

  chmod +x *.sh

- Feel free to customize the training and evaluation flags within the respective scripts or via command-line arguments to suit your specific data and experimental setup.
- The provided scripts are modular and can be extended or modified to accommodate more advanced usage scenarios or specific requirements.

7. Conclusion

Through this project, I systematically collected a specialized dataset (20 NEET-based English questions plus a small synthetic corpus of 50 sentences) and performed fine-tuning (LoRA) on an IndicTrans2 model to translate from English to Hindi, Gujarati, Marathi, and Tamil within an educational domain. Despite time and resource constraints, the results demonstrate that adapter-based fine-tuning can significantly improve translation quality—particularly for languages like Hindi and Gujarati—while still presenting opportunities for refinement of other under-represented languages.

Key Outcomes

- **Enhanced Domain Accuracy:** LoRA fine-tuning elevated chrF scores by capturing technical terms and improving sentence coherence.
- **Revealed Gaps:** Marathi performance underscores that even with advanced techniques, low-resource languages can require more targeted data, hyper-parameter tuning, or iterative refinement.

Importance of Accurate Educational Translations:

Accurate, domain-specific translations are crucial in educational settings: they ensure learner comprehension of complex scientific principles, maintain technical fidelity for advanced subjects (e.g., physics, chemistry), and promote language inclusivity. Misinterpretations—such as confusing “magnetic” with “kinetic”—can severely hinder conceptual understanding, underscoring the need for robust translation strategies.

Lessons Learned and Future Applications

- **Domain-Centric Focus:** Concentrating on NEET-style questions proved effective for scientific accuracy; however, broader subject integration (literature, social sciences) would amplify real-world applicability.
- **Iterative Fine-Tuning:** Adapter-based approaches like LoRA allow rapid adaptation while keeping computational costs manageable. Additional cycles of error analysis and post-editing can further refine outputs.
- **Significance of Resource Balancing:** Over- or under-representation of certain languages influences final results, reinforcing the importance of carefully distributing data and fine-tuning parameters.
- **Scalability:** Strategies such as back-translation, advanced transformer models, and reinforcement learning pave the way for more extensive and generalizable solutions in educational and beyond.

In essence, this project illustrates how purposeful data curation, parameter-efficient fine-tuning, and iterative validation can deliver meaningful gains in educational translation quality.

8. References

- **Gala, J., Chitale, P. A., Raghavan, A. K., Gumma, V., Doddapaneni, S., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V., Kumar, P., Khapra, M. M., Dabre, R., & Kunchukuttan, A. (2023). IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. *Transactions on Machine Learning Research*, *0*. Retrieved from https://openreview.net/forum?id=vfT4YuzAYA**
- **AI4Bharat. (n.d.). *IndicTrans2 GitHub Repository*. Retrieved from https://github.com/AI4Bharat/IndicTrans2/tree/main**
- **The IndicTrans2 model, as described in Gala et al. (2023), is available on Hugging Face.**

  @article{gala2023indictrans,

  title={IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages},

  author={Jay Gala and Pranjal A Chitale and A K Raghavan and Varun Gumma and Sumanth Doddapaneni and Aswanth Kumar M and Janki Atul Nawale and Anupama Sujatha and Ratish Puduppully and Vivek Raghavan and Pratyush Kumar and Mitesh M Khapra and Raj Dabre and Anoop Kunchukuttan},

  journal={Transactions on Machine Learning Research},

  issn={2835-8856},

  year={2023},

  url={https://openreview.net/forum?id=vfT4YuzAYA},

  note={}

  }
- **National Testing Agency. (2020). *NEET (UG) 2020 - Archive*. Retrieved from https://neet.nta.nic.in/archive/**