EVALUATING LARGE LANGUAGE MODELS FOR MACHINE TRANSLATION ON INDIAN LANGUAGES.

by

Geetha Syam Sai Akula

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science in

Computer Science

at

The University of Wisconsin-Milwaukee

December 2024

ABSTRACT

EVALUATING LARGE LANGUAGE MODELS FOR MACHINE TRANSLATION ON INDIAN LANGUAGES.

by

Geetha Syam Sai Akula

The University of Wisconsin-Milwaukee, 2024
Under the Supervision of Professor Rohit J. Kate

This study assesses how well Large Language Models (LLMs), such as LLaMA-v3 and GPT-3.5, perform while translating English into Indian languages. For three Indian languages, translations from English were evaluated by humans and were found to be fairly good. Automated measures like BLEU, METEOR, and BERTScore were then evaluated by comparing them to human evaluation scores. Translations from English obtained by LLMs were then automatically evaluated on eleven Indian languages using the Samanantar dataset. The results show that while LLaMA has significant advantages in terms of fluency and semantic accuracy, LLMs are prone to errors related to language specific conventions. As part of the study, the impact of prompt engineering on improving translation quality was also examined.

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES vii

LIST OF ABBREVIATIONS viii

ACKNOWLEDGEMENTS ix

Introduction 1

1.1 Background and Motivation 1

1.2 Advantages of Machine Translation: 1

1.3 Large Language Models for Machine Translation: 2

1.4 Advantages of LLMs for Translation: 2

1.4.1 Adaptability: 2

1.4.2 Lower Dependence on Bilingual Data: 2

1.6 Objectives: 3

1.6.1 Translation Performance: 3

1.6.2. Evaluation Metrics: 3

Literature Overview 4

2.1 LLMs for MT 4

2.2 Automatic Evaluation Metrics for MT 6

Materials and Methods 8

3.1 Samanantar Dataset 8

3.1.1 LaBSE (Language -agnostic BERT Sentence Embedding) Scores 9

3.2 Evaluation Metrics 9

3.2.1 BLEU SCORE 9

3.2.2 Modified n-gram Precision 10

3.2.3 Brevity Penalty 11

3.2.4 BLEU Score 11

3.2.5 METEOR SCORE 11

3.2.6 BERTScore 12

3.3 Correlation Coefficient Metrics: 14

3.4 Python libraries: 15

3.5 Large Language Models: 16

3.5.1 Examples 16

3.6 Experimental Methodology 18

3.6.1 Dataset and Filtering Criteria 18

3.6.2 Human Evaluation on Three Indian Languages 18

3.6.3 Evaluation of Automatic Metrics 18

3.6.4 Automatic Evaluation on Eleven Indian Languages 19

3.6.5 Prompt Refinement 19

3.6.5.1 Example Prompt 20

Results and Discussion 21

4.1 Human Evaluation of Translation for Hindi, Telugu and Marathi 21
4.1.1 Evaluation of Automatic MT Metrics 22
4.2 Automated Evaluation of Translations for Eleven Indian Languages 24
4.2.1 Observations for Sample Size of 100 Sentences: 25
4.2.2 Observations for Sample Size of 1000 Sentences 28
4.3 Impact of Prompt Engineering on LLM Translation Performance 29
4.4 Future Work 30
Conclusion **32**
Bibliography **33**
APPENDIX A **36**
Interacting with LLMs using APIs: 36
1. **Example API for GPT** 36
2. **Example Response for GPT-4o** 37
3. **Example API Call for LLaMA-v3** 38
4. **Examples Response for LLaMA-v3** 39
APPENDIX B **41**
Merging Datasets: A Sample Demonstration 41

LIST OF FIGURES

Figure 1 idx: A unique identifier, src: The source sentence in English,..... 41

LIST OF TABLES

Table 3.1 Comparison of different Large Language Models (LLMs) for the Study..... 16

Table 4. 1 Human Evaluation Scores (0-100 scale) for GPT, LLaMA output on 100 sentences in each language. 21

Table 4. 2 Human Evaluation Scores (0-100 scale) for Reference Sentences 22

Table 4. 3 Results of Correlation Measures- Spearman, Cohen's Kappa and Pearson..... 23

Table 4. 4 Observations for Sample Size of 100 Sentences..... 25

Table 4. 5 Varying BLEU Scores for different Models..... 26

Table 4. 6 Varying METEOR Scores for different Models..... 27

Table 4. 7 Average Automated Evaluations of LLAMA scores for 11Indian language..... 28

Table 4. 8 Performance Metrics of LLaMA Model for 1000 sentences with Prompt. 30

LIST OF ABBREVIATIONS

LLM	Large language Model
MT	Machine Translation
GPT	Generative Pre-trained Transformer
LLaMA	Large Language Model Meta AI
LaBSE	Language-Agnostic BERT Sentence Embedding
BLEU	Bilingual Evaluation Understudy
BERT	Bidirectional Encoder Representations from Transformers
METEOR	Metric for Evaluation of Translation with Explicit ORdering
API	Application programming interface
FAISS	Facebook AI Similarity Search
OCR	Optical Character Recognition

ACKNOWLEDGEMENTS

I would like to thank Professor. Rohit J. Kate for helping me throughout the thesis preparation process and providing valuable insights with much patience.

I would like to thank Dr. Susan Mcroy and Professor. Jun Zhang for agreeing to join the thesis committee.

I would also like to thank my Family for the care and all support they provided to me in my education. I would also like to thank my friends Sai Yerni Akhil Madabattula and Sri Adi Narayana Repudi for the support.

Chapter 1

Introduction

1.1 Background and Motivation

There are about 7000 languages around the world, including 22 main Indian languages with most states of the country having a unique language. Since a lot of the information on the Internet is available in English, people that are used to their own language might struggle with material published in English. Moreover, travelers visiting these places cannot communicate with locals in their language. This highlights how crucial it is to develop technologies like machine translation in India, so that information is readily available in their various languages and that everyone has equal access to it. Communicating to people in their native language is essential to creating a sense of comfort and connection.

Machine translation (MT), often known as automated translation, is the process of translating text across languages using computer software without the need for human intervention.

1.2 Advantages of Machine Translation:

Machine Translation enables businesses to communicate effectively with customers and partners across various languages. This is crucial for globally opening companies, as it allows them to interact with their clients in their native languages, strengthening bonds and improving partnerships. Machine translation also enables travelers to communicate with locals in their language.

Machine translation provides a significant advantage with its fast turnaround time. Traditional human translation can be time-consuming, but Machine translate algorithms can quickly process large volumes of text in just seconds or minutes. This increase in translation speed

allows users to access translated content almost instantly making it perfect for time- sensitive projects or urgent communication requirements.

1.3 Large Language Models for Machine Translation:

Traditionally, machine translation was done using statistical methods [1]. But recently, the advent of Large Language Models (LLMs) has enhanced the state-of-the-art in almost every natural language processing task, including MT. LLMs for MT has significantly advanced the capabilities of machine translation systems superseding the traditional statistical MT methods. Leveraging cutting-edge techniques like self-supervised learning, LLMs can learn from large datasets without requiring manual labeling. This feature allows LLMs to generate translations that are context-appropriate and logically consistent, addressing linguistic details that traditional machine learning systems frequently struggle with. Notably, LLMs are capable of handling complex translations across different languages.

1.4 Advantages of LLMs for Translation:

1.4.1 Adaptability:

A key strength of LLMs is that their capability to handle a wide range of topics and linguistic styles. This adaptability makes them ideal for a wide range of content, including policy documents and casual text messages [10].

1.4.2 Lower Dependence on Bilingual Data:

Compared to standard machine translation systems, LLMs require fewer bilingual data for pre-training major languages. This reduced dependence on bilingual data is especially advantageous for languages with scarce bilingual resources, which make LLMs a more flexible and widely applicable tool for translation [10].

1.6 Objectives:

This study's main goal is to evaluate the performance of Large Language Models (LLMs) in translating English to Indian languages. To achieve this, the research is guided by the following research questions:

1.6.1 Translation Performance:

- How effectively LLMs can translate English sentences into Indian languages?

1.6.2. Evaluation Metrics:

- How reliable are traditional Machine Translation (MT) metrics, such as BLEU, in automatically evaluating LLM-generated translations?
- Are there more efficient alternative automatic evaluation metrics for assessing translation quality?

1.6.3. Prompt Engineering:

- Can the performance of LLM-based Machine Translation be improved by providing helpful prompts?

In this work, above research questions are answered using state-of-the-art LLMs – GPT-3.5, GPT-4 and LLaMA, and a large recent MT dataset of Indian languages.

Chapter 2

Literature Overview

2.1 LLMs for MT

In recent years, the use of Large Language Models (LLMs) in machine translation (MT) has gained increasing attention. This literature review presents the current research progress in the field of LLMs in MT.

The study [23], "Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis" conducted by Wenhao Zhu et al.(2023), examined the benefits and difficulties of LLMs in multilingual machine translation. LLMs were evaluated by the authors based on their performance in 102 languages and 606 translation directions, including translations that were mostly English, French, or Chinese. The authors observed that LLMs' translation capabilities are constantly evolving, and GPT-4 sets a new performance standard. They generally perform better when translating into English than from English. They suggest future research should investigate more translation directions, such as Russian-centric and Arabic-centric translation.

The research paper "Efficient Large Language Models: A Survey" by Wan et al. [21] describes prompt engineering is an essential data-centric technique for improving the efficiency of Large Language Models (LLMs). This process involves crafting effective prompts to guide LLMs in generating desired outputs, thereby boosting inference efficiency. The authors emphasize that prompt engineering can be especially useful for straightforward tasks like semantic classification, where it can even replace fine-tuning and achieve high accuracy. Here the author discussed the techniques of Few-Shot Prompting, Prompt Compression and Prompt Generation and emphasize

the effectiveness of few-shot prompting in enabling LLMs to handle diverse tasks without the need for extra training or fine-tuning.

The study conducted by Vandan et al. [3] focuses on exploring the multilingual capabilities of large language models by using machine translation for Indian languages. The author experiments with various pre-trained models to test their language abilities and evaluates the performance of raw LLMs as well as fine-tuned versions of selected models, including LLaMA-7B, Falcon-7B, Bloom-7B, LLaMA-2-7B, and LLaMA-2-13B. The authors concentrate on models based on LLaMA-2, specifically LLaMA-2-13B, which performs better in both zero-shot and fine-tuned scenarios. They highlight the need for future LLMs to address the limited representation of Indian languages in their vocabulary. The author compares the translation quality of GPT-3.5 to the LLaMA models using the BLEUscore and chrF (character F-score). This study is limited to evaluating machine translations using just two metrics, with a particular emphasis on the LLaMA models.

The study "Multilingual Tourist Assistance using ChatGPT: Comparing Capabilities in Hindi, Telugu, and Kannada" by Kolar et al. [25] evaluates the effectiveness of ChatGPT as a translation tool for tourists in India, focusing on translations from English to Hindi, Kannada, and Telugu using gpt-3.5-turbo to translate 50 questions about travel, eating, and general knowledge. On a scale of 1 to 5, native speakers rated the correctness and fluency of translations in subjective evaluations, while the BLEU score was used for objective evaluation. The study recognizes limitations, which include the subjective nature of native speaker assessments, the fact that it only looks at three language pairs, and the absence of comparisons with other translation tools.

2.2 Automatic Evaluation Metrics for MT

Human evaluation of machine translation is costly and comprehensive. Human assessments can take months to complete and require work that cannot be reproduced. In their study, Papineni et al. [18] provided BLEU score which is a language-independent, fast, and low-cost automatic machine translation approach with a low marginal cost per run that has a strong correlation with human evaluation. It is simply a measure of how similar a machine translation is to one or more reference human translations (described in Chapter 3).

In another study, Banerjee et al. [20] proposed the METEOR score, an automated evaluation metric, for which the basis is the generalized unigram matching between machine-generated translations and human references. The matching makes use of complex tactics involving surface forms, stems, and synonyms.

Zhang et al. [19] in their study proposed an automated assessment called BERTScore was created to evaluate the quality of produced text, especially in tasks like image captioning and machine translation. BERTScore makes use of the contextual embeddings to provide more accurate and sophisticated assessments of the generated text. The key features of the BERTScores are the contextual similarity - allows for a more complex comprehension of the semantic relationships between words, token-level comparison - every token in the candidate sentence is compared to every token in the reference sentence to determine its similarity score.

The study done by Graham et. Al. and Chatzikoumi [17] explore the application of continuous rating scales for evaluating the machine translation outputs. They mention that the usage of continuous scales gives us the degree of difference between translations. They also

mention the importance of considering a combination of automated and human evaluation methods, to provide a well-rounded assessment of the quality of machine translation.

Chapter 3

Materials and Methods

3.1 Samanantar Dataset

The Samanantar v0.2 Dataset, released on 15th May 2021 which includes the largest collection of parallel corpora specifically designed for Indic languages, serves as the main source of data for this study. It includes 49.7 million English-to-11 Indian language sentence pairings, with 23,000 pairs were used for this study. Existing parallel corpora were combined with fresh data mined from several sources, including scanned documents, web-crawled monolingual corpora, and multilingual content from websites, to build the dataset [11].

Both English-to-Indic and Indic-to-English translation activities are supported by its design. Also, each language has its own script to represent its letters. The researchers employed various techniques to mine parallel sentences to build the Samanantar corpus. They took parallel sentences from news sources and educational platforms that publish content in multiple languages. They used OCR and alignment-based similarity scores to determine parallel sentences from scanned texts, such as speeches from legislative assemblies and government processes and Web Scale Monolingual Corpora where they mined parallel sentences from IndicCorp, a large monolingual corpus for Indic languages, using FAISS for efficient nearest neighbor search and LaBSE for similarity scoring (described below). They used this method on Wikipedia documents as well [11].

3.1.1 LaBSE (Language -agnostic BERT Sentence Embedding) Scores

LaBSE [12] generates sentence embeddings, which are used to measure the semantic similarity using the embeddings of two sentences. This is done by applying cosine similarity and arc cosine distance as similarity metrics.

Higher cosine similarity scores reflect greater similarity between sentences, with score of 1 being the maximum score, indicating perfect similarity, and score of -1 indicating the least score, indicating completely opposite meanings.

3.2 Evaluation Metrics

In this thesis, I have chosen three evaluation measures: BLEU [18], METEOR [20], and BERT [19]. Each metric offers distinct benefits that fit the objectives of my research on the quality of machine translation.

3.2.1 BLEU SCORE

BLEU is one of the popular metrics that demonstrates a strong alignment with human quality judgements. By using BLEU, I aim to evaluate the n-grams of human-translated target sentences to the n-grams of machine-translated sentences. This metric assesses translations on a scale of 0 to 1, measuring both fluency and adequacy. It assesses whether the output conveys the same meaning as the input sentence and whether the translation is logical and fluent in target language [18]. Better translation quality is indicated by a score nearer to 1, which shows more overlap with the human reference translation. Higher the BLEU scores represent higher the overlap of words. Here are the key components and formulas for calculating BLEU scores:

The BLEU score calculation involves comparing candidate translations C and correct references R.

3.2.2 Modified n-gram Precision

We can calculate it by using the below formula:

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n\text{ gram} \in C} \text{Count}_{\text{clip}}(n\text{ gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n\text{ gram}' \in C'} \text{Count}(n\text{ gram}')}$$

Where:

- *C* = candidate translation
- n gram = n-gram in the candidate translation
- Count = normal count
- $\text{Count}_{\text{clip}}$ = min ($\text{Count}_{\text{candidate}}$, Max_Ref_Count)

The $\text{Count}_{\text{clip}}$ function limits each word's count in the candidate translation to the maximum number of times that word appears in any reference translation. This prevents inflated scores from overuse of certain words.

3.2.2.1 Example

Here's an example from the paper[18] demonstrating the calculation of modified unigram precision:

Candidate**:** the the the the the the the

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Modified Unigram Precision = 2/7

In this case:

- The candidate has 7 words, all "the"
- The maximum count of "the" in reference is 2
- The clipped count is min(7, 2) = 2

- The modified precision is therefore 2/7

This example illustrates how BLEU penalizes excessive repetition of words, even if they appear in the reference translations.

3.2.3 Brevity Penalty

To penalize short translations, a brevity penalty is introduced:

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

Where:

- c = length of candidate translation
- r = effective reference corpus length

3.2.4 BLEU Score

This can be calculated as

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where:

- N = maximum n-gram order (typically 4)
- w_n = weights (typically uniform, i.e., 1/N)

3.2.5 METEOR SCORE

METEOR is selected because to its capacity to take word order, stemming, and synonyms into account throughout the evaluation process . METEOR[20] is an automatic metric for evaluating translation quality. Here the score ranges between 0 to 1, where 0 represents that there is no word

overlap, no alignment in grammar structure whereas 1 represents the high fluency, correct word order and semantic Equivalence.

The following formula is used by the METEOR metric to determine scores:

$$\text{METEOR score} = F_{mean} * (1 - \text{Penalty})$$

F_{mean} calculation

$$F_{mean} = \frac{10PR}{R+9P}$$

Where:

P = Unigram precision (matched unigrams / total unigrams in translation)

R = Unigram recall (matched unigrams / total unigrams in reference)

$$\text{Penalty} = 0.5 * \left(\frac{\text{chunks}}{\text{unigrams matched}}\right)^3$$

This penalty favors longer contiguous matches, which lessens translation fragmentation.

3.2.6 BERTScore

BERTScore[19] is included to deepen the translation quality analysis by assessing the semantic similarity rather than lexical similarity. Unlike BLUE score which are based on exact word matches, BERTScore can accommodate synonyms.

Bert outputs the result in the form of precision, recall and F1 scores

Where:

$$F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

- P_{BERT} is the precision score
- R_{BERT} is the recall score

BERTScore consists of three components which are *recall, precision, and F1 Scores.* The formulas are discussed below:

Recall: $R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$ (Cosine similarity)

Precision: $P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$ (Greedy matching)

Where:

Reference sentence, $x = (x_1, x_2, \ldots, x_n)$

Candidate sentence, $\hat{x} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)$

F1 represents the balanced measure of precision and recall which are calculated using token-level similarity checks between reference and candidate sentences. These scores range from 0 to 1, with 0 representing no semantic similarity. If the score is 1, it means a perfect semantic match between machine generated translation and reference text (original or expected version). Here are the formulas for calculating the F1 measure and BERTSCORE. Here are the relevant formulas. The F1 measure is calculated using precision and recall scores:

This score is calculated using the same formula as the F1 measure with the addition of weights using the inverse document frequency(*idf), shown below:*

$$R_{\text{BERT}} = \frac{\sum_{x_i \in x} \text{idf}(x_i) \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\sum_{x_i \in x} \text{idf}(x_i)}$$

Where *idf(w)* is calculated as:

$$\text{idf}(w) = -\log \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\left[w \in x^{(i)}\right]$$

3.3 Correlation Coefficient Metrics:

Statistical measures such as Spearman, Pearson, and Cohen's Kappa are employed to assess various kinds of correlations and agreements among evaluation scores over a dataset. Every one of these measures has a distinct function and works best with certain kinds of data.

The Spearman rank correlation coefficient rank measures the strength and direction of the association between two ranked variables. It evaluates how well a monotonic function captures the relationship between two variables. It is especially helpful for analyzing ranked relationships in ordinal or non-normally distributed continuous data. It is applied when the data do not satisfy the assumptions necessary for Pearson's correlation, particularly when the data are not normally distributed or are ordinal. A perfect positive rank correlation, where both variables improve as one rises, is indicated by a maximum score of +1. The worst score is -1, which denotes a perfect negative rank correlation, meaning that as one variable rises, the other falls.

The Pearson correlation coefficient quantifies the linear relationship between two continuous variables indicating how much one variable tends to increase or decrease as the other variable does. It works best when evaluating relationships between normally distributed interval data and where a linear relationship is anticipated. It helps in measuring the linear relationship between the two variables' strength and direction, whether it be positive or negative. A complete positive linear relationship is indicated with a maximum value of +1. A perfect negative linear relationship is indicated by a minimum score of -1. There is absolutely no linear relationship when the score is 0.

Cohen's Kappa, often known as for categorical items measures inter-rater reliability, or agreement. It evaluates the degree of agreement between two raters who independently classify items into mutually exclusive categories. It is used extensively in the social sciences, psychology, and medicine to ascertain the degree of agreement between two observers' classifications. It is

specifically made for categorical data. Perfect agreement is indicated with a maximum score of +1. The minimum value can be less than zero, indicating that the agreement is worse. A value of 0 indicates no agreement.

3.4 Python libraries:

In this study, some Python packages are used for data processing, analysis, and translation quality assessment:

- Pandas: It is a tool for data analysis and manipulation that offers effective data structures, such as Data Frames, for managing structured data.
- NumPy: Essential for numerical computing, enabling operations on large multi-dimensional arrays and matrices.
- SciPy: SciPy builds on NumPy, adds more features for scientific computing and statistical analysis.
- NLTK: A natural language processing toolkit that includes linguistic analysis, categorization, and text processing capabilities.
- Fireworks AI: An SDK for interacting with the LLaMA model via a RESTful API, facilitating the generation and evaluation of translations.
- Scikit-learn (sklearn): Offers metrics and machine learning tools for evaluating models, such as correlation coefficients for judging the quality of translations.
- Evaluate Library: Includes functionalities for model evaluation, specifically for calculating the BLEU score to compare machine-generated translations against reference translations.

3.5 Large Language Models:

Probabilistic models of natural language which are used to comprehend, produce and evaluate human language are called language models. These language models are trained on large textual datasets to predict the probability of word sequences. Large Language Models (LLMs) in comparison to the traditional Language Models are massive in scale with billions of parameters and trained over datasets with billions of words, whereas traditional language models are simple and rely on statistical and rule-based approaches.

For this study, among popular easily available LLMs, GPT-3.5, LLaMA-v3 and GPT-4o are selected.

Table 3.1 Comparison of different Large Language Models (LLMs) for the Study

Model	Context Window	Max Output Tokens	Training Data
GPT-3.5-turbo	16,385 tokens	4,096 tokens	Up to September 2021
GPT-4o	128,000 tokens	16,384 tokens	Up to December 2023
LLaMA-v3-70b-instruct	8,000 tokens	2,048 tokens	Up to December 2023

While there is no publicly available information about the number of parameters for the GPT-3.5-turbo and GPT-4o, there are 70 billion parameters for the LLaMA-v3-70b-instruct model as the name suggests.

3.5.1 Examples

LLMs can be used for MT by asking them to translate a sentence using a prompt as follows:

When using the model GPT-3.5-turbo,

Prompt: Translate this sentence into Hindi and just give me the translation only

Input: He said case has been registered and investigations are on.

Output उसने कहा कि मामला दर्ज हो गया है और जांच जारी है।

(Usne kaha ki maamla darj ho gaya hai aur jaanch jaari hai)

Note: English transliterations are shown above for the readability for a non-Hindi speaking reader, they are not part of the LLM output.

When using the model GPT-4o,

Prompt: Translate this sentence into Hindi and just give me the translation only

Input: He said case has been registered and investigations are on.

Output केस दर्ज कर दिया गया है और जांच जारी है।

(Case darj kar diya gaya hai aur jaanch jaari hai)

When using the model LLaMA-V3,

Prompt: Translate this sentence into hindi and just give me the translation only

Input: He said case has been registered and investigations are on.

Output उन्होंने कहा कि मामला दर्ज कर लिया गया है और जांच चल रही है।

(Unhone kaha ki maamla darj kar liya gaya hai aur jaanch chal rahi hai)

3.6 Experimental Methodology

3.6.1 Dataset and Filtering Criteria

The study utilizes the Samanantar v0.2 dataset [11], enhanced with LaBSE scores from the v0.2.1 update. This combined dataset includes fields for the source sentence (src), target (tgt) and LaBSE similarity scores (las). To guarantee high semantic similarity and translation quality, only sentence pairs with a LaBSE score higher than 0.9 are included. LaBSE score greater than 0.9 signifies that the sentences from the pair are nearly identical in meaning. Sentences with less than 8 words don't usually represent real-world usage and result in an overestimation of the quality of MT. Shorter sentences with less than 5 words have a higher error rate in the study conducted by Matić [14] in English-Croatian machine translation. So, this study assumes the cut-off for the length of the sentence to be greater than 8 words.

3.6.2 Human Evaluation on Three Indian Languages

Once the dataset is obtained with the former-mentioned filters, a sample size of 100 sentence pairs (of source and target) for the Hindi, Telugu, and Marathi were selected and the translations were performed by GPT and LLaMA LLMs using the APIs that are available. We used these three languages for human evaluation because of our own fluency in them. Those translations were then manually evaluated by using simple and continuous scales [17]: Perfect (100 - 90), Good (80-79), Fair (69 - 51), Poor (50 - 30), Bad (29 - 11), Worst (10 - 0) - based on overall translation accuracy and fluency.

3.6.3 Evaluation of Automatic Metrics

Each LLM's translations are evaluated using two distinct scoring methods as mentioned previously using human and automated evaluation methods. This study examines the correlation between

these two approaches. For instance, GPT3.5 translations of Marathi are assessed based on human evaluation scores (using a continuous scale) and their corresponding automated scores. The correlation between these scores in analyzed using standardized correlation measures, including Cohen's Kappa, Spearman's rank correlation coefficient and Pearson Correlation coefficient. The Cohen's kappa correlation coefficient differs from other metrics employed in this study due to its requirement for continuous variables. To accommodate this constraint and enable the application of Cohen's kappa, we implemented a conversion process for the scores. The discrete evaluation scores are converted into a continuous scale through the procedure described in the preceding sections. This analysis is extended to translations generated by GPT4o and LLaMA and is repeated for other languages – Hindi and Telugu.

3.6.4 Automatic Evaluation on Eleven Indian Languages

Apart from analyzing translations in languages like Telugu, Hindi, and Marathi, we also thoroughly evaluated LLaMA translations in eight additional Indian languages using a dataset of 1,000 sentences. BLEU, BERTScore, and METEOR scores were used in the evaluation to guarantee a thorough examination of translation quality.

3.6.5 Prompt Refinement

With these translation outputs by the LLMs, there was a pattern that we noticed and wanted to further improve the translation quality by giving cues to the LLMs by using the technique of "prompt engineering" [16]. Now the machine translation was carried out for all 11 languages with and without any prompts, to see the effect of prompt engineering on the quality of machine translation.

3.6.5.1 Example Prompt

Prompt Input: Translate this sentence into Telugu- On his complaint, the police has registered a case.

Output: అతని ఫిర్యాదు మీద, పోలీసు ఒక కేసును నమోదు చేసింది. **(singular word చేసింది = chesindi)**

(Athani Piryadhu meedha policu oka casunu namodhu chesindhi)

Enhanced Prompt:

"Translate with respectful language for authorities (e.g., police, ministers) by using plural forms in the target language rather than singular forms to indicate respect". Now translate this sentence into Telugu: On his complaint, the police has registered a case.

Output: అతని ఫిర్యాదు మీద, పోలీసులు ఒక కేసు నమోదు చేసారు. **(plural respectful word చేసారు = chesaru)**

(Athani Piryadhu meedha policulu oka casunu namodhu chesaru)

The above example shows improvement in translation due to the use of a respectful word for police obtained by providing the enhanced prompt.

Chapter 4

Results and Discussion

4.1 Human Evaluation of Translation for Hindi, Telugu and Marathi

From the pool of 11 languages, a subset of three languages– Hindi, Marathi and Telugu were chosen because of our expertise in them.

Table 4. 1 Human Evaluation Scores (0-100 scale) for GPT, LLaMA output on 100 sentences in each language.

Language	**Human Evaluation Scores**	
	GPT 3.5	**LLaMA**
Hindi	91.48	94.65
Telugu	64.6	91.91
Marathi	99.23 gpt-4o	96.45

As evident from the results above **Table 4.1**, for Hindi and Marathi, all the LLMs performed comparably well. However, for Telugu, GPT-3.5 showed the lowest performance among the LLMs, while LLaMA emerged as the most effective model for translating English into all three languages. The manual evaluation revealed a number of issues with the language model's performance. These included the use of unnecessary English words in translations, improper recognition of honorific plurals, and inconsistent adherence to the tense of the source sentence. Furthermore, the model occasionally generated terms that were illogical or nonsensical and occasionally produced translations that contained missing or unnecessary words.

We also evaluated the reference translations in the Samanantar dataset for the languages Hindi, Marathi and Telugu, which were scored using a continuous scale as outlined in Section 3.4.2 for the sample size of 100 sentences. The average human evaluation scores for reference sentences in Hindi, Marathi and Telugu are shown below **Table 4.2**.

Table 4. 2 Human Evaluation Scores (0-100 scale) for Reference Sentences

Language	**Human Evaluation Scores for References**
Hindi	94.6
Telugu	93.8
Marathi	92.33

According to the Human Evaluation scores for the reference sentences in the three languages mentioned above, we can infer that the dataset's sample sentences selected are of high quality.

4.1.1 Evaluation of Automatic MT Metrics

The Correlation analysis aims to assess the relationship between human evaluation scores of GPT-3.5, LLAMA and Machine Evaluation scores (BLEU and BERTScore) of GPT-3.5 and LLAMA. For the set of 100 sentences selected at the beginning of the study, the results for the three Correlation measures are as follows.

Table 4. 3 Results of Correlation Measures- Spearman, Cohen's Kappa and Pearson

Correlation Type	Metric	Model	Marathi	Telugu	Hindi
Spearman	GPT	Human eval vs BERT	0.1163	0.1822	0.2900
		Human eval vs BLEU	0.0819	0.2052	0.2246
	LLAMA	Human eval vs BERT	0.1984	0.2287	0.2711
		Human eval vs BLEU	0.0877	0.2242	0.0960
Cohen's Kappa	GPT	Human eval vs BERT	0.0000	0.0308	0.1139
		Human eval vs BLEU	0.0000	0.0088	-0.0002
	LLAMA	Human eval vs BERT	0.0619	0.0246	0.0571
		Human eval vs BLEU	-0.0014	-0.0051	0.0094
Pearson	GPT	Human eval vs BERT	0.0847	0.1331	0.3266
		Human eval vs BLEU	0.0979	0.2246	0.1950

	LLAMA	Human eval vs BERT	0.1509	0.1806	0.2600
		Human eval vs BLEU	0.1285	0.2018	0.0917

The agreement between automated metrics (BERT and BLEU) and human evaluation scores is revealed by the Spearman coefficient. Stronger agreement between the scores is indicated by a higher positive Spearman's Rank Correlation value, whereas values near zero or negative indicate less agreement or even disagreement, as discussed in Section 3.2.1.

Spearman correlation scores from **Table 4.3** reveal that there is somewhat strong agreement between the automated BERTScores and human evaluation metrics for GPT-3.5, for Hindi (0.2900) and a relatively weak agreement for GPT-3.5 for Marathi (0.0819).

We find a moderately positive correlation between human evaluation and BERTScores for GPT-3.5, especially for Hindi (0.1139) and a weak disagreement between human evaluation and BERTScores for LLaMA-V3 for Marathi (-0.0051), according to Cohen's Kappa correlation scores. This implies that human judgment and BERT's evaluation of translation quality are somewhat in agreement.

Similarly, according to Pearson Correlation coefficient scores, a moderately strong correlation was found between human evaluation and BERTScores for GPT, for Hindi (0.3266) and a relatively weak agreement between human evaluation and BERTScores for Marathi (0.0847)

4.2 Automated Evaluation of Translations for Eleven Indian Languages

We show results of automatic evaluation on all 11 Indian languages in the dataset.

4.2.1 Observations for Sample Size of 100 Sentences:

The approach here is to evaluate with smaller set of sentences(e.g.,100) for initial testing to ensure efficiency, quicker iterations and adjustments to the evaluation metric or system. Additionally, it offers a chance to improve methods and evaluate model performance before to moving on to more resource-intensive research.

Table 4. 4 Observations for Sample Size of 100 Sentences for BERT

Language	**BERTScores (F1 Measure)**		
	GPT3.5	**GPT4o**	**LLAMAv3**
Assamese	0.7958	0.8389	0.8264
Bengali	0.8388	0.8744	0.8596
Gujarati	0.8379	0.8629	0.8566
Kannada	0.8302	0.8712	0.8631
Malayalam	0.7743	0.8352	0.8284
Oriya	0.963	0.9647	0.9743
Punjabi	0.8345	0.8634	0.8601
Tamil	0.8027	0.8496	0.8396
Hindi	0.8748	0.8964	0.8822
Telugu	0.8072	0.8593	0.8458
Marathi	0.7858	0.8657	0.8498

From **Table 4.4**, it is evident that, in most languages, GPT4o continuously outperforms GPT-3.5 and LLaMA, demonstrating its improved translation ability, mostly because of its access to the latest training data. Although GPT-4o seems to outperform LLaMA, it is not a significant

change in most cases. LLaMA V3 has the highest score for Oriya (0.963), although it is not that significantly higher, which indicates that the target and translation are semantically similar.

Significant differences between the tested models were revealed when the translation was evaluated using BLEU.

Table 4. 5 Observations for Sample Size of 100 Sentences for BLEU

Language	**BLEU Scores**		
	GPT3.5	**GPT4o**	**LLAMAv3**
Assamese	0.052	0.1102	0.0766
Bengali	0.0845	0.1714	0.1418
Gujarati	0.107	0.1354	0.1292
Kannada	0.0507	0.1193	0.1045
Malayalam	0.0300	0.0738	0.0675
Oriya	0.0126	0.0551	0.0316
Punjabi	0.0705	0.1547	0.1304
Tamil	0.0795	0.1163	0.1072
Hindi	0.2240	0.3434	0.2858
Telugu	0.0346	0.0991	0.0778
Marathi	0.0365	0.1661	0.1129

The consistently low BLEU scores show suboptimal performance for all models in the translation task as can be seen in **Table 4.5**.

GPT4o performed better than the other LLMs, especially when translating into Hindi language (0.3434). This implies that GPT-4o is much more reliable and has better generation ability for Hindi.

On the other hand, out of all models, GPT-3.5 performed the worst. This poor performance draws attention to possible shortcomings in its architecture or training data for these particular languages. One noteworthy finding is that all models translated Oriya with consistently subpar results.

Table 4. 6 Observations for Sample Size of 100 Sentences for METEOR

Language	**METEOR Scores**		
	GPT3.5	**GPT4o**	**LLAMAv3**
Assamese	0.2542	0.3921	0.3421
Bengali	0.3479	0.4613	0.4175
Gujarati	0.3876	0.4647	0.4448
Kannada	0.3108	0.4245	0.4047
Malayalam	0.2159	0.3309	0.3289
Oriya	0.1837	0.3002	0.3006
Punjabi	0.3736	0.4905	0.4665
Tamil	0.2909	0.4042	0.368
Hindi	0.5493	0.6362	0.5769
Telugu	0.2442	0.4095	0.3733
Marathi	0.2396	0.4858	0.4209

From **Table 4.6**, with the highest METEOR scores, GPT-4o continuously outperformed other models in all languages. This implies that, in comparison to GPT-3.5 and LLaMA, GPT-4o has more robust understanding and superior capabilities for generating Indic languages.

On Hindi, all models did reasonably well; GPT-4o scored the highest (0.6362) and followed by LLaMA (0.5769) and GPT-3.5 (0.5493). On almost all languages, LLaMA and GPT-4o performed equally well.

4.2.2 Observations for Sample Size of 1000 Sentences

In subsequent phase of our research, the sample size was increase to 1000 sentences in order to broaden the scope of our investigation. The LLaMA was then used to machine translate them. After the translation process, the quality of the translation was then automatically evaluated using the aforementioned metrics. The results of this evaluations are presented in **Table 4.7**, which provides a comprehensive overview of our findings.

Table 4. 7 Average Automated Evaluations of LLAMA scores for 11Indian language

Language	Avg. BLEU Score	Avg. BERT F1 Score	Avg. Meteor Score
Hindi	0.28844	0.8877	0.5883
Marathi	0.09427	0.8458	0.4056
Telugu	0.0821	0.8458	0.3824
Assamese	0.10345	0.8273	0.3571
Bengali	0.13238	0.8625	0.4100
Gujarati	0.09583	0.8458	0.4146
Kannada	0.09764	0.8573	0.3905
Malayalam	0.06290	0.8250	0.3101
Oriya	0.04011	0.9660	0.2877
Punjabi	0.14628	0.8602	0.4548
Tamil	0.09622	0.8361	0.3549

An examination of **Table 4.7** shows trends that are generally in line with those found in the 100-sentence sample size.

The BLEU scores indicate that LLaMA demonstrated superior performance in Hindi translation (0.2884), while its efficacy was comparatively lower for Oriya (0.04011). This pattern suggests a persistent language-specific performance across sample sizes. It's interesting to note that the BERTScore assessments offer an alternative viewpoint, while still being in line with what was observed with the sample size of 100. According to this metric, Oriya (0.9660) translations exhibited the highest quality, while Malayalam translations were assessed as the least accurate (0.8250). This departure from the BLEU score results emphasizes how crucial it is to use a variety of assessment measures in order to obtain a thorough grasp of translation quality.

On the other hand, METEOR ratings seem to support the patterns found in BLEU assessments. Hindi (0.5883) translations once again emerged as the most proficient, while Oriya (0.2877) translations were rated as the least accurate.

4.3 Impact of Prompt Engineering on LLM Translation Performance

Manual evaluation of 100 translated sentences revealed systematic errors in the model's output, particularly in singular-plural distinctions and the appropriate use of honorific plurals for addressing dignitaries. To address these issues, we formulated a targeted prompt for the LLM, aiming to mitigate these specific linguistic inaccuracies in subsequent translations. The revised prompt was designed to help the Large Language Model (LLM) address these persistent problems, which could result in higher-quality translations.

To isolate the effect of prompt engineering on translation efficiency, this study introduced an enhanced prompt while keeping sample sizes constant with the earlier trials. This approach

allows for a direct comparison between the baseline LLaMA performance and its capability when augmented with targeted prompting strategies.

Table 4.8 presents the performance metrics of the LLaMA model in translating a corpus of 1000 sentences into Hindi, Marathi, and Telugu. An artificial prompt intended to improve translation quality is included in this assessment. The results provide insights into the model's efficacy across these languages when guided by targeted prompting strategies.

Table 4. 8 Performance Metrics of LLaMA Model for 1000 sentences with Prompt.

Language	Without Prompt	With Prompt
Hindi	0.8877	0.8734
Marathi	0.8458	0.8367
Telugu	0.8457	0.8369

Unexpected findings are seen when the 1000-sentence sample's results with and without prompts are compared. In contrast to what was anticipated, the BLEU Score show that the addition of prompts had no positive effect on any of the languages; on the contrary, performance declined, although the decrease was not drastic. This trend was consistently mirrored in both the BERT and METEOR scores, suggesting that the implemented prompting strategy did not yield the anticipated enhancements in translation quality across these metrics.

4.4 Future Work

In future, human evaluation could be done at a larger scale and with all the eleven Indian languages available in the Samanantar dataset. In this study, manual translation evaluations were carried out using subjective judgments of translation quality. Future assessments will concentrate on a methodical approach using predefined rules to have the scoring much more consistent. Additional

metrics like BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) can also be used to test the translation quality. The prompt engineering technique aimed to improve the machine translation quality in the areas of commonly made mistakes did not yield the expected results. The findings highlight the significance of customized prompting techniques and the need for a balanced and context-aware approach to prompt engineering in machine translation, particularly when addressing specific linguistic phenomena such as plurality and honorific forms. So, to increase accuracy and fluency, especially for Indian languages, future research should concentrate on a more comprehensive examination of various prompts, possibly investigating the chaining of prompts and evaluating them more extensively.

Chapter 5

Conclusion

The following conclusions can be drawn from this study. Through human evaluation we found that translations obtained using LLMs from English to three Indian languages (Hindi, Marathi and Telugu) were good.

LLaMA consistently performed better than GPT-3.5 in all three languages, indicating that LLaMA has a better grasp of nuances in these Indic languages that are not fully captured by automated metrics. By comparing human evaluation scores to automatic MT metrics, we found that the automatic metrics are not satisfying metrics, but BERTScore was found to be better than BLEU and METEOR. For most Indian languages, when evaluated using automated metrics, GPT-4o performed the best. This implies that GPT-4o has a more robust understanding and generating capacity due to its access to more up-to-date and extensive training data.

Bibliography

[1] Koehn, Philipp. Statistical machine translation. Cambridge University Press, 2009.

[2] Chatzikoumi, Eirini. "How to Evaluate Machine Translation: A Review of Automated and Human Metrics." Natural Language Engineering, Sept. 2019, https://doi.org/10.1017/s1351324919000469

[3] Mujadia, Vandan, et al. "Assessing Translation Capabilities of Large Language Models Involving English and Indian Languages." ArXiv, Nov. 2023, https://www.semanticscholar.org/paper/088617a8862cfa372a62070916e88a5f10e7690b.

[4] Contributors to Wikimedia projects. BLEU- Wikipedia. Wikimedia Foundation, Inc., 16 Sept. 2024, BLEU

[5] Maria Stasimioti. "Meet the First Machine Translation Model Supporting All 22 Scheduled Indian Languages." Slator, June 29, 2023. https://slator.com/first-machine-translation-model-supporting-22-scheduled-indian-languages/

[6] Emmanuel Mark Ndaliro. *The Advent of Large Language Models (LLM).* Medium, 14 Mar. 2024, https://medium.com/@kram254/the-advent-of-large-language-models-llm-7c940dee7d83

[7] The History, Timeline, and Future of LLMs. toloka.ai, 21 Nov. 2024, https://toloka.ai/blog/history-of-llms/

[8] https://reverieinc.com/blog/machine-translation-for-indian-language-localisation/

[9] ATLTranslate. June 2023. Machine Translation: Navigating the Future of Language Technology. Retrieved from https://www.atltranslate.com/blog/machine-translate

[10] Advantages and Disadvantages of LLM's for Translation https://www.pairaphrase.com/llm-translation-advantages-disadvantages/

[11] Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., Mahalakshmi, J., Kakwani, D., Kumar, N., Pradeep, A., Nagaraj, S., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., & Khapra, M. S. (2021). Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *arXiv preprint arXiv:2104.05596*.

[12] Feng, Fangxiaoyu, et al. "Language-agnostic BERT sentence embedding." *arXiv preprint arXiv:2007.01852* (2020).

[13] https://blog.modernmt.com/understanding-mt-quality-BLeU-scores/

[14] Matić, Katharina. *Impact of sentence length on machine translation quality*. Diss. University of Zagreb. Faculty of Humanities and Social Sciences. Department of English language and literature, 2021, https://repozitorij.unizg.hr/islandora/object/ffzg:5057

[15] Pouget-Abadie, Jean, et al. "Overcoming the curse of sentence length for neural machine translation using automatic segmentation." *arXiv preprint arXiv:1409.1257* (2014). https://arxiv.org/abs/1409.1257.

[16] Zhang, Biao, Barry Haddow, and Alexandra Birch. "Prompting large language models for machine translation: A case study." International Conference on Machine Learning. PMLR, 2023. https://arxiv.org/abs/2301.07069.

[17] Graham, Yvette, et al. "Continuous measurement scales in human evaluation of machine translation." *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. 2013. https://aclanthology.org/W13-2305.pdf.

[18] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002. https://aclanthology.org/P02-1040.pdf.

[19] Zhang, Tianyi, et al. "BERTScore: Evaluating text generation with Bert." *arXiv preprint arXiv:1904.09675* (2019). https://arxiv.org/abs/1904.09675.

[20] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005. https://aclanthology.org/W05-0909.pdf.

[21] Wan, Z., Wang, X., Liu, C., Alam, S., Zheng, Y., Liu, J., Qu, Z., Yan, S., Zhu, Y., Zhang, Q., Chowdhury, M., & Zhang, M. (2023). Efficient Large Language Models: A Survey. *ArXiv, abs/2312.03863*.

[22] Brown, George, and Alexandrina Florescu. "Integrating Error Analysis and Prompting for Improved Translation Evaluation in Large Language Models." *Eastern European Journal for Multidisciplinary Research* 3.2 (2024): 162-166. https://snmzpublisher.com/index.php/eejmr/article/view/61

[23] Zhu, Wenhao, et al. "Multilingual machine translation with large language models: Empirical results and analysis." arXiv preprint arXiv:2304.04675 (2023).

[24] https://docs.kolena.com/metrics/bertscore/

[25] Kolar, Sanjana, and Rohit Kumar. "Multilingual tourist assistance using chatgpt: Comparing capabilities in hindi, telugu, and kannada." *arXiv preprint arXiv:2307.15376* (2023) https://doi.org/10.48550/arXiv.2307.15376

APPENDIX A

Interacting with LLMs using APIs:

For interacting with the LLMs I have used the REST APIs provided by the ChatGPT and FireworksAI platform

1. **Example API for GPT**

```
import requests
OPENAI_APIKEY = os.getenv('OPENAI_API_KEY')
url = "https://api.openai.com/v1/chat/completions"
headers = {
  'Content-Type': 'application/json',
  'Authorization': f'Bearer {OPENAI_APIKEY}'
}
payload = json.dumps({
      "model": "GPT-4o",
      "messages": [
        {
          "role": "user",
          "content": f"translate this sentence into hindi and just give me the translation only: He said case has been registered and investigations are on."
        }
      ]
})
response = requests.request("POST", url, headers=headers, data=payload)
```

2. Example Response for GPT-4o


```
{
  "id": "chatcmpl-AXDWWt10qMy86PCGoxbq32YQjMnsW",
  "object": "chat.completion",
  "created": 1732480172,
  "model": "gpt-3.5-turbo-0125",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "उसने कहा कि मामला दर्ज किया गया है और जांच जारी है।",
        "refusal": "None"
      },
      "logprobs": "None",
      "finish_reason": "stop"
    }
  ],
  "usage": {
    "prompt_tokens": 31,
    "completion_tokens": 51,
    "total_tokens": 82,
    "prompt_tokens_details": {
      "cached_tokens": 0,
      "audio_tokens": 0
```

```
    },
    "completion_tokens_details": {
      "reasoning_tokens": 0,
      "audio_tokens": 0,
      "accepted_prediction_tokens": 0,
      "rejected_prediction_tokens": 0
    }
  },
  "system_fingerprint": "None"
}
```


3. Example API Call for LLaMA-v3

```
from fireworks.client import Fireworks


FIREWORKS_APIKEY = os.getenv('FIREWORKS_API_KEY')
client = Fireworks(api_key=FIREWORKS_APIKEY)


response = client.chat.completions.create(
    model="accounts/fireworks/models/LLaMA-v3-70b-instruct",
    messages=[{
      "role": "user",
      "content": f”translate this sentence into hindi and just give me the translation only: He said case has been registered and investigations are on.",
    }],
  )
```

4. Examples Response for LLaMA-v3


```
{
  "id": "362e73ff-8ed4-4115-bd76-1fab9fcdea01",
  "object": "chat.completion",
  "created": 1732480281,
  "model": "accounts/fireworks/models/LLaMA-v3-70b-instruct",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "उन्होंने कहा कि मामला दर्ज किया गया है और जांच जारी है.",
        "tool_calls": null,
        "tool_call_id": null,
        "function": null,
        "name": null
      },
      "finish_reason": "stop",
      "logprobs": null,
      "raw_output": null
    }
  ],
  "usage": {
    "prompt_tokens": 34,
    "total_tokens": 65,
```

```
  "completion_tokens": 31
}}
```

APPENDIX B

Merging Datasets: A Sample Demonstration

A sample of five sentences is used to illustrate the merging of v0.2 dataset and the LabSE scores v0.2.1 dataset. This sample demonstrates how the src (source sentence) and tgt (target sentence) fields from the v0.2 dataset are combined with the corresponding LabSE scores (las) from the v0.2.1 dataset. The figure3, the two datasets were merged using the idx column as the common key. This was implemented using the pandas library in Python, as follows:

```
sents_las_500 = pd.merge(sents_500, las_score_500, on='idx', how='inner')
```

The merge () function combines the rows from both datasets where the idx value match. After merging, the resulting dataset contains the following fields

Figure 1 idx: A unique identifier, src: The source sentence in English, tgt: The target sentence in Telugu , las: The LabSE similarity score.

	idx	src	tgt	las
0	0	Rise again.	మళ్లీ ఉదయిస్తాడు.	[[0.8289942]]
1	1	How do we glorify Jehovahs undeserved kindness?	యెహోవా కృపను మనమెలా మహిమపరచవచ్చు?	[[0.80769086]]
2	2	India also continues to push back economically.	ఆర్థికంగా కూడా భారత్ వే గంగా పయనిస్తున్నది.	[[0.8420545]]
3	3	I remember my childhood days.	'విద్యార్థులను చూస్తుంటే నాకు చిన్నప్పటి రోజల...	[[0.751739]]
4	4	All transactions are made online.	ఆర్థిక లావాదేవలన్నీ ఆన్‌లైన్‌లోనే	[[0.79096377]]