Towards Large Language Model driven Reference-less Translation Evaluation for English and Indian Languages

Vandan Mujadia , Pruthwik Mishra , Arafat Ahsan and Dipti Misra Sharma

LTRC, IIIT Hyderabad, India
{vandan.mu, pruthwik.mishra}@research.iiit.ac.in,
{arafat.ahsan, dipti}@iiit.ac.in

Abstract

With the primary focus on evaluating the effectiveness of large language models for automatic reference-less translation assessment, this work presents our experiments on mimicking human direct assessment to evaluate the quality of translations in English and Indian languages. We constructed a translation evaluation task where we performed zero-shot learning, in-context example-driven learning, and fine-tuning of large language models to provide a score out of 100, where 100 represents a perfect translation and 1 represents a poor translation. We compared the performance of our trained systems with existing methods such as COMET, BERT-Scorer, and LABSE, and found that the LLM-based evaluator (LLaMA-2-13B) achieves a comparable or higher overall correlation with human judgments for the considered Indian language pairs (Refer figure 1).

1 Introduction

The field of natural language processing (NLP) and artificial intelligence (AI) has been transformed by the rapid advancements in Large Language Models (LLMs), as they have demonstrated their capabilities in a wide range of natural language processing tasks, including open/close question answering, summarization (Chang et al., 2023; Min et al., 2023), code completion, and code debugging (Wang et al., 2023; Zan et al., 2023; Surameery and Shakor, 2023), etc. These models have significantly impacted NLP applications by enhancing various aspects of language understanding, generation, and analysis (Zhao et al., 2023). As LLMs continue to evolve and improve, they hold immense potential for further advancements in NLP, paving the way for more sophisticated and intelligent solutions (Hadi et al., 2023).

Translation, on the other hand, plays a vital role in bridging the gap between different languages, enabling effective communication. With the advancement of machine translation systems (Ranathunga et al., 2023; Gala et al., 2023), there has been a significant shift in how translations are produced (Arivazhagan et al., 2019). Translation engines have made great progress in understanding languages and generating translations, but they still face numerous challenges in accurately conveying the intended meaning (Al Sharou and Specia, 2022; Wan et al., 2022; Hayakawa and Arase, 2020). While these systems have made remarkable progress, their evaluation remains an indispensable component in ensuring the accuracy and quality of translations produced by them (Sangal, 2022).

To date, human involvement has been considered one of the most reliable and effective ways to evaluate translations (Freitag et al., 2021;

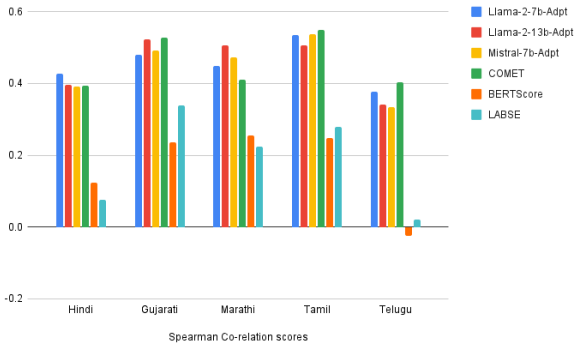

Figure 1: Spearman co-relation: Human translation evaluation vs different reference-less translation evaluation metrics. Llama-2-7b-Adapt (lora), Llama-2-13b-Adapt (lora), Mistral-7b-Adpt (lora), COMET-QE (https://github.com/Unbabel/COMET) are fine-tuned models using WMT-23 MT-QE corpora and evaluated on development corpora. BERTScore (https://github.com/Tiiiger/bert_score) and LABSE (https://huggingface.co/sentence-transformers/LaBSE) represents direct cosine-similarity scores.

Rivera-Trigueros, 2022; Guzmán et al., 2015). However, this requires a significant investment of time, expertise, and financial resources. The cost and effort involved can limit the frequency and scale at which human involved evaluation can be conducted. Therefore, it is crucial to develop a scalable, replicable, and efficient automatic method that mimics human evaluation to ensure reliable and effective translation assessment.

Recent developments in the field of automatic translation evaluation have demonstrated that techniques utilizing multilingual embeddings have a tendency to outperform other traditional approaches and display the strongest correlation with human assessments (Zerva et al., 2022). Notable instances of these techniques include BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020, 2022; Kocmi et al., 2022). To further improve the effectiveness of automatic translation evaluation, it is reasonable to investigate methodologies that leverage large language models, considering their notable capability for comprehension.

In this work, we aim to assess the capability of LLMs and utilize them for **reference-less translation evaluation** involving English and Indian languages. The research questions that we pose are as follows:

- Do LLMs possess zero-shot or in-context translation evaluation capabilities?
- How do fine-tuned LLMs compare with existing state-of-the-art translation evaluation methods such as COMET, BertScore, and LABSE?
  - Does translation evaluation fine-tuning for LLMs enhance their translation evaluation capabilities? Does fine-tuning improve their performance when applied to other unseen Indian languages in fine-tuning?
  - Does fine-tuning LLMs for both translation and translation evaluation increase their evaluation capabilities?

In this study, our objective is to gain a clearer understanding of the reference-less translation evaluation capabilities of several popular large language models. Specifically, we focus on Opt (Zhang et al., 2022), bloom (Scao et al., 2022), LLaMA-1 [1], MPT [2], Falcon (Penedo et al., 2023), LLaMA-2 (Touvron et al., 2023), and Mistral (Jiang et al., 2023) for translating English to five Indian languages: Hindi, Gujarati, Marathi, Tamil, and Telugu.

Reference-less translation evaluation generally involves assessing the translation of a given source language sentence without a specific reference translation. In this evaluation, an automatic reference-less evaluation system assigns a score to the translation, which is similar to the direct assessment (DA) score (Akhbardeh et al., 2021), commonly used in human translation evaluation. This score ranges from 1 to 100, with 1 representing a poor translation and 100 representing a perfect translation. In our work, we initially evaluate the reference-less translation evaluation capabilities of the raw large language models mentioned above in both zero-shot (Kojima et al., 2022) and in-context learning [3] (Brown et al., 2020) scenarios. Subsequently, we employ the LoRa parameter-efficient fine-tuning technique (Hu et al., 2021), as well as full fine-tuning, to refine the selected base LLM models. Additionally, considering that LLMs are known for their multi-task learning abilities (Radford et al., 2019), we explore the fine-tuning of LLMs on translation corpora alongside translation evaluation corpora to determine if translation-based fine-tuning improves overall translation evaluation performance.

The key findings of our study, presented in Figure 1, highlight the performance of our fine-tuned LLM-based reference-less translation evaluation models compared to various well-known reference-less evaluation methods, such as COMET, LABSE, and Bert-scorer, based on their Spearman correlation with human judgments.

Our findings emphasize the significant potential of large language models for reference-less translation evaluation tasks involving English and Indian languages. Raw LLMs do not inherently possess the capabilities for translation evaluation as they do not provide a score as an evaluation outcome. However, our multi-lingual LLM

[1] https://huggingface.co/decapoda-research/llama-7b-hf

[2] https://huggingface.co/mosaicml/mpt-7b

[3] https://ai.stanford.edu/blog/understanding-incontext/

based LORa-fine-tuned models (LLaMA-2-7b, LLaMA-2-13b and Mistral-7b) demonstrate competitive or superior correlation with human judgments compared to existing reference-less methods like COMET under same training and evaluation configurations. We did not observe any additional benefits when we perform translation evaluation fine-tuning with translation fine-tuning under the multi-task setting.

The results suggest that fine-tuned LLMs hold promise for translation evaluation in the targeted reference-less translation evaluation task. Our study represents an essential and pioneering milestone in assessing and enhancing the reference-less translation evaluation capabilities of LLMs, involving English and Indian languages.

2 Related Work

There are two main categories of automatic machine translation (MT) metrics: string-based metrics and pretrained models based metrics. In the following sub-section, we discuss them briefly one by one.

2.1 String-based metrics:

String-based metrics involve comparing the coverage of various substrings between the human-generated reference and MT translations. This includes metrics such as ChrF (Popović, 2015), BLEU (Papineni et al., 2002), METEOR (Gupta et al., 2010), or TER (Snover et al., 2009). String-based methods heavily rely on the quality of reference translations. However, they have the advantage of predictable and faster performance, as it is computationally easy to identify which substrings have the most impact on the score.

2.2 Pretrained models based metrics:

This category consists of metrics that utilize pretrained models to evaluate the quality of machine translation (MT) outputs given the source sentence, the human reference, or both. Evaluation metrics in this category include COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2019). These models are based on pretrained models such as XLM-RoBERTa (Conneau et al., 2020) and MBERT (Devlin et al., 2018). These metrics are not strictly dependent on reference quality and can better evaluate synonyms or paraphrases. Several studies (Mathur et al., 2020; Kocmi et al., 2022; Akhbardeh et al., 2021) have demonstrated their superiority over string-based metrics. However, it is important to note that the performance of these metrics is influenced by the data they have been trained on, which can introduce bias. Pretrained model-based metrics can be further categorized as reference-based and reference-free metrics. As the name suggests, reference-free metrics (also known as quality estimation metrics) do not require a human reference for evaluation. COMET-QE (Chimoto and Bassett, 2022; Rei et al., 2020), BERTScore (Zhang et al., 2019), and LABSE (Feng et al., 2022) fall under this category of metrics. In this direction, we develop a reference-less metric utilizing the large language model (LLM) and human-labeled data for fine-tuning. In the following section, we discuss our considered LLM models briefly.

3 Large Language Models

Language modeling, a well-established task in the field of natural language processing, has garnered significant attention over the years (Bellegarda, 2004; Bengio et al., 2000). This task involves predicting the probability of the next token in a sequence of words. Transformers have emerged as the fundamental architecture underlying many existing Large Language Models (Vaswani et al., 2017). Transformers based auto-regressive models like GPT (Brown et al., 2020; Radford et al., 2018, 2019) have played a crucial role in advancing Natural Language Processing (NLP). In this work, we used following base LLM models to check how we can utilise them for machine translation evaluation involving English and Indian Languages.

opt-6.7b[4] : The OPT-6.7b (Zhang et al., 2022) model has been extensively trained on the objective of causal language modeling (CLM) using English text with 6.7 billion parameters.

Bloom-7B[5] : BLOOM (Scao et al., 2022) was the first largest multilingual large language model with causal language modeling objective and supports 46 languages and 13 programming languages. It has 7,069,016,064 parameters.

LLaMA-7B[6]: LLaMA is a collection of foundation language models ranging from 7B to 65B parameters. In our experiments we evaluated LLaMA model with 7B parameters where 4096 is the em-

[4] https://huggingface.co/facebook/opt-6.7b
[5] https://huggingface.co/bigscience/bloom-7b1
[6] https://huggingface.co/decapoda-research/llama-7b-hf

bedding dimensions and 32 layers and 32 attention head.

MPT-7B[7] : Similar to above models MPT-7B model is trained on a large amount of data 1T tokens on causal language modeling objective.

Falcon[8] : Falcon (Penedo et al., 2023) is another large language model trained on causal language modeling (CLM) objective. Here, we utilised Falcon-7B model which is a 7B parameters for our experiments.

LLaMA-2-7B[9] **and LLaMA-2-13B**[10] : LLaMA 2 based models (Touvron et al., 2023) are also trained on causal language modeling (CLM) objective and pretrained on 2 trillion tokens of data from publicly available sources. In our experiments we have experimented with 7B and 13B LLaMA-2 models. LLaMA-2-7B network has 32 layers and 32 attention heads while LLaMA-2-13B has 40 layers and 40 attention heads.

Mistral-7B[11] : Mistral-7B Large Language Model (LLM) (Jiang et al., 2023) is a pre-trained on causal language modeling (CLM) objective with 7 billion parameters. It uses Sliding Window Attention (SWA) to handle longer sequences at smaller cost and Grouped-query attention (GQA) for faster inference which reduces the memory requirement during decoding. It has 4096 embedding dimension, 32 layers and 32 attention heads with context length of 8192.

4 Reference-less Translation Evaluation on Raw LLM

To evaluate the effectiveness of the LLMs mentioned above for translation evaluation tasks without reference, we conducted two different experiments. The first involved assessing the performance of the pre-trained (raw) LLM for reference-less translation evaluation. In the second experiment, we performed example-based in-context learning for the same purpose. Both experiments were carried out using translation evaluation data mentioned in the section 5.1.

As part of our experimental setup, we configured a prompting pipeline depicted in Figure 2. This pipeline involved using a Prompt Generator to generate specific prompts for the source and target language pairs for source and translation text. Subsequently, an LLM call is triggered to generate a response, which was then processed by a reply parser to obtain the actual translation. To ensure high-throughput and memory-efficient inference and serving for LLMs, we utilized the vLLM library[12] (Kwon et al., 2023). We conducted all experiments using a temperature parameter of 0, which ensures that the model behaves deterministically. By setting the temperature to 0, the model is constrained to select the word with the highest probability, effectively limiting its choice to the most likely option (Aksitov et al., 2023). All of our experiments are conducted using this library, and the models are deployed on A100, 40GB GPUs.

4.1 Zero-shot

We performed manual trials to determine the optimal prompt for zero-shot reference-less translation evaluation. These trials revealed that instructing an LLM to mimic a human evaluator and provide scores out of 100, combined with presenting the text in JSON format, resulted in better results (prompt presented below).

Zero-shot Translation Evaluation Prompt Example:

You are an experienced translation evaluator and you need to evaluate a translation for <Source Language> language to <Target Language> language.

<Source Language>: <Source Language Text>

<Target Language>: <Target Translated Text>

The evaluation score out of 100 is

4.2 Example-based (In-Context Learning - ICL)

In a similar manner, we identified and adjusted the prompt for example-based in-context learning with LLM. This specific prompt is outlined in the Example above (ICL Translation Prompt). In all of our experiments, specific to language pairs, we used a single and identical human translation evaluation with a score as a contextual learning

[7] https://huggingface.co/mosaicml/mpt-7b

[8] https://huggingface.co/tiiuae/falcon-7b

[9] https://huggingface.co/meta-llama/Llama-2-7b-hf

[10] https://huggingface.co/meta-llama/Llama-2-13b-hf

[11] https://huggingface.co/mistralai/Mistral-7B-v0.1

[12] https://github.com/vllm-project/vllm

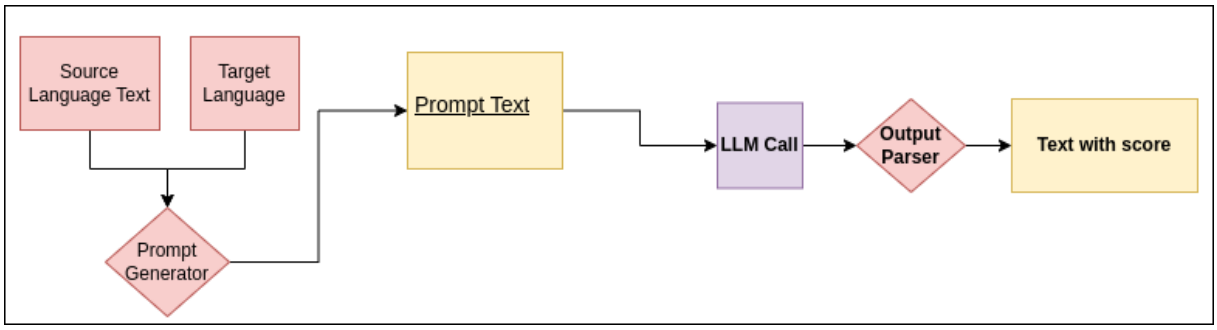

Figure 2: Prompting Mechanism for reference-less translation evaluation

example before executing the actual translation evaluation command using the training data mentioned in sub-section 5.1.

Example-based ICL Translation Evaluation Prompt Example:

If the <Source Language> to <Target Language> translation score by human for '<Source Language Text>' to '<Target Language Translation>' is <Human Score> from 100 then following that, if you are an experienced translation evaluator and you need to evaluate a translation for <Source Language> language to <Target Language> language.

<Source Language>: <Source Language Text>

<Target Language>: <Target Translated Text>

The evaluation score out of 100 is

5 Fine-tuning LLM for Reference-less Translation Evaluation

To assess the potential enhancement in reference-less translation evaluation performance of LLMs beyond the zero-shot LLM baseline, we performed fine-tuning using the training corpus and method described below.

5.1 Human Judgement Train and Development Corpora

For this task, we take the publicly available datasets released under the quality estimation shared task [15] (MTQE) in WMT for Indian languages. English to five Indian languages are included in the shared task this year. The dataset comprises of 2 predominant language families: 3 from Indo-Aryan family and 2 Dravidian languages. The details of the train and development sets are presented in the table 2.

Method	Hyper-para	Value
LoRA/Full	LoRA	PEFT[13]; FSDP[14]
	rank	8
	dropout	0.05
	learning rate	1e-4
	batch size	4
	epochs	5

Table 1: Hyper-parameter configurations of LoRA based and full fine-tuning for 4*A100 40GB GPUs

Lang Pair	#Train_Snts	#Dev_Snts
English-Hindi	7000	1000
English-Gujarati	7000	1000
English-Marathi	26000	1000
English-Tamil	7000	1000
English-Telugu	7000	1000

Table 2: Train and Development Set Details for Reference-less Translation Evaluation

5.1.1 Using Min-Max scaling on z-scores

In order to fine-tune large language models on the aforementioned training data, we utilized the mean of z-scores[16], followed by language-specific min-max-based re-scaling [17] from 1 to 100 for the entire corpora.

[15] https://wmt-qe-task.github.io/

[16] https://www.investopedia.com/terms/z/zscore.asp

[17] https://en.wikipedia.org/wiki/Feature_scaling

Model/Language Pair EN-	Hindi			Gujarati			Marathi			Tamil			Telugu		
Co-relation scores	S	P	K	S	P	K	S	P	K	S	P	K	S	P	K
Llama-2-7b-Adpt	**0.4268**	**0.5303**	0.3056	0.4795	0.585	0.3757	0.4481	0.4939	0.3182	0.5345	**0.6632**	0.3967	**0.3772**	0.3324	0.2893
Llama-2-13b-Adpt	0.3967	0.587	0.2824	**0.5226**	**0.604**	**0.419**	**0.5062**	**0.5485**	**0.3689**	0.5057	0.6501	0.3774	0.3408	0.2976	0.2708
Mistral-7b-Adpt	0.3923	0.5295	0.2822	0.4908	0.5724	0.3994	0.4736	0.5518	0.3364	0.5372	0.615	0.3969	0.333	0.2685	0.2662
Llama-2-7b-Adpt-Full	0.4003	0.5249	**0.3068**	0.4959	0.5503	0.3958	0.4524	0.4938	0.3402	**0.5413**	0.6535	**0.4138**	0.2940	0.2500	0.2360
Llama-2-13b-Adpt-Full	0.3004	0.2164	0.2144	0.3619	0.3294	0.2994	0.4424	0.4229	0.3245	0.4408	0.484	0.3454	0.2849	0.1690	0.2271
Llama-2-7b-Adpt-Trans	0.3273	0.4602	0.2316	0.4717	0.5389	0.3824	0.4196	0.4554	0.2993	0.4921	0.6404	0.359	0.2800	0.2266	0.2137
Llama-2-13b-Adpt-Trans	0.4011	0.5165	0.2888	0.5185	0.6039	0.4088	0.4451	0.4956	0.3146	0.5234	0.6287	0.3892	0.3264	0.2714	0.2544
Mistral-7b-Adpt-Trans	0.3507	0.4611	0.2525	0.4638	0.5105	0.3671	0.4308	0.4849	0.3174	0.468	0.5796	0.3425	0.3748	**0.3403**	**0.3042**

Table 3: Correlation between Human judgement and Fine-tuned LLM with reference-less translation evaluation task for English to 4 Indian Languages is shown. Here, S, P, and K represent Spearman's Rank Correlation Coefficient, Pearson Correlation Coefficient, and Kendall's Rank Correlation Coefficient, respectively.

5.2 Training Data for Translation Fine-Tuning

To fine-tune LLMs for translation along with translation evaluation task under multi-task setting, we used the publicly available Bharat Parallel Corpus Collection (BPCC) designed for English to 22 Indic languages. We used BPCC-Human dataset, containing 2.2 million English-Indic pairs for this fine-tuning purpose.

5.3 LLM Fine-tuning Details

Considering the raw LLM performance, model parameters, and resource constraints, we selected a subset of LLMs for the fine-tuning process. Specifically, we opted for LLaMA-2-7b, LLaMA-2-13b, and Mistral-7B for the fine-tuning experiment. For these selected LLMs, we decided to perform low-rank adaptation-based fine-tuning (Hu et al., 2021) as well as full fine-tuning. In the fine-tuning process, we conducted a multi-lingual approach by considering corpora from all languages. Additionally, we conducted another experiment under multi-task settings, where we fine-tuned both LLaMA-2-7b and LLaMA-2-13b for reference-less translation evaluation and translation tasks. For the translation task, we utilized the translation dataset mentioned earlier (in subsection-5.2), along with a simple JSON prompt instructing the model to translate a given sentence from the source language to the target language.

For both types of fine-tuning LLMs, we utilized the llama-recipes codebase[18] which provides an efficient implementation for LoRa-based adaptor fine-tuning with PEFT (Mangrulkar et al., 2022). For more details, please refer to the llama-recipes documentation [19]. The hyperparameters for the fine-tuning process are specified in Table 1.

Score Bin	Hindi	Gujarati	Marathi	Tamil	Telugu
99	2	0	0	1	2
95	5	43	32	9	0
90	943	788	940	941	930
80	50	155	27	45	68
70-80	0	14	1	4	0

Table 4: LLAMA-2-7B zero-shot performance on reference-less translation evaluation for English to 5 languages. Here, the 'Score Bin' represents the score (out of 100) produced by LLAMA-2-7B. The corresponding language columns represent the respective sentence pairs that got that particular score bin. A total of 1000 samples were involved per language pair.

6 Results and Discussion

In order to determine the performance of the LLM models, we examined the correlation between human judgments and metric output scores. This correlation served as the primary evaluation factor for identifying the most effective fine-tuned model. To conduct this analysis on the aforementioned development corpora, we utilized Spearman's Rank Correlation Coefficient, Pearson Correlation Coefficient, and Kendall's Rank Correlation Coefficient. To calculate these correlations, we employed the SciPy library[20].

6.1 Zero shot vs ICL based Reference-less Translation Evaluation over Raw LLMs

Overall, the correlation scores with human evaluation for both Raw LLMs and In Context Learning (ICL) based LLMs in reference-less translation evaluation are substandard. All LLMs exhibited poor performance, which could be attributed to a lack of understanding or knowledge of the translation evaluation task. In our manual analysis, we observed that LLMs consistently provide a common number close to 100, regardless of the transla-

[18] https://github.com/facebookresearch/llama-recipes/

[19] https://github.com/facebookresearch/llama-recipes/blob/main/docs/LLM_finetuning.md

[20] https://scipy.org/

English to INs	Average Across Languages		
Co-relation scores	**S**	**P**	**K**
COMET-QE	0.4568	0.5236	0.32042
Llama-2-7b-Adpt	0.45322	0.52096	0.3371
Llama-2-13b-Adpt	**0.4574**	**0.53744**	**0.3437**
Mistral-7b-Adpt	0.44538	0.50744	0.33622
Llama-2-7b-Adpt-Full	0.43678	0.4945	0.33852
Llama-2-13b-Adpt-Full	0.36608	0.32434	0.28216
Llama-2-7b-Adpt-Trans	0.39814	0.4643	0.2972
Llama-2-13b-Adpt-Trans	0.4429	0.50322	0.33116
Mistral-7b-Adpt-Trans	0.41762	0.47528	0.31674

Table 5: The correlation between Human judgement and Fine-tuned LLM and COMET-QE with the reference-less translation evaluation task is shown, averaged across English to 4 Indian Languages (Hindi, Gujarati, Marathi, Tamil and Telugu). Here, S, P, and K represent Spearman's Rank Correlation Coefficient, Pearson Correlation Coefficient, and Kendall's Rank Correlation Coefficient, respectively.

tion quality, for zero-shot evaluation (refer Table 4). For in-context learning, LLMs simply mimic the example translation evaluation scores. Therefore, in response to our initial question, it can be concluded that **Raw LLMs do not possess inherent translation evaluation capabilities, either with zero-shot or example-driven contextual learning**. Refer to the appendix for examples and correlation details.

6.2 Fine-Tuned LLM driven Reference-less Translation Evaluation

We conducted an evaluation to compare the performance of our Fine-Tuned LLM models. The comparison results for English to 5 Indian language reference-less translation evaluation correlation with human judgments are presented in Table 3. It is evident that the highest correlations are achieved with LoRa-based fine-tuning methods for LLaMa-2-7b and LLaMa-2-13b models (indicated by -Adpt). On the other hand, Full Fine-tuning (indicated by -Adpt-Full) demonstrates maximum correlation scores for Tamil. However, overall, there is a low correlation for Telugu, indicating the need for further exploration, as Telugu language may have less representation compared to other languages in base LLM models. Additionally, it is worth noting that multi-task learning-based fine-tuning (indicated by -Adpt-Trans) does not lead to performance improvement compared to single-task fine-tuning for translation evaluation.

Table 5 displays the average performance across all English to 4 language pairs for different fine-tuned LLMs and COMET-QE. We trained COMET-QE following the COMET architecture as described in COMET by Unbabel[21], using the same training data configuration as our LLM fine-tuning. It is worth noting that the Llama-2-13b-Adpt model, which is an adapted model using the LORa method, achieves a high overall average human correlation across language pairs. This indicates the superior performance of LLM-driven reference-less translation evaluation. Hence, it highlights the potential of LLMs for the translation evaluation task involving English and Indian languages.

In answer to the question posed in the introduction, **Fine-tuning LLMs does indeed improve translation reference-less evaluation capabilities. However, when it comes to multi-task fine-tuning, including the translation task does not result in better performance compared to fine-tuning focused solely on the translation evaluation task.**

Based on our manual analysis of outputs, it became evident that the scores provided by the Llama-2-13b-Adpt model indicate superior quality. The model demonstrates the ability to detect translation differences, which is indeed reflected in its scores. Refer to the appendix for examples and correlation details.

7 Limitations

We conducted all our experiments using high-performance GPUs, specifically the A100-40GB, which may not be readily available for everyone to reproduce these experiments and obtain the same results due to compute limitations. To address this constraint, our goal is to make all outputs, including model outputs and results, openly accessible [22] for further research.

8 Conclusion and Future Work

Our experiments and results indicate that fine-tuned LLMs show promise for translation evaluation in the targeted reference-less translation evaluation task. The findings call for further analysis and understanding of translation evaluation tasks

[21] https://github.com/Unbabel/COMET

[22] https://github.com/vmujadia/LLMT-Eval

involving large language models.

As part of future work, we plan to incorporate additional Indian languages for the reference-less translation evaluation task. Furthermore, we aim to utilize large language models for reference-driven translation evaluation, encompassing English and a wider range of Indian languages.

Overall, our study represents a significant milestone in assessing and enhancing the reference-less translation evaluation capabilities of LLMs, specifically in the context of English and Indian languages. For the future of automatic translation evaluation, we strongly advocate for solutions based on large language models as the primary method of evaluation. We have also released our models and predictions for future research.

Acknowledgement

We would like to extend our sincere appreciation to Palash Gupta and Khoushik Ananth for their invaluable contributions at various stages of this project. The Ministry of Electronics and In- formation Technology, Government of India, has generously funded this endeavor, as part of the Sanction Order : 11(1)/2022-HCC(TDIL)-Part(2)/A/B/C and the Administrative Approval: 11(1)/2022-HCC(TDIL)-Part(2).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. 2023. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models. *arXiv preprint arXiv:2302.05578*.

Khetam Al Sharou and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Jerome R Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech communication*, 42(1):93–108.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Everlyn Chimoto and Bruce Bassett. 2022. COMET-QE and active learning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4735–4740, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021.

Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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

Ankush Gupta, Sriram Venkatapathy, and Rajeev Sangal. 2010. Meteor-hindi: automatic mt evaluation metric for hindi as a target. In *Proceedings of ICON-2010: 8th international conference on natural language processing, Macmillan Publishers. India*.

Francisco Guzmán, Ahmed Abdelali, Irina Temnikova, Hassan Sajjad, and Stephan Vogel. 2015. How do humans evaluate machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 457–466, Lisbon, Portugal. Association for Computational Linguistics.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects.

Takeshi Hayakawa and Yuki Arase. 2020. Fine-grained error analysis on English-to-Japanese machine translation in the medical domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 155–164, Lisboa, Portugal. European Association for Machine Translation.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

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

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Irene Rivera-Trigueros. 2022. Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, 56(2):593–619.

Rajeev Sangal. 2022. Evaluating MT Systems: A Theoretical Framework. *arXiv preprint arXiv:2202.05806*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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

Matthew G Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23:117–127.

Nigar M Shafiq Surameery and Mohammed Y Shakor. 2023. Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290*, 3(01):17–22.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Yu Wan, Baosong Yang, Derek Fai Wong, Lidia Sam Chao, Liang Yao, Haibo Zhang, and Boxing Chen. 2022. Challenges of Neural Machine Translation for Short Texts. *Computational Linguistics*, 48(2):321–342.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Wang Yongji, and Jian-Guang Lou. 2023. Large language models meet NL2Code: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7443–7464, Toronto, Canada. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Appendix

A.1 Examples (English-Hindi)

Example: Translation Evaluation Output-1 (Llama-2-13b; Zeroshot; Human - 88)

You are an experienced translation evaluator and you need to evaluate a translation for English language to Hindi language.

English: Scheduled Castes numbered 4,641 and Scheduled Tribes numbered 1,535.
Hindi: अनुसूचित जातियों की संख्या 4,641 और अनुसूचित जनजातियों की संख्या 1,535 है।

The evaluation score out of 100 is **80**.

Example: Translation Evaluation Output-2 (Llama-2-13b; Zeroshot; Human - 84.5)

You are an experienced translation evaluator and you need to evaluate a translation for English language to Hindi language.

English: The region then came under British influence.
Hindi: यह क्षेत्र तब ब्रिटिश प्रभाव में आ गया था।

The evaluation score out of 100 is **50**.

Example: Translation Evaluation Output-3 (Llama-2-13b; ICL (In-context Learning); Human - 84.5)

If the English to Hindi translation score by human for 'We shall first make a plot and then view the history and save it, for that we can type x is equal to linspace within brackets minus 2 into pi comma 2 into pi comma 100 plot within bracket x comma xsinx cosx and hit enter.' to 'हम पहले एक प्लॉट बनाएंगे और फिर इतिहास देखेंगे और इसे सहेजेंगे, इसके लिए हम टाइप कर सकते हैं x कोष्ठक के भीतर लाइनस्पेस के बराबर है−2 में pi कोमा 2 में pi कोमा 100 प्लॉट के भीतर x कोमा xsinx cosx और प्रवेश करें।' is 77.67 from 100 then following that, if you are an experienced translation evaluator and you need to evaluate a translation for English language to Hindi language.

English: The region then came under British influence.
Hindi: यह क्षेत्र तब ब्रिटिश प्रभाव में आ गया था।

The evaluation score out of 100 is **77.67**.

Example: Translation Evaluation Output-3 (Llama-2-13b; Fine-Tuned; Human - 84.5)

You are an experienced translation evaluator and you need to evaluate a translation for English language to Hindi language.

English: The region then came under British influence.
Hindi: यह क्षेत्र तब ब्रिटिश प्रभाव में आ गया था।

The evaluation score out of 100 is **85.21**.

Example: Translation Evaluation Output-4 (Llama-2-13b; Full Fine-Tuned; Human - 84.5)

You are an experienced translation evaluator and you need to evaluate a translation for English language to Hindi language.

English: The region then came under British influence.
Hindi: यह क्षेत्र तब ब्रिटिश प्रभाव में आ गया था।

The evaluation score out of 100 is **84.78**.

Example: Translation Evaluation Output-5 (Llama-2-13b; Fine-Tuned with Translation; Human - 84.5)

You are an experienced translation evaluator and you need to evaluate a translation for English language to Hindi language.

English: The region then came under British influence.
Hindi: यह क्षेत्र तब ब्रिटिश प्रभाव में आ गया था।

The evaluation score out of 100 is **85.67**.

Example: (*Misjudgment*) Translation Evaluation Output-6 (Llama-2-13b; Fine-Tuned; Human - 56.25)

You are an experienced translation evaluator and you need to evaluate a translation for English language to Hindi language.

English: The central Sahara is hyperarid, with sparse vegetation.
Hindi: मध्य सहारा हाइपररिड है, जिसमें बहुत कम वनस्पति है।

The evaluation score out of 100 is **89.02**.

A.2 Results

Model/Language Pair: English-	Hindi			Gujarati			Marathi			Tamil			Telugu		
Co-relation scores	**S**	**P**	**K**	**S**	**P**	**K**	**S**	**P**	**K**	**S**	**P**	**K**	**S**	**P**	**K**
bloom-7b1	-0.0058	-0.0116	-0.0047	0.0154	0.0216	0.0129	0.0122	0.0222	0.0098	-0.0256	-0.015	-0.0208	0.0268	0.0091	0.0222
falcon-7b	0.0211	0.0466	0.0162	-0.0108	0.0072	-0.0079	-0.1052	-0.0083	-0.0811	0.0651	0.0787	0.0496	0.1151	0.0775	0.0882
llama-7b	0.02	0.019	0.0178	-0.035	-0.014	-0.0276	-0.0239	-0.0243	-0.019	0.053	0.0311	0.0441	0.0519	0.0292	0.0473
Llama-2-7b	0.022	0.0223	0.0177	-0.0301	-0.0135	-0.0249	-0.0222	-0.0212	-0.018	0.0674	0.0328	0.0547	0.0606	0.0304	0.0504
Mistral-7B-v0.1	-0.0474	-0.0582	-0.0382	0.0741	0.063	0.0615	0.077	0.0766	0.0628	-0.0487	-0.054	-0.0367	-0.0143	-0.0062	-0.0116
opt-6.7b	NAN	NAN	NAN	NAN	NAN	NAN	0.0826	0.0229	0.0673	NAN	NAN	NAN	NAN	NAN	NAN
mpt-7b	-0.0168	-0.0003	-0.0137	0.0098	0.0204	0.0082	0.0234	-0.0095	0.0191	NAN	NAN	NAN	0.0405	0.0312	0.0337
Llama-2-13b	0.0347	-0.0038	0.0267	0.0862	0.0965	0.0685	-0.0691	-0.0698	-0.0551	0.0748	0.0276	0.0591	0.0407	-0.0077	0.0323

Table 6: Correlation between Human judgement and Zero-shot LLM with reference-less translation evaluation task for English to 4 Indian Languages is shown. Here, S, P, and K represent Spearman's Rank Correlation Coefficient, Pearson Correlation Coefficient, and Kendall's Rank Correlation Coefficient, respectively.

Model/Language Pair: English-	Hindi			Gujarati			Marathi			Tamil			Telugu		
Co-relation scores	**S**	**P**	**K**	**S**	**P**	**K**	**S**	**P**	**K**	**S**	**P**	**K**	**S**	**P**	**K**
bloom-7b1	-0.0496	-0.0398	-0.0405	0.0093	0.0075	0.0078	0.1191	0.1094	0.0973	0.0233	-0.0143	0.0191	-0.0461	-0.0381	-0.0384
falcon-7b	0.1374	0.0175	0.1112	0.03496	0.0497	0.0284	0.0008	0.0009	0.0008	0.127	0.1015	0.102	-0.1161	-0.0825	-0.0938
llama-7b	-0.0492	-0.0615	-0.0402	NAN	NAN	NAN	NAN	NAN	NAN	-0.0416	-0.0262	-0.034	NAN	NAN	NAN
Llama-2-7b	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	-0.0416	-0.0262	-0.034	NAN	NAN	NAN
Mistral-7B-v0.1	0.0382	0.0144	0.0312	0.0287	0.0382	0.0239	NAN	NAN	NAN	-0.0023	-0.0117	-0.0018	0.008	0.0348	0.0066
opt-6.7b	-0.0576	-0.0436	-0.0471	-0.0095	0.0005	-0.0079	0.0278	0.0129	0.0227	0.1141	0.0936	0.0933	0.0918	0.0791	0.0764
mpt-7b	0.0661	0.0454	0.054	-0.0539	-0.0498	-0.0449	0.0639	0.0431	0.0522	0.1546	0.137	0.1263	-0.1268	-0.0826	-0.1054
Llama-2-13b	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	NAN	-0.0006	-0.0167	-0.0005	0.0159	0.0214	0.0132

Table 7: Correlation between Human judgement and In Context Learning LLM with reference-less translation evaluation task for English to 4 Indian Languages is shown. Here, S, P, and K represent Spearman's Rank Correlation Coefficient, Pearson Correlation Coefficient, and Kendall's Rank Correlation Coefficient, respectively.