Codemix Generation

Lost in Context team -

Sahasra V. 2022102024

Sreeja P. 2022101081

Shravya P. 2022102077

Machine translation from English to Hinglish

Machine translation from English to Hinglish involves automatically converting English text into Hindi transliterated using the Roman script. It requires natural language processing techniques such as sequence-to-sequence modeling, transliteration, and handling code-mixed data. The system must preserve semantic accuracy while generating syntactically and phonetically correct Hinglish output using Romanized Hindi.

The datasets used are HinGE dataset, and on findnitai/english-to-hinglish dataset from huggingface.

Baseline models

N-gram model

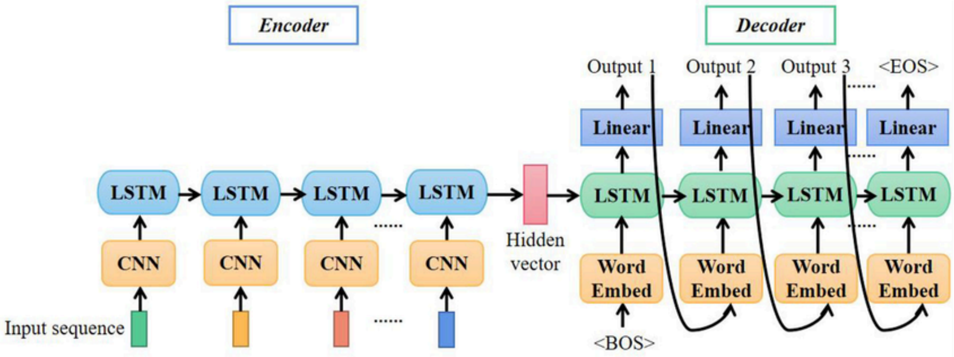
We implemented a n-gram based (n=3) statistical machine translation model for English-to-Hinglish translation using a simple n-gram alignment approach. Instead of relying on complex neural networks, it leverages statistical co-occurrence patterns to learn word- and phrase-level correspondences from a synthetic parallel corpus. The model builds a translation dictionary by identifying frequently co-occurring n-gram pairs in aligned English-Hinglish sentences. During inference, English input sentences are broken into trigrams, and each trigram is replaced by its most probable Hinglish counterpart using the learned mappings. To ensure fluency, overlapping repetitions from consecutive trigrams are merged. This rule-based and interpretable approach offers a lightweight and transparent baseline for machine translation compared to more resource-intensive neural models like transformers.

The core hyperparameter of this model is **n (n-gram size)**:

This determines the size of contiguous word groups used to learn and align patterns between English and Hinglish sentences. In this project, we used trigrams(n = 3), which strike a balance between capturing meaningful phrases (more contextual than unigrams or bigrams) and avoiding data sparsity issues that can occur with higher-order n-grams (like 4-grams or 5-grams).

- Smaller n values (e.g., 1 or 2) may result in more general but less expressive translations, often missing multi-word expressions.
- Larger n values (e.g., 4 or 5) can capture richer context but require more data to avoid sparsity and noise, which can lead to poor generalization.

LSTM Model



The LSTM-based architecture is used for English-to-Hinglish translation. It follows a standard sequence-to-sequence (Seq2Seq) model with attention, allowing the decoder to focus on relevant parts of the input sequence during translation. Pre-trained GloVe embeddings are used for input word representation.

The model architecture consists of an encoder-decoder structure enhanced with attention. The **encoder** uses two stacked LSTM layers, each with 300 hidden units. Input sequences are first passed through an embedding layer initialized with pre-trained GloVe vectors of dimension 300, capturing rich semantic representations. The **decoder** has its own embedding layer and a single LSTM layer, also with 300 hidden units, which generates output tokens sequentially.

To improve translation quality, an A**ttention mechanism** is applied, allowing the decoder to dynamically focus on relevant parts of the input sequence at each timestep. Finally, the decoder outputs are concatenated with attention context vectors and passed through a **time-distributed dense layer** with a softmax activation to predict token probabilities across the vocabulary.

Hyperparameter Tuning:

- latent_dim = 300
- batch_size = 64

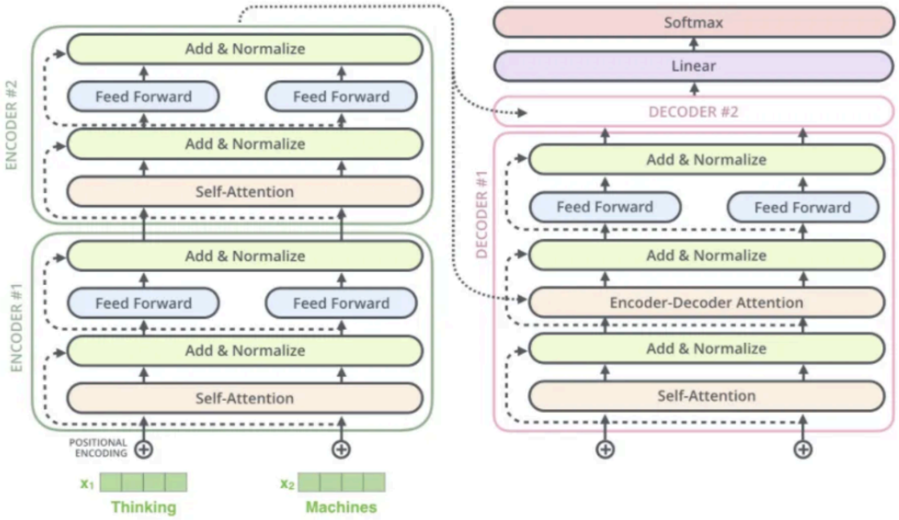
- optimizer = RMSProp
- loss = sparse_categorical_crossentropy
- Pre-trained embeddings used from GloVe with 300 dimensions.

Results and Analysis: Loss and Repeated Translations

The model initially resulted in repetitive translations, as challenges arose from applying the softmax function over a large vocabulary. This often caused the model to favor high-frequency, less informative tokens, which contributed to poor translation quality.

Transformer-based Models

T5



Feature	Original Transformer	T5
Training Objective	Machine translation objective	Unified text-to-text format for all NLP tasks
Positional Encoding	Sinusoidal positional encoding	Relative positional encoding (no fixed sinusoidal)
LayerNorm Placement	Post-activation	Pre-activation (LayerNorm before self-attn & FFN)
Feed-forward Activation	ReLU	GELU

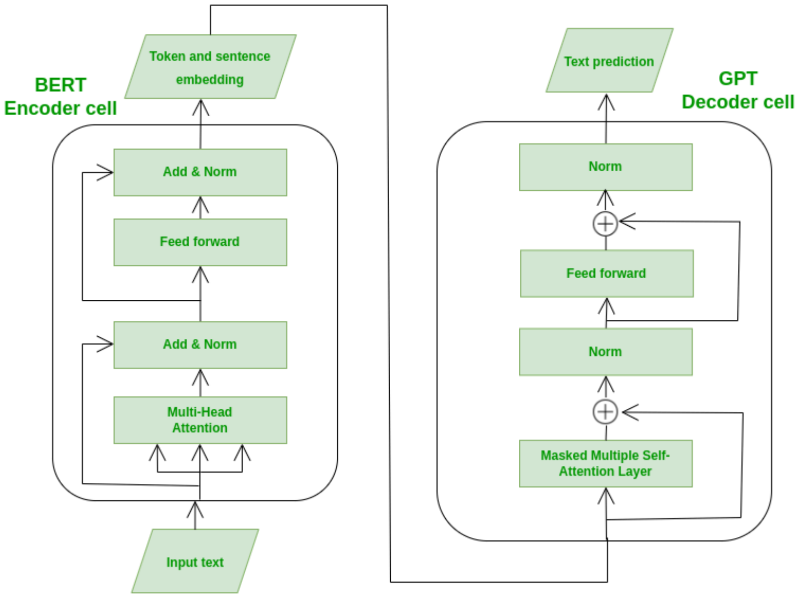
The T5 model was employed for the English-to-Hinglish translation task using a parameter-efficient fine-tuning technique known as QLoRA. The base model chosen was t5-small, a lightweight variant of the Text-to-Text Transfer Transformer, which reformulates all tasks into a unified text-to-text format.

To make training memory-efficient and accessible on lower-end hardware, the model was loaded with 4-bit quantization using the BitsAndBytes library. This significantly reduced GPU memory consumption without sacrificing performance. On top of that, Low-Rank Adaptation (LoRA) layers were injected into the T5 attention modules, specifically targeting the query and value projections. These adapter layers were trained while the rest of the model remained frozen, allowing efficient adaptation with fewer trainable parameters.

Hyperparameters used:

- model: t5-small
- batch_size: 32
- num_train_epochs: 3
- learning_rate: 1e-5
- max_input_length: 128
- max_output_length: 128
- quantization: 4-bit using BitsAndBytes
- adapter_type: LoRA with r = 16, α = 32, dropout = 0.1
- optimizer: AdamW (via Trainer)

IndicBART



IndicBART is a multilingual sequence-to-sequence language model developed specifically for Indian languages. It is based on BART architecture (Bidirectional and Auto-Regressive Transformers), but trained on a large corpus of Indian language data, primarily for tasks such as machine translation, summarization, and text generation involving Indian languages. It is more flexible than either BERT or GPT alone for encoder-decoder tasks.

Architecture -

- Encoder: Fully bidirectional like BERT.
- Decoder: Autoregressive like GPT (uses causal masking).
- Pretraining Objective: Denoising autoencoder — it corrupts the input text (e.g., by masking, shuffling, deleting tokens) and learns to reconstruct the original.

We fine-tuned the IndicBART model (ai4bharat/IndicBART) for the task of English-to-Hinglish translation. The model is based on the MBart architecture, pre-trained on multiple Indian languages. It uses an encoder-decoder transformer structure, making it well-suited for sequence-to-sequence tasks like translation.

The loss steadily decreased over epochs, indicating that the model was learning the structure and vocabulary of Hinglish effectively and the validity loss suggested good generalization to unseen examples.

Here's the configuration for the **IndicBART model**:

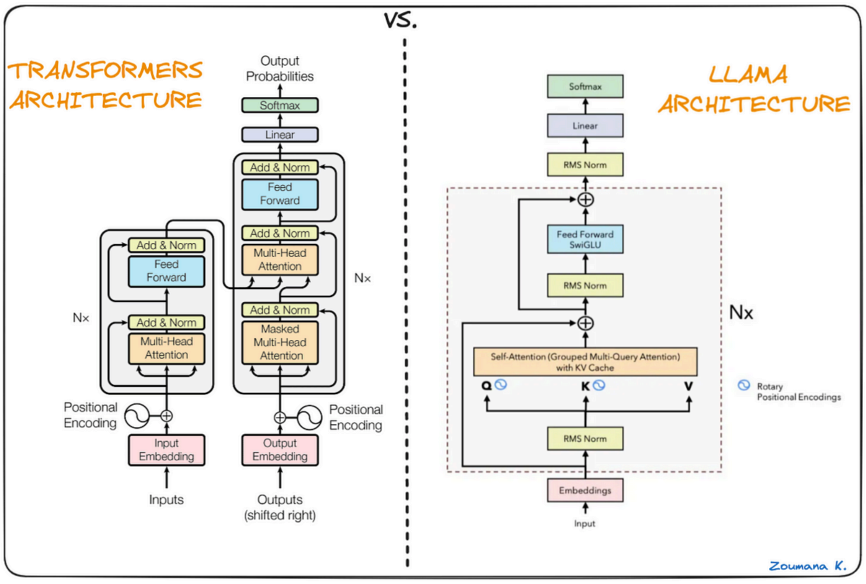
Hyperparameters used:

- model: indic-bart

- batch_size: 16
- num_train_epochs: 1
- learning_rate: 1e-3
- max_input_length: 128
- max_output_length: 128
- optimizer: AdamW (via Trainer)
- max_steps: 80000

Epoch	Training Loss	Validation Loss
1	0.090400	0.085314

Llama



The Llama series of models are autoregressive decoder-only Transformers, but there are some minor differences:

- SwiGLU activation function instead of GeLU
- Rotatory Positional Embeddings (RoPE) instead of absolute positional embedding
- RMSNorm instead of layer normalization

LLaMA's developers focused their effort on scaling the model's performance by increasing the volume of training data, rather than the number of parameters.

Hyperparameters used -

- `max_seq_length` : `100`
- `batch_size` : `8`
- `grad_accum_steps` : `2`
- `learning_rate` : `4e-4`
- `log_interval` : `100`
- `bnb_config` , `LoRAConfig`
- `optimizer` : `torch.optim.AdamW`

Epoch 1/1: 100%| 11250/11250 [1:51:47<00:00, 1.68it/s, loss=0.186, lr=1.98e-06]

mT5

mT5 has been trained on a large corpus that includes data from multiple languages. This allows it to handle code-switching and translations between languages like English and Hindi, which are essential for Hinglish. While T5 was primarily trained on English, making it less suited for tasks involving languages other than English, such as Hindi. Final training loss is 0.186,

Hyperparameters used are -

model_name = "google/mt5-small"
dataset_name = "findnitai/english-to-hinglish"
batch_size = 4
num_epochs = 1
learning_rate = 5e-4
max_length = 64

Evaluation Metrics

GLUECoS (General Language Understanding Evaluation for Code-Switched Languages)

It serves as a set of metrics for a wide range of tasks including both classification and generation. It is used for code switching datasets like Hindi- English, English- Telugu, etc.

Metric	Used In	Description

Accuracy	Sentiment Analysis, NLI, Language Identification	The proportion of correct predictions out of total predictions. Suitable for balanced classification tasks.
Macro F1-score	Sentiment Analysis, NLI, Intent Classification	F1-score computed independently for each class and then averaged. Useful when classes are imbalanced.
Micro F1-score	Token-level tasks (e.g., POS Tagging)	F1-score computed globally by counting total true positives, false negatives, and false positives.
Entity-level F1-score	Named Entity Recognition (NER), Slot Filling	Considers precision and recall at the entity span level rather than individual tokens, ensuring exact span matches are rewarded.
Token-level Accuracy	POS Tagging	Measures how many individual token labels (e.g., POS tags) are correctly predicted.
Exact Match (EM)	Question Answering (QA)	The percentage of predictions that exactly match the ground truth span. Strictest QA metric.
Perplexity	Language Modeling	Exponential of the negative average log-likelihood. Lower values indicate better fluency in language modeling tasks.

MIPE (Metrics for Indic Performance Evaluation)

It was designed to assess the performance of NLP models (especially generation models) in Indic languages. It was introduced to address the unique challenges in evaluating Indic NLP tasks due to script diversity (e.g., Devanagari, Bengali, Tamil scripts), lack of consistent tokenization tools, code-mixed and transliterated texts, and the morphological richness of Indic languages.
The term MIPE generally encompasses a standard set of evaluation metrics and preprocessing protocols used in Indic NLP generation tasks.

Metric	Description
BLEU	Measures n-gram overlap; tokenization affects its reliability in Indic langs
chrF	Character n-gram F-score, works better for morphologically rich languages
ROUGE-L	Longest common subsequence based metric (used in summarization and gen)
SacreBLEU	A standard BLEU variant ensuring reproducibility
Exact Match (EM)	Fraction of predictions that match references exactly

Evaluation of Transformer models

IndicBART

```
GLUECoS

English Sentence: My smartwatch just died in the middle of a workout.
Hinglish Sentence: meri smartwatch just died in the middle of a workout.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'negative'}
===========================================================
=========================================
English Sentence: The Wi-Fi router is acting up again.
Hinglish Sentence: Wi-Fi router fir se chal raha hai.
NER Comparison: {'NER Consistency': False, 'English Entities': {'MISC'}, 'Hi
nglish Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'neutral'}
===========================================================
=========================================
English Sentence: I need to clear my browser history.
Hinglish Sentence: mujhe mere browser history clear karna hai.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'ORG', 'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'neutral'}
===========================================================
=========================================
English Sentence: I forgot to cancel my subscription to that streaming servi
ce.
Hinglish Sentence: I forgot to cancel my subscription to that streaming serv
ice.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'negative'}
===========================================================
=========================================
English Sentence: The game crashed right before I reached the final boss.
Hinglish Sentence: The game crashed right before I reached the final boss.
```

```
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'negative'}
===========================================================
=========================================
English Sentence: My phone froze when I was about to check an important
message.
Hinglish Sentence: mera phone froze kab tha jab mujhe important message
check karne ke liye.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'ORG'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'neutral'}
===========================================================
=========================================
English Sentence: I need to fix the bug in my code before the deadline.
Hinglish Sentence: mujhe deadline se pehle mere code me bug fix karna ha
i.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '3 stars', 'Hinglish Sentiment': 'neutral'}
===========================================================
=========================================
English Sentence: I'm trying to get my hands on the new gaming console.
Hinglish Sentence: mai new gaming console par apne hands laana chahta h
oon.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'ORG', 'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '3 stars', 'Hinglish Sentiment': 'neutral'}
===========================================================
=========================================
English Sentence: I ordered food online, but they gave me the wrong item.
Hinglish Sentence: I ordered food online, but they gave me the wrong item.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
```

```
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'negative'}
============================================================
========================================
English Sentence: I need a caffeine boost to survive this meeting.
Hinglish Sentence: mujhe is meeting me caffeine boost chahiye.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'neutral'}
============================================================
========================================
English Sentence: Has my timer started?
Hinglish Sentence: Kya mera timer shuru hoga?
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'ORG'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'neutral'}
============================================================
========================================
English Sentence: Set an alarm for me.
Hinglish Sentence: Mere liye ek alarm set karen.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'neutral'}
============================================================
========================================
English Sentence: Did I get new messages?
Hinglish Sentence: Kya maine naye messages milgaye hai?
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'ORG', 'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'neutral'}
============================================================
========================================
English Sentence: What is the time right now?
Hinglish Sentence: Abhi ka time kya hai?
```

```
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'neutral'}
============================================================
==========================================
English Sentence: It will be sunny today.
Hinglish Sentence: Aaj dhoop hogi.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '4 stars', 'Hinglish Sentiment': 'negative'}
============================================================
==========================================

Average NER Consistency: 0.26666666666666666
Average Sentiment Consistency: 0.0


MIPE

Average BLEU (English): 1.0000
Average BLEU (Hinglish): 0.2587
Average chrF (English): 14.8172
Average chrF (Hinglish): 1.9422
Average ROUGE-L (English): 1.0000
Average ROUGE-L (Hinglish): 0.4172
Average SacreBLEU (English): 100.0000
Average SacreBLEU (Hinglish): 33.0704
Average Exact Match (English): 100.0000
Average Exact Match (Hinglish): 20.0000
```

mT5

```
GLUECoS

English Sentence: My smartwatch just died in the middle of a workout.
Hinglish Sentence: meri workout ke middle me ek workout mein hein.
```

```
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'neutral'}
============================================================
==========================================
English Sentence: The Wi-Fi router is acting up again.
Hinglish Sentence: Wi-Fi ka acting up kiya gaya tha.
NER Comparison: {'NER Consistency': False, 'English Entities': {'MISC'}, 'Hi
nglish Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'negative'}
============================================================
==========================================
English Sentence: I need to clear my browser history.
Hinglish Sentence: mujhe mere browser history ko clear karna chahiye.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'ORG', 'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'negative'}
============================================================
==========================================
English Sentence: I forgot to cancel my subscription to that streaming servi
ce.
Hinglish Sentence: mai apne subscription ko cancel karna chahta hoon.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'ORG'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'negative'}
============================================================
==========================================
English Sentence: The game crashed right before I reached the final boss.
Hinglish Sentence: maine shuru kiya tha uske baad game crash kiya tha.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'negative'}
============================================================
========================================
English Sentence: My phone froze when I was about to check an important
message.
Hinglish Sentence: mere phone ko booze kiya tha jab maine important mes
sage check kiya tha.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'neutral'}
==========================================================
========================================
English Sentence: I need to fix the bug in my code before the deadline.
Hinglish Sentence: mujhe mere code se pehle fix karna chahiye.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '3 stars', 'Hinglish Sentiment': 'negative'}
==========================================================
========================================
English Sentence: I'm trying to get my hands on the new gaming console.
Hinglish Sentence: mai naya gaming device par milna chahta hoon.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '3 stars', 'Hinglish Sentiment': 'neutral'}
==========================================================
========================================
English Sentence: I ordered food online, but they gave me the wrong item.
Hinglish Sentence: maine ek naya item kharidne ka hein.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'neutral'}
==========================================================
========================================
English Sentence: I need a caffeine boost to survive this meeting.
Hinglish Sentence: mujhe is meeting ko ek caffeine ka accha laga.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'negative'}
===========================================================
=========================================

Average NER Consistency: 0.4
Average Sentiment Consistency: 0.0


MIPE

Average BLEU (English): 1.0000
Average BLEU (Hinglish): 0.0000
Average chrF (English): 11.1822
Average chrF (Hinglish): 0.0000
Average ROUGE-L (English): 1.0000
Average ROUGE-L (Hinglish): 0.1951
Average SacreBLEU (English): 100.0000
Average SacreBLEU (Hinglish): 6.3217
Average Exact Match (English): 100.0000
Average Exact Match (Hinglish): 0.0000
```

Llama

```
GLUECoS

English Sentence: My smartwatch just died in the middle of a workout.
Hinglish Sentence: My smartwatch ko workout mein ek din mein mar diya.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'neutral'}
===========================================================
=========================================
English Sentence: The Wi-Fi router is acting up again.
Hinglish Sentence: The Wi-Fi router is acting up again.
NER Comparison: {'NER Consistency': False, 'English Entities': {'MISC'}, 'Hi
nglish Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'neutral'}
============================================================
====================================
English Sentence: I need to clear my browser history.
Hinglish Sentence: mujhe apna browser history clear karna chahiye.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'ORG', 'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'negative'}
============================================================
====================================
English Sentence: I forgot to cancel my subscription to that streaming servi
ce.
Hinglish Sentence: mujhe wo streaming service mei apna subscription canc
el karne ke liye yaad nahi bacha.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'ORG'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'negative'}
============================================================
====================================
English Sentence: The game crashed right before I reached the final boss.
Hinglish Sentence: The game crashed right before I reached the final boss.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'negative'}
============================================================
====================================
English Sentence: My phone froze when I was about to check an important
message.
Hinglish Sentence: mujhe mere phone ko pankh milne par froze hoga.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'negative'}
```

```
==========================================================
========================================
English Sentence: I need to fix the bug in my code before the deadline.
Hinglish Sentence: mujhe deadline se pehle code mei bug fix karna chahiy
e.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '3 stars', 'Hinglish Sentiment': 'neutral'}
==========================================================
========================================
English Sentence: I'm trying to get my hands on the new gaming console.
Hinglish Sentence: mujhe new gaming console ke paas mujhe haan.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '3 stars', 'Hinglish Sentiment': 'neutral'}
==========================================================
========================================
English Sentence: I ordered food online, but they gave me the wrong item.
Hinglish Sentence: I ordered food online, but they gave me the wrong item.
NER Comparison: {'NER Consistency': True, 'English Entities': set(), 'Hinglis
h Entities': set()}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '1 star', 'Hinglish Sentiment': 'negative'}
==========================================================
========================================
English Sentence: I need a caffeine boost to survive this meeting.
Hinglish Sentence: mujhe is meeting me survive karne ke liye caffeine boos
t chahiye.
NER Comparison: {'NER Consistency': False, 'English Entities': set(), 'Hingli
sh Entities': {'PER'}}
Sentiment Comparison: {'Sentiment Consistency': False, 'English Sentimen
t': '5 stars', 'Hinglish Sentiment': 'neutral'}
==========================================================
========================================

Average NER Consistency: 0.4
Average Sentiment Consistency: 0.0


MIPE

Average BLEU (English): 1.0000
Average BLEU (Hinglish): 0.3000
Average chrF (English): 11.1822
Average chrF (Hinglish): 4.5780
Average ROUGE-L (English): 1.0000
Average ROUGE-L (Hinglish): 0.4534
Average SacreBLEU (English): 100.0000
Average SacreBLEU (Hinglish): 35.9602
Average Exact Match (English): 100.0000
Average Exact Match (Hinglish): 30.0000
```

Evaluating on larger test sets

Translations for about 200 samples were generated by trained Llama and mT5 models which were then evaluated using GLUECoS and MIPE.

GLUECoS Scores

Model	NER Consistency	Sentiment Consistency
LLaMA	0.425	0.0
mT5	0.36	0.0

MIPE scores

Metric	LLaMA (English)	LLaMA (Hinglish)	mT5 (English)	mT5 (Hinglish)
BLEU	0.9450	0.0534	0.9600	0.0291
chrF	15.2936	4.4407	15.8106	4.2856
ROUGE-L	1.0000	0.2878	1.0000	0.2726
SacreBLEU	94.5000	14.0348	96.0000	12.2755
Exact Match	100.0000	0.0000	100.0000	0.0000

Analysis

- Sentiment Analysis is poor across all models, with consistency scores at 0.0. This likely stems from the informal, ambiguous nature of Hinglish and limited sentiment-specific fine-tuning on code-switched data.
- LLaMA performs best overall on both GLUECoS (NER: 0.425) and MIPE tasks, especially for English outputs (BLEU: 0.945, chrF: 15.29, ROUGE-L: 1.0, SacreBLEU: 94.5, Exact Match: 100%). This can be attributed to its larger size and stronger pretraining, making it more robust at both classification and generation tasks.
- mT5 shows decent GLUECoS NER performance (0.36) but fails on Hinglish generation (BLEU: 0.0291, ROUGE-L: 0.2726), despite strong English scores. This suggests it handles classification reasonably but struggles with generating fluent Hinglish text—likely due to a lack of domain-specific fine-tuning.
- IndicBART's performance is not included in the larger test set results above, but from earlier findings, it underperformed in both NER and sentiment tasks. Its weakness could stem from insufficient Hinglish data exposure or weaker adaptation to code-switched, informal text styles.
- LLaMA also dominates generation in English, with perfect Exact Match and ROUGE scores, and does relatively better in Hinglish generation than mT5, though all models perform poorly in this area.
- Overall, LLaMA is the most balanced and capable model, followed by mT5 for classification, while IndicBART lags behind, possibly due to limited fine-tuning or smaller scale.

Challenges faced and drawbacks

N-gram Model

- Due to its fixed context window (n = 3), the model failed to capture complex sentence structures and long-range dependencies.
- It heavily relied on surface-level patterns, resulting in incorrect or unnatural translations, especially when encountering unseen or infrequent trigrams in the test data.

LSTM-based Seq2Seq

- The model often produced repetitive or generic outputs. This issue was attributed to the softmax operation over a large vocabulary, which biased the model towards high-frequency words.

- Although attention mechanisms were implemented, the model struggled with longer sentences and exhibited exposure bias during inference.

T5-small with QLoRA

- Despite the advantages of 4-bit quantization and LoRA adapters, which made training more resource-efficient, the model significantly underperformed in sentence formation and grammar.
- Many translations consisted of repetitive or disjointed phrases.
- The large dataset used for training posed computational challenges, which were only partially mitigated through the use of QLoRA.

IndicBART

- The model demonstrated better alignment with Hindi linguistic structures due to its pretraining; however, initial training was constrained by computational limits, resulting in a smaller effective dataset.
- Performance improved after training on the larger dataset, but the model still failed to generalize to longer or less common sentence structures.

mT5

- Represented an improvement over T5 in certain respects but continued to suffer from weak sentence formation and grammar.
- When encountering sentences dissimilar to its training data, the model often produced translations that were semantically incorrect or meaningless.

LlaMA

- The model's large parameter size led to frequent memory overflows and slow training/inference times.
- For niche or structurally unique input sentences, it often resorted to outputting the English sentence verbatim, indicating a lack of robust generalization in those cases.

Conclusion

For Codemix generation, we began with basic models like N-gram and LSTM, which performed poorly. We then moved to transformer-based models like mT5 and IndicBART, which produced better translations but struggled to generalize to new or uncommon sentences. Next, we used the LLaMA model, which gave the best results overall, showing stronger generalization based on our observations. To support this, we evaluated the transformer models using GLUECoS and MIPE scores, which confirmed that LLaMA was the most effective. Based on both observation and metrics, the models ranked as follows for English to Hinglish translation: LLaMA > mT5 > IndicBART > T5 > LSTM > N-gram.

References:

Llama 2

LIMA paper (Less Is For More Aligment)

QLORA parameter efficient finetunning

GLUECoS: https://aclanthology.org/2020.acl-main.329/

MIPE: https://arxiv.org/abs/2107.11534

https://huggingface.co/docs/transformers/model_doc/t5

https://huggingface.co/docs/transformers/en/model_doc/mt5

https://huggingface.co/ai4bharat/IndicBART