**Slide 1 - Nav**

Hi, We are group 35 and we will talk about. ROMANSETU: Efficiently unlocking multilingual capabilities of Large Language Models models via Romanization.

**Slide 2 - Nav**

Large Language Models are typically trained on massive amounts of text data, but the vast majority of this data is in English. This leads to challenges when using LLMs for tasks involving non-English languages. To resolve this issue, ROMANSETU proposes leveraging romanization, a technique used for converting text from a non-Latin script to the Latin script.

**Slide 3 - Shreyas**

Authors have highlighted two main problems processing of non latin scripts and tokenization. These are challenging tasks because of different character sets, grammatical structures and high fertility. Therefore the authors have used romanization that is widely used in informal communication and shares tokens with english suggesting cross lingual alignment. The authors have performed sentiment analysis in hindi and hindi to english machine translation in this paper.

**Slide 4 - Anusha/Nav**

The first core task that the authors addressed was choosing a romanization scheme to convert the Devanagari text to it’s romanized form. They experimented with two of these, extended ITRANS and IndicXlit, and chose the latter because of its performance.

**Slide 5 - Anusha/Nav**

The authors utilized a different dataset for each of their sub-tasks. Here, they picked up existing datasets, and romanized the Hindi text using the scheme we just discussed.

**Slide 6 - SKIP**

**Slide 7 - Sebastian**

The authors propose the following design with three main sections

* Data preprocessing
* Continual Pre-training
* Supervised fine-tuning

**Slide 8 - Sebastian**

They prepare the data in different forms

* Hindi
* Transliterated (Roman Hindi)
* Hindi + Transliterated (Roman Hindi)

**Slide 9 & 10 - Sebastian**

For the all the tasks. They use Llama2 which is mostly trained on english data from the internet.

The main objective of of CPT is to familiarize the Base model with hindi and romanized hindi script without forgetting its english capabilities.

Thus, they use the same amount of hindi + english data

**Slide 11 & 12 - Nav**

During the supervised fine tuning phase the model is trained to perform specific tasks.

The paper explored 4 different strategies for fine-tuning. These strategies involved training the LLM with different input configurations:

1. romanized examples,
2. native script examples,
3. multi-script examples,
4. aggregating examples of all three types

In this example the LLM is taking multi-script, that is hindi text and it’s romanized text as an input to get target english translation.

**Slide 13 - Nav**

Here we can see that the task is sentimental analisis and the input is the same and the target now is positive or negative sentiment.

**Slide 14 - Paul**

Let’s take a look at the results

For the translation task, the evaluation metric used was BLEU scores of the output

Notably, overall highest BLEU score was achieved when the LLM was continually pretrained and fine-tuned with the multi-script dataset (Hindi+RH).

**Slide 15 & 16 - Paul**

In addition to an improvement in quality of the output, the authors noted an improvement in efficiency at the time of inference when using the romanized input.

They reported that the fertility of the romanized text is 2x times lower than the native text

What this means is that after the LLM’s tokenization of the input, the length of the tokens for the pure hindi script is two times larger than the length of the tokens for the romanized input

Thus, Processing romanized text is significantly more efficient than native text at the time of inference

Now let's look at the sentiment analysis task

The accuracies of sentiment predicted is presented in the table here

Once again, the highest accuracy was achieved when the LLM was continually pre-trained with the multiscript dataset

**Slide 17 - Paul**

To conclude we have both an improvement in performance by the LLMs and a significant improvement in efficiency at the time of inference when training and prompting the LLM with the romanized representation of Hindi

Future work in this area includes extending the ROMANSETU project to more non-Latin languages, to perform diverse NLP tasks, training the LLM with a Larger text corpora, and exploring cross lingual transfer to broaden impact of research.

Thank you.