Alif الف: An Urdu Generative Pre-trained Language Model

Ali Muhammad Asad, Hammad Sajid
Syeda Haya Fatima, Zainab Haider
Syed Muhammad Ali Naqvi

Supervisor: Dr. Abdul Samad

QR Code

FYP FINAL YEAR PROJECT SHOWCASE 2025

Introduction

Alif is a revolutionary pre-trained Urdu large language model.

Motivation

- Urdu is a *Low-Resource Language*
- Bias in LLMs due to Lack of Urdu Training Data
- Existing Models: High-Resource & Lack Security

Solution

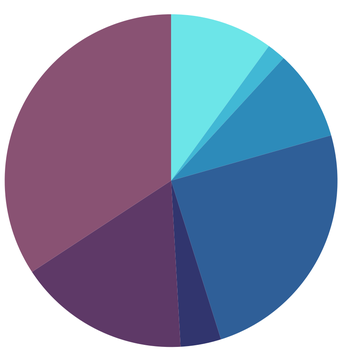
- Develop an SLM which is smaller, efficient and performs well with low-computational resources for Urdu Language.
- Incurs minimum costs to fine-tune and used for various use cases like banks, customer support, etc in Pakistan.
- Sets a foundational benchmark for LLMs in Urdu Language for future Urdu-NLP research.

Dataset Curation

Collection: Urdu text gathered from web articles (Selenium/Beautiful Soup), scanned books (Google Vision OCR), Existing Datasets and Translation (Google Translate API), covering diverse domains like news, blogs, and literature.

Cleaning & Deduplication: Removed non-Urdu content, normalized text (spacing, encoding, Urdu numerals), eliminated noise (links, emails, numbers), and used Jaccard Similarity with Min-Hashing for deduplication.

- **News Sites - 3.3GB**
- **Blog Sites - 0.6 GB**
- **Existing Datasets - 2.9 GB**
- **CC Dump 1 - 8.1 GB**
- **Books - 1.3 GB**
- **Translation - 5.5GB**
- **CC Dump 2 - 11.3 GB**



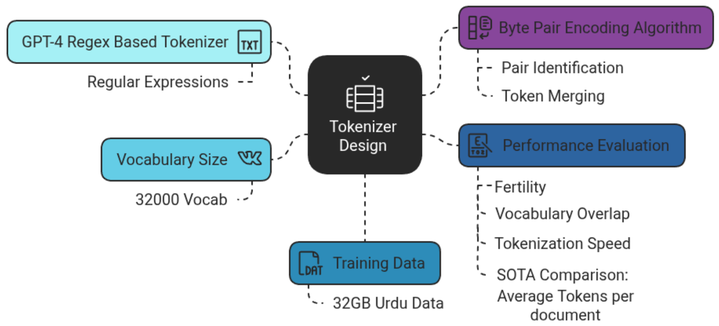
Final ALIF Pre-training Dataset:

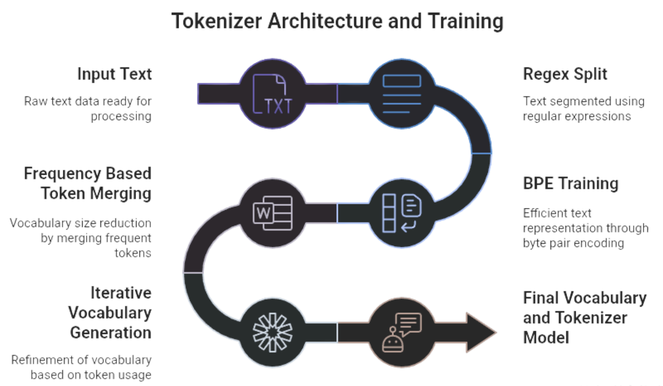
Size: **33GB**

Number of Rows:**~13 Million**

Number of Tokens: **~5-6 Billion**

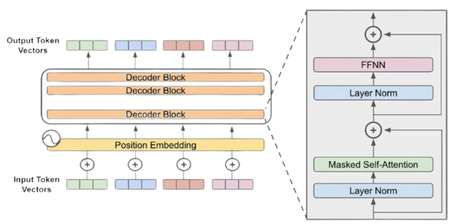
Tokenizer Architecture & Training

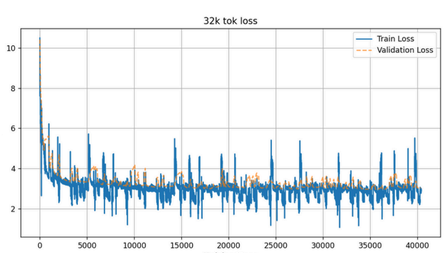




Model Architecture & Training

	Alif-Small	Alif-Large
Model Size	116M	1B
Block Size	1024	4096
Epochs	5	5
N. Heads	12	24
N.Layers	12	24

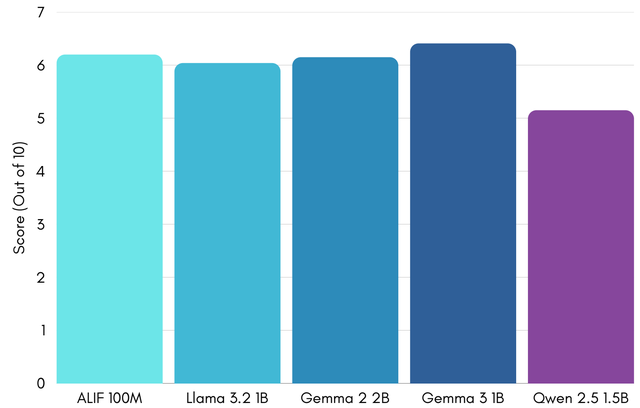




Evaluation

LLM-as-a-judge was used to score text generation on coherence, fluency, and relevance against SOTA models

- Our model demonstrated comparable performance to SOTA models 10× its size.



Results

- The model performance indicates that monolingual Urdu model surpasses Multilingual Models
- The model has inherent understanding of the language and can be finetuned on urdu downstream tasks with minimal fine-tuning
- In Future, ALIF can be scaled to (8B, 16B) for commercial use cases.