For all purpose

INTRO TO NLP

PROJECT

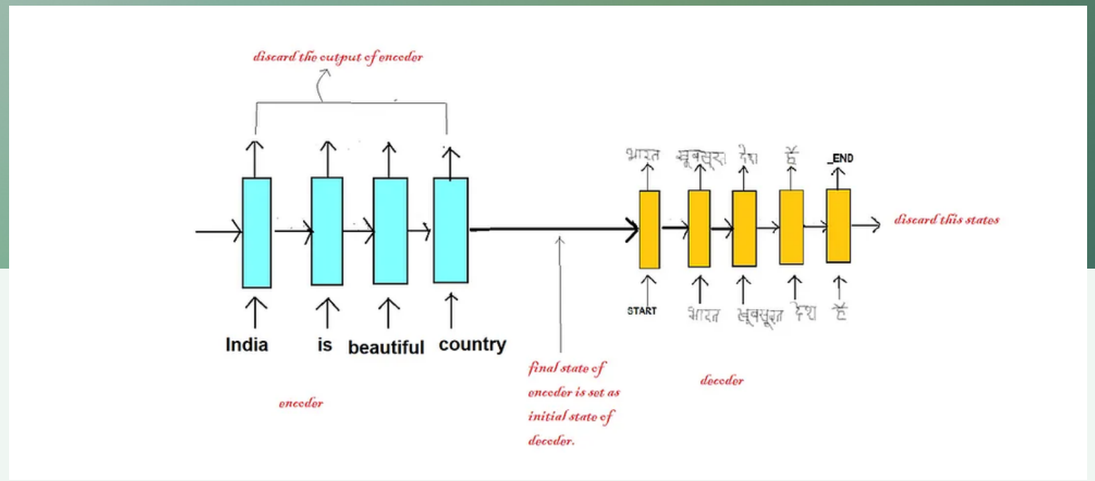
Akshit Kumar
Aryan Chandramania
Lakshmipathi Balaji

OVERVIEW

Our project aims to generate a Hindi-English (Hinglish) code-mixed sentence given a sentence in English. To this end, we used 3 different models - an LSTM encoder-decoder, the mT5-small model, and the IndicBART model.

Interim Submission

Our approach in generating the codemixed sentences was unconditioned in our interim submission. After a meeting with the mentors, the plans got changed and we went for conditioned code-mixed generation basically translating a given English sentence to Hinglish code mixed sentence and finding bleu_scores.



LSTM

The model is trained using a teacher-forcing approach, which involves feeding the correct output sequence from the previous time step as input to the decoder during training. Using teacher-forcing can lead to faster convergence and better model accuracy.

FINETUNING

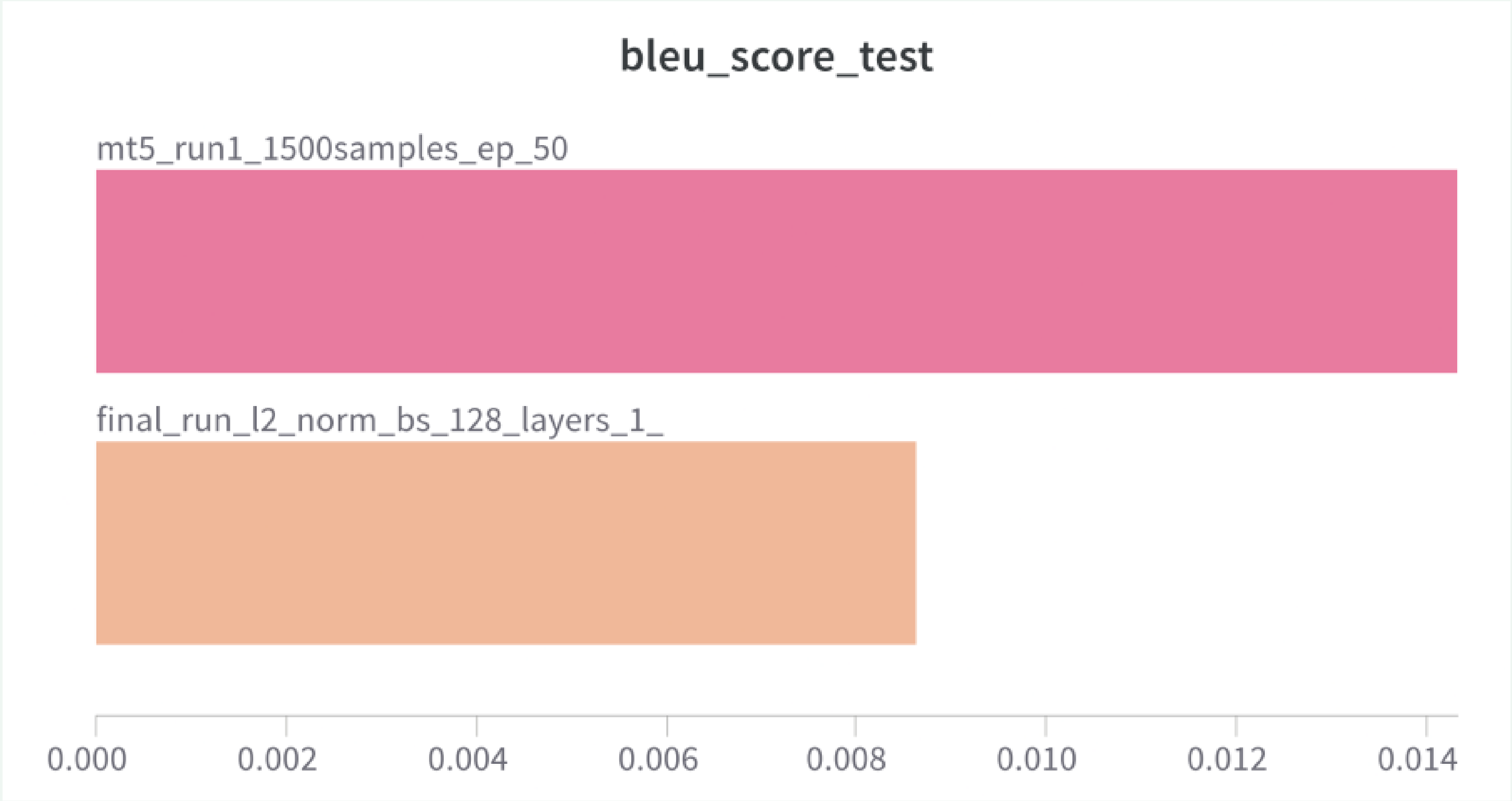
mT5 - small

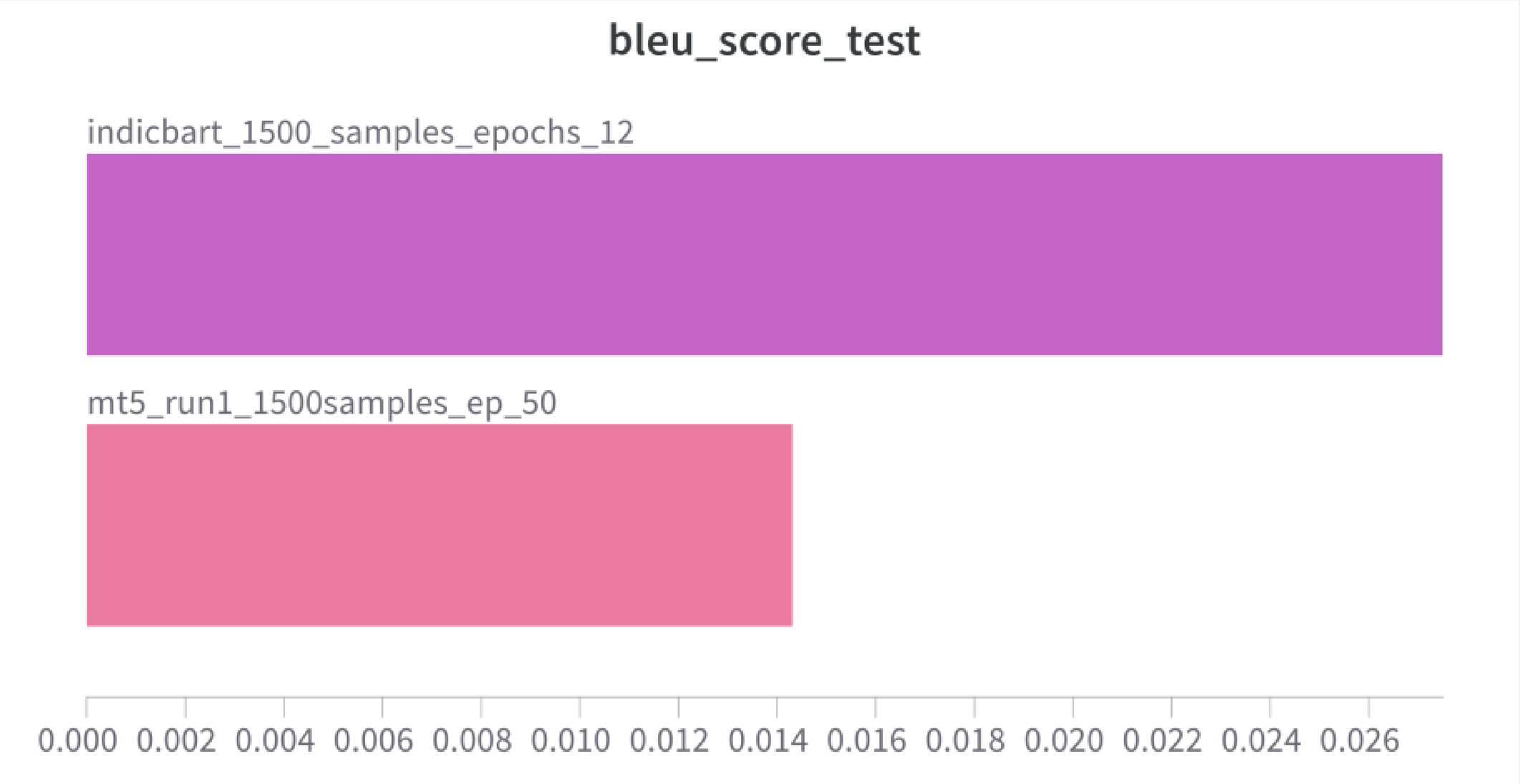
A multilingual variant of T5 that was pre-trained on a new Common Crawl-based dataset covering 101 languages.

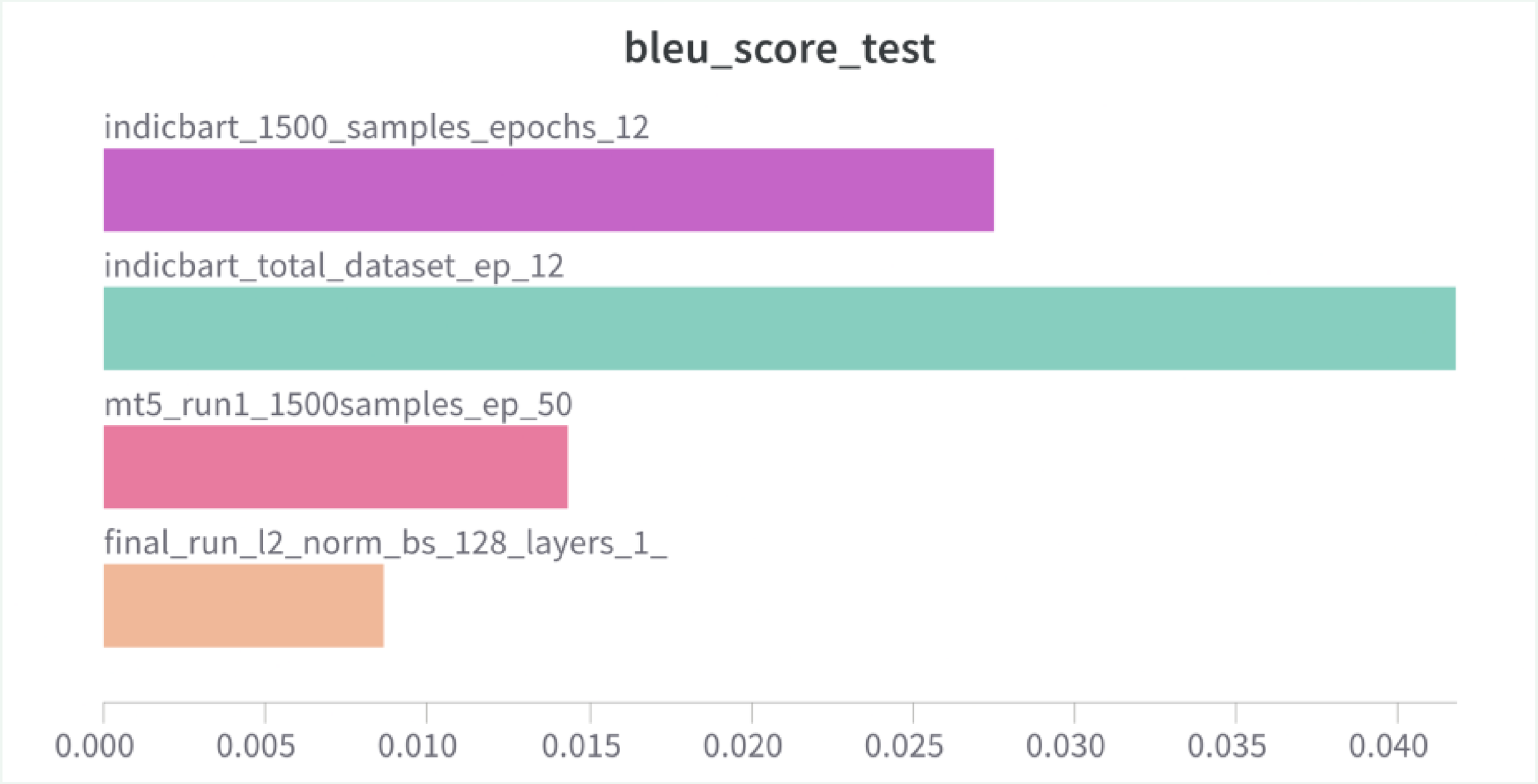
Especially useful for machine translation and summarization tasks.

IndicBART

A multilingual, sequence-to-sequence pre-trained model focusing on Indic languages and English. It currently supports 11 Indian languages and is based on the mBART architecture.


bleu_score_test
mt5_run1_1500samples_ep_50
final_run_l2_norm_bs_128_layers_1_
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014


bleu_score_test
indicbart_1500_samples_epochs_12
mt5_run1_1500samples_ep_50
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
0.018
0.020
0.022
0.024
0.026


bleu_score_test
indicbart_1500_samples_epochs_12
indicbart_total_dataset_ep_12
mt5_run1_1500samples_ep_50
final_run_l2_norm_bs_128_layers_1_
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040

WHAT ELSE WE TRIED

- A seq2seq transformer based encoder-decoder model which had unreliable results thus was not included in the report.

- indic-bert, a multilingual ALBERT model pretrained exclusively on 12 major Indian languages. later found out it is not suited to seq2seq tasks.

CHALLENGES

Lack of compute power

We were restricted to using either local systems or Google Colab, since none of us even had access to the Ada cluster or the like. Thus we had to compromise in a lot of places, such as size of data, size of model, training parameters, etc.

Scarcity of good datasets

It proved very hard to find some good, sizeable datasets for what we wanted to do; what we did find was mediocre.

References

- Sentence Bleu-Score in NLTK
- A guide for seq2seq encoder-decoder LSTM-based model for machine translation
- mT5
- IndicBart [https://doi.org/10.18653%2Fv1%2F2022.findings-acl.145]

ACKNOWLEDGEMENTS

- Prof. Manish Shrivastava, for giving us the opportunity to work on such a project.
- Teaching Assistants Ekansh Chauhan and Ankita Maity, for guiding us through the process.
- Prashant Kodali at LTRC, for providing help and insight.