How Does a Multilingual LM Handle Multiple Languages?

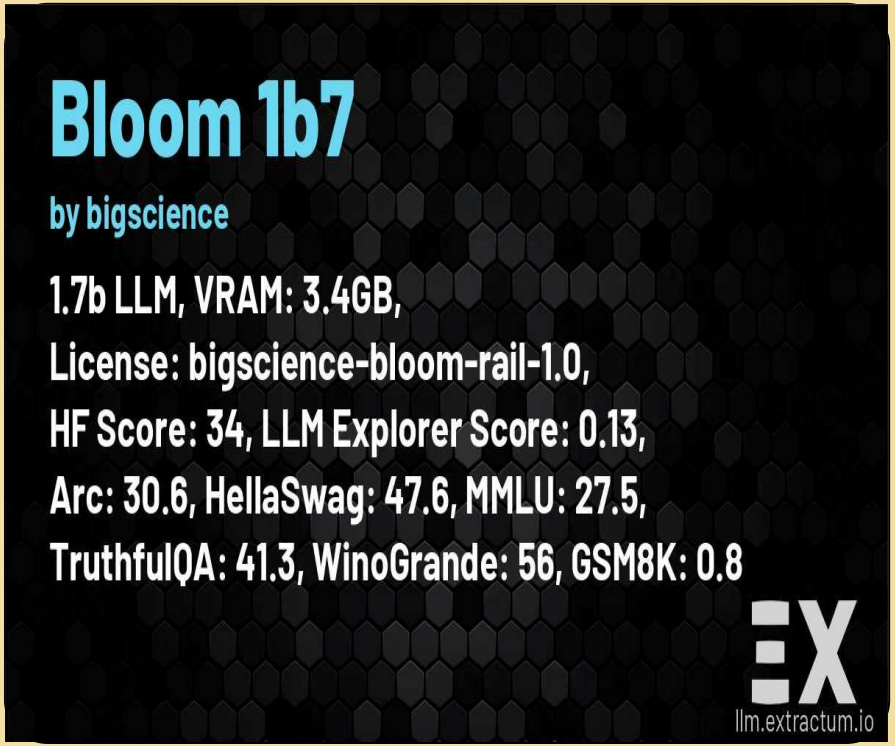
Shashwat Bhardwaj
2023AIY7528
Tony Stark

AIL821 | Large Language Models (LLMs): Introduction and Recent Advances | IIT-Delhi

TABLE OF CONTENTS

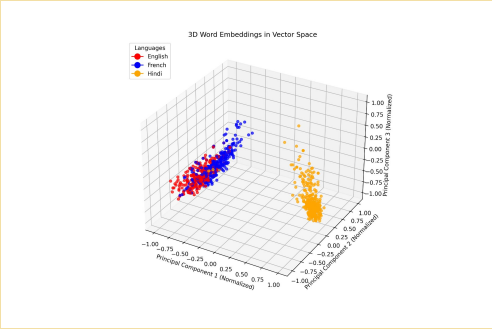
INTRODUCTION

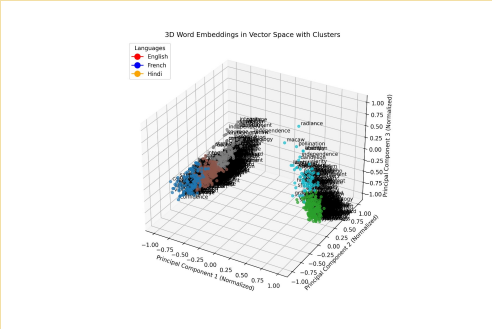
The project was aimed to understand how multiple languages are processed by a multilingual language model BLOOM 1.7B in our case and whether they have the ability to transfer knowledge across languages or not



Tony Stark

Task-1

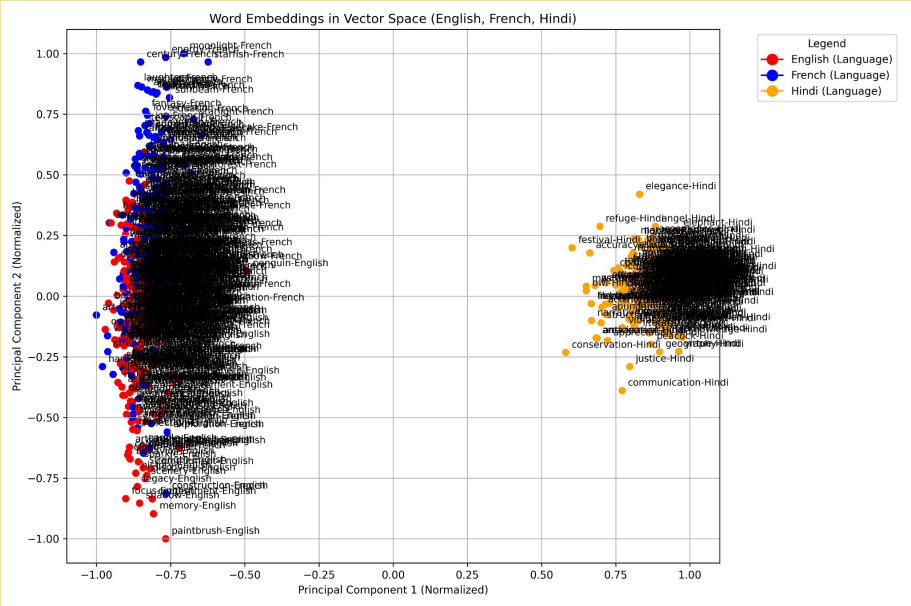


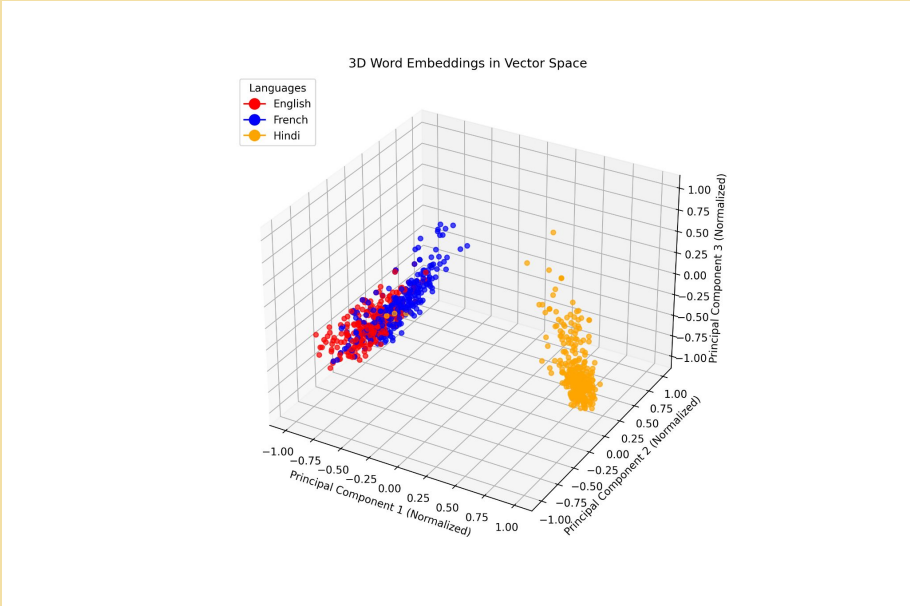


We investigate the multilingual semantic alignment in BLOOM-1.7B by examining word embeddings of semantically identical words across three languages: English, French, and Hindi.

Tony Stark

Task-1





2D PCA Plots and 3D PCA Plots

Tony Stark

Task-1

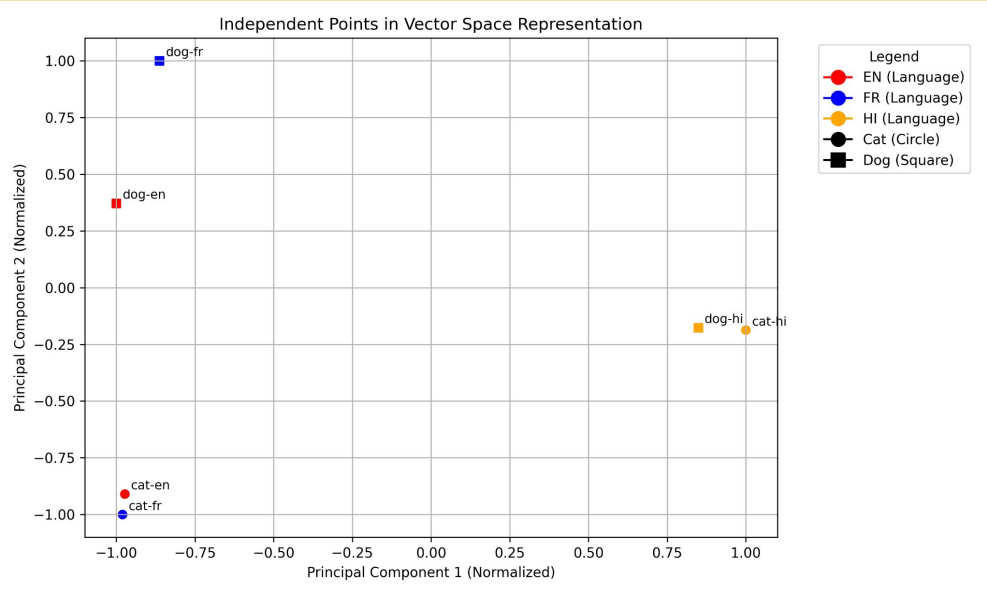
Findings: 

- English and French are closer in embedding space
- Hindi being from an Indic class of languages has a different embedding space

2D PCA Plots and 3D PCA Plots

Tony Stark

Task-1



Independent points in vector space

Tony Stark

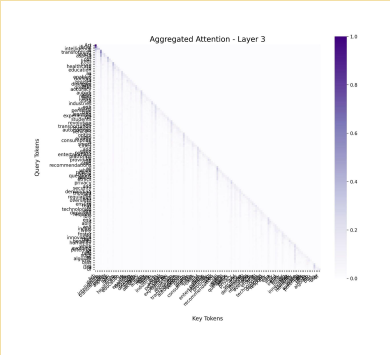
Task-1

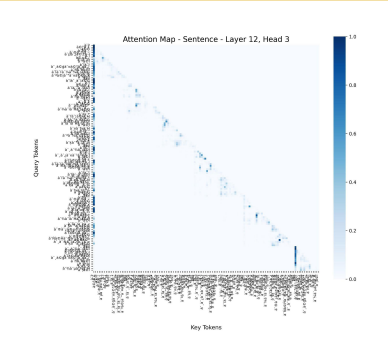
Findings: 

- English and French embeddings for "cat" and "dog" are close, with "cat" being particularly similar.
- Hindi embeddings lie far from English and French, reflecting its distinct language family.

Tony Stark

Task-2





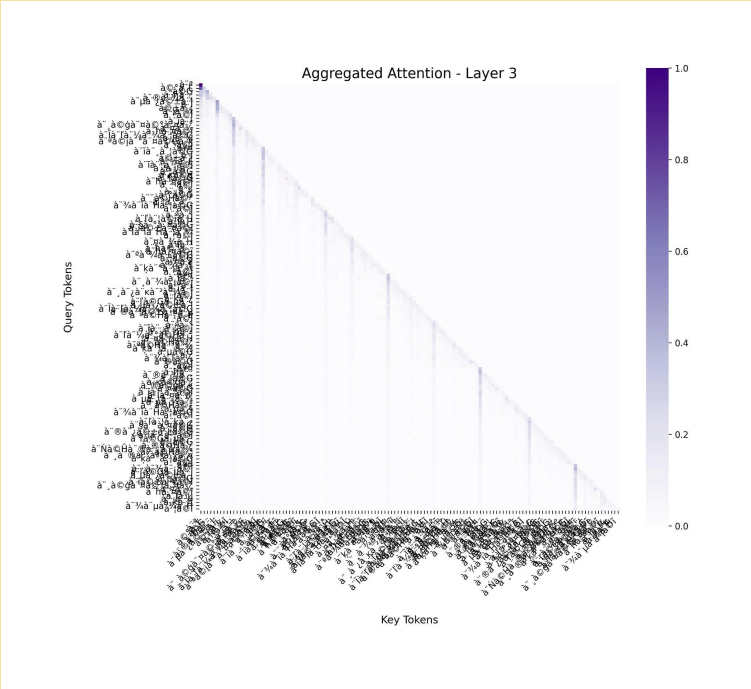
This task explores the internal workings of BLOOM-1.7B by visualizing attention maps and employing hooking techniques to understand the role of individual layers and their contribution to the model's overall performance.

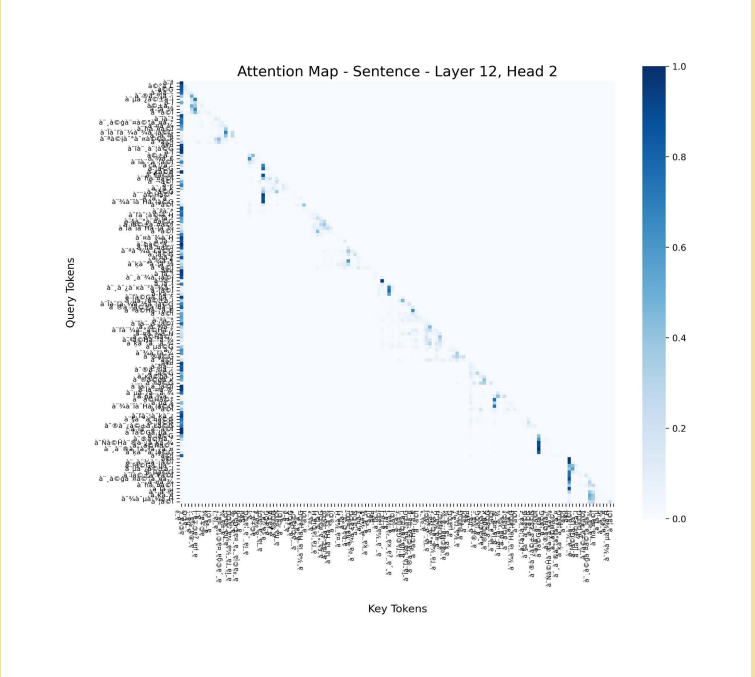
Two types of attention visualizations were produced:

- **Single-Head Attention Maps**: These maps illustrate the attention weights for individual heads, highlighting token-to-token relationships.
- **Aggregated Attention Maps**: These maps combine attention weights across all heads in a layer, providing a holistic view of the layer's focus

Tony Stark

Task-2

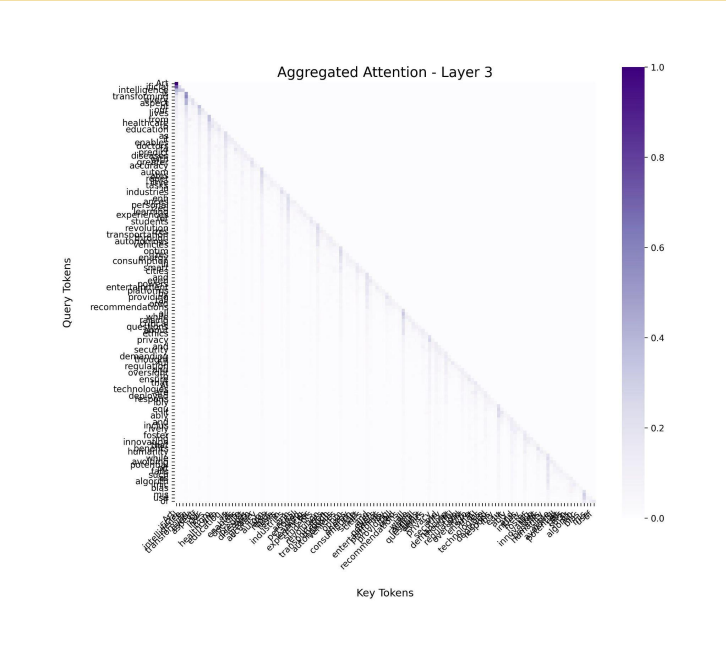


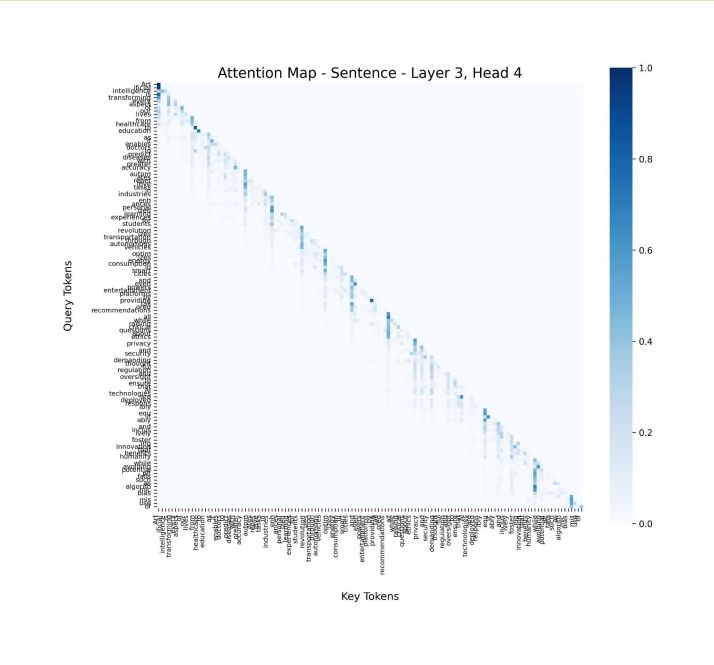


Attention Maps:=> Punjabi (Gurumukhi) Agg Attn and Layer-Headwise

Tony Stark

Task-2

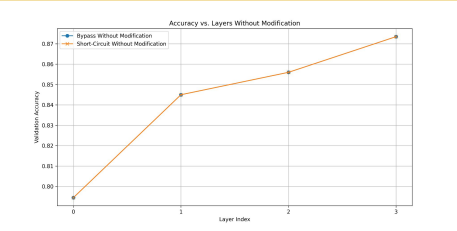


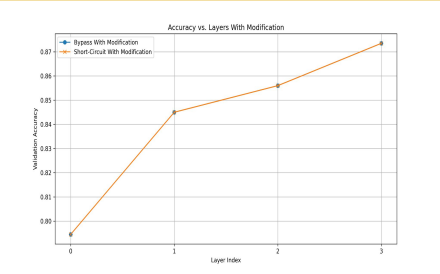


Attention Maps:=> English Agg Attn and Layer-Headwise

Tony Stark

Task-2





To understand the role of individual layers in BLOOM-1.7B, probing techniques were employed using transformer hooking mechanisms.

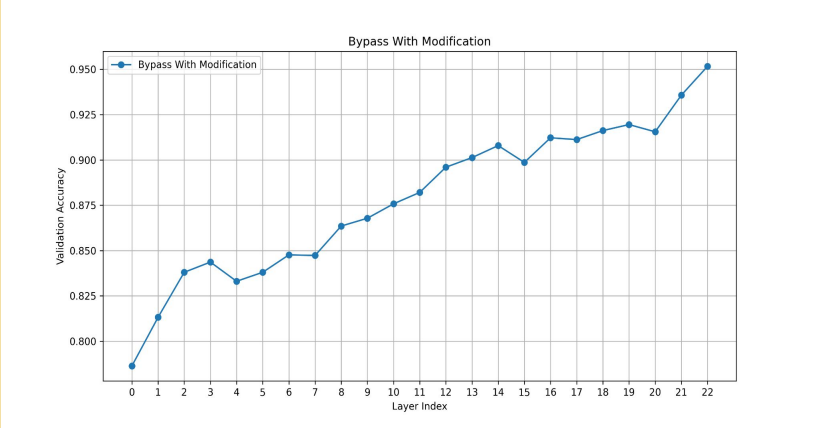
Two types of attention visualizations were produced:

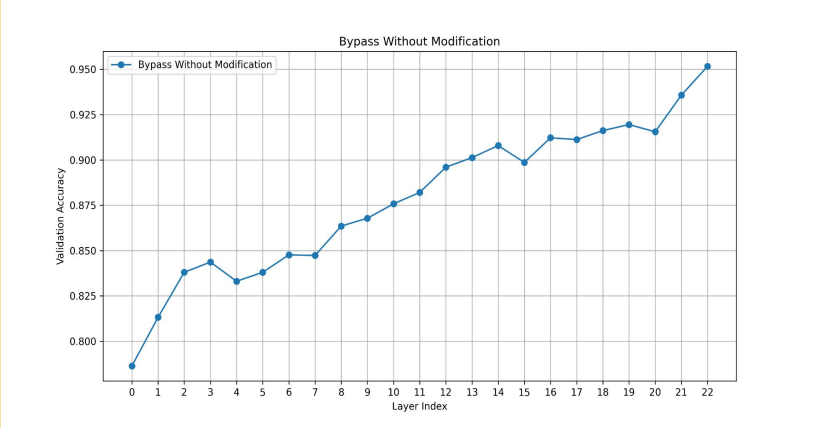
- **Bypassing Layers**: It involves bypassing a specific layer by equating its output to the previous layer's output. It helps analyze the contribution of the bypassed layer to the model's performance.

- **Short-Circuiting Layers**: It involves zeroing out the outputs of a specific layer to measure its impact on downstream tasks. It allows us to identify the importance of the layer for specific tasks.

Tony Stark

Task-2

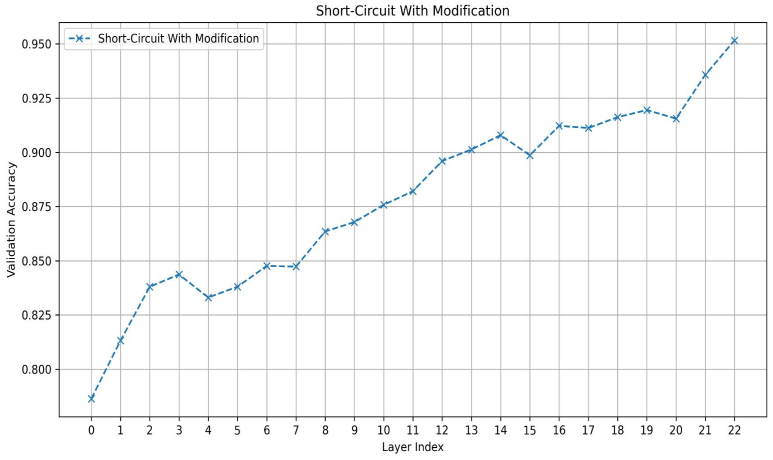


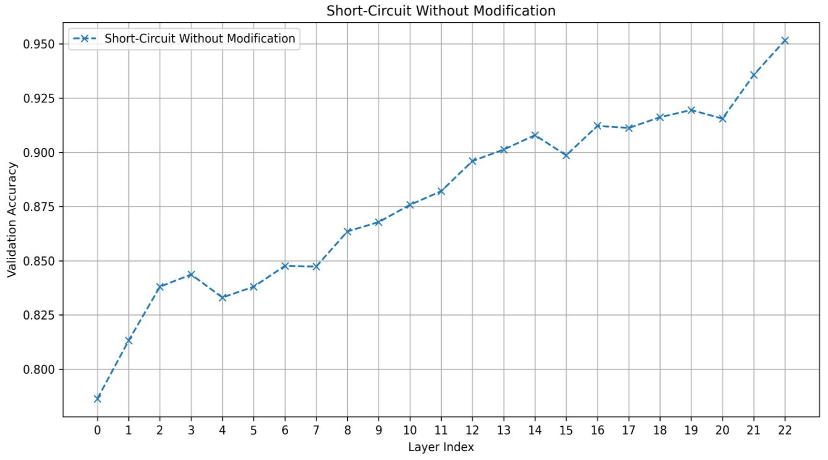


By passing:=> With and Without Modification

Tony Stark

Task-2





Short Circuiting:=> With and Without Modification

Tony Stark

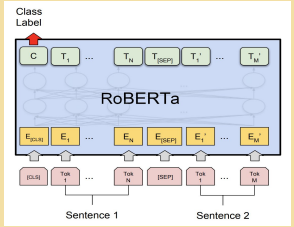
Task-2

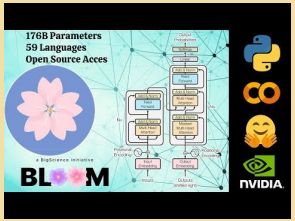
Findings: 

One critical observation is that as we hook into deeper layers of the model, the hidden states become more and more enriched which is evident from increasing accuracies across layers hooks.

Tony Stark

Task-3





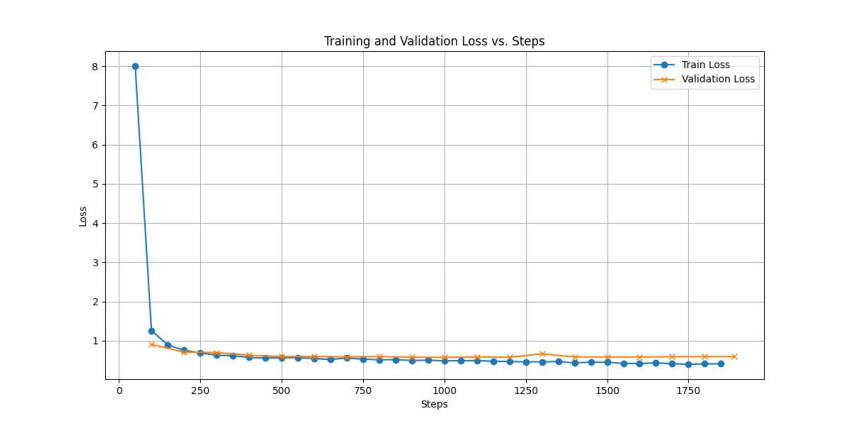
In this task, three models—BLOOM-1.7B, RoBERTa, and IndicBERT—were fine-tuned on English and evaluated on Bengali and Gujarati datasets.

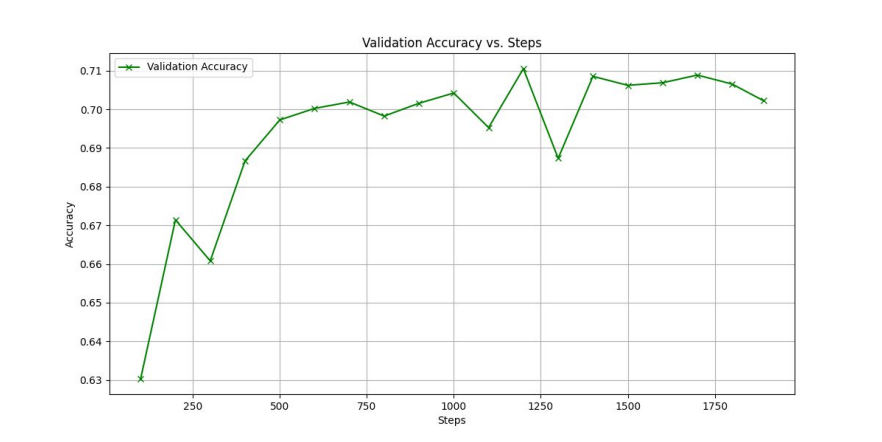
Three models were fine-tuned in as experiments

- **Bloom 1.7B [with QLoRA]**
- **RoBERTa**
- **IndicBERT**

Tony Stark

Task-3

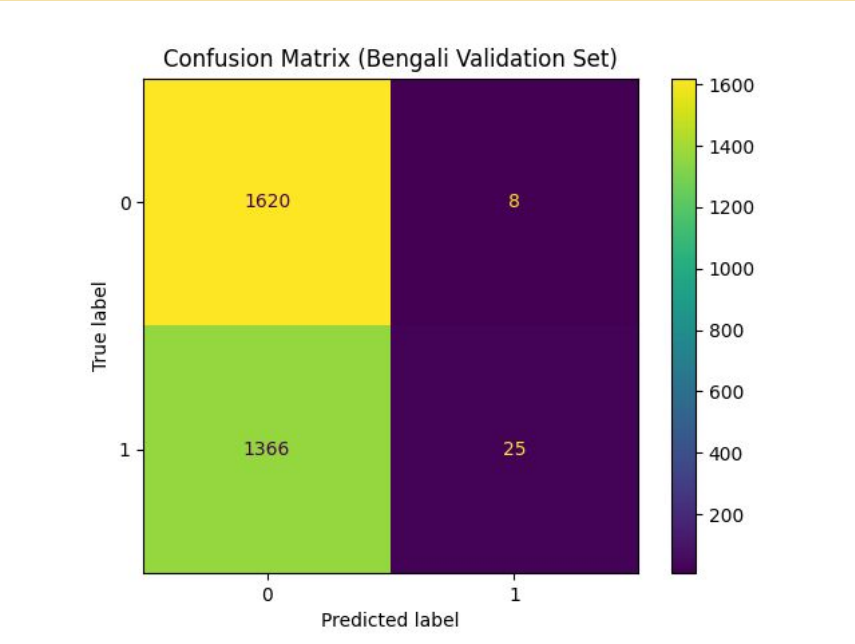


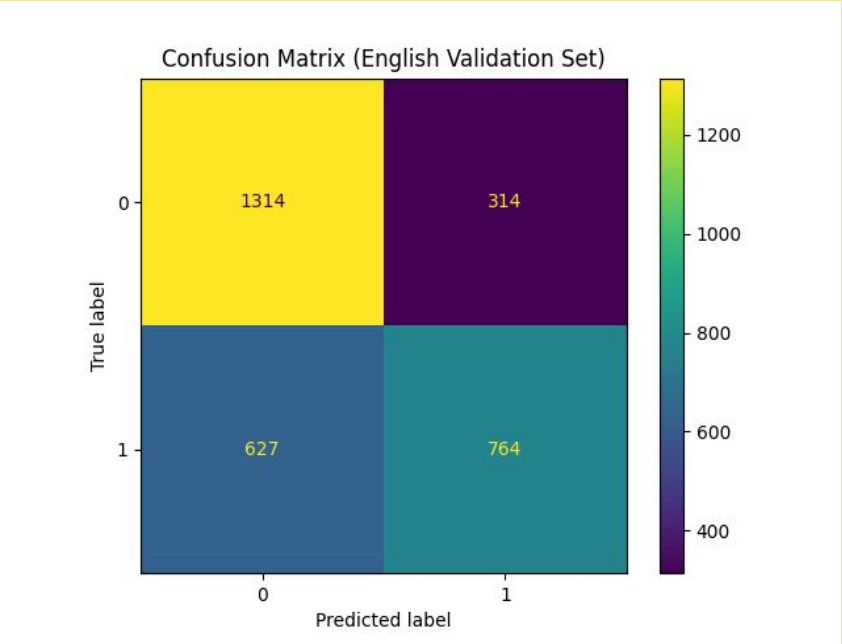


Loss and Accuracy Plots Bengali (Bloom 1.7B+QLoRA)

Tony Stark

Task-3

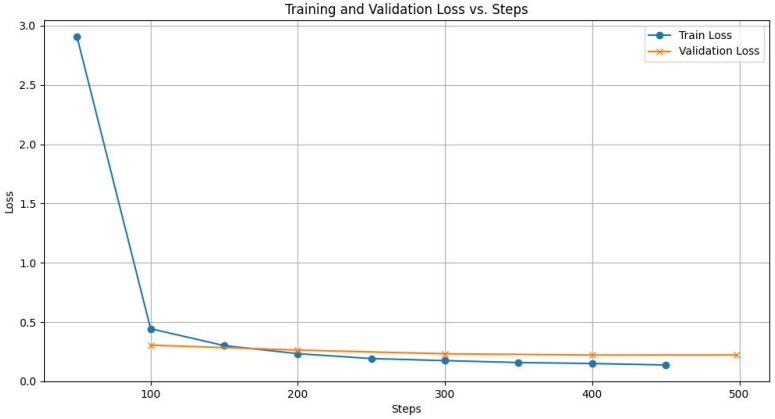


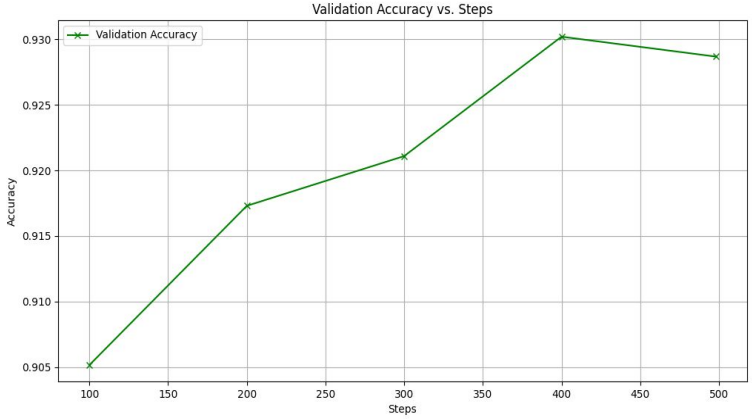


Confusion Matrix Bengali (Bloom 1.7B+QLoRA)

Tony Stark

Task-3

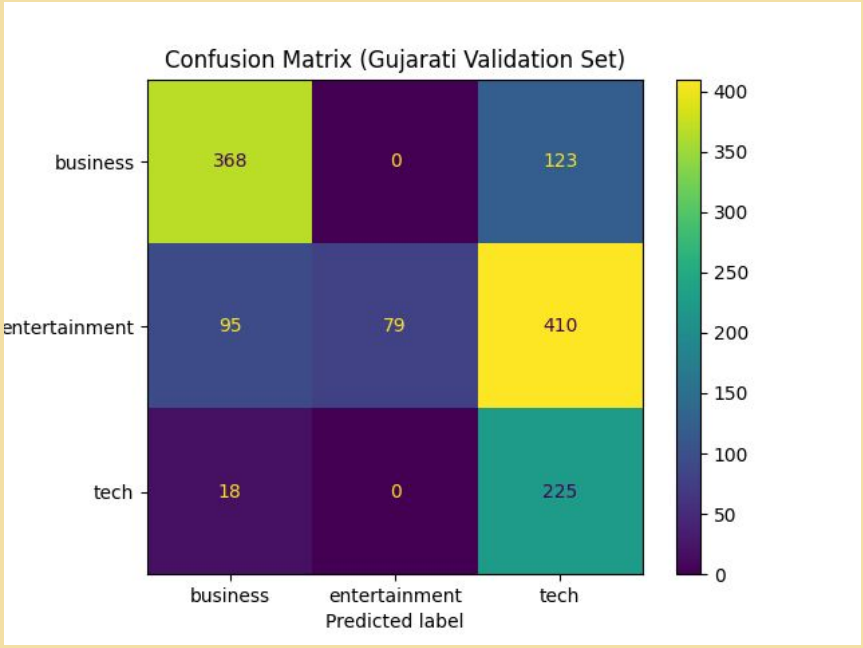


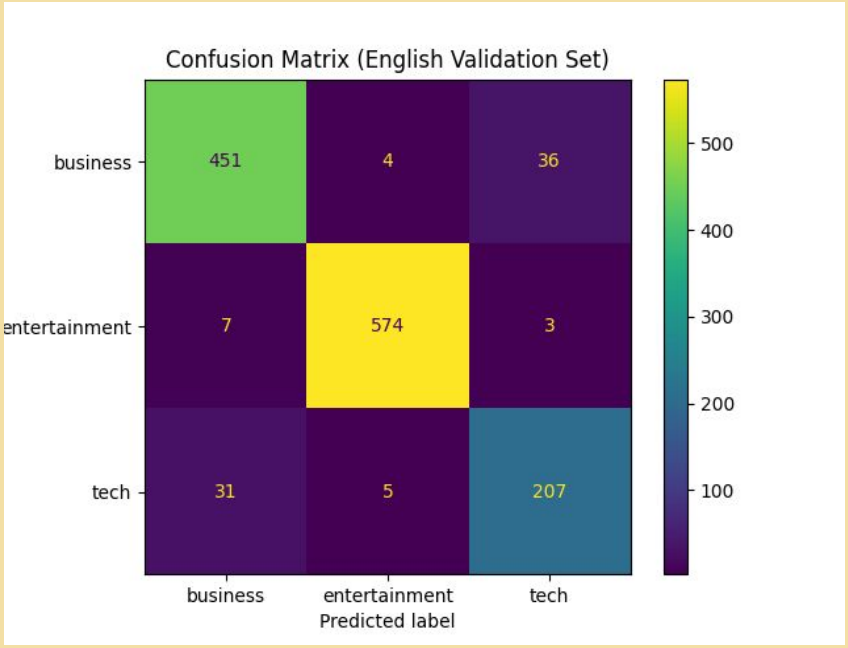


Loss and Accuracy Plots - Gujarati(Bloom 1.7B+QLoRA)

Tony Stark

Task-3

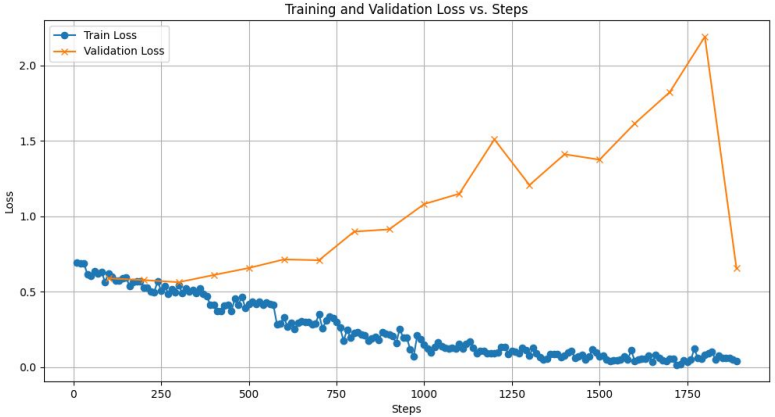


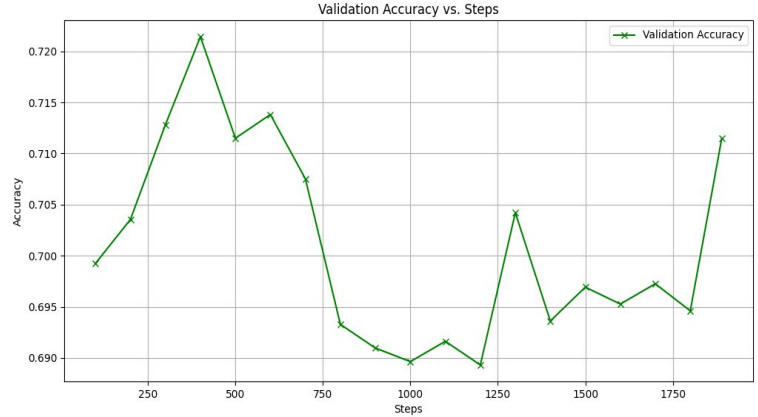


Confusion Matrix Bengali (Bloom 1.7B+QLoRA)

Tony Stark

Task-3

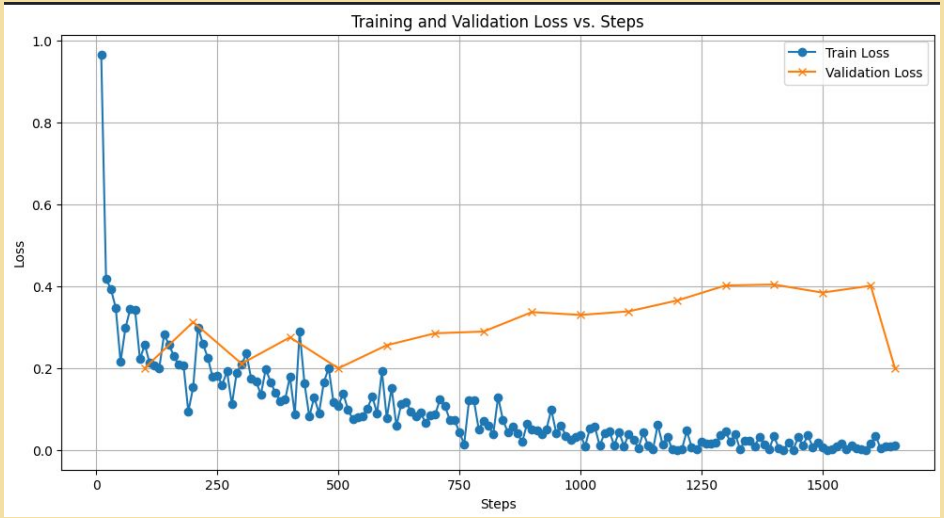


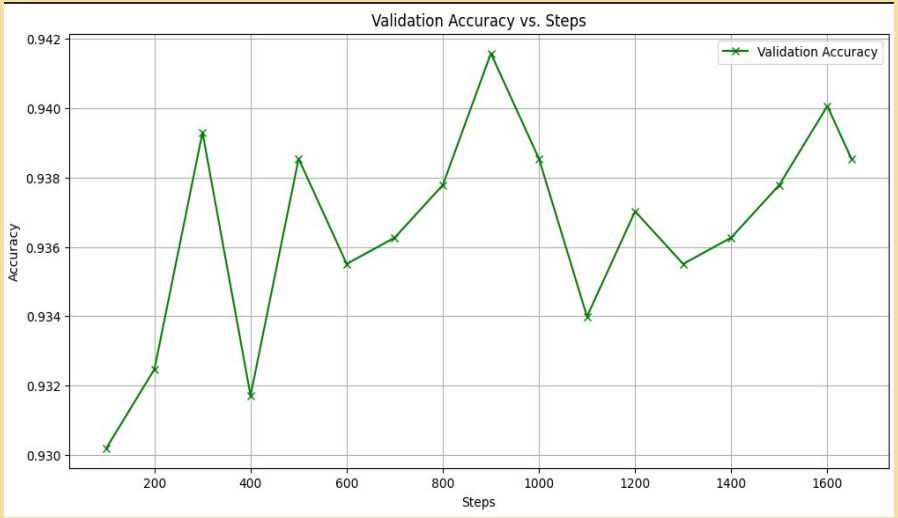


Loss and Accuracy Plots Bengali (RoBERTa)

Tony Stark

Task-3

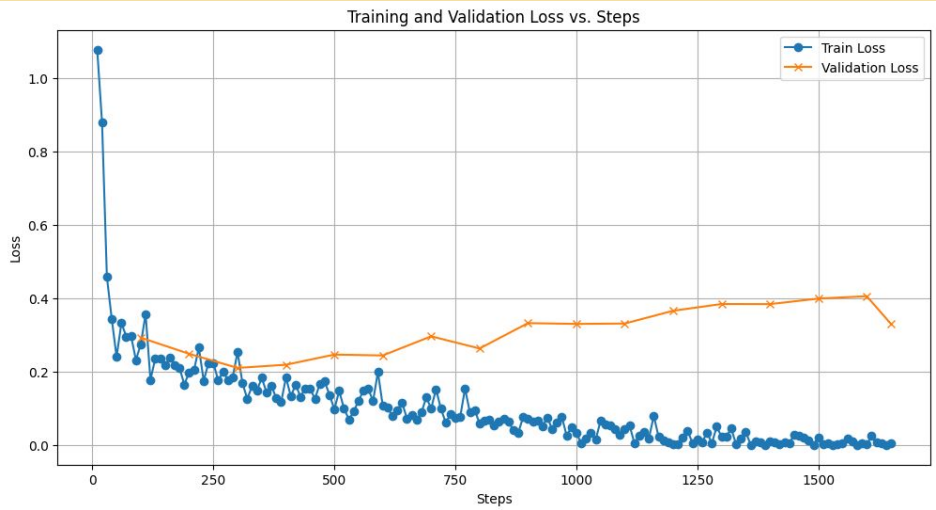


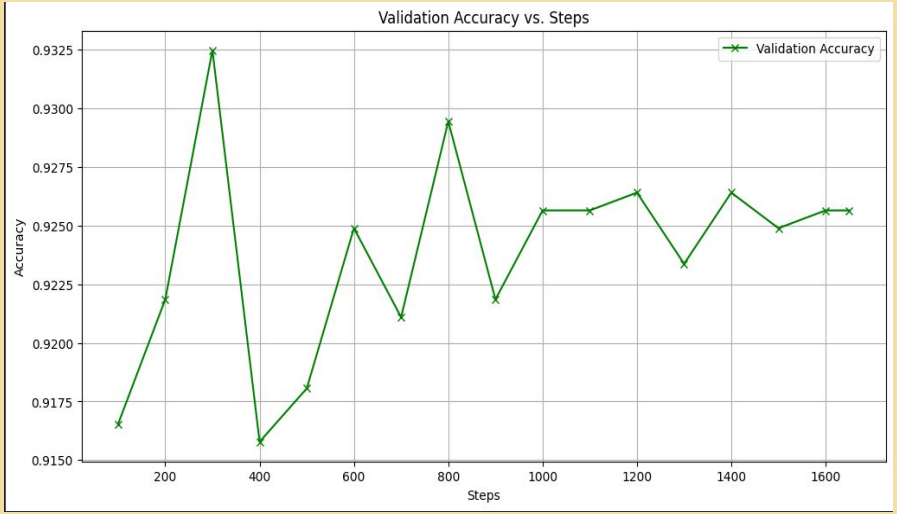


Loss and Accuracy Plots Gujarati (RoBERTa)

Tony Stark

Task-3

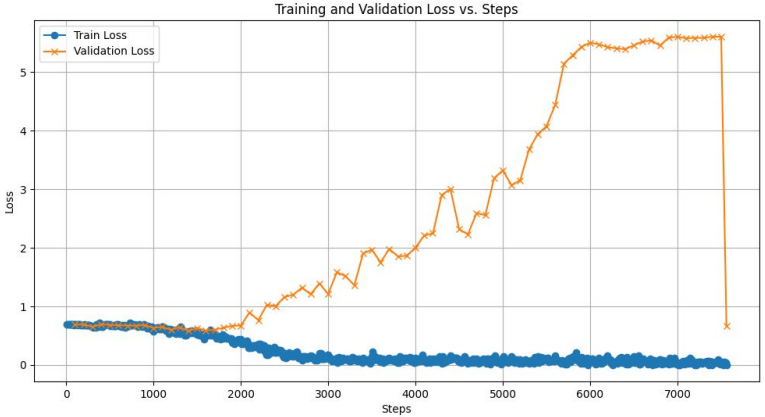


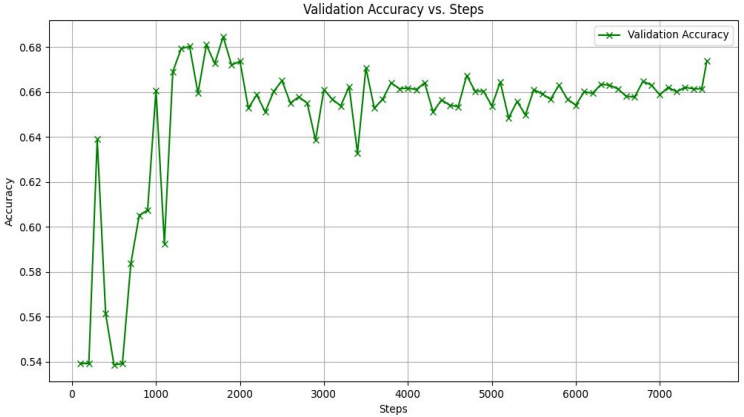


Loss and Accuracy Plots – Gujarati (IndicBERT)

Tony Stark

Task-3





Loss and Accuracy Plots Bengali (IndicBERT)

Tony Stark

Task-3

Model	Dataset	English Accuracy	Native Accuracy
RoBERTa	Bengali	0.9385	0.4355
IndicBERT	Bengali	0.9256	0.6851
BLOOM-1.7B	Bengali	0.6883	0.5449
RoBERTa	Gujarati	0.9385	0.4355
IndicBERT	Gujarati	0.9256	0.6851
BLOOM-1.7B	Gujarati	0.9347	0.5099

Table 1: Accuracy Results for Bengali and Gujarati Datasets [0-1]

Final Accuracy Results

Tony Stark

Task-3

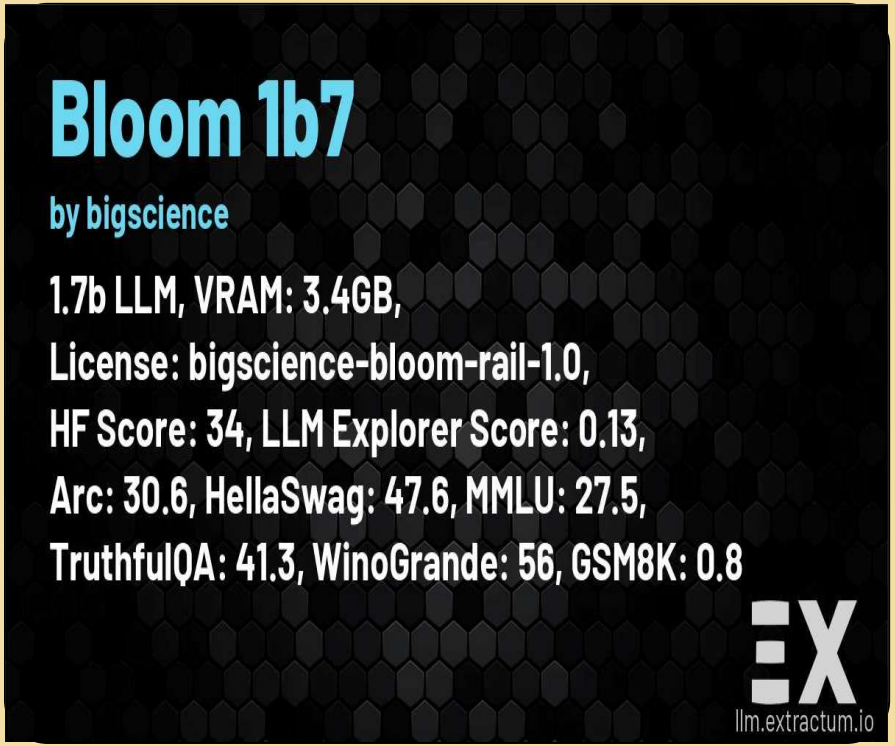
Findings: 

- **Training Data Bias**: Richer English corpora vs. low-resource Bengali and Gujarati datasets.
- **Language Complexity**: Morphological richness and distinct scripts (Bengali, Gujarati).
- **Model Pretraining**: BLOOM/RoBERTa favor high-resource languages; IndicBERT excels in Indic languages.
- **Cross-Lingual Limitations**: Structural differences hinder transferability.
- **Dataset Size**: Smaller fine-tuning datasets for Bengali/Gujarati.
- **Task Complexity**: Gujarati's 3-class task more challenging than Bengali's binary classification.

Future Scope and Improvements

Potential Areas of Improvements:

- In Task-2, my probing currently is giving same outputs whether I short circuit it or bypass it. Will try to look it.
- Also will try to visualize eigenvectors of attention maps so as to uncover more structural patterns.
- Will also train Bloom 1.7B model for more epochs if the val loss decreases more



Thank You
Shashwat Bhardwaj
2023AIY7528
Team: Tony Stark