
# Project Overview: Neural Machine Translation — Arabic to English

Build and fine-tune a Transformer model translating Arabic text into English. Utilizing Hugging Face tools with deployment via a Gradio GUI.

# Model Architecture: MarianMT

## Architecture

Based on Transformer encoder-decoder architecture.

Encoder uses multi-layer self-attention for Arabic input.

## Pretrained Model

Starts with Helsinki-NLP/opus-mt-ar-en checkpoint.

Decoder generates English output autoregressively.

# Dataset: OPUS Project

## OPUS Corpus

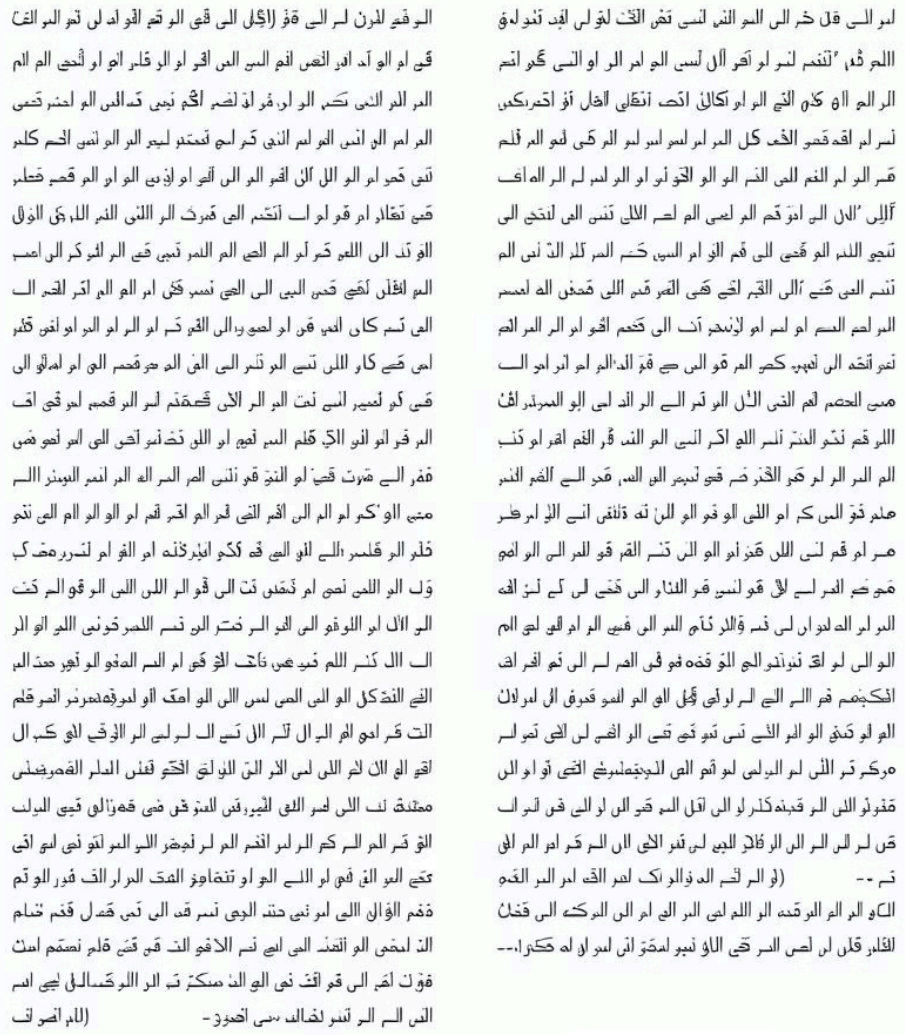
Large multilingual corpus covering diverse domains.

## Arabic-English Subset

Contains millions of parallel sentence pairs for training.

## Data Quality

Carefully curated to ensure alignment and relevance.



# Data Preprocessing

### Tokenization

Uses SentencePiece or BPE tokenizer for consistency.

### Cleaning

Removes noise and handles special characters.

### Vocabulary

Builds separate vocabularies for Arabic and English.

### Padding & Truncating

Ensures uniform sequence lengths for efficient batching.

# Training Setup

## Frameworks

Uses PyTorch or TensorFlow with Transformers library.

## Hardware

Utilizes NVIDIA Tesla V100 GPUs for acceleration.

## Optimization

Adam optimizer applied with cross-entropy loss function.

# Fine-Tuning Details

## 1 Dataset

Fine-tunes on ar-en parallel corpus.

## 2 Batch Size

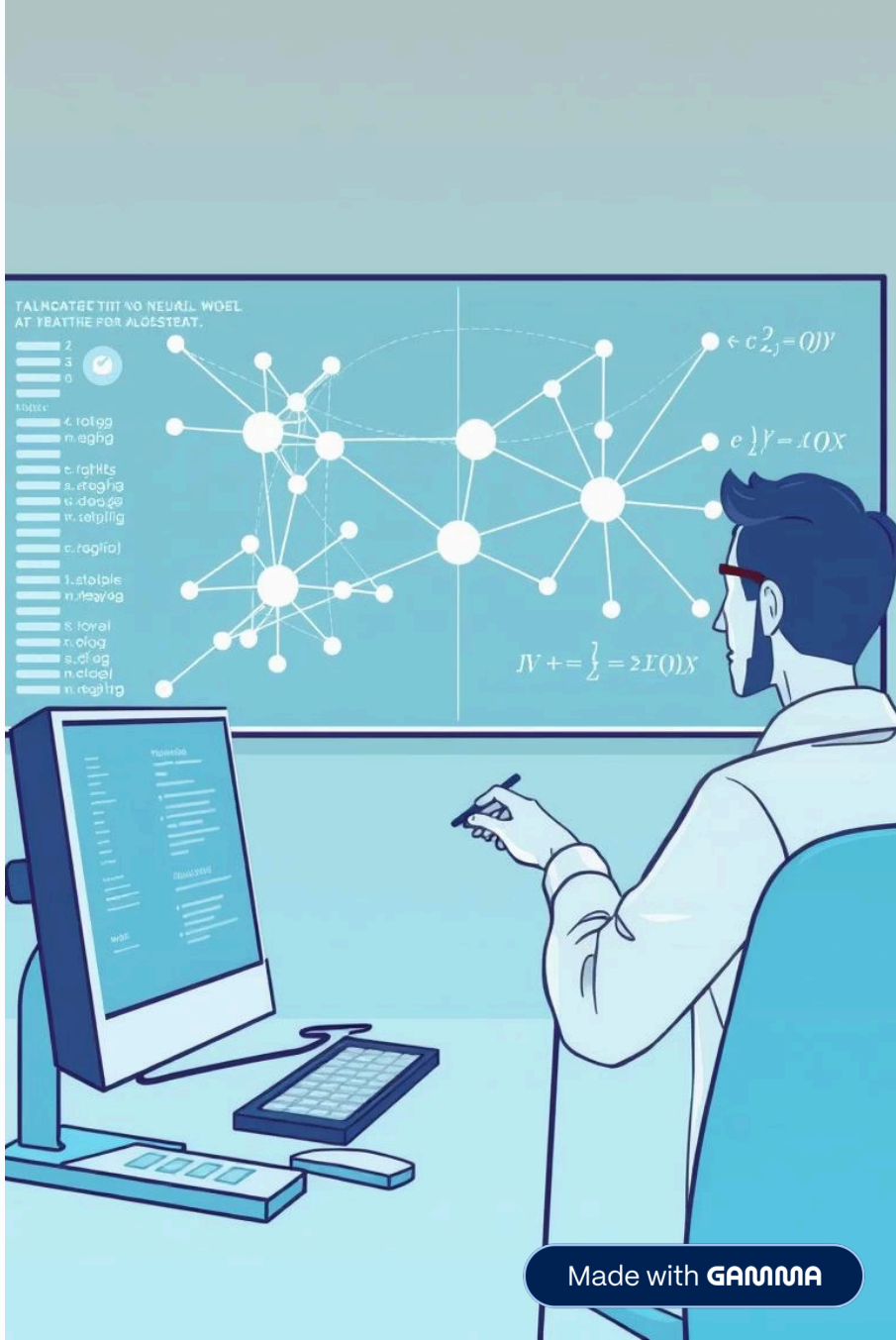
Tuned between 32 and 64 based on GPU memory.

## 3 Learning Rate

Adjusted using a scheduler for optimal convergence.

## 4 Epochs

Training runs for 10-20 epochs to ensure performance.




# Evaluation Metrics

## BLEU Score

Primary metric measuring translation accuracy.

## Perplexity

Assesses model uncertainty on test data.

## Human Evaluation

Qualitative assessment of translation quality.

# Results and Analysis

## Performance

Achieved BLEU score approx. 35 on test set.

## Successes & Failures

Shows examples of good and problematic translations.

## Error Analysis

Identifies common mistakes and improvement areas.


# Deployment with Gradio

## Interface

Arabic input, English output.

## Accessibility

Easy testing and demo.

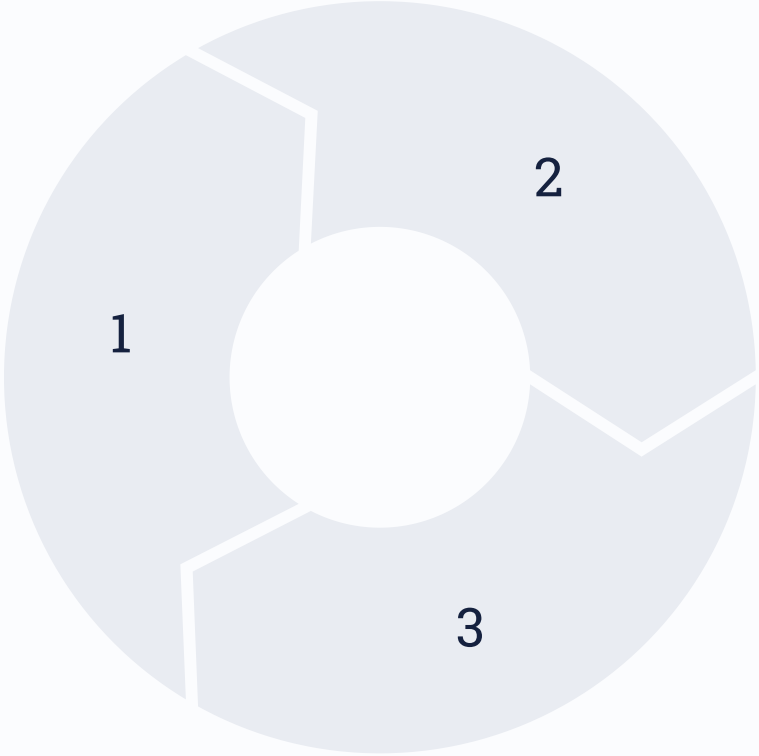
## Code

Quick Gradio setup.

# Conclusion and Future Work

## Summary

Successfully built and fine-tuned Arabic-English translation model.




## Improvements

Plan to apply data augmentation and back-translation.

## Future Directions

Explore advanced Transformer models and multilingual translation.