# Fine-Tuned Model Overview

## Model Purpose

Fine-tuning a language model on Egyptian Arabic text for improved understanding and generation in dialectal Arabic.

## Dataset Details

- **Dataset Name**: oscar/unshuffled\_deduplicated\_arz  
- **Language**: Arabic (Egyptian dialect)  
- **Total Samples**: 71,940  
- **Train/Test Split**:

- **Training Set**: 57,552 samples (80%)  
 - **Test Set**: 14,388 samples (20%)

## Model Training Configuration

- **Base Model**: aragpt2-base  
- **Tokenizer**: GPT2Tokenizer  
- **Epochs**: 3  
- **Batch Size**: 4  
- **Learning Rate**: 5e-5  
- **Max Sequence Length**: 128 tokens  
- **Optimizer**: AdamW  
- **Framework**: Hugging Face Transformers with PyTorch backend

## Model Limitations

- **Dialect Specificity**: Focuses on Egyptian Arabic; may not generalize well to other Arabic dialects or MSA.

- **Informal Language Bias**: Dataset sourced from web text; may contain informal or noisy data.

- **Resource Intensive**: Fine-tuning large models like GPT-2 requires significant computational power.  
- **Tokenization Limits**: GPT-2 tokenizer may not perfectly handle complex Arabic morphology and orthographic variations.