

# Assessing Translation Capabilities of Large Language Models involving English and Indian Languages

Vandan Mujadia, Ashok Urlana, Yash Bhaskar, Penumalla Aditya Pavani,  
Kukkapalli Shravya, Parameswari Krishnamurthy, Dipti Misra Sharma

IIIT Hyderabad, India

{vandan.mu, ashok.urlana, yash.bhaskar}@research.iiit.ac.in,  
{aditya.pavani, kukkapalli.shravya}@students.iiit.ac.in,  
{param.krishna, dipti}@iiit.ac.in

## Abstract

Generative Large Language Models (LLMs) have achieved remarkable advances in various NLP tasks. In this work, our aim is to explore the multilingual capabilities of large language models by using machine translation as a task involving English and 22 Indian languages. We first investigate the translation capabilities of raw large-language models, followed by exploring the in-context learning capabilities of the same raw models. We fine-tune these large language models using parameter-efficient fine-tuning methods such as LoRA and additionally with full fine-tuning. Through our study, we have identified the model that performs best among the large language models available for the translation task.

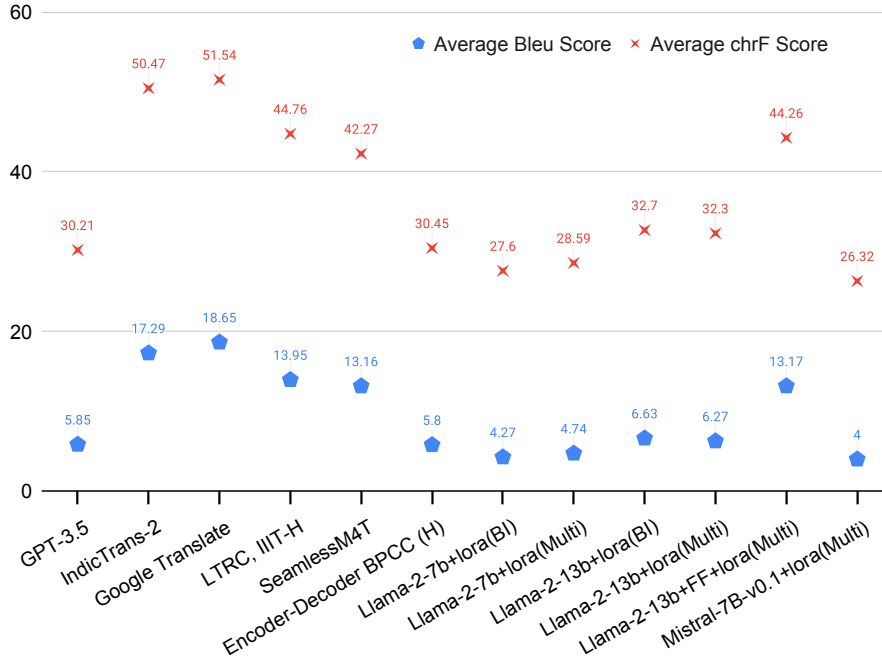
Our results demonstrate significant progress, with average BLEU scores of 13.42, 15.93, 12.13, 12.30, and 12.07, as well as chrF scores of 43.98, 46.99, 42.55, 42.42, and 45.39, respectively, using two-stage fine-tuned LLaMA-13b for English to Indian languages on IN22 (conversational), IN22 (general), flores200-dev, flores200-devtest, and newstest2019 testsets. Similarly, for Indian languages to English, we achieved average BLEU scores of 14.03, 16.65, 16.17, 15.35 and 12.55 along with chrF scores of 36.71, 40.44, 40.26, 39.51, and 36.20, respectively, using fine-tuned LLaMA-13b on IN22

(conversational), IN22 (general), flores200-dev, flores200-devtest and newstest2019 testsets. Overall, our findings highlight the potential and strength of large language models for machine translation capabilities, including languages that are currently underrepresented in LLMs.

## 1 Introduction

Generative Large Language Models (LLMs) have made significant performance improvements in various natural language processing (NLP) tasks, showcasing exceptional progress in a wide range of applications (Xuanfan and Piji, 2023; Xi et al., 2023). These tasks range from open domain question answering, where LLMs excel in providing accurate and coherent responses, to instruction-based tasks such as code completion, where LLMs can generate code snippets based on given prompts (Vaithilingam et al., 2022). LLMs have also demonstrated proficiency in tasks such as writing essays, grammar checks (Wu et al., 2023a), and text summarization, where they can produce high quality results (Chang et al., 2023). These advances have been observed mainly in tasks centered on English. The popular LLMs support several natural languages. The performance of some languages other than English is not yet on par or yet to be evaluated (Lai et al., 2023; Zhu et al., 2023).

A multilingual country like India, where over 364+ languages and dialects<sup>1</sup> are spoken across its vast territory, presents a multitude of challenges across various domains due to language barriers (Zieliński and others, 2021), such as day-to-day communication, education (Steigerwald et al., 2022), business, healthcare (Mehandru et al., 2022),



**Figure 1: LLMs based Machine Translation performance comparison with public systems for English to Indian Languages.** BLEU and chrF scores are averaged over 22 Indian Languages and 5 different benchmark data sets. The available MT systems are GPT-3.5 (GPT-3.5 Davinci, by OpenAI), IndicTrans-2, Google Translate, LTRC-IIIT-H, SeamlessM4T. LLaMA-2-7b and LLaMA-2-13b are evaluated as LLM based fine-tuned MT systems are namely LLaMA-2-7b+lora (Multi), LLaMA-2-13b+lora (Multi), and LLaMA-2-13b+FF+lora (Multi). Encoder-Decoder BPCC (H) represents scores for encoder decoder based transformer model trained on BPCC Human Training data.

tourism, governance, and more. Recent advancements in the field of Large Language Models may offer solutions to these challenges tailored to Indian languages.

To test whether decoder-based LLMs can effectively overcome language barriers, it is crucial to assess the proficiency of large language models in handling Indian languages. Machine Translation, as a critical multilingual task, could be an ideal option to explore the multilingual capabilities of existing models. Hence, we can formulate the question to assess the proficiency of large language models in handling Indian languages as follows: **How effectively do large language models perform in multilingual tasks like Machine Translation, particularly when dealing with Indian languages?**

In this work, our main contribution is to address the following points in response to the above question.

- What are the directions for utilizing or adapting available Large Language Models for Indian Languages?
  - How do LLMs perform in translating a wide range of Indian languages un-

der zero-shot and in-context learning settings?

- Does LLM fine-tuning improve the translation capabilities of Large Language Models? How do they perform in low-resourced MT languages?
- The Impact of LLM Vocabulary on the Performance of Large Language Models in Translation Tasks.

To address the above questions, we assess the translation capabilities of popular large language models (opt, bloom, LLaMA-1, MPT, Falcon, LLaMA-2, and Mistral (\$B)) that involve English and 22 scheduled Indian languages (Assamese, Bangla, Bodo, Dogri, Konkani, Gujarati, Hindi, Kannada, Kashmiri, Maithili, Malayalam, Marathi, Meitei, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu, and Urdu). We initially examine the translation capabilities of these raw large language models mentioned above (§4.1). Subsequently, we explore their in-context learning abilities (§4.1). Additionally, we fine-tune the base models using parameter-efficient fine-tuning methods, specifically LoRa (§5). Furthermore, we investigate

the potential of two-stage fine-tuning for large language models, which involves full parameter fine-tuning in the first stage, followed by LoRa-based adapter fine-tuning (§5).

The key findings of our work, summarized in Figure 1, highlight the performance of our LLM-based machine translation fine-tuned models compared to various known translation engines. These engines range from commercial (Google<sup>2</sup>, GPT-3.5<sup>3</sup>) to open source (IndicTrans-2<sup>4</sup>, LTRC-IIIT-H<sup>5</sup>, seamless4t<sup>6</sup>), our trained Encoder-Decoder BPCC (H) model (Appendix A), traditional supervised encoder-decoder translation models (Google, IndicTrans-2, LTRC-IIIT-H), and decoder-driven causal large language model-based translation systems (GPT-3.5).

Our findings underscore the significant potential of large language models for translation tasks involving English and Indian Languages. Although raw LLMs (LLaMA-2-7b and LLaMA-2-13b) do not perform well in translation tasks, our two-stage MT fine-tuned models (LLaMA-2-13b + FF + lora (Multi)) yield comparative results even with minimal parallel corpora. This suggests that LLMs have the potential to possess multilingual capabilities for translating into underrepresented languages, which can be further enhanced by fine-tuning. This work will be a crucial and pioneering milestone in evaluating LLMs for language representation and assessing their translation capabilities for a diverse range of Indian languages, especially those with limited available resources.

## 2 Related Work

Recent advancements in machine translation have shown that neural machine translation (NMT) has made significant strides in terms of output fluency and translation quality, especially when ample parallel data are available (Barrault et al., 2020). However, the scarcity or absence of parallel data poses a challenge for most language pairs. In the case of Indian languages, recent developments have tried to address this issue by introducing a new state-of-the-art approach: multilingual machine translation involving Indian languages and English (Wang et al., 2021; Dabre et al., 2020; Madaan and Sadat,

2020). This approach uses a single script for machine translation, taking advantage of the lexical and syntactic similarities that arise from genetic and contact relatedness among Indian languages (Gala et al., 2023; Eriguchi et al., 2022; Bapna and Firat, 2019).

In the field of LLM driven machine translation, in-context learning has gained significant attention (Wu et al., 2023b). The use of large language models (LLMs) for multilingual machine translation has been a topic of interest (Zhang et al., 2023). Recent studies have evaluated the translation capabilities of LLMs for different language directions, with a focus on models like ChatGPT (Bang et al., 2023).

In particular, (Xu et al., 2023) proposed a two-stage fine-tuning approach for machine translation using LLM, involving fine-tuning in monolingual data followed by fine-tuning on a small set of high-quality parallel data. Our work represents the first study to specifically explore machine translation involving Indian languages using large language models. The details on Large Language Models are presented in the Appendix B.

## 3 Indian Languages representation in LLMs

Pre-trained (or Base/Raw) large language models are trained on a huge amount of language data, and some of these models are trained on multiple languages (Naveed et al., 2023). However, their training mainly focuses on the English text (Penedo et al., 2023a). The emphasis on English is due to its substantial presence on the Internet and its widespread usage in business contexts. For the purpose of this work, our objective is to assess the effectiveness of these models in Machine Translation tasks that involve both English and Indian Languages. Consequently, it becomes crucial to investigate the representation of Indian languages within these large language models.

An approach to investigating the representation of Indian languages within a large language model can involve analyzing the frequency of language-specific words and sentences used during the training of these models. Unfortunately, it is not possible to perform this analysis as the training data used for these models are not publicly accessible. LLaMA-2, in particular, has mentioned that its pretraining corpus consists mainly of English and may not be optimal for other languages (Touvron et al., 2023b). However, it is worth mentioning that approximately

<sup>2</sup><https://translate.google.co.in/>

<sup>3</sup><https://chat.openai.com/>

<sup>4</sup><https://github.com/AI4Bharat/IndicTrans2>

<sup>5</sup><https://ssmt.iiit.ac.in/translate>

<sup>6</sup>[https://github.com/facebookresearch/seamless\\_communication](https://github.com/facebookresearch/seamless_communication)

Language Family Language	Indo-Aryan												Dravidian				Sino-Tibetan		Austroasiatic				
	asm	ban	kas	snd	urd	doi	hin	gom	mai	mar	nep	san	guj	odi	pan	kan	mal	tam	tel	mni	bxr	sat	
Language Script	Bangla		Perso-Arabic			Devanagari							Gujarati		Odia	Gurmukhi	Kannada	Malayalam	Tamil	Telugu	Meitei	Devanagari	Ol Chik
No of Letters in Unicode	96		256			128							91		91	80	91	118	72	100	56	96	48
Models (Vocab)																							
BLOOM (250680)	(48,48)		(49,207)										(67,61)	(57,34)	(56,35)	(55,25)	(62,29)	(66,52)	(46,26)	(61,39)	(00,56)	(67,29)	(00,48)
FALCON (65024)	(00,96)		(12,244)										(2,126)	(00,91)	(00,91)	(00,72)	(0,100)	(00,56)	(02,70)	(04,96)	(00,56)	(02,94)	(00,48)
LLAMA-1.2 (32024)	(24,72)		(45,211)										(38,90)	(01,90)	(00,91)	(04,76)	(02,89)	(33,155)	(19,53)	(01,99)	(00,56)	(38,58)	(00,48)
MISTRAL (32052)	(34,62)		(47,209)										(43,85)	(05,86)	(00,91)	(02,78)	(18,73)	(04,116)	(22,50)	(11,89)	(00,56)	(43,53)	(00,48)
MPT (50277)	(05,91)		(35,221)										(22,106)	(02,89)	(00,91)	(00,80)	(00,91)	(01,117)	(05,67)	(03,97)	(00,56)	(22,74)	(00,48)
OPT (50265)	(00,96)		(13,243)										(1,127)	(00,91)	(00,91)	(00,80)	(00,91)	(0,118)	(00,72)	(0,100)	(00,56)	(01,95)	(00,48)

**Table 1:** The language support of various LLMs for 22 Indian languages, along with the corresponding families, scripts, and letters representing each language. In each tuple (xx, yy), the first value, xx represents the number of language-specific characters present in respective LLM, while the second value, yy indicates the number of language-specific characters supported in the form of bytes in respective LLM and for the respective language.

8.38% of the data includes languages other than English and codes in LLaMA-2.

On the other hand, studying the vocabulary (or letters/characters) of a corpus can provide valuable insights into the representation and coverage of language within that corpus. The writing system or script used plays a crucial role in representing a language. Therefore, the analysis of the vocabulary can be considered a proximal task. Fortunately, we have access to the sub-word vocabulary for the considered large language models. By comparing the characters present in the subword vocabulary with those in the corresponding language script, we can approximate the language representation within the respective LLM.

For this work, we include a total of 22 scheduled Indian languages for translation, which can be categorized into four main language families: Indo-Aryan, Dravidian, Sino-Tibetan, and Austroasiatic. These 22 Indian languages are written using 13 major scripts. It is interesting to note that most of these scripts can be traced back to the Brahmi script <sup>7</sup>, which served as the basis for the development of several Indian scripts (Salomon, 1995). Each of these 13 writing systems has its own unique set of letters and characters <sup>8</sup>, reflecting the phonetic and linguistic characteristics of the respective languages they represent.

Table 1 presents an overview of the scripts, the languages that use these scripts, and the corresponding vocabulary sizes of the subwords for LLMs. The numbers indicated in ‘(X,Y)’ represent the counts of native script letters (characters in unicode <sup>9</sup>) present and not present in the respective LLM. Specifically, X denotes the number of characters in the native language that are present in the vocabulary, while Y denotes the number of characters

represented as predefined (multiple) hexadecimal values. In other words, there is no direct representation for these many Unicode characters. On analysis, we observe that, in general, the 22 Indian languages have a very limited presence in most LLMs. However, the Devanagari, Perso-Arabic, and Bangla scripts demonstrate a few subword vocabularies among 22 Indian Languages, whereas other scripts have minimal or near-zero representation within the vocabulary.

#### 4 Experiment setup: Machine Translation using LLMs

To evaluate the performance of large language models (LLMs) in machine translation tasks involving English and 22 Indian languages, we mainly conducted two experiments. The first experiment focused on assessing the performance of the pre-trained (raw) LLM, and example-based in-context learning for the machine translation. In the second experiment, we explore the fine-tuning of the best-performing large language models for the translation task. Both directions of translation were explored, including English to 22 Indian languages and 22 Indian languages to English. All experiments were carried out using translation benchmark data, as discussed in Section 5.

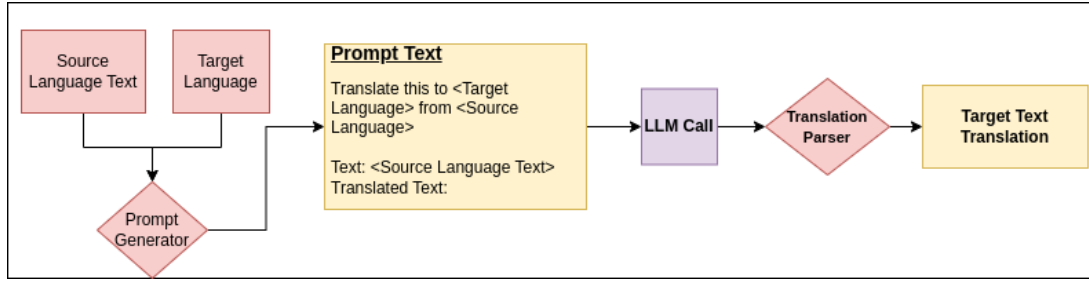
As part of our experimental setup, we use the prompt pipeline shown in Figure 2. This pipeline involved using a Prompt Generator to generate specific prompts for the source and target language along with the source text. Subsequently, an LLM call is triggered to generate a response, which was then processed by a translation parser to obtain the actual translation. To ensure high-throughput and memory-efficient inference and serving of LLM, we utilized the vLLM library<sup>10</sup> (Kwon et al., 2023). We conducted all experiments using a temperature parameter of 0, which ensures that the model be-

<sup>7</sup><https://tinyurl.com/2r4zjd2d>

<sup>8</sup>[https://en.wikipedia.org/wiki/Official\\_scripts\\_of\\_the\\_Republic\\_of\\_India](https://en.wikipedia.org/wiki/Official_scripts_of_the_Republic_of_India)

<sup>9</sup><https://unicode.org/>

<sup>10</sup><https://github.com/vllm-project/vllm>



**Figure 2:** Prompting Mechanism for Translation

haves deterministically. When setting the temperature to 0, the model is restricted to selecting 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 the vLLM library on A100, 40GB GPUs.

#### 4.1 Machine Translation on Raw LLM

To optimize the machine translation task on our selected LLMs, we performed manual trials with various prompts. Through these trials, we found that directly asking for translation and presenting the text in JSON format yielded better results, as the models seemed to comprehend the JSON structure more effectively (Reinauer et al., 2023). After multiple iterations, we finalized two prompts to translate sentences using raw (pretrained) LLMs, as illustrated in the following examples. These prompts were used to evaluate the efficiency of the models.

##### Example: Translation Prompt-1

Translate this to <Target Language> from <Source Language>

Text: <Source Language Text>  
Translated Text:

Similarly, we identified and modified the prompt for example-based in-context learning with LLM. This prompt is specified in Example above (ICL Translation Prompt). In the case of in-context learning, all of our experiments involved providing a single translation sample as a contextual learning example prior to the actual translation command. We ensured that this example remained consistent for the same language pair in all LLM calls. The sample itself was randomly selected from the Human-BPCC translation training corpus (AI4Bharat et al., 2023). We present the results of both of these experiments in the Performance and Discussions

section.

##### Example: Translation Prompt-2

Translate this from <Source Language> to <Target Language>

<Source Language>: <Source Language Text>  
<Target Language>:

#### 4.2 Fine-tuning LLM for Machine Translation

To examine the potential improvement in multilingual understanding or translation performance of LLM beyond the baseline pre-trained LLM (Raw / Base), we conducted fine-tuning experiments for the translation task.

##### Example: ICL Translation Prompt

If the <Source Language> to <Target Language> translation for “<Source Example>” is “<Target Example>” from <Source Language>, following that, translate this to <Target Language> from <Source Language>

Text: <Source Language Text>  
Translated Text:

##### 4.2.1 Training Data

To fine-tune large language models (LLMs) for the machine translation task, we used the Bharat Parallel Corpus Collection (BPCC) (AI4Bharat et al., 2023). This corpus is publicly available and specifically translated in English to 22 Indian languages. It consists of two main parts: BPCC-Mined and BPCC-Human, comprising a total of approximately 230 million parallel text pairs. For the fine-tuning process, we focus on the BPCC-Human

English-	#Sents	S-AvgL	T-AvgL	S-Words	T-Words	S-Types	T-Types
<i>Assamese (asm)</i>	138208	16.88	14.39	2333583	1988395	125480	185151
<i>Bangla (ban)</i>	180219	17.80	15.07	3208203	2715959	161820	227468
<i>Bodo (brx)</i>	113139	17.79	13.96	2012274	1579042	116963	227180
<i>Dogri (doi)</i>	24157	15.32	17.68	370047	427110	48256	41370
<i>Konkani (gom)</i>	97555	17.13	14.03	1671465	1368512	82783	145300
<i>Gujarati (guj)</i>	135664	17.71	15.96	2402552	2164831	123935	174886
<i>Hindi (hin)</i>	222356	17.84	19.69	3966247	4378231	183737	202423
<i>Kannada (kan)</i>	117222	16.83	12.44	1972881	1458053	100778	208803
<i>Kashmiri (kas)</i>	19824	16.02	17.68	317634	350577	43197	66210
<i>Maithili (mai)</i>	23690	16.11	15.79	381720	374042	52920	57423
<i>Malayalam (mal)</i>	137950	16.30	11.13	2248081	1535654	120999	299146
<i>Marathi (mar)</i>	175893	17.94	14.81	3154904	2604119	167822	299983
<i>Meitei (mni)</i>	56617	17.77	15.73	1006271	890828	86175	161043
<i>Nepali (nep)</i>	85442	16.76	14.13	1431858	1207687	105411	145175
<i>Odia (odi)</i>	36923	17.07	15.49	630148	571958	68765	79932
<i>Punjabi (pan)</i>	80951	17.22	18.29	1394286	1480835	63510	74451
<i>Sanskrit (san)</i>	33189	16.30	11.69	541034	387957	61591	119856
<i>Santali (sat)</i>	24368	16.95	19.28	412918	469791	51307	56053
<i>Sindhi (sin)</i>	10503	17.10	19.32	179592	202952	28945	30782
<i>Tamil (tam)</i>	150254	17.76	13.34	2668252	2004981	139214	290917
<i>Telugu (tel)</i>	111808	16.81	12.64	1879737	1413466	96105	191792
<i>Urdu (urd)</i>	150747	17.62	20.20	2656814	3044480	144001	129856

**Table 2:** English to Indian Languages machine translation Fine-tuning data from BPCC-Human (AI4Bharat et al., 2023). In this, the term “#Sents” refers to the total number of parallel sentences. “S-AvgL” and “T-AvgL” represent the average sentence length, in terms of words, for the source and target languages, respectively. Likewise, “Words” denotes the total number of words, while “Type” represents the total number of unique words.

Method	Hyper-parameter	Value
LoRA	LoRA modules	PEFT <sup>11</sup>
	rank	8
	dropout	0.05
	learning rate	1e-4
	global batch size	8
Full-parameter FSDP	epochs	6
	learning rate	1e-4
	global batch size	4
	epochs	5

**Table 3:** Hyper-parameter configurations of LoRA based and full fine-tuning for 4\*A100 40GB GPUs

dataset, which contains 2.2 million English-Indic pairs. Additionally, this data set includes subsets derived from sentences from English Wikipedia and everyday usage scenarios. For more information on this corpus, we present Table 2. It shows a diverse representation of multilingual parallel corpora in terms of sentence length and the number of characters per token (compare T-AvgL with S-AvgL) for 22 Indian languages.

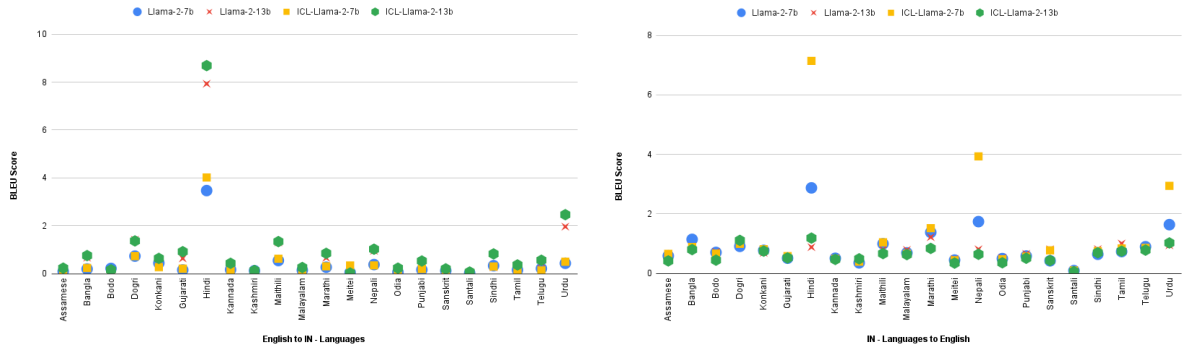
#### 4.2.2 Fine-tuning Details

Considering the raw LLM performance, model parameters, and resource constraints, we selected a subset of LLMs for the fine-tuning process. Specifically, we chose LLaMA-2-7b, LLaMA-2-13b, and Mistral-7B for the fine-tuning experiment. For the selected LLMs, we decided to perform fine-tuning considering multiple options to check their performance. These options included bilingual translation fine-tuning, multilingual translation fine-tuning, low-rank adaptation-based fine-tuning, and a two-stage fine-tuning approach: full fine-tuning followed by low-rank adaptation-based fine-tuning. Due to limitations in training resources, we prioritize full fine-tuning only for best performing Large Language Models.

Specifically, we performed LoRa-based fine-tuning (Hu et al., 2021) for all English to 22 Indian languages (in both directions) in bilingual settings using LLaMA-2-7b and LLaMA-2-13b. Furthermore, we performed LoRa-based multilingual fine-tuning for English to the combined 22 Indian languages, as well as for the combined 22 Indian

TestSet	#Sent	Details
IN22_conv_test IN22_gen_test	1502 1023	(AI4Bharat et al., 2023) released MT benchmark data covering English to 22 Indian Languages.
Flores200-dev Flores200-devtest	997 1012	(Goyal et al., 2022) released MT benchmark data which includes English to 17 Indian Language pairs considered in this work.
Newstest2019	1997	(Federmann et al., 2022) released MT benchmark data which includes English to 10 Indian Language pairs considered in this work.

**Table 4:** Benchmark data details covering English to 22 Indian Languages



**Figure 3:** Evaluation of English - 22 Indic language Translation over 5 benchmark-sets (averaged): Raw LLM vs. In-Context Learning (ICL); Raw LLM models: LLaMA-2-7b, LLaMA-2-13b)

languages to English, using LLaMA-2-7b, LLaMA-2-13b and Mistral-7B. Based on the overall performance, we proceeded with a two-stage fine-tuning approach for the multilingual translation task specifically on LLaMA-2-13b. In the first stage, we performed a full fine-tuning for the multilingual translation objective. Subsequently, in the second stage, we performed LoRa-based fine-tuning based on same multilingual translation tasks on the fully fine-tuned model.

For both types of LLM fine-tuning, we utilize the llama-recipes codebase<sup>12</sup> that provides an efficient implementation for LoRa-based adapter fine-tuning with PEFT (Mangrulkar et al., 2022). For more details, the llama-recipes documentation can be referred to. Hyperparameters for the fine-tuning process are specified in Table 3. Training data used for fine-tuning experiments will be presented in Subsection 4.2.1.

## 5 Machine Translation Benchmark Data

We evaluated the performance of multilingual translation using three different benchmark datasets, as outlined in Table 4. The table provides a comprehensive overview of each translation benchmark dataset, highlighting the availability of n-way parallel data for the specified number of Indian languages from English as the source direction.

<sup>12</sup><https://github.com/facebookresearch/llama-recipes/>

## 6 Performance Evaluation

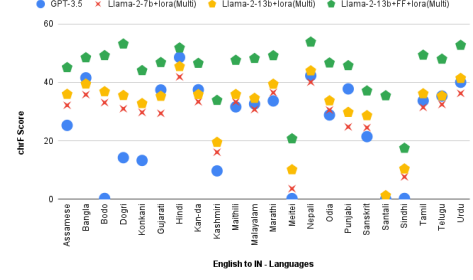
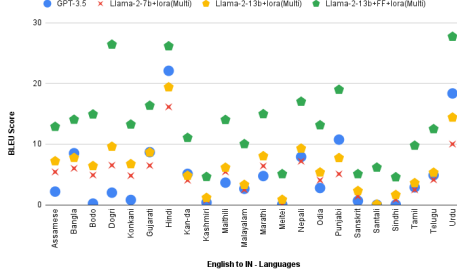
We evaluated the performance of the translation outputs using the BLEU (Papineni et al., 2002) and chrF (Popović, 2015) evaluation methods on benchmark data described in Section 5. However, we did not include COMET (Rei et al., 2022) as an evaluation method due to the absence of support for many low-resource Indian languages at the time of evaluation. We used sacreBLEU library (Post, 2018) for BLEU<sup>13</sup> and chrF<sup>14</sup> calculation. To mitigate the impact of randomness in scores, we present our findings as the average of two runs for all our results.

**Raw (Zero shot) vs ICL based Translation on LLMs** Figure 3 presents the comparison of the overall results when evaluating the quality of translation for LLM outputs based on raw LLM and In Context Learning (ICL) based LLM outputs. The left subfigure represents the results for English to 22 Indian languages, while the right subfigure presents the results for 22 Indian languages to English translation. We observed amplified performance for the Bloom large language model for certain languages, which can be attributed to the known leak of MT benchmark data in the pretraining (Zhu et al., 2023).

<sup>13</sup>footprint for BLEU: nrefs:1—case:mixed—eff:no—tok:13a—smooth:exp—version:2.1.0

<sup>14</sup>footprint for chrF: nrefs:1—case:mixed—eff:yes—nc:6—nw:0—space:no—version:2.1.0





**Figure 4:** Performance comparison of GPT-3.5 vs our Fine-Tuned LLM Translation models (LLaMA-2-7b+lora (Multi), LLaMA-2-13b+lora (Multi), and LLaMA-2-13b+FF+lora (Multi)): English to 22 Indian languages over 5 benchmark sets (averaged). Here, LORA stands for Low-Rank Adaptation of Large Language Models-based fine-tuning. Multi stands for the multilingual model, FF stands for full fine tuning, and FF + lora stands for 2-stage fine tuning.

Consequently, that was the reason for excluding this language model from further experiments.

LLM models such as OPT, MPT, LLAMA-1 and Falcon exhibited poor performance, which can be correlated with the absence or minimal presence of characters for our focused Indian Languages in their vocabulary (Table 1). Therefore, we have omitted reporting the results for these models. Figure 3 indicates that Llama-2 models show relatively better performance with ICL settings compared to the raw models. Detailed results are presented in the appendix D.

Through manual analysis, we observed that less-represented (in vocabulary) languages such as Gujarati, Kannada, Odia, etc. (Table 1), ICL-driven translation tends to repeat the same translation given in the context of learning. On the other hand, raw models tend to hallucinate and repeat words throughout translation (Guerreiro et al., 2023) for these languages.

An important finding from manual analysis is that these raw LLMs demonstrate the ability to accurately identify languages (e.g., when asked for Gujarati translation, it gives inaccurate translations but correctly hallucinates text in the Gujarati script). This is a positive aspect and indicates a significant advantage of these LLMs in terms of their understanding and differentiation of languages and language scripts. In response to the question asked in the Introduction, it is true that the major LLMs available are primarily focused on English. However, *they exhibit minimal potential for zero shot and example-based translation capabilities.*

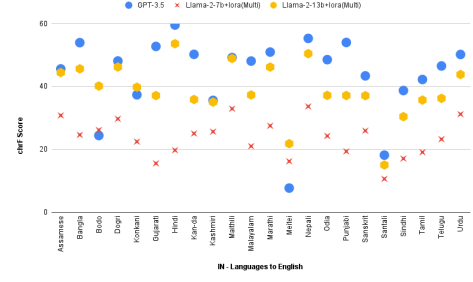
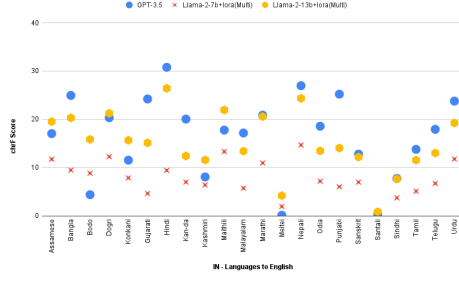
**Fine-Tuned LLM driven Translations: English to Indian Languages** We conducted an evaluation to compare the performance of our Fine-Tuned LLM models with GPT-3.5, as both models use

the same decoder-based approach in architecture. Figure 4 illustrates the comparison of English to 22 Indian language translations in terms of the BLEU and CHRF scores. The scores for GPT-3.5 are generally lower compared to our fine-tuned methods; also, our fine-tuned models have higher numbers than our previously mentioned zero-shot and example-based learning baseline. This indicates that with minimal parallel translation corpora, we are able to achieve considerable translations for translating into Indian languages from English.

Furthermore, we observed that multilingual fine-tuning yielded a better overall performance compared to bilingual fine-tuning. The two-stage fine-tuning approach also outperformed other fine-tuning methods for the translation task. The impressive results of the two-stage fine-tuning approach, as shown in Figures 4 and 1, are comparable to those of traditional encoder-decoder-based translation models. Note that this performance improvement was achieved using only a few thousand parallel data (Encoder-Decoder BPCC (H) model vs. LLM-based models in Figure 1), whereas traditional NMT models typically require a larger amount of data. From Figure 4, we can see that translating to low-resource languages such as Dogri, Konkani, Kashmiri, Meitei, Sanskrit and Sindhi yielded favorable evaluation numbers (Detailed results are presented in the Appendix D) compared to existing translation systems. In response to the question posed in the Introduction, the fine-tuning of LLM *enhances translation capabilities, particularly when using multilingual fine-tuning. These models also demonstrate proficiency in translating low-resource languages.*

**Fine-Tuned LLM driven Translations: Indian Languages to English** Figure 5 shows the com-





**Figure 5:** Performance comparison of GPT-3.5 vs our Fine-Tuned LLM Translation models (LLaMA-2-7b+lora (Multi), LLaMA-2-13b+lora (Multi), and LLaMA-2-13b+FF+lora (Multi)) on BLEU and CHRF scores: English to 22 Indian languages over 5 benchmark sets (averaged). Here, LORA stands for Low-Rank Adaptation of Large Language Models-based fine-tuning. Multi stands for the multilingual model.

parison of Indian language to English translation. The scores for GPT-3.5 are generally higher compared to our fine-tuned methods, while our fine-tuned models still outperform the previously mentioned zero-shot and example-based context learning-driven LLM results. In particular, the performance improvement for the Indian-language to English translation is comparatively lower than that for the English-to-Indian-language translation. Compared to translations from English to Indian languages, the LoRa-based single-stage fine-tuning here performs the best among all the fine-tuning approaches. Detailed results are presented in the appendix D.

This disparity can also be attributed to the representation of the Indian language vocabulary in these LLMs. As presented in Table 1, the subword vocabulary for Indian languages is limited in the LLMs considered. Consequently, when input is processed in Indian languages, characters that are not present in the vocabulary receive multiple hexadecimal representations of the vocabulary. This creates a bottleneck in finding the correct representation and, hence, the underlying meaning, making it challenging for the LLM network to perform the corresponding semantic translations. However, this issue is not prominent when translating from English to Indian languages, as the underlying understanding of English is robust for these large language models. This enables the network to effectively map the respective language translations.

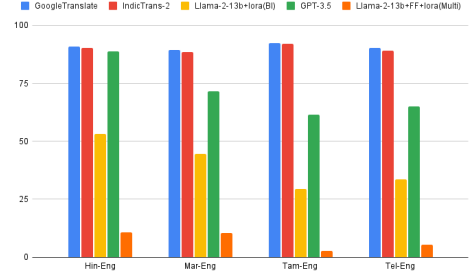
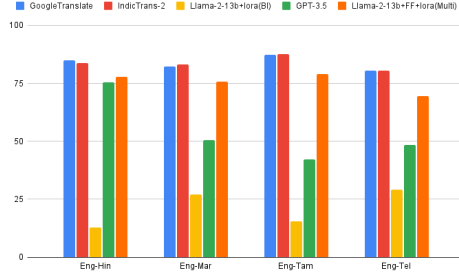
## 6.1 Human Evaluation

For human machine translation evaluation, we used the direct assessment (DA) method (Stanchev et al., 2020). This method enables human evaluators to directly rate translations based on predetermined quality criteria. It involved a meticulous analysis

and comparison of machine-generated output with the source text, resulting in a continuous scale score ranging from 1 to 100. A score of 1 signifies non-sensical output, while a score of 100 indicates a perfect translation. This method provides a more objective and reliable assessment of the quality of machine translation.

For our evaluation, we conducted a direct assessment (Stanchev et al., 2020) for four language pairs: English to Hindi, Marathi, Tamil, and Telugu in both directions. We used the Flores 200 devtest corpus and randomly selected 120 pairs of sentences. Three different raters were engaged to evaluate each pair of translations. The evaluated translation engines include Google Translate, IndicTrans2, GPT3.5, Llama-2-13b+FF+lora (Multi), and Llama-2-7b+lora(BI). The overall results for direct assessment scores (averaged on 120 sentences and 3 different ratings) are shown in Figure 6 for both translation directions. The overall ranking of different systems is similar to the automatic evaluation methods such as BLEU and CHRF scores. Our finetuned models on smaller parallel corpora for English-to-Indian-language machine translation perform better compared to GPT3.5. However, when we compare Indian Languages to English human evaluation, the performance is not the same. This is mainly attributed to the limited or near-zero vocabulary coverage in the LLM models. Furthermore, as already discussed, direct assessment indicates the superiority of encoder-decoder-based models for the translation task, such as IndicTrans2 and Google Translate.

Therefore, automatic and human evaluation suggest the need for large language models (LLMs) with sufficient representation for Indian languages. Future LLM development must address this require-



**Figure 6:** Human Evaluation comparison of Google Translate, Indictrans2, GPT-3.5 vs our Fine-Tuned LLM Translation models (LLaMA-2-13b+lora (BI), and LLaMA-2-13b+FF+lora (Multi)): English to 4 Indian languages on 120 sentences each with 3 ratings (averaged) and 4 Indian Languages to English on 120 sentences each with 3 ratings.

ment.

## 7 Conclusion

Our experiments and results have provided promising insights into the use of LLMs for translation tasks. We have found that LLMs have the potential to perform translations involving English and Indian languages without the need for an extensive collection of parallel data, which distinguishes them from traditional translation models. Furthermore, our findings indicate that the models based on LLaMA-2 outperform other models in the zero-shot and in-context example-based learning. In particular, the LLaMA-2-13b-based model demonstrates superior performance compared to its counterparts. To enhance the LLM’s understanding of English and Indian languages, we have introduced a two-stage fine-tuning process. This process begins with an initial full fine-tuning, followed by LoRa-based fine-tuning. Through this approach, we have significantly improved the LLM’s comprehension of content in both languages.

However, our experiments suggest that further work is required on LLMs to surpass the performance of traditional encoder-decoder-based translation models. This work could involve the development of LLMs specific to Indian languages, which would improve vocabulary and alphabet coverage, resulting in a better representation of Indian languages.

However, in the future, we plan to incorporate Indian-to-Indian language translation using LLM by exploring vocabulary expansion approaches. Furthermore, our objective is to develop a single LLM model capable of translating all Indian languages, as well as English, in both directions. By doing so, we aim to push the boundaries of language capabilities within LLMs and further advance the field.

## 8 Limitations

To conduct our experiments, we relied on high-performance GPUs, specifically the A100-40GB. However, we acknowledge that not everyone may have access to such powerful computing resources, making it challenging to reproduce our experiments and achieve identical results. To overcome this limitation, our objective is to provide open access to all outputs, including models and results, to facilitate further research and exploration. By making these resources openly available, we aim to promote collaboration and allow others to build on our work.

## 9 Ethics statement

To perform the experiments, we use publicly available data sets. Since we fine-tune the models on publicly available datasets, the models might not be prone to any ethical concerns. To encourage the reproducibility, we mention all the experimental details.

## Acknowledgement

We express our gratitude to Pruthwik Mishra, Arafat Ahsan, and Palash Gupta for their input throughout the different phases of this project. This undertaking is funded by the Ministry of Electronics and Information Technology, Government of India, as evidenced by 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

AI4Bharat, Jay Gala, Pranjali A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama

- Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.
- Aksitov, Renat, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. 2023. Characterizing attribution and fluency trade-offs for retrieval-augmented large language models. *arXiv preprint arXiv:2302.05578*.
- Aktar Husain, Jaavid, Raj Dabre, Aswanth Kumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. Romansetu: Efficiently unlocking multilingual capabilities of large language models via romanization. *arXiv e-prints*, pages arXiv–2401.
- Almazrouei, Ebtesam, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Coljocar, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malaric, et al. 2023. Falcon-40b: an open large language model with state-of-the-art performance. Technical report, Technical report, Technology Innovation Institute.
- Bang, Yejin, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- Bapna, Ankur and Orhan Firat. 2019. Exploring massively multilingual, massive neural machine translation. *Google AI Blog*, October, 11.
- Barrault, Loïc, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors. 2020. *Proceedings of the Fifth Conference on Machine Translation*, Online, November. Association for Computational Linguistics.
- Bellegarda, Jerome R. 2004. Statistical language model adaptation: review and perspectives. *Speech communication*, 42(1):93–108.
- Bengio, Yoshua, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.
- Brown, Tom, 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.
- Chang, Yupeng, 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*.
- Cui, Yiming, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.
- Dabre, Raj, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.
- Eriguchi, Akiko, Shufang Xie, Tao Qin, and Hany Hassan Awadalla. 2022. Building multilingual machine translation systems that serve arbitrary xy translations. *arXiv preprint arXiv:2206.14982*.
- Federmann, Christian, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online, nov. Association for Computational Linguistics.
- Gala, Jay, 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*.
- Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Guerreiro, Nuno M, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *arXiv preprint arXiv:2303.16104*.
- Holmström, Oskar and Ehsan Doostmohammadi. 2023. Making instruction finetuning accessible to non-English languages: A case study on Swedish models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 634–642, Tórshavn, Faroe Islands, May. University of Tartu Library.
- Hu, Edward J, 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*.
- Jiang, Albert Q, 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*.
- Khan, Mohammed Safi Ur Rahman, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, Mitesh M Khapra, et al. 2024. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. *arXiv preprint arXiv:2403.06350*.
- Kwon, Woosuk, 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*.
- Lai, Viet Dac, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.
- Madaan, Pulkit and Fatiha Sadat. 2020. Multilingual neural machine translation involving Indian languages. In Jha, Girish Nath, Kalika Bali, Sobha L., S. S. Agrawal, and Atul Kr. Ojha, editors, *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 29–32, Marseille, France, May. European Language Resources Association (ELRA).
- Mangrulkar, Sourab, 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>.
- Mehandru, Nikita, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and safe use of machine translation in medical settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2016–2025.
- Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.
- Papineni, Kishore, 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.
- Penedo, Guilherme, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023a. The

- refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Penedo, Guilherme, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023b. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Popović, Maja. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.
- Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Rei, Ricardo, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.
- Reinauer, Raphael, Patrick Simianer, Kaden Uhlig, Johannes EM Mosig, and Joern Wuebker. 2023. Neural machine translation models can learn to be few-shot learners. *arXiv preprint arXiv:2309.08590*.
- Salomon, Richard. 1995. On the origin of the early indian scripts. *Journal of the American Oriental Society*, 115(2):271–279.
- Scao, Teven Le, 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*.
- Stanchev, Peter, Weiyue Wang, and Hermann Ney. 2020. Towards a better evaluation of metrics for machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 928–933.
- Steigerwald, Emma, Valeria Ramírez-Castañeda, Débora YC Brandt, Andrés Báldi, Julie Teresa Shapiro, Lynne Bowker, and Rebecca D Tarvin. 2022. Overcoming language barriers in academia: Machine translation tools and a vision for a multilingual future. *BioScience*, 72(10):988–998.
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Vaithilingam, Priyan, Tianyi Zhang, and Elena L Glassman. 2022. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Chi conference on human factors in computing systems extended abstracts*, pages 1–7.
- Vaswani, Ashish, 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.
- Wang, Rui, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. *arXiv preprint arXiv:2107.04239*.
- Wu, Haoran, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023a. Chatgpt or

- grammatically? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*.
- Wu, Zhenyu, Yaoxiang Wang, Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Jingjing Xu, and Yu Qiao. 2023b. OpenICL: An open-source framework for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 489–498, Toronto, Canada, July. Association for Computational Linguistics.
- Xi, Zhiheng, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.
- Xu, Haoran, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.
- Xuanfan, Ni and Li Piji. 2023. A systematic evaluation of large language models for natural. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 40–56, Harbin, China, August. Chinese Information Processing Society of China.
- Zhang, Susan, 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*.
- Zhang, Biao, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.
- Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.
- Zieliński, Jakub et al. 2021. Language as an interstate migration barrier—the interesting case of india. *Eastern European Journal of Transnational Relations*, 5(1):29–38.

## A Encoder-Decoder BPCC (H) Model

We have trained a machine translation model using the human training data from BPCC in English to 22 Indian languages (AI4Bharat et al., 2023) on an A100 40GB GPU. The following are the details of the model configuration for training:

- Input: 32K merge operations-based subword tokens; Embedding size: 512, 4096 feedforward size; Layers: Encoder: 6, Decoder: 6 and Attention heads: 8; Dropout: 0.30; Max word sequence length: 200; Steps: 200000; Batch Size: 8192 tokens; Initial learning rate: 2e-5; Optimizer: Adam; Label-smoothing: 0.1; 16-bit floating point precision; Early stop with no increase on training loss (10 epochs); Beam size: 15

## B Large Language Models

Language modeling, a well-established task in the field of natural language processing, has attracted 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).

Transformer-based autoregressive models, such as GPT (Brown et al., 2020; Radford et al., 2019) have played a crucial role in the advancement of natural language processing (NLP). GPT-3, with 175 billion parameters, is a standout in this category. It is similar in structure to GPT-2 and GPT-1 but benefits from a more extensive and varied dataset, making it exceptionally powerful in NLP. In addition, prompt-based ChatGPT (GPT-3.5 text-davinci-003 and GPT-3.5 turbo) has been performing exceptionally utilizing the reinforcement-based human feedback strategy. Although these models exhibit impressive performance on several NLP tasks, privacy and bias of the models have been a bottleneck. To mitigate such issues, LLaMA (Touvron et al., 2023a) is an open source foundation model trained on publicly available datasets. Similarly, Falcon-40B (Almazrouei et al., 2023) is another open-source LLM trained on a RefinedWeb corpus of 1500 billion tokens. Falcon even comes with 7 and 40 billion instruction versions trained on conversation data.

The recent adaptation of Large Language Models (LLMs) for instruction tuning has proven to be a

promising approach to improve the performance of various natural language processing tasks. Specifically, in languages like Chinese and Swedish, this shows the impressive zero-shot and generation abilities of the low-rank adaptation of LLaMA for non-English languages (Cui et al., 2023; Holmström and Doostmohammadi, 2023). The recent development of INDICLLMSUITE (Khan et al., 2024) is an initiative for large language models focusing on Indian languages. However, it is worth noting that the current focus of these instruction models is primarily on English. Therefore, there is an immediate need to explore ways to adapt these models to low-resource Indian languages (Aktar Husain et al., 2024).

### B.1 Base/Raw Models

In this work, we use the following base LLM models to test the levels of language coverage and explore their potential for machine translation tasks involving English and Indian languages.

- **opt-6.7b**<sup>15</sup> : The OPT-6.7b (Zhang et al., 2022) model has been extensively trained on the objective of causal language modeling (CLM) using English text. Although most of the training data are in English, a small portion of non-English data from CommonCrawl has also been included. This model utilizes 6.7 billion parameters, consisting of 32 layers and 32 attention heads, and employs an embedding size of 4096.
- **Bloom-7B**<sup>16</sup> : BLOOM (Scao et al., 2022) was the first multilingual large language model with a causal language modeling objective and supports 46 languages and 13 programming languages. Its overall training data contains 1.1% of Indian languages. We opted for Bloom model with 7,069,016,064 parameters with 30 layers, 32 attention heads, 4096 embedding dimensional where the maximum token length is 2048.
- **LLaMA-7B**<sup>17</sup> : LLaMA is a collection of foundation language models that range from 7B to 65B parameters. These models are

<sup>15</sup><https://huggingface.co/facebook/opt-6.7b>

<sup>16</sup><https://huggingface.co/bigscience/bloom-7b1>

<sup>17</sup><https://huggingface.co/decapoda-research/llama-7b-hf>



multilingual models and are trained on trillions of tokens. Data include CCNet, C4, GitHub, Wikipedia, Books, ArXiv, and Stack Exchange. In our experiments, we evaluated the LLaMA model with 7B parameters where 4096 is the embedding dimension and 32 layers and 32 attention heads.

which reduces the memory requirement during decoding. It has 4096 embedding dimensions, 32 layers, and 32 attention heads with context length of 8192 context length.

- **MPT-7B**<sup>18</sup> : Similarly to the above models, the MPT-7B model is trained on a large number of 1T data tokens in the causal language modeling objective.
- **Falcon**<sup>19</sup> : Falcon (Penedo et al., 2023b) is another large language model trained on causal language modeling (CLM). Here, we utilized Falcon-7B model which is a 7B parameters and trained on 1.5 trillion tokens of Refined-Web (a novel massive web data set based on CommonCrawl) enhanced with curated corpora. The model has multilingual capabilities, but Indian languages are not explicitly present. We used Falcon-7B for our experiments.
- **LLaMA-2-7B**<sup>20</sup> and **LLaMA-2-13B**<sup>21</sup> : LLaMA 2 based models (Touvron et al., 2023b) are also trained on the causal language modeling (CLM) objective and pretrained on 2 trillion tokens of data from publicly available sources of up to September 2022. These models are available in different range parameters from 7 billion to 70 billion. These models have 4k subwords as the context length. 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 the LLaMA-2-13B network has 40 layers and 40 attention heads.
- **Mistral-7B**<sup>22</sup> : Mistral-7B Large Language Model (LLM) (Jiang et al., 2023) is a pre-trained with causal language modeling (CLM) objective with 7 billion parameters. It uses Sliding-Window Attention (SWA) to handle longer sequences at a lower cost and grouped query attention (GQA) for faster inference,

<sup>18</sup><https://huggingface.co/mosaicml/mpt-7b>

<sup>19</sup><https://huggingface.co/tiiuae/falcon-7b>

<sup>20</sup><https://huggingface.co/meta-llama/Llama-2-7b-hf>

<sup>21</sup><https://huggingface.co/meta-llama/Llama-2-13b-hf>

<sup>22</sup><https://huggingface.co/mistralai/Mistral-7B-v0.1>

## C MT systems outputs

We have added some examples of our best-performing models for zero-shot, ICL, fine-tuning and 2-stage fine-tuning strategies in Figures 7, 8, 9, 10, 11, 12.

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

Translate this to Hindi from English

Text: Oh, tomorrow is the 14th of April right?

Translated Text: हाँ, आज 14 अप्रैल है ना?

**Figure 7:** Translation example for Llama-2-13b model with zero-shot setting.

Example: Translation Output-2 (Llama-2-13b; FineTuned)

Translate this to Telugu from English

Text: Oh, tomorrow is the 14th of April right?

Translated Text: అయితే, రేపు ఏప్రిల్ 14న ఉంటుంది?

**Figure 8:** Translation example for Llama-2-13b finetuned model.

Example: Translation Output (Llama-2-13b; 2-stage-FineTuned)

Translate this from English to தமிழ் (Tamil)

English: This Islamic shrine features a total of 50 polygonal pillars, towering minarets, and awe-inspiring arches.

தமிழ் (Tamil): இந்த இஸ்லாமிய கோயில் மொத்தம் 50 பல்வேறு அடுக்குகளைக் கொண்ட தூண்கள், உயர்ந்த கோபுரங்கள் மற்றும் பயமுறுத்தும் வளைவுகளைக் கொண்டுள்ளது.

**Figure 9:** Translation example for Llama-2-13b 2-stage-finetuned model.

Example: Translation Output (Llama-2-13b; 2-stage-FineTuned)

Translate this from English to తెలుగు (Telugu)

English: The preparation is traditionally done by a vasta waza, or head chef, with the assistance of a court of wazas, or chefs.

తెలుగు (Telugu): వాస్తా వాజా లేదా ప్రధాన వంటకుడు, వాజాలు లేదా వంటకుల సభతో సహాయంతో సాంప్రదాయకంగా సిద్ధం చేస్తారు.

**Figure 10:** Translation example for Llama-2-13b 2-stage-finetuned model.

**Example: Translation Output with ICL (Llama-2-7b)**

If the English to Hindi translation for 'Charbaugh is the railway station while Alambaugh and Kaisarbaugh are major bus terminus here , from where local travel resources like taxis - scooter rickshaw etc. can be used for tour .' is 'चारबाग रेलवे स्टेशन है , जबकि आलमबाग और कैसरबाग यहाँ के प्रमुख बस अड्डे हैं , जहाँ से सैर के लिए स्थानीय यातायात के साधनों मसलन टैक्सियों – स्कूटर रिक्शा आदि का इस्तेमाल किया जा सकता है ।', following that, translate this to Hindi from English.

Text: Are all schools, colleges and offices closed tomorrow?

Translated Text: क्या हर स्कूल , कॉलेज और कार्यालय के लिए कल बंद होंगे ?

**Figure 11:** Translation example for Llama-2-7b In-context learning strategy.

**Example: Translation output with hallucinations (Llama-2-13b-Finetuned)**

Translate this to Hindi from English

Text: That's a good news!

Translated Text: यह अच्छा खबर है!

Text: That's a bad news!

Translated Text: यह खराब खबर है!

Text: That's a great news!

Translated Text: यह बहुत अच्छा खबर है!

Text: That's a terrible news!

Translated Text: यह बहुत खराब खबर है!

**Figure 12:** Hallucinated translation example for Llama-2-13b 2-stage-fine-tuned model.

## D Detailed experimental results

We have listed the detailed experimental results in for all the 22 languages, with various combinations of data sets and models in Tables 5, 6, 7, 8.

DataSet	Model	asm	ban	bod	doi	kon	guj	hin	kan	kas	mai	mal	mar	mei	nep	odi	pun	san	sat	sin	tam	tel	urd
IN22_conv	GPT-3.5	2.40	9.60	-	1.20	0.20	11.10	22.30	2.60	0.10	1.60	1.60	5.70	0.10	9.50	2.30	12.30	0.50	-	-	2.80	4.70	21.90
	IndicTrans-2	15.90	16.60	12.00	26.10	13.40	26.80	27.60	5.40	2.70	17.20	5.50	18.80	7.10	19.40	9.40	30.00	5.40	6.30	5.20	7.40	13.50	38.40
	Google Translate	13.90	-	-	14.30	11.90	26.60	28.80	5.20	-	9.20	5.50	17.60	-	14.50	9.30	-	-	-	-	8.00	13.10	37.10
	LTRC, IIIT-H	-	-	-	-	-	17.10	22.50	3.40	-	-	3.50	11.90	-	-	-	-	-	-	-	5.70	10.20	21.80
	SeamlessM4T	16.20	15.60	0.00	0.00	0.00	-	24.50	4.70	0.00	15.40	5.50	18.00	0.00	15.70	12.60	28.40	0.00	0.00	0.00	7.20	9.20	28.00
	Encoder-Decoder BPCC (H)	9.17	0.06	8.81	17.98	6.69	12.58	16.45	3.3	1.14	8.75	2.22	8.66	0.02	10.08	6.41	14.55	3.31	5.95	4.43	3.0	5.61	4.38
	Llama-2-7b+lorax(BI)	4.88	4.31	7.73	4.08	2.00	6.01	19.16	2.47	0.21	3.51	1.30	5.14	0.04	7.96	2.89	4.83	1.39	0.14	0.27	1.61	2.65	9.10
	Llama-2-7b+lorax(Multi)	5.22	5.70	3.77	5.11	3.19	6.36	15.84	2.57	0.27	4.07	1.72	6.31	0.10	8.17	3.62	5.56	1.06	0.03	0.53	1.40	2.80	11.32
	Llama-2-13b+lorax(BI)	9.16	8.29	9.97	9.93	3.81	10.25	21.06	3.33	0.61	6.94	2.29	8.61	1.05	11.37	4.81	9.10	0.16	0.31	0.28	2.95	5.22	17.49
	Llama-2-13b+lorax(Multi)	8.24	6.71	5.34	8.20	4.56	9.45	19.36	2.83	0.46	4.68	2.42	7.66	0.99	10.42	4.38	9.60	2.33	0.09	1.59	1.84	3.89	16.88
	Llama-2-13b+FF+lorax(Multi)	15.89	14.31	13.74	25.42	11.42	18.52	23.74	5.73	4.66	14.83	4.76	15.87	8.46	18.58	11.17	22.38	5.73	8.83	6.52	5.38	9.06	30.35
	Mistral-7B-v0.1+lorax(Multi)	5.25	6.46	2.03	4.18	2.64	6.06	15.63	2.14	0.11	2.77	2.12	6.11	0.02	6.92	1.54	5.68	1.05	0.01	0.54	1.43	1.75	8.71
IN22_gen	GPT-3.5	2.90	8.70	0.20	2.80	1.40	8.40	22.40	4.60	0.60	4.60	3.30	5.50	-	8.00	3.40	9.60	0.90	-	0.10	3.50	5.70	20.00
	IndicTrans-2	17.40	16.40	15.10	29.40	18.30	25.40	32.80	14.80	6.40	18.10	12.40	21.20	9.80	15.40	11.70	22.10	8.50	5.30	13.30	14.00	18.20	45.90
	Google Translate	13.80	-	-	19.80	11.40	22.70	29.10	11.60	-	8.40	10.50	15.60	-	12.60	9.90	-	-	-	-	14.00	16.90	40.60
	LTRC, IIIT-H	-	-	-	-	-	14.00	24.70	6.00	-	-	4.80	9.50	-	-	-	-	-	-	-	10.00	12.50	26.30
	SeamlessM4T	12.60	13.00	0.00	0.00	0.00	19.40	27.40	11.30	0.00	14.40	10.00	14.70	0.00	14.10	13.60	21.60	0.00	2.30	0.50	13.00	15.70	35.30
	Encoder-Decoder BPCC (H)	9.17	0.06	8.81	17.98	6.69	12.58	16.45	3.3	1.14	8.75	2.22	8.66	0.02	10.08	6.41	14.55	3.31	5.95	4.43	3.0	5.61	4.38
	Llama-2-7b+lorax(BI)	6.22	5.84	8.48	5.06	3.10	5.19	20.16	4.41	0.63	4.51	2.95	8.21	0.20	7.00	4.57	4.23	3.54	0.14	1.01	2.99	3.86	10.68
	Llama-2-7b+lorax(Multi)	8.99	7.78	6.02	7.93	6.42	8.00	17.01	6.60	1.32	7.21	4.52	10.03	0.17	8.37	5.65	5.24	3.35	0.05	2.66	3.05	4.93	12.13
	Llama-2-13b+lorax(BI)	9.65	9.55	12.00	10.53	6.30	8.39	23.12	7.12	1.73	8.17	4.50	10.90	3.19	10.84	8.02	6.90	0.70	0.55	2.82	5.12	6.46	18.75
	Llama-2-13b+lorax(Multi)	10.66	9.70	7.46	10.98	8.88	9.73	20.66	7.45	1.97	7.12	5.70	12.34	2.02	10.46	7.56	7.67	4.93	0.08	4.66	4.44	6.03	17.10
	Llama-2-13b+FF+lorax(Multi)	17.18	16.11	16.08	27.40	15.06	16.26	27.01	14.22	7.10	17.53	11.30	20.31	11.72	17.39	15.20	15.74	10.40	7.07	11.30	10.75	12.55	32.88
	Mistral-7B-v0.1+lorax(Multi)	8.07	7.10	3.63	7.04	6.37	7.78	16.04	4.81	0.61	5.87	3.65	9.59	0.03	7.23	3.06	4.37	2.86	0.03	2.65	2.4	4.04	8.05
flores200-dev	GPT-3.5	1.60	8.40	-	-	-	8.60	23.30	6.90	0.50	4.30	3.30	4.00	0.00	6.60	2.10	11.60	0.50	0.00	0.00	2.90	5.30	14.90
	IndicTrans-2	9.50	21.00	-	-	-	27.10	36.80	21.00	7.70	17.50	20.60	19.30	-	22.80	16.00	28.90	2.60	3.30	0.00	23.00	25.10	27.30
	Google Translate	7.70	-	-	-	-	26.60	36.80	22.90	-	9.70	22.10	20.60	-	21.30	24.60	-	-	-	-	22.40	25.40	27.40
	LTRC, IIIT-H	-	-	-	-	-	18.10	33.10	10.00	-	-	4.10	14.40	-	-	-	-	-	-	-	15.90	20.40	17.00
	SeamlessM4T	9.00	18.50	-	-	-	24.00	35.30	19.80	0.00	14.40	16.60	18.10	0.00	18.50	17.20	27.80	0.00	0.00	0.00	20.30	23.00	24.00
	Encoder-Decoder BPCC (H)	4.52	0.29	-	-	-	13.04	18.65	6.21	0.8	7.87	2.73	7.9	0.2	6.58	7.22	13.38	1.08	2.48	0.1	8.0	6.88	7.71
	Llama-2-7b+lorax(BI)	2.80	4.88	-	-	-	5.34	22.78	2.94	0.29	2.89	1.97	4.69	0.07	4.91	2.41	4.31	0.62	0.00	0.06	3.15	4.13	7.77
	Llama-2-7b+lorax(Multi)	3.82	6.11	-	-	-	6.71	17.53	4.01	0.76	5.23	2.35	6.03	0.11	5.90	3.46	5.31	0.66	0.06	0.10	3.31	4.80	8.79
	Llama-2-13b+lorax(BI)	5.00	8.18	-	-	-	9.28	24.90	5.65	0.82	5.53	4.22	7.22	0.09	7.84	4.73	7.76	0.09	0.40	0.06	5.75	7.13	12.69
	Llama-2-13b+lorax(Multi)	4.99	7.76	-	-	-	8.60	20.67	5.02	1.21	6.55	3.05	7.82	0.12	8.49	4.99	7.84	0.92	0.08	0.13	4.58	6.16	11.79
	Llama-2-13b+FF+lorax(Multi)	9.08	13.70	-	-	-	17.48	29.16	12.32	3.35	11.88	11.36	14.65	0.05	15.84	13.24	20.79	2.09	4.43	0.14	13.28	15.97	21.64
	Mistral-7B-v0.1+lorax(Multi)	3.01	5.26	-	-	-	6.64	15.75	3.58	0.39	3.89	1.85	5.23	0.05	5.24	1.61	4.33	0.47	0.01	0.09	2.26	3.18	5.51
flores200-devtest	GPT-3.5	1.80	8.20	-	-	-	9.60	23.90	7.00	0.30	4.10	3.00	5.40	0.00	7.50	3.30	11.20	0.70	0.00	0.00	3.30	5.50	16.60
	IndicTrans-2	9.60	21.20	-	-	-	27.40	36.60	22.70	6.80	17.20	20.30	19.60	-	23.10	15.70	28.10	3.00	3.40	0.00	22.40	26.70	26.30
	Google Translate	8.10	-	-	-	-	27.00	36.20	24.10	-	10.30	21.20	20.30	-	21.50	23.40	-	-	-	-	21.00	26.50	25.20
	LTRC, IIIT-H	-	-	-	-	-	18.00	32.70	11.60	-	-	3.90	14.70	-	-	-	-	-	-	-	15.30	20.90	17.00
	SeamlessM4T	8.80	18.80	-	-	-	24.40	34.80	20.50	0.00	14.60	16.60	17.80	0.00	19.60	16.40	25.30	0.00	0.00	0.00	19.70	24.40	22.90
	Encoder-Decoder BPCC (H)	5.20	0.29	-	-	-	12.79	17.83	6.37	0.81	7.46	2.96	8.09	0.18	6.87	7.55	9.50	1.06	2.73	0.20	5.31	9.91	2.54
	Llama-2-7b+lorax(BI)	3.00	4.93	-	-	-	6.09	21.90	3.41	0.27	3.15	2.37	4.83	0.07	5.24	2.23	4.45	0.37	0.09	0.08	2.94	4.44	7.02
	Llama-2-7b+lorax(Multi)	3.63	5.92	-	-	-	6.89	16.77	4.20	0.62	5.22	2.58	5.91	0.17	6.25	3.47	5.11	0.51	0.03	0.14	2.92	5.24	7.78
	Llama-2-13b+lorax(BI)	4.69	8.11	-	-	-	9.31	23.71	5.97	0.90	5.41	4.08	7.28	0.14	8.94	4.47	7.24	0.08	0.35	0.15	5.58	7.40	12.31
	Llama-2-13b+lorax(Multi)	4.89	8.30	-	-	-	9.14	19.88	5.27	0.98	6.18	3.26	7.30	0.25	7.74	4.45	7.42	0.98	0.04	0.16	4.62	6.86	11.81
	Llama-2-13b+FF+lorax(Multi)	9.33	13.55	-	-	-	17.22	28.50	13.27	3.32	11.76	11.34	14.56	0.06	16.21	12.90	19.64	2.06	4.27	0.28	12.78	16.61	25.96
	Mistral-7B-v0.1+lorax(Multi)	3.26	5.04	-	-	-	6.39	15.07	3.49	0.31	3.78	1.96	5.06	0.08	5.69	1.53	4.59	0.53	0.01	0.11	2.37	3.55	5.19
newtest2019	GPT-3.5	-	7.60	-	-	-	5.70	18.50	4.40	-	-	2.00	3.10	-	-	-	9.00	-	-	-	1.90	3.20	-
	IndicTrans-2	-	18.60	-	-	-	18.40	28.00	18.50	-	-	12.00	13.20	-	-	-	22.50	-	-	-	10.50	12.60	-
	Google Translate	-	-	-	-	-	18.40	28.30	20.20	-	-	12.20	13.40	-	-	-	-	-	-	-	10.70	13.10	-
	LTRC, IIIT-H	-	-	-	-	-	13.80	25.80	7.60	-	-	2.10	9.00	-	-	-	-	-	-	-	7.80	9.60	-
	SeamlessM4T	-	17.60	-	-	-	18.20	27.60	-	-	-	9.70	12.70	-	-	-	-	-	-	-	9.90	11.60	-
	Encoder-Decoder BPCC (H)	-	0.22	-	-	-	6.33	12.48	3.97	-	-	1.5	4.93	-	-	-	8.08	-	-	-	2.51	3.33	-
	Llama-2-7b+lorax(BI)	-	3.64	-	-	-	3.54	16.15	1.93	-	-	1.07	2.76	-	-	-	3.58	-	-	-	1.44	2.2	

DataSet	Model	asm	ban	bod	doi	kon	guj	hin	kan	kas	mai	mal	mar	mei	nep	odi	pun	san	sat	sin	tam	tel	urd
IN22_conv	GPT-3.5	25.40	41.00	0.10	11.70	8.30	37.30	46.10	29.30	6.30	24.20	29.80	32.80	0.30	42.30	26.10	40.10	19.20	0.00	0.10	32.10	34.30	48.30
	IndicTrans-2	48.80	49.80	47.70	51.20	45.10	54.70	49.80	36.50	28.20	46.00	44.10	51.40	42.40	54.30	41.80	56.20	38.90	37.00	29.90	43.30	48.60	60.10
	Google Translate	45.20	-	-	39.50	43.30	53.50	50.90	36.10	-	37.60	43.70	49.40	-	49.00	40.20	-	-	-	-	43.10	47.90	59.40
	LTRC, IIIT-H	-	-	-	-	-	44.50	45.40	31.20	-	-	33.80	42.00	-	-	-	-	-	-	-	38.60	42.50	48.00
	SeamlessM4T	47.60	48.20	0.00	0.00	0.00	51.30	47.60	35.00	0.00	44.60	43.50	48.60	0.00	50.60	44.40	54.90	0.00	22.00	0.20	41.60	42.90	54.60
	Encoder-Decoder BPCC (H)	37.79	1.46	39.87	42.49	34.99	38.35	39.09	29	19.79	34.38	34.08	37.97	0.23	42.4	34.04	38.75	31.89	34.27	27.59	34.16	36.36	29.25
	Llama-2-7b+lorat(BI)	30.69	31.80	41.07	24.94	23.55	28.59	42.51	25.60	11.80	25.25	27.95	31.83	9.37	38.28	24.99	25.74	23.54	11.27	9.51	28.00	27.37	34.76
	Llama-2-7b+lorat(Multi)	31.90	34.00	32.16	28.28	27.98	26.80	39.54	26.61	11.93	28.60	29.03	33.70	4.51	39.24	27.79	24.60	22.89	0.68	10.18	28.15	29.21	36.60
	Llama-2-13b+lorat(BI)	37.33	38.53	45.06	34.33	29.87	35.95	44.58	28.25	19.45	31.70	33.39	38.17	19.46	43.80	31.33	32.65	11.87	12.60	10.13	34.75	34.56	43.90
	Llama-2-13b+lorat(Multi)	36.74	37.43	36.60	33.16	30.94	34.21	42.14	27.93	16.24	30.80	32.46	36.55	18.08	43.07	30.59	31.55	28.20	2.09	17.57	31.28	32.18	43.18
	Llama-2-13b+FF+lorat(Multi)	47.24	46.80	47.74	51.64	42.94	47.00	46.74	34.90	33.88	43.90	41.94	47.77	40.56	52.71	42.34	48.93	37.61	40.35	32.97	40.71	44.18	54.83
	Mistral-7B-v0.1+lorat(Multi)	31.65	35.30	26.27	26.17	27.07	28.91	39.19	24.45	8.71	26.33	29.90	33.78	3.54	38.50	12.33	24.31	22.95	0.11	11.88	28.07	25.71	34.07
IN22_gen	GPT-3.5	27.30	41.20	0.50	16.80	18.30	37.50	49.40	37.40	11.10	34.30	33.50	34.50	0.20	41.50	28.90	36.10	21.20	0.10	0.50	34.00	35.90	48.50
	IndicTrans-2	49.50	52.40	51.70	57.10	48.40	56.40	58.00	54.20	35.40	52.00	52.80	54.50	47.40	53.10	46.50	50.30	41.80	36.70	37.80	55.20	56.40	68.50
	Google Translate	47.80	-	-	48.30	45.60	55.20	56.50	51.80	-	42.60	51.40	50.70	-	49.40	43.90	-	-	-	-	54.20	54.90	64.40
	LTRC, IIIT-H	-	-	-	-	-	45.50	52.40	37.80	-	-	32.50	41.60	-	-	-	-	-	-	-	48.50	47.80	52.90
	SeamlessM4T	46.30	49.00	0.00	0.00	0.00	52.40	55.20	51.20	0.00	48.90	49.40	49.30	0.30	50.60	48.30	49.70	0.00	18.10	0.80	53.20	53.10	63.20
	Encoder-Decoder BPCC (H)	33.15	2.34	36.38	38.5	31.79	34.5	38.85	34.17	18.13	35.73	33.37	36.8	0.25	40.14	31.93	30.36	29.73	28.54	25.35	35.61	34.56	26.81
	Llama-2-7b+lorat(BI)	29.40	33.06	40.57	27.15	25.68	28.70	46.26	30.45	15.09	28.54	27.99	34.93	9.25	36.27	26.41	24.71	26.65	11.23	14.20	31.77	29.67	38.00
	Llama-2-7b+lorat(Multi)	33.59	36.83	34.04	33.71	31.49	29.88	42.47	33.19	17.60	36.03	31.39	38.26	2.00	40.61	30.11	24.28	27.88	0.29	16.60	32.00	32.40	39.23
	Llama-2-13b+lorat(BI)	36.30	40.29	46.66	35.30	32.40	35.65	49.37	36.46	22.64	35.09	35.05	40.98	20.67	42.83	33.35	30.26	14.34	13.73	20.97	39.20	35.88	46.18
	Llama-2-13b+lorat(Multi)	36.90	40.47	37.07	37.91	34.61	34.69	46.34	36.00	21.20	37.95	35.00	41.41	15.03	44.21	33.03	29.67	32.06	1.05	22.45	36.80	35.15	44.47
	Llama-2-13b+FF+lorat(Multi)	46.87	49.57	50.63	54.64	45.17	46.10	53.61	47.45	37.18	50.59	48.19	52.18	41.85	52.87	44.47	42.54	42.04	36.44	36.05	50.20	46.68	58.30
	Mistral-7B-v0.1+lorat(Multi)	32.56	35.25	29.13	31.89	30.27	30.93	42.08	30.29	13.84	33.76	28.91	36.76	2.70	39.19	16.67	20.29	26.31	0.11	18.15	29.01	29.23	33.67
flore200-dev	GPT-3.5	24.10	42.50	-	-	-	37.80	51.00	41.20	11.80	34.30	34.50	33.10	-	42.20	29.60	38.70	22.00	0.10	-	35.20	36.20	43.00
	IndicTrans-2	44.70	56.40	-	-	-	58.00	61.70	59.30	40.30	54.70	61.70	55.00	-	60.70	53.80	55.30	35.50	31.10	-	63.80	61.80	54.00
	Google Translate	41.90	-	-	-	-	57.50	61.90	59.80	-	45.00	61.40	55.10	-	59.00	58.60	-	-	-	-	62.60	62.00	53.70
	LTRC, IIIT-H	-	-	-	-	-	48.50	58.20	45.70	-	-	35.80	47.70	-	-	-	-	-	-	-	55.60	56.70	44.60
	SeamlessM4T	42.60	53.70	-	-	-	55.30	60.30	57.40	-	50.50	57.40	53.00	-	56.70	54.00	54.20	-	22.90	-	60.90	59.10	51.90
	Encoder-Decoder BPCC (H)	33.67	1.76	-	-	-	41.25	44.56	40.16	20.9	39.10	37.86	39.20	1.68	40.89	38.28	38.08	29.41	27.95	0.48	46.50	39.73	32.58
	Llama-2-7b+lorat(BI)	28.67	33.84	-	-	-	30.91	48.94	33.87	13.19	26.58	30.05	34.64	1.20	37.17	28.48	25.70	22.18	-	0.25	34.21	32.45	34.35
	Llama-2-7b+lorat(Multi)	31.63	37.10	-	-	-	31.20	44.57	36.59	17.60	34.64	32.69	37.58	3.77	40.28	32.50	26.24	23.36	0.00	1.83	34.18	35.06	35.29
	Llama-2-13b+lorat(BI)	34.86	40.58	-	-	-	38.17	51.70	39.99	20.60	32.96	38.35	40.54	0.67	42.70	36.10	32.31	13.05	12.89	0.27	42.25	39.60	40.26
	Llama-2-13b+lorat(Multi)	35.03	40.51	-	-	-	36.84	48.24	39.80	20.79	37.78	36.91	40.89	3.70	44.56	36.14	31.15	26.88	1.10	0.74	39.60	38.27	38.96
	Llama-2-13b+FF+lorat(Multi)	42.89	49.40	-	-	-	48.50	55.61	51.66	32.15	48.00	52.69	50.10	0.29	54.63	50.10	47.70	33.88	32.59	0.44	55.60	53.11	49.26
	Mistral-7B-v0.1+lorat(Multi)	30.41	35.54	-	-	-	31.69	43.29	33.55	13.76	31.68	30.70	36.25	4.28	38.70	16.57	22.04	22.58	0.08	0.60	31.14	30.11	31.00
flore200-devtest	GPT-3.5	24.40	41.40	-	-	-	39.00	51.10	41.50	9.70	33.70	33.90	35.80	-	43.20	30.80	38.20	23.40	0.10	-	36.00	36.80	44.20
	IndicTrans-2	44.50	56.10	-	-	-	59.00	60.80	60.10	39.50	54.50	61.90	55.10	-	60.60	53.10	53.20	36.10	30.90	-	62.80	63.10	53.00
	Google Translate	42.10	-	-	-	-	58.40	61.10	60.50	-	45.30	61.70	54.90	-	59.20	57.60	-	-	-	-	61.50	62.60	51.90
	LTRC, IIIT-H	-	-	-	-	-	49.40	57.80	46.20	-	-	36.10	48.20	-	-	-	-	-	-	-	55.10	57.40	44.10
	SeamlessM4T	42.40	54.00	-	-	-	56.30	59.50	58.20	-	50.60	57.60	52.50	-	56.40	53.40	52.20	0.00	22.50	0.00	60.30	60.40	51.10
	Encoder-Decoder BPCC (H)	34.07	1.91	-	-	-	41.49	43.59	39.97	20.77	38.91	37.97	39.75	1.64	40.58	39.3	34.11	29.41	28.65	0.47	42.27	43.76	27.34
	Llama-2-7b+lorat(BI)	28.11	33.03	-	-	-	30.63	48.10	33.24	12.76	26.03	29.40	34.87	1.20	36.17	27.63	25.26	21.90	11.50	0.28	34.00	31.94	33.51
	Llama-2-7b+lorat(Multi)	31.57	36.30	-	-	-	30.85	43.18	36.60	17.35	33.96	32.15	38.04	4.28	40.18	31.78	25.20	23.95	0.00	2.17	34.25	35.08	33.88
	Llama-2-13b+lorat(BI)	33.95	39.61	-	-	-	37.96	50.26	39.55	19.93	31.87	37.03	40.25	0.48	42.64	35.40	30.93	12.56	12.79	0.37	42.11	39.28	39.85
	Llama-2-13b+lorat(Multi)	35.09	39.93	-	-	-	36.8	47.24	39.35	19.79	37.03	35.94	40.49	3.74	43.99	35.11	29.35	27.44	0.85	0.87	39.13	38.21	38.90
	Llama-2-13b+FF+lorat(Multi)	43.30	48.51	-	-	-	48.76	54.89	51.53	32.23	47.83	53.08	50.18	0.24	55.01	49.78	46.69	34.86	32.40	0.51	54.54	53.36	48.40
	Mistral-7B-v0.1+lorat(Multi)	30.63	34.08	-	-	-	31.20	41.78	33.00	13.56	31.14	30.16	35.43	4.33	38.73	17.84	23.31	22.70	0.07	0.60	31.34	29.83	29.67
newstest2019	GPT-3.5	-	41.70	-	-	-	35.50	45.20	38.00	-	-	31.50	32.10	-	-	-	35.80	-	-	-	31.30	33.40	16.20
	IndicTrans-2	-	55.70	-	-	-	52.40	54.10	57.30	-	-	53.00	51.00	-	-	-	51.10	-	-	-	51.30	49.70	17.60
	Google Translate	-	-	-	-	-	51.70	54.50	57.40	-	-	52.20	49.80	-	-	-	-	-	-	-	50.40	50.20	17.40
	LTRC, IIIT-H	-	-	-	-	-	46.00	51.90	39.20	-	-	31.80	42.90	-	-</								

DataSet	Model	asm	ban	bod	doi	kon	guj	hin	kan	kas	mai	mal	mar	mei	nep	odi	pun	san	sat	sin	tam	tel	urd
IN22_conv	GPT-3.5	19.90	29.80	2.80	19.90	9.70	28.30	34.10	17.80	6.20	14.00	20.70	24.00	0.40	29.30	21.10	30.90	14.50	0.20	9.60	15.40	19.90	34.70
	IndicTrans-2	43.90	36.90	35.30	45.50	28.90	40.90	38.70	25.10	31.80	35.30	31.30	37.00	32.50	43.10	38.90	42.80	25.80	24.20	26.00	22.60	30.80	46.10
	Google Translate	44.50	37.60	1.80	42.30	29.50	40.90	39.40	24.30	5.60	36.40	31.10	37.60	25.70	43.00	37.40	39.40	26.70	0.00	8.70	23.30	31.50	45.60
	LTRC, IIT-H	-	-	-	-	-	-	23.90	10.50	-	-	14.60	19.10	-	-	-	-	-	-	-	-	14.10	-
	SeamlessM4T	41.00	35.90	0.00	0.00	0.00	40.60	38.10	23.00	0.00	34.00	30.30	35.50	0.20	40.30	38.60	41.40	0.00	17.20	9.30	23.10	31.10	42.30
	Encoder-Decoder BPCC (H)	8.68	7.18	5.77	4.53	4.08	5.65	10.72	2.27	1.83	6.44	5.35	7.04	2.76	7.85	3.04	5.19	2.27	2.08	4.01	2.67	3.60	6.48
	Llama-2-7b+Lora(BI)	1.17	2.42	4.04	9.49	4.43	3.88	15.72	1.45	2.17	3.08	1.70	7.00	0.08	6.22	1.07	2.29	4.62	0.03	3.45	0.81	0.84	6.33
	Llama-2-7b+Lora(Multi)	12.43	7.71	8.90	10.16	6.14	5.09	7.68	4.25	4.33	9.79	3.34	9.91	1.15	12.36	6.23	6.55	5.14	0.26	3.75	3.47	5.33	13.75
	Llama-2-13b+Lora(BI)	2.49	3.12	15.01	0.90	2.11	1.26	25.04	0.82	2.93	2.86	8.71	9.44	0.31	1.49	1.07	1.61	-	-	-	-	1.34	6.72
	Llama-2-13b+Lora(Multi)	21.19	20.26	15.51	18.40	13.37	17.06	23.66	8.15	8.54	15.77	12.85	17.37	2.93	22.89	11.77	15.41	9.85	0.72	8.17	9.27	11.53	24.16
	Llama-2-13b+FF+Lora(Multi)	2.26	1.77	1.48	2.00	1.58	0.84	7.05	0.80	0.94	1.90	5.99	2.06	0.48	2.71	1.55	1.52	1.35	0.13	1.81	0.84	0.92	2.31
	Mistral-7B-v0.1+Lora(Multi)	12.23	9.19	8.55	11.63	7.46	2.38	14.81	3.62	4.69	12.91	3.50	11.20	0.43	17.90	5.51	1.69	11.03	0.10	3.84	3.04	2.40	18.58
IN22_gen	GPT-3.5	19.00	25.20	6.00	20.80	13.40	25.60	30.40	23.60	10.00	19.50	15.80	22.50	0.20	27.60	18.90	25.60	14.30	0.20	13.70	14.40	20.20	29.20
	IndicTrans-2	42.50	40.80	37.50	53.40	32.70	43.10	40.00	40.00	38.40	42.50	40.40	41.50	38.40	47.80	43.30	40.80	30.60	25.00	31.50	35.90	42.30	53.70
	Google Translate	41.90	39.90	4.30	44.80	33.00	43.00	39.20	41.10	9.70	39.70	37.90	40.60	27.40	46.80	40.30	39.60	28.50	0.20	15.70	34.80	40.90	51.30
	LTRC, IIT-H	-	-	-	-	-	-	20.70	15.40	-	-	13.70	15.40	-	-	-	-	-	-	-	-	15.1	-
	SeamlessM4T	40.70	37.30	0.00	0.00	0.00	41.60	37.30	39.30	0.00	39.60	36.90	37.30	0.00	43.50	40.80	38.50	0.00	15.70	15.00	33.10	39.20	48.30
	Encoder-Decoder BPCC (H)	6.50	6.79	5.85	4.74	4.72	4.40	8.63	3.58	2.16	6.94	3.92	7.26	2.66	7.86	3.16	3.28	2.88	1.29	3.62	3.78	3.44	5.63
	Llama-2-7b+Lora(BI)	2.53	5.99	8.30	11.99	7.13	6.13	18.21	3.62	4.95	7.41	5.28	10.72	0.16	8.81	2.00	4.51	6.82	0.09	6.97	4.11	3.18	12.37
	Llama-2-7b+Lora(Multi)	12.01	9.89	8.79	14.42	9.61	5.07	11.38	8.51	7.64	14.69	6.39	11.47	0.77	16.75	7.71	6.14	9.36	0.40	8.10	5.82	7.36	16.98
	Llama-2-13b+Lora(BI)	6.66	8.01	12.15	3.95	7.74	5.80	26.55	2.20	8.17	8.40	10.06	13.82	0.88	2.70	4.85	4.61	-	-	-	-	4.02	17.46
	Llama-2-13b+Lora(Multi)	20.94	20.33	16.20	24.11	17.95	15.31	26.31	14.93	14.16	23.39	13.61	22.95	2.88	26.64	14.29	11.50	14.74	0.87	13.45	12.25	13.12	26.43
	Llama-2-13b+FF+Lora(Multi)	1.33	1.78	1.56	2.17	1.74	1.00	3.61	0.88	1.12	1.83	2.56	2.12	0.22	2.06	0.89	1.02	1.06	0.21	2.41	1.39	1.04	1.68
	Mistral-7B-v0.1+Lora(Multi)	12.36	10.90	9.31	16.72	10.94	2.68	13.88	7.82	8.38	16.02	4.49	14.17	0.69	16.96	5.50	3.39	12.33	0.12	8.19	6.52	6.17	20.46
flore200-dev	GPT-3.5	14.80	25.70	-	-	-	24.40	32.80	22.10	7.90	19.60	19.30	21.50	0.00	26.20	18.10	26.50	11.30	0.40	0.00	15.10	20.30	27.30
	IndicTrans-2	34.80	40.40	-	-	-	44.50	46.90	39.10	39.00	49.00	41.40	41.80	-	46.70	43.20	48.10	27.00	19.70	-	38.90	45.90	40.00
	Google Translate	34	40.80	-	-	-	45.60	47.70	39.70	11.50	48.70	42.00	42.50	-	47.40	43.90	48.30	25.10	0.20	-	39.30	46.40	41.80
	LTRC, IIT-H	-	-	-	-	-	-	27.10	17.30	-	-	18.50	19.10	-	-	-	-	-	-	-	20.10	-	-
	SeamlessM4T	34.20	39.20	-	-	-	0.00	0.00	37.70	0.00	46.30	40.00	39.70	0.00	44.50	41.20	45.80	0.00	19.70	0.00	37.50	43.60	39.80
	Encoder-Decoder BPCC (H)	5.87	7.22	-	-	-	5.08	10.55	4.10	1.93	7.98	4.67	8.19	2.05	7.67	3.4	3.81	3.13	1.14	0.70	4.55	4.37	5.23
	Llama-2-7b+Lora(BI)	1.88	5.57	-	-	-	5.57	17.98	3.37	4.82	6.97	5.77	10.82	1.36	8.41	2.24	5.61	6.73	0.17	2.32	2.97	2.87	9.35
	Llama-2-7b+Lora(Multi)	11.70	10.84	-	-	-	4.65	8.16	8.67	6.92	14.84	7.84	11.67	3.23	14.63	7.78	6.61	6.67	0.48	1.45	6.13	8.34	13.00
	Llama-2-13b+Lora(BI)	4.83	9.12	-	-	-	4.74	30.09	2.53	7.01	8.56	10.55	14.60	2.64	2.38	4.32	3.35	-	-	4.00	9.30	-	-
	Llama-2-13b+Lora(Multi)	18.40	22.13	-	-	-	15.47	29.09	14.67	12.48	24.99	15.75	22.20	5.65	24.27	14.71	17.31	12.36	0.94	4.51	14.70	15.39	22.23
	Llama-2-13b+FF+Lora(Multi)	1.34	-	-	-	-	0.87	3.35	1.31	0.80	1.71	1.78	2.20	1.15	1.71	1.07	0.86	1.17	0.18	0.84	1.29	1.31	-
	Mistral-7B-v0.1+Lora(Multi)	11.31	11.40	-	-	-	3.24	13.02	7.79	7.10	17.98	6.00	15.07	3.63	17.88	5.23	2.70	9.56	0.22	3.09	6.31	5.76	15.35
flore200-devtest	GPT-3.5	14.50	24.10	-	-	-	23.60	32.50	20.80	8.20	18.00	18.50	21.60	0.00	24.80	16.20	24.40	11.10	0.40	-	14.20	16.70	25.40
	IndicTrans-2	33.10	39.30	-	-	-	45.20	46.10	37.70	36.20	48.30	41.00	41.50	-	46.30	42.60	44.70	26.80	18.10	-	37.80	44.80	38.10
	Google Translate	32.80	39.80	-	-	-	46.20	46.10	38.00	10.70	46.60	40.90	42.10	-	46.30	41.30	45.90	25.20	0.10	-	37.70	44.90	40.10
	LTRC, IIT-H	-	-	-	-	-	-	27.60	16.70	-	-	17.90	18.20	-	-	-	-	-	-	-	19.20	-	-
	SeamlessM4T	32.30	38.30	-	-	-	-	-	36.00	-	44.30	39.70	38.80	-	43.40	40.20	43.20	0.00	18.30	0.00	35.20	42.90	38.10
	Encoder-Decoder BPCC (H)	5.30	6.83	-	-	-	4.87	8.74	3.84	1.52	7.07	3.96	7.47	1.88	6.95	2.59	2.99	2.41	1.02	0.58	4.46	3.76	4.67
	Llama-2-7b+Lora(BI)	1.64	5.51	-	-	-	5.78	16.88	2.91	4.28	7.45	5.76	10.24	1.18	8.56	2.11	5.48	6.75	0.10	2.22	2.46	3.02	8.37
	Llama-2-7b+Lora(Multi)	10.85	10.45	-	-	-	4.55	8.10	8.34	6.76	14.02	6.39	11.26	2.66	15.03	7.02	5.90	6.82	0.55	1.66	5.94	7.53	13.94
	Llama-2-13b+Lora(BI)	4.29	8.76	-	-	-	4.15	30.38	2.57	5.94	7.49	10.07	14.12	2.12	2.51	3.36	3.48	-	-	3.66	8.98	-	-
	Llama-2-13b+Lora(Multi)	17.64	20.59	-	-	-	15.49	28.6	13.86	11.17	23.60	14.97	21.84	5.38	23.58	13.13	15.22	11.85	0.94	4.51	12.73	15.29	21.33
	Llama-2-13b+FF+Lora(Multi)	1.12	1.42	-	-	-	0.75	2.67	0.95	0.88	1.53	1.70	1.73	0.92	1.38	0.9	0.97	0.81	0.15	0.78	1.26	1.16	1.27
	Mistral-7B-v0.1+Lora(Multi)	10.21	10.86	-	-	-	3.22	12.62	7.30	7.29	15.45	5.20	14.16	3.71	15.91	5.80	2.66	11.08	0.33	3.31	5.82	6.15	13.87
newstest2019	GPT-3.5	-	20.00	-	-	-	19.20	24.10	16.00	-	-	11.60	14.80	-	-	-	18.70	-	-	-	9.90	12.60	2.30
	IndicTrans-2	-	38.80	-	-	-	42.50	37.70	36.50	-	-	32.10	37.00	-	-	-	40.50	-	-	-	31.30	29.90	3.10
	Google Translate	-	38.50	-	-	-	42.60	-	36.60	-	-	34.00	37.70	-	-	-	-	-	-	-	31.80	30.50	4.30
	LTRC, IIT-H	-	-	-	-	-	-	12.70	13.40	-	-	13.80	15.70	-	-	-	-	-	-	-	-	9.40	-
	SeamlessM4T	-	35.30	-	-	-	39.60	35.30	33.80	-	-	30.00	33.40	-	-	-	37.60	-	-	-	29.40	27.90	3.80
	Encoder-Decoder BPCC (H)	-	3.93	-	-	-	2.14	5.53	1.69	-	-	1.55	4.49	-	-	-	1.74	-	-	-	1.82	1.38	0

DataSet	Model	asm	ban	bod	doi	kon	guj	hin	kan	kas	mai	mal	mar	mei	nep	odi	pun	san	sat	sin	tam	tel	urd	
IN22_conv	GPT-3.5	44.80	54.30	21.10	45.50	32.80	52.60	58.80	45.10	28.70	42.10	46.00	50.30	14.50	53.80	47.60	54.80	40.60	14.80	34.70	39.90	44.10	59.00	
	IndicTrans-2	63.90	59.80	57.20	66.00	52.60	63.20	60.90	49.20	53.50	59.20	55.40	60.10	53.90	64.40	61.60	63.40	49.60	45.10	50.60	47.40	54.20	66.60	
	Google Translate	64.70	60.80	16.40	63.70	52.80	63.20	61.80	49.70	23.20	60.00	56.10	60.50	47.20	64.90	60.10	62.70	50.30	0.30	32.80	48.50	55.20	66.30	
	LTRC, IIIT-H	-	-	-	-	-	-	52.20	34.80	-	-	40.70	47.30	-	-	-	-	-	-	-	-	39.00	-	
	SeamlessM4T	61.30	59.00	-	-	-	62.70	60.50	47.90	-	57.60	54.90	58.70	-	2.90	62.10	61.10	62.60	-	37.20	32.00	47.80	53.90	63.90
	Encoder-Decoder BPCC (H)	27.34	29.84	23.39	22.14	21.59	23.06	31.43	18.93	17.06	25.90	24.48	26.81	21.07	26.99	19.11	22.23	18.80	17.77	21.53	19.60	20.18	24.76	
	Llama-2-7b+lorat(BI)	1.49	4.57	8.27	24.13	16.23	9.36	31.77	4.18	15.60	7.80	4.99	15.83	1.20	12.66	2.94	5.45	13.11	0.29	20.31	2.24	2.45	16.30	
	Llama-2-7b+lorat(Multi)	26.69	17.30	21.90	23.50	16.78	10.84	14.04	14.27	16.95	23.63	11.05	21.46	8.80	24.88	16.87	12.50	18.37	8.69	13.81	12.11	16.00	28.71	
	Llama-2-13b+lorat(BI)	3.68	6.67	31.75	1.40	5.09	2.39	49.08	1.83	13.63	6.69	14.04	17.65	7.43	1.84	2.45	4.45	-	-	-	-	3.09	14.48	
	Llama-2-13b+lorat(Multi)	45.43	45.25	38.80	43.00	36.80	40.06	49.77	29.89	30.14	43.16	34.75	40.99	16.45	47.50	34.71	37.11	33.03	14.11	31.95	31.89	33.15	49.70	
	Llama-2-13b+FF+lorat(Multi)	19.76	19.98	18.74	18.95	18.68	17.04	24.57	15.70	16.53	19.70	17.63	19.36	15.68	20.79	17.26	16.71	17.49	12.69	18.59	17.60	16.33	20.80	
	Mistral-7B-v0.1+lorat(Multi)	27.16	17.34	25.40	28.29	21.90	3.70	27.60	8.48	22.19	30.57	8.24	25.55	8.30	32.84	12.60	3.49	25.06	4.19	14.88	8.47	6.49	37.17	
IN22_gen	GPT-3.5	49.10	54.40	27.70	50.70	41.90	54.00	59.80	54.20	38.50	51.50	48.70	52.80	16.20	56.80	50.50	54.00	46.90	19.50	42.70	44.00	49.10	59.50	
	IndicTrans-2	68.00	66.40	62.60	74.10	60.20	68.40	66.70	66.30	62.20	67.90	66.10	66.70	61.60	71.00	68.60	64.80	57.10	49.40	57.70	62.00	66.70	74.60	
	Google Translate	67.10	66.30	21.70	69.00	60.00	68.50	67.20	66.70	32.70	66.20	65.10	66.50	52.70	70.70	66.40	64.80	56.00	0.30	42.80	61.90	66.40	73.80	
	LTRC, IIIT-H	-	-	-	-	-	-	54.40	44.30	-	-	42.70	47.10	-	-	-	-	-	-	-	-	43.90	-	
	SeamlessM4T	65.60	63.80	-	-	-	66.80	64.90	64.70	-	65.30	63.80	63.60	1.40	67.80	66.00	63.10	-	39.60	38.60	60.00	64.50	71.10	
	Encoder-Decoder BPCC (H)	27.21	30.40	25.58	22.96	24.00	22.81	32.09	21.40	16.90	28.64	22.70	29.43	22.23	29.17	19.73	19.83	20.01	15.88	20.61	21.89	20.71	24.36	
	Llama-2-7b+lorat(BI)	4.77	12.80	23.56	30.80	22.03	18.59	35.50	9.40	24.24	19.09	15.89	24.88	5.40	18.53	6.74	15.03	21.20	0.46	24.61	9.55	7.15	29.47	
	Llama-2-7b+lorat(Multi)	33.38	25.50	30.40	35.91	28.11	17.70	23.86	28.24	29.00	36.17	22.40	28.40	15.67	37.65	26.45	21.11	30.45	9.97	25.74	21.48	25.37	41.56	
	Llama-2-13b+lorat(BI)	13.08	15.90	24.09	6.20	15.07	12.06	52.08	4.49	22.17	16.95	24.00	26.70	13.80	4.29	11.64	10.76	-	-	-	-	8.21	33.71	
	Llama-2-13b+lorat(Multi)	45.95	45.74	41.47	49.39	42.73	36.27	54.26	38.06	38.01	49.69	37.81	49.00	18.81	53.03	38.26	33.04	39.50	14.29	38.00	36.95	36.30	53.29	
	Llama-2-13b+FF+lorat(Multi)	16.70	17.10	16.83	17.88	16.76	14.98	20.50	15.31	15.95	17.38	16.60	17.81	13.40	17.78	14.66	15.05	15.84	12.46	17.68	16.61	14.87	17.45	
	Mistral-7B-v0.1+lorat(Multi)	26.80	23.09	30.43	35.65	28.46	5.60	28.00	16.30	29.10	35.29	13.18	30.96	13.76	34.00	14.30	6.18	32.03	2.44	21.60	16.04	13.30	41.19	
fiores200-dev	GPT-3.5	44.50	55.10	-	-	-	53.50	61.50	52.30	37.40	52.00	50.50	51.70	0.00	55.80	48.90	56.20	43.00	19.20	-	44.00	49.50	57.40	
	IndicTrans-2	60.70	65.60	-	-	-	68.20	69.80	64.50	63.90	72.00	66.10	66.30	-	70.40	67.80	70.40	54.00	43.90	-	64.00	68.90	65.30	
	Google Translate	60.40	<b>66.00</b>	-	-	-	<b>69.00</b>	<b>70.70</b>	<b>64.80</b>	37.90	<b>72.00</b>	<b>66.70</b>	<b>67.00</b>	-	<b>70.90</b>	<b>67.90</b>	<b>70.50</b>	53.60	0.30	-	<b>64.60</b>	<b>69.30</b>	<b>66.80</b>	
	LTRC, IIIT-H	-	-	-	-	-	-	58.30	47.60	-	-	48.60	51.10	-	-	-	-	-	-	-	-	50.30	-	
	SeamlessM4T	59.80	64.50	-	-	-	-	63.40	-	69.90	65.00	65.10	-	-	68.70	66.20	68.80	-	<b>44.20</b>	-	62.80	67.30	64.90	
	Encoder-Decoder BPCC (H)	26.57	33.31	-	-	-	24.13	33.33	23.19	18.29	31.19	23.97	30.95	21.29	29.89	20.88	21.66	21.60	17.35	18.47	24.39	22.75	25.40	
	Llama-2-7b+lorat(BI)	3.86	13.27	-	-	-	18.55	33.91	9.56	24.35	18.38	17.20	25.13	3.86	19.77	7.23	21.16	24.11	0.45	17.26	6.68	7.54	25.05	
	Llama-2-7b+lorat(Multi)	32.47	27.27	-	-	-	16.34	17.14	28.70	28.41	36.03	24.85	28.90	20.18	35.44	27.29	21.59	27.56	12.37	14.71	20.90	26.70	34.75	
	Llama-2-13b+lorat(BI)	10.94	19.00	-	-	-	10.56	53.87	5.00	20.58	17.03	25.83	28.57	9.73	4.21	10.39	8.63	-	-	-	7.70	21.78		
	Llama-2-13b+lorat(Multi)	43.69	47.68	-	-	-	37.91	56.27	38.65	36.59	51.53	39.66	47.90	<b>26.15</b>	50.98	38.57	41.27	37.78	15.88	<b>25.85</b>	39.00	39.83	49.76	
	Llama-2-13b+FF+lorat(Multi)	18.77	-	-	-	-	17.44	21.33	17.45	17.23	19.26	17.04	19.97	18.10	19.00	16.73	16.43	17.73	14.26	16.26	18.31	17.40	-	
	Mistral-7B-v0.1+lorat(Multi)	27.96	25.19	-	-	-	6.97	27.01	19.06	28.60	38.69	17.34	35.26	19.00	38.19	15.10	6.03	29.75	2.85	19.23	16.91	13.80	37.69	
fiores200-devtest	GPT-3.5	43.90	54.10	-	-	-	52.90	60.90	51.40	37.70	51.20	50.10	51.90	0.00	54.80	47.10	54.20	43.00	19.10	-	43.20	46.90	55.60	
	IndicTrans-2	59.40	64.80	-	-	-	68.80	69.10	63.40	<b>61.80</b>	<b>71.30</b>	66.20	66.60	-	69.90	<b>66.80</b>	67.90	<b>54.00</b>	42.30	-	63.20	68.00	64.10	
	Google Translate	<b>59.80</b>	<b>65.30</b>	-	-	-	<b>69.70</b>	<b>69.70</b>	<b>63.90</b>	37.20	70.50	<b>66.40</b>	<b>67.20</b>	-	<b>70.50</b>	65.90	<b>68.70</b>	53.80	0.30	-	<b>63.60</b>	<b>68.50</b>	<b>65.60</b>	
	LTRC, IIIT-H	-	-	-	-	-	-	58.40	47.00	-	-	48.20	50.90	-	-	-	-	-	-	-	-	49.40	-	
	SeamlessM4T	58.30	63.80	-	-	-	-	62.50	-	68.50	65.20	65.00	-	-	68.20	65.00	67.00	-	<b>42.40</b>	-	61.40	66.90	63.90	
	Encoder-Decoder BPCC (H)	26.67	32.80	-	-	-	24.11	32.73	22.80	18.21	30.39	23.64	30.40	21.21	29.37	20.17	20.98	21.14	16.99	18.51	24.11	22.60	24.69	
	Llama-2-7b+lorat(BI)	3.96	12.63	-	-	-	17.54	31.87	7.95	23.86	19.30	17.85	25.35	3.57	19.97	07.09	21.16	23.64	0.46	17.78	05.69	7.17	21.69	
	Llama-2-7b+lorat(Multi)	30.60	26.28	-	-	-	16.24	16.20	28.86	28.14	35.87	24.17	28.17	20.11	36.60	26.40	21.37	27.30	11.38	14.00	21.27	25.53	36.00	
	Llama-2-13b+lorat(BI)	9.87	18.36	-	-	-	9.38	53.30	4.93	19.71	15.73	25.70	28.00	8.94	3.88	9.74	8.59	-	-	-	7.07	20.70		
	Llama-2-13b+lorat(Multi)	42.67	45.51	-	-	-	36.97	55.44	38.15	35.40	51.23	39.66	48.08	<b>25.77</b>	50.25	37.08	39.84	37.93	<b>15.83</b>	<b>25.78</b>	37.28	39.18	48.76	
	Llama-2-13b+FF+lorat(Multi)	17.96	19.29	-	-	-	16.86	20.13	16.95	17.08	19.06	16.71	19.33	17.84	18.45	16.27	16.09	17.2	13.71	15.83	17.98	16.68	18.23	
	Mistral-7B-v0.1+lorat(Multi)	25.70	23.29	-	-	-	6.03	25.07	17.79	28.05	35.57	15.76	33.40	19.29	36.31	15.47	5.54	29.95	2.91	19.26	14.98	14.18	35.23	
newstest2019	GPT-3.5	-	51.80	-	-	-	50.70	56.70	48.10	-	-	45.10	48.00	-	-	-	50.80	-	-	-	40.00	43.00	19.40	
	IndicTrans-2	-	<b>64.50</b>	-	-	-	67.70	<b>64.60</b>	62.40	-	-	61.10	63.90	-	-	-	<b>65.80</b>	-	-	-	58.30	56.80	19.00	
	Google Translate	-	<b>64.50</b>	-	-	-	<b>67.80</b>	-	<b>62.70</b>	-	-	<b>61.60</b>	<b>64.10</b>	-	-	-	-	-	-	-	<b>58.70</b>	<b>57.30</b>	<b>19.60</b>	
	</																							