Alif الف: A Pre-trained Urdu Generative Language Model

Ali Muhammad Asad[1], Hammad Sajid[1]
Syeda Haya Fatima[1], Zainab Haider[1]
Syed Muhammad Ali Naqvi[1]
Supervisors: Dr. Abdul Samad[1], Dr. Inaya Ullah[*,2]
[1] Habib University, [2] BUITEMS, [*]External Support

Introduction

- **High Cost & Energy Use**: SOTA models are large, expensive, and carbon-intensive.
- **Language Bias**: Optimized for English, underperform on low-resource languages like Urdu.
- **Accessibility Barriers**: Proprietary models require powerful infrastructure and raise privacy issues.
- **Lack of Urdu Resources**: Few benchmarks, datasets, or pretraining efforts exist.

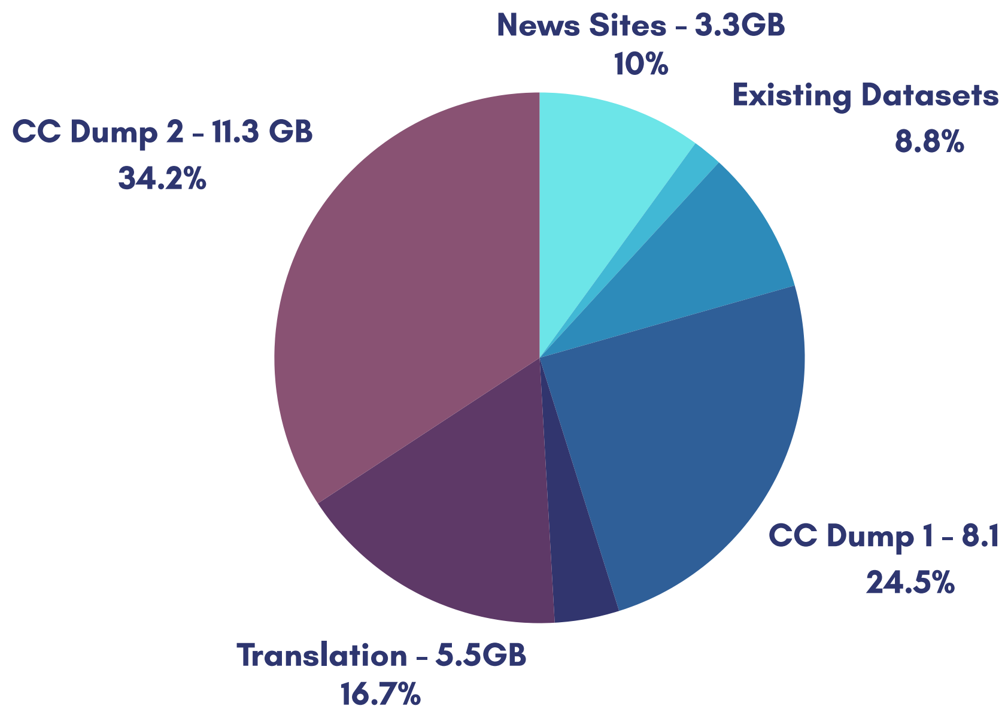
Objective

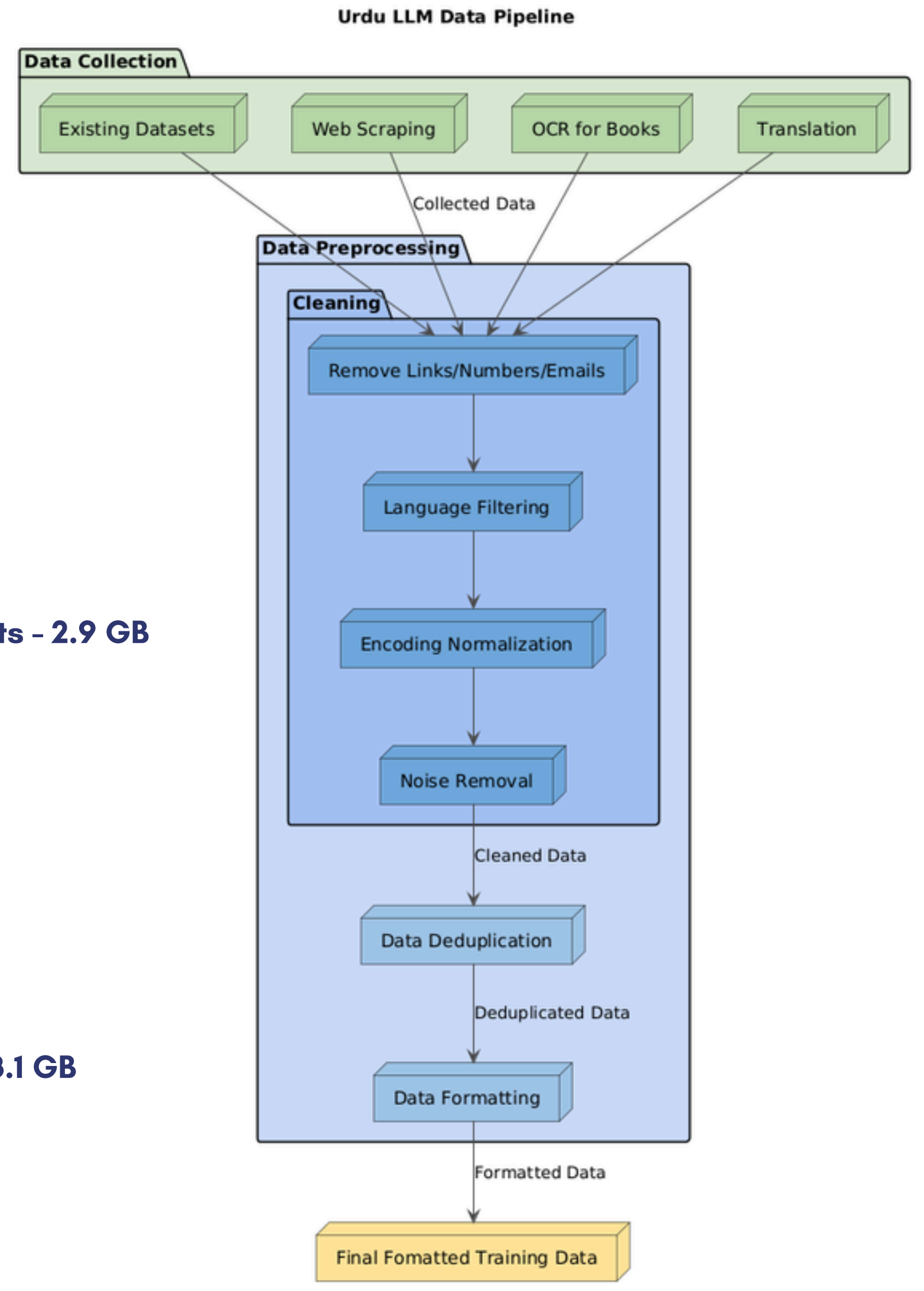
To develop a compact, efficient Urdu Foundation Model that performance well on low-resource infrastructure, enables real-world applications in Pakistan, and establishes benchmarks for future Urdu NLP research.

Pretraining Dataset

ALIF Pre-training Dataset:

- **33GB**
- **~13 Million Rows**
- **~5-6 Billion Tokens**

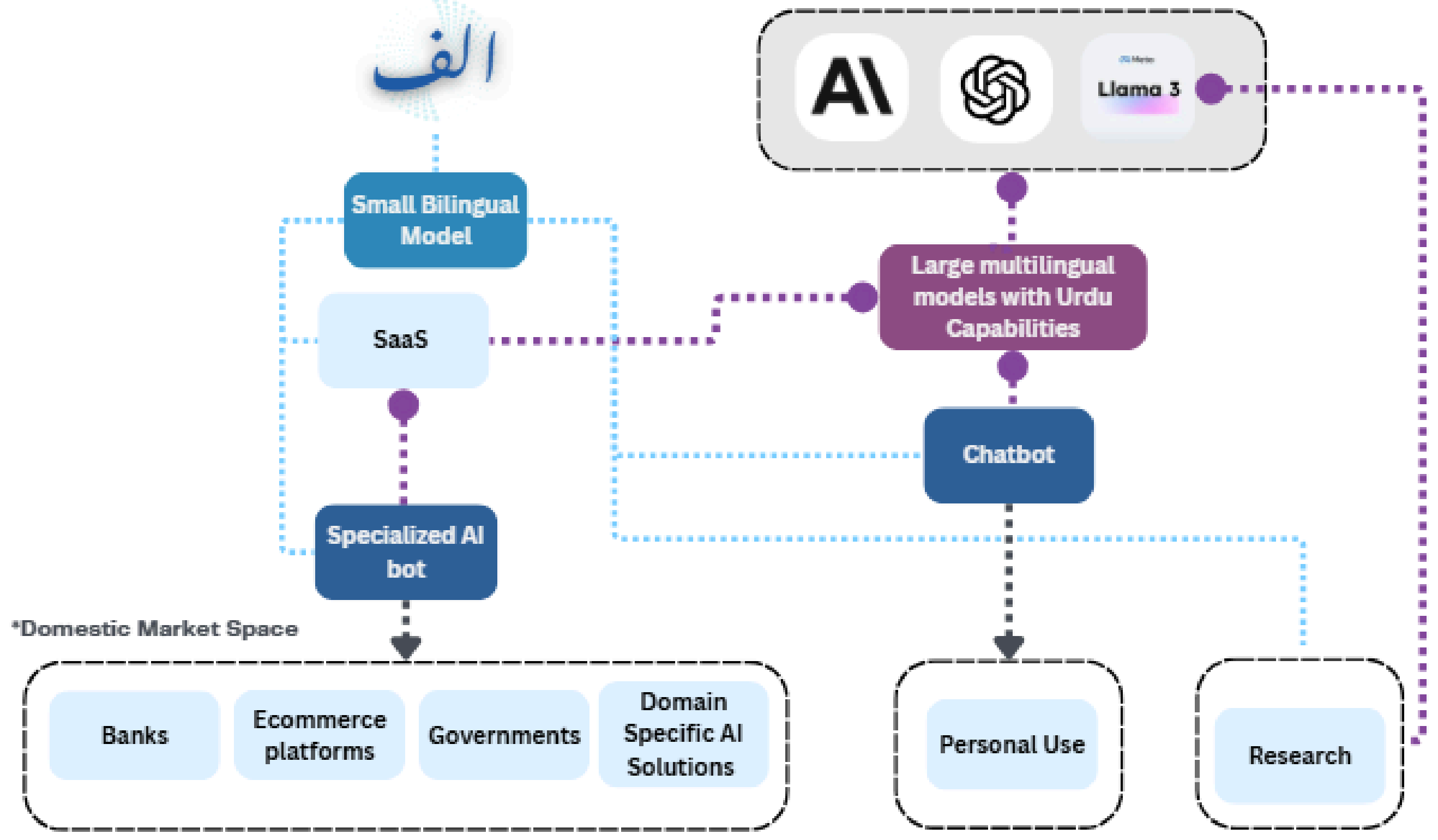




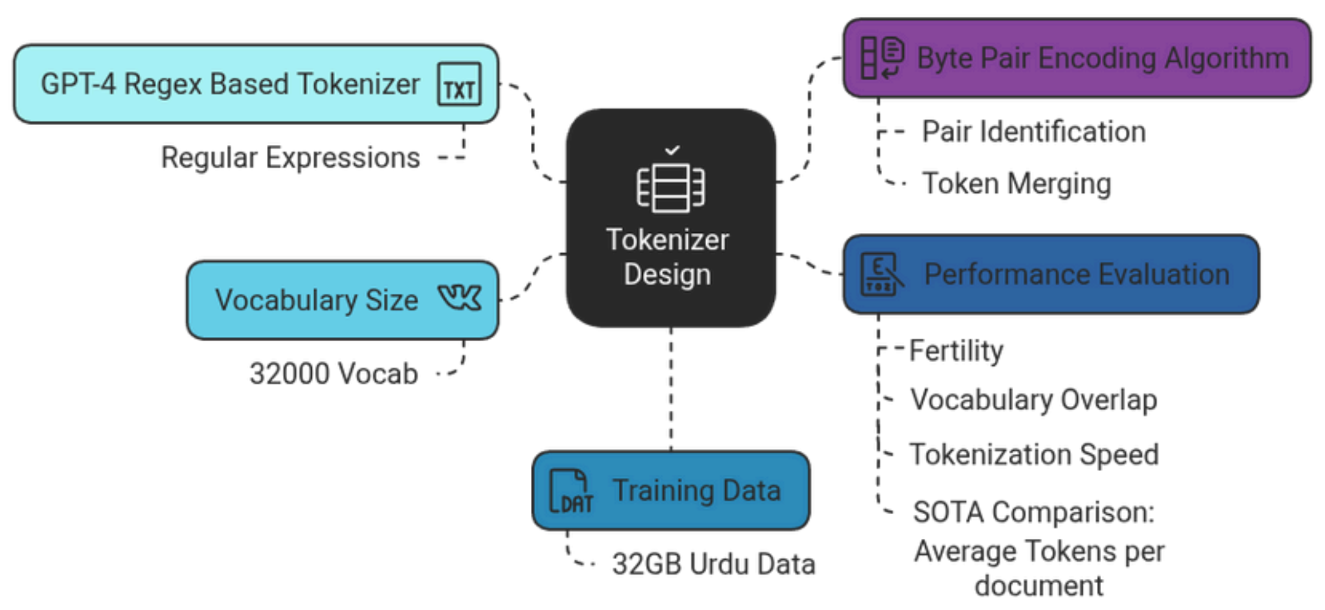
Instruct Dataset

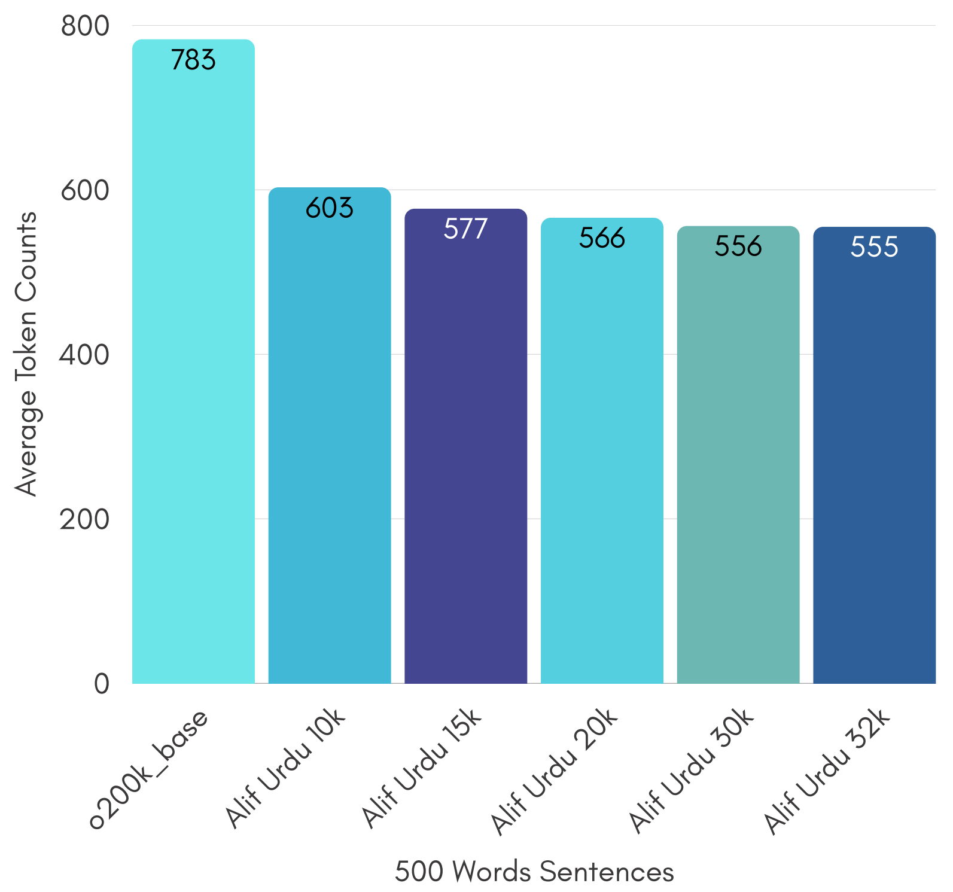
Dataset	Rows	Description
UrduDolly	15k	Logic-heavy, culturally adapted
UrduAlpaca	50K	General tasks (QA, summaries)
UrduStanford	85K	Reading comprehension
Literary QnA	10K	Poetry, grammar, literature

Market Fit & Deployment



Tokenizer Architecture & Training





ALIF Model Series

Alif 100M
Base Pretrained Model
Compact Model for research and low resource environments

Alif 1B
Base Pretrained Model
Standard model with enhanced generation capabilities

Alif Instruct
Instruct-tuned model
Optimized for dialogue and following instructions

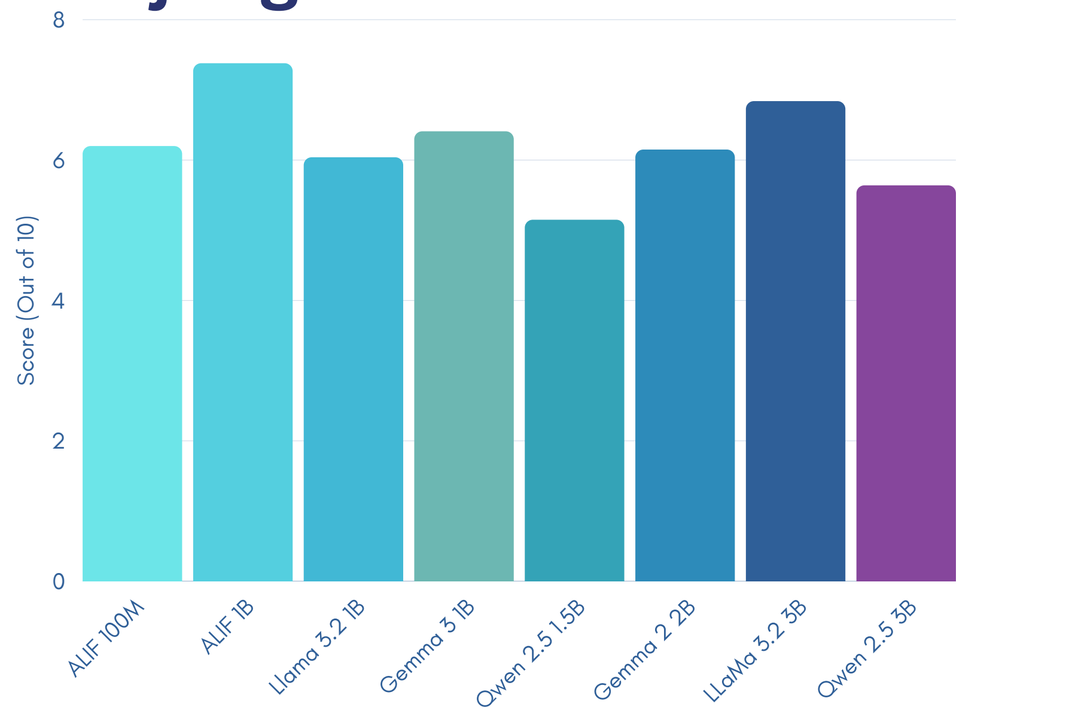
	Alif-Small	Alif-Large
Model Size	116M	1B
Block Size	1024	4096
Emb Dim	768	3072
N. Heads	12	24
N.Layers	12	24

Evaluation

ALIF outperforms SOTA multilingual models 3× to 10× its size.

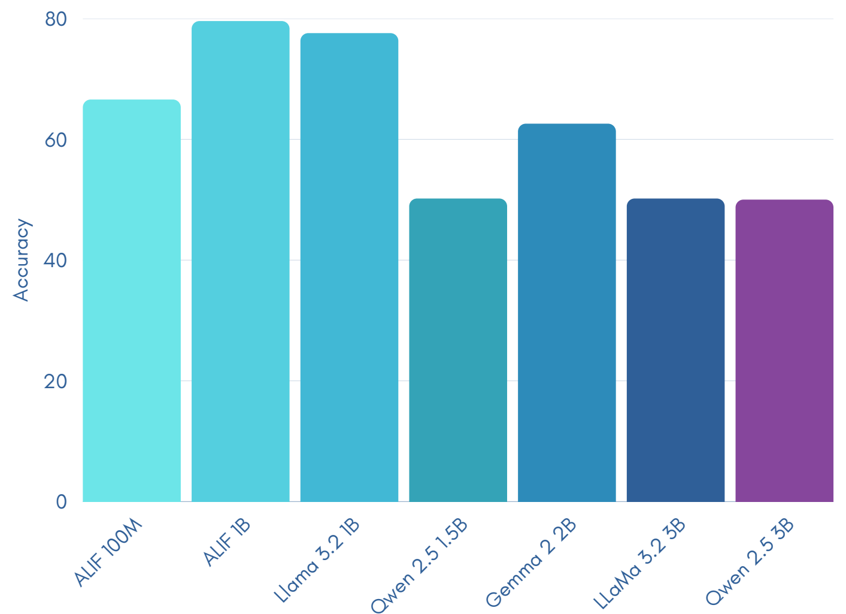
LLM-as-a-judge

- **Criteria:**
  - Cohesion
  - Fluency
  - Relevance
- Evaluator: **Gemini 2.0-Flash**

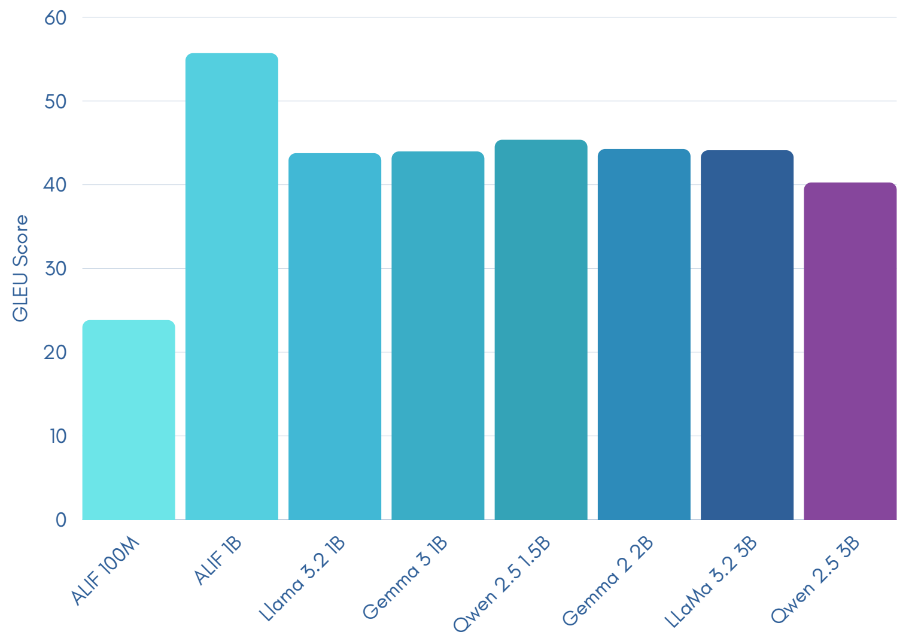


Text Generation

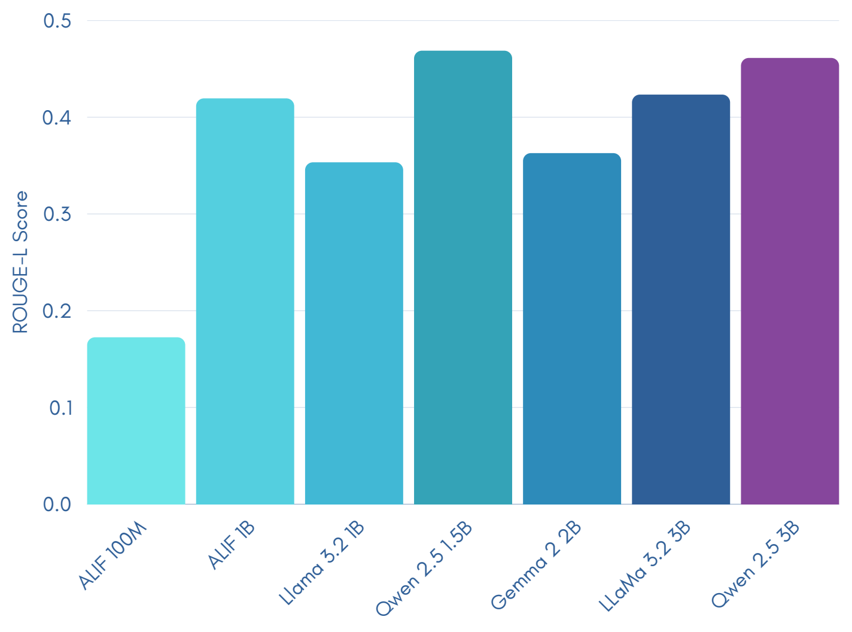
Few Shot Testing



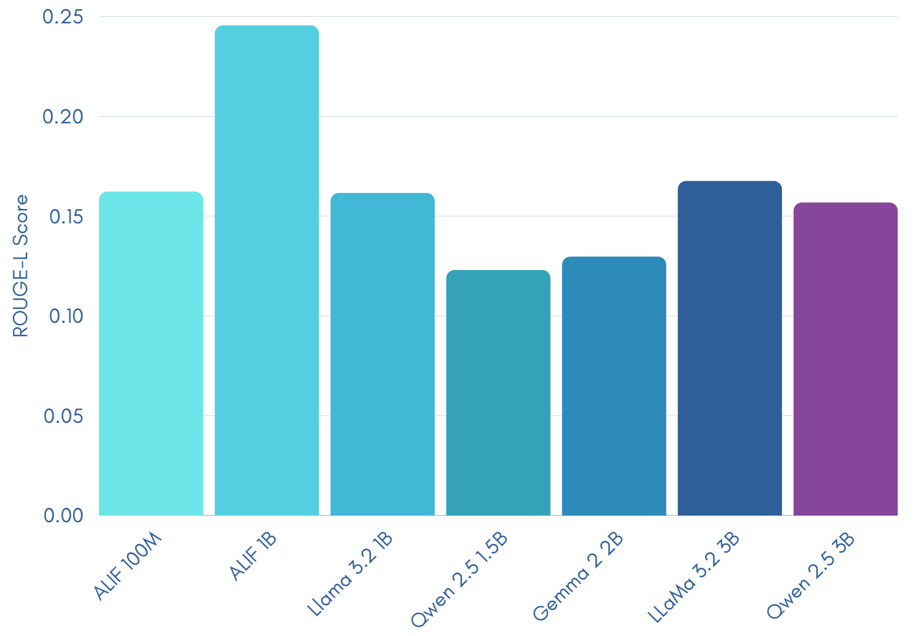
Sentiment Classification



Grammar Error Correction



QA With Context



QA Without Context

Cost Comparison

1GB Finetuning Urdu-Eng Dataset

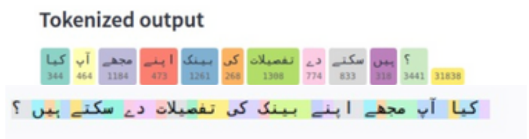
Model	No of Train Tokens	Training Computational Cost (PFLOPS)	Time (h)	Energy Consumption (KWh)	Carbon Emission (kgCO2e/kWh)	Train Cost (PKR)
Alif 1B Instruct	~180 Million	1080	2.678	1.606	1.48	56.21
Llama 3.2 3B	~504 Million	9720	23.17	13.9	12.81	486.5

Inferencing

Input:

کیاآپ مجھے اپنے بینک کی تفصیلات دے سکتے ہیں؟

Tokenized output



Model	No of Inferencing Tokens	Training Computational Cost (TFLOPS)	Time (h)	Energy Consumption (KWh)	Carbon Emission (kgCO2e/kWh)	Train Cost (PKR)
Alif 1B Instruct	12	0.09	1.47 e-6	8.82 e-7	8.13 e-7	3.085 e-5
Llama 3.2 3B	33	0.594	8e-4	2.4 e-4	2.2128 e-4	8.4 e-3