Translation	Model	BLEU Score	ROUGE-1 (R)	ROUGE-1 (P)	ROUGE-1 (F)	ROUGE-2 (R)	ROUGE-2 (P)	ROUGE-2 (F)	ROUGE-L (R)	ROUGE-L (P)	ROUGE-L (F)
Eng to Hin	NLLB	0.6326	0.5621	0.5856	0.5684	0.3192	0.3310	0.3213	0.5193	0.5410	0.5248
Eng to Hin	IndicTrans	0.6508	0.5768	0.5913	0.5782	0.3311	0.3375	0.3306	0.5457	0.5598	0.5470
Eng to Hin	ChatGPT	0.5858	0.5231	0.5477	0.5311	0.2844	0.3016	0.2884	0.4778	0.4993	0.4844
Hin to Eng	NLLB	0.7084	0.6267	0.6184	0.6160	0.3757	0.3686	0.3671	0.5782	0.5704	0.5679
Hin to Eng	IndicTrans	0.7199	0.6733	0.6389	0.6494	0.4372	0.4134	0.4193	0.6283	0.5967	0.6060
Hin to Eng	ChatGPT	0.6718	0.5825	0.5823	0.5764	0.3355	0.3372	0.3324	0.5342	0.5327	0.5278
Hin to Guj	NLLB	0.7084	0.6267	0.6184	0.6160	0.3757	0.3686	0.3671	0.5782	0.5704	0.5679
Hin to Guj	IndicTrans	0.6086	0.4727	0.4694	0.4668	0.2421	0.2367	0.2364	0.4565	0.4522	0.4502
Hin to Guj	ChatGPT	0.4947	0.3025	0.3104	0.3025	0.1190	0.1208	0.1187	0.2951	0.3035	0.2954
Guj to Hin	NLLB	0.6379	0.5648	0.5791	0.5660	0.3477	0.3629	0.3515	0.5362	0.5497	0.5392
Guj to Hin	IndicTrans	0.6358	0.5657	0.5804	0.5682	0.3348	0.3531	0.3395	0.5329	0.5498	0.5368
Guj to Hin	ChatGPT	0.5686	0.4674	0.4781	0.4688	0.2485	0.2507	0.2475	0.4479	0.4581	0.4492

In the above table you can see the result that has been performed on same 50 sentences by NLLB ,IndicTrans and ChatGPT

Translation	Model	BLEU Score	ROUGE-1 (R)	ROUGE-1 (P)	ROUGE-1 (F)	ROUGE-2 (R)	ROUGE-2 (P)	ROUGE-2 (F)	ROUGE-L (R)	ROUGE-L (P)	ROUGE-L (F)
Eng to Hin	NLLB	0.6678	0.5902	0.6094	0.5960	0.3525	0.3639	0.3558	0.5528	0.5712	0.5584
Eng to Hin	IndicTrans	0.6997	0.6249	0.6296	0.6239	0.3911	0.3936	0.3902	0.5897	0.5948	0.5891
Hin to Eng	NLLB	0.7135	0.6241	0.6305	0.6236	0.3944	0.3964	0.3927	0.5896	0.5952	0.5889
Hin to Eng	IndicTrans	0.7525	0.6618	0.6631	0.6590	0.4482	0.4454	0.4442	0.6288	0.6297	0.6261
Hin to Guj	NLLB	0.5935	0.4540	0.4718	0.4587	0.2133	0.2212	0.2157	0.4425	0.4593	0.4468
Hin to Guj	IndicTrans	0.6354	0.4937	0.4970	0.4917	0.2484	0.2487	0.2466	0.4697	0.4731	0.4679
Guj to Hin	NLLB	0.6749	0.5988	0.6093	0.6000	0.3703	0.3755	0.3703	0.5638	0.5735	0.5649
Guj to Hin	IndicTrans	0.6722	0.5895	0.5882	0.5847	0.3549	0.3540	0.3519	0.5546	0.5541	0.5505

In the above table you can see the result that has been performed on same 1000 sentence by NLLB and IndicTrans Model.

BLEU(Bilingual Evaluation Understudy) Score:

- Measures the similarity between a machine-generated translation and one or more human reference translations.
- Scores range from 0 to 1, with higher scores indicating better translation quality.
- Calculates precision of n-grams (sequences of n words) and applies a brevity penalty. BLEU score is often reported as a cumulative score, which combines the precision of multiple n-gram lengths (1-gram, 2-gram, 3-gram, etc.) into a single score. This provides a holistic evaluation of the translation quality across different n-gram lengths
- Limitations include not considering semantic equivalence or fluency.

ROUGE(Recall-Oriented Understudy for Gisting Evaluation) Score:

- Evaluates the quality of summaries or translations by comparing them to reference summaries or translations.
- Measures recall of n-grams and other units like word sequences or longest common subsequences.
- Provides insights into the overlap between the generated and reference texts.
- Commonly used in summarization tasks and ranges from 0 to 1, with higher scores indicating better quality.
- ROUGE-1: This metric calculates the overlap of unigrams (individual words) between the generated summary or translation and the reference text. It measures the precision, recall, and F1 score based on the number of overlapping unigrams.
- ROUGE-2: ROUGE-2 calculates the overlap of bigrams (sequences of two adjacent words) between the generated summary or translation and the reference text. Similar to ROUGE-1, it measures precision, recall, and F1 score based on the number of overlapping bigrams.
- ROUGE-L: ROUGE-L measures the longest common subsequence (LCS) between the generated summary or translation and the reference text. It considers not only individual word overlap but also the longest sequence of words that appears in both the generated and reference texts. This helps capture the fluency and coherence of the generated text.
- ROUGE scores provide a quantitative assessment of how well the machine-generated translation captures the content, fluency, and coherence of the reference translation(s). Higher ROUGE scores indicate greater similarity and, ideally, better quality translations. These scores are commonly used in machine translation research to evaluate and compare different translation systems or approaches.

Key Learnings:

1. As we can see in first table Both Model Performs better then ChatGPT on 50 sentences.

2. In most cases, the IndicTrans model performs better then NLLB and ChatGPT models in terms of both BLEU and ROUGE scores across various translation directions. This suggests that IndicTrans may be more effective for translation tasks involving these language pairs.But we are not very sure because in case of Gujarati to Hindi NLLB performs better then IndicTrans in both tables.

3. The performance of models can vary depending on the translation direction. For instance, in Hindi to English translation, all models generally perform better compared to Hindi to Gujarati translation, indicating potential challenges specific to the Gujarati language pair.

4. While BLEU scores provide a measure of overall translation quality, ROUGE scores offer insights into the precision, recall, and F1-score of generated translations, particularly at the unigram, bigram, and longest common subsequence levels.

5. Despite the differences in performance among models, there is still room for improvement in translation quality across all translation directions.

6. All models seem to struggle more with translating from Hindi to Gujarati compared to other translation directions, as evidenced by lower BLEU and ROUGE scores in this direction.

7. The performance of translation models varies depending on the translation direction. For example, NLLB and IndicTrans generally perform better when translating from Hindi to English compared to translating from English to Hindi.

8. Longer sentences may have lower Rouge and BLEU scores due to a lower chance of exact n-gram matches.