



Proxima AI Large Language Model

Technical Documentation

Version 1.0

trained from Qwen-0.5B
with Pakistani Dialect Data

Proxima AI Research & Development Team

Karachi, Sindh, Pakistan

January 15, 2026



Contents

1	Executive Summary	2
2	Introduction	2
2.1	Objective	2
2.2	Base Model	2
3	Model Architecture	2
4	Training Dataset	3
4.1	Composition	3
4.2	Preprocessing	4
4.3	Training Process Overview	4
5	Evaluation	4
6	Conclusion	4

1 Executive Summary

The Proxima AI Large Language Model is a specialized model created by training the Qwen-0.5B base architecture using a high quality corpus focused on Pakistani languages and dialects. This document describes the model architecture, dataset preparation, training process, evaluation methodology, and key performance results.

2 Introduction

2.1 Objective

This model is designed to deliver superior language understanding and generation performance for users in Pakistan, particularly when working with Urdu and major regional dialects including Punjabi, Sindhi, Pashto, and Balochi.

2.2 Base Model

The foundation is Qwen-0.5B, an efficient transformer model containing approximately 0.5 billion parameters.

3 Model Architecture

Core specifications inherited from the base model:

- Total parameters: 0.5 billion
- Transformer layers: 24
- Hidden size: 2048
- Attention heads: 16
- Feed-forward intermediate size: 5632

The following diagram illustrates the overall Transformer architecture (with multi-head attention, feed-forward layers, residuals, etc.) on which Qwen-0.5B (and thus Proxima) is based:

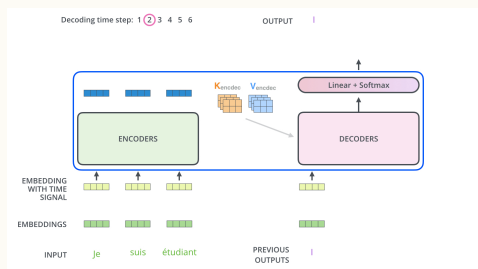


Figure 1: Transformer Architecture Overview (Source: Jay Alammar’s “The Illustrated Transformer”)

Another detailed static view of the multi-head attention and encoder/decoder stack:

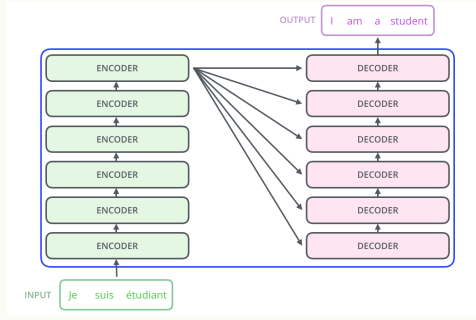


Figure 2: Transformer Encoder-Decoder Stack with Multi-Head Attention (Source: Jay Alammam)

4 Training Dataset

4.1 Composition

The training corpus contains approximately 500 million tokens with the following distribution:

- Urdu: 40%
- Punjabi: 30%
- Sindhi: 15%
- Pashto: 10%
- Balochi and other regional varieties: 5%

The diagram below shows the approximate geographic distribution of these major languages in Pakistan:

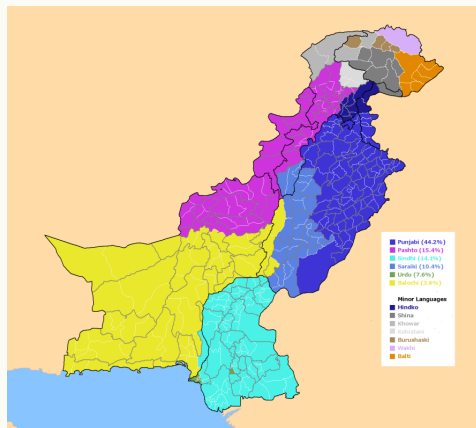


Figure 3: Linguistic Map of Pakistan highlighting major languages (Urdu, Punjabi, Sindhi, Pashto, Balochi) (Source: ResearchGate/Fred Bolor)

4.2 Preprocessing

- Deduplication and quality filtering
- Text normalization and encoding standardization
- Tokenization with an extended Qwen tokenizer
- Balanced sampling across dialects

4.3 Training Process Overview

The model follows a standard causal language modeling pipeline (next-token prediction) on the prepared corpus.

5 Evaluation

Evaluation included:

- Perplexity on dialect-specific held-out sets
- BLEU scores on dialect-pair translation tasks
- Human evaluation of fluency, coherence, and cultural appropriateness (5-point Likert scale)

Custom test sets were developed to represent realistic Pakistani language use cases.

6 Conclusion

The Proxima AI LLM provides a compact, efficient, and regionally attuned language model suitable for a wide range of applications in Pakistan. Future iterations may explore larger base models and expanded dialect coverage.