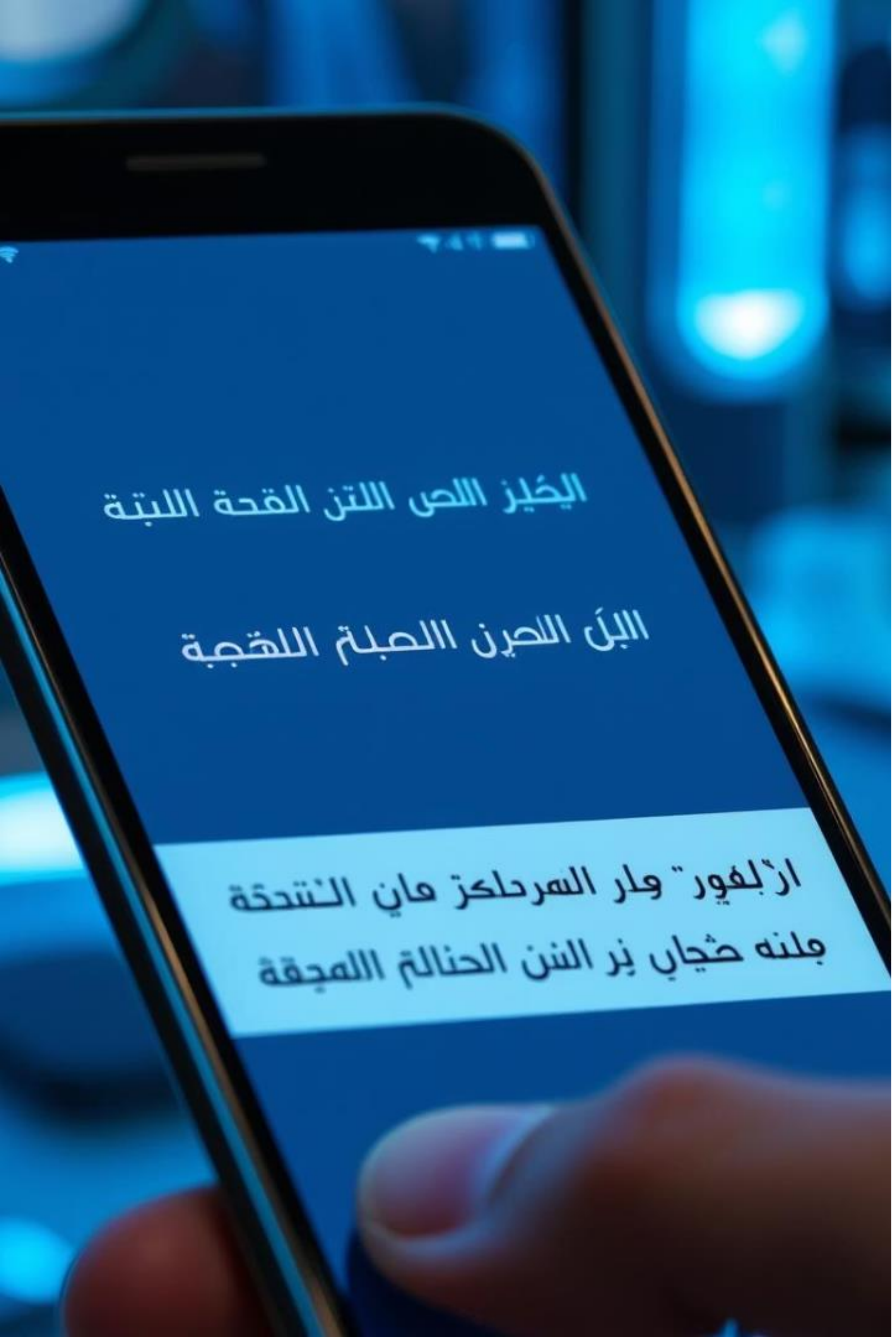

الخيز اللص اللتن القحة اللبتة
البَل اللحزن المبلة اللهجة
ازْلفور" ولر المرحلكز ماي النتحقة
جلنه حجاي بز النن الحنالك المجقة

Arabic Autocomplete system with Transformers

NLP project presentation

Understanding Autocomplete & Arabic Language

Autocomplete Purpose

Predict next word or phrase while typing

Enhances typing speed and accuracy

Why Focus on Arabic?

Rich morphological structure

Limited existing autocomplete solutions

Project Objectives

Model Development

Build Arabic autocomplete language model

Fine-Tuning

Use transformer-based models for better accuracy

Evaluation

Measure performance with NLP metrics

Dataset & Preprocessing

Data Source

Dataset Summary: **OSCAR (Arabic Subset**)

Source: Extracted from Common Crawl and classified by language (Arabic).

Size Used: 1% of the full Arabic dataset for faster training.

Content: Raw Arabic web data (news, blogs, forums); includes Modern Standard Arabic and dialects.

Preprocessing: Deduplicated and unshuffled; truncated to 128 tokens during tokenization.

Strengths: Large-scale, diverse, suitable for general Arabic language modeling.

Preprocessing Steps

- Tokenization using Hugging Face tokenizer
- Text cleaning and normalization

Model Architecture

- Pretrained GPT-2 Model
- Fine-Tuned on Arabic Data
- Trainer API for Training

Parameters and Training Setup

Parameters

Trainer API from Hugging Face Transformers.Training Arguments:

- Batch size: 4
- Learning rate: 5e-5
- Epochs: 2
- Weight decay: 0.01
- Save strategy: Every 500 steps
- Logging: every 100 step

Optimization

AdamW optimizer with learning rate scheduler

Tracking

Weights & Biases (wandb.ai) for experiment monitoring

Hardware

Google Colab GPU (14GB)

Evaluation Metrics

0.5

eval loss

87.3%

eval Accuracy

5.52

Perplexity

Perplexity is the preferred performance measure for generative tasks.

Challenges Encountered

Limited GPU Memory

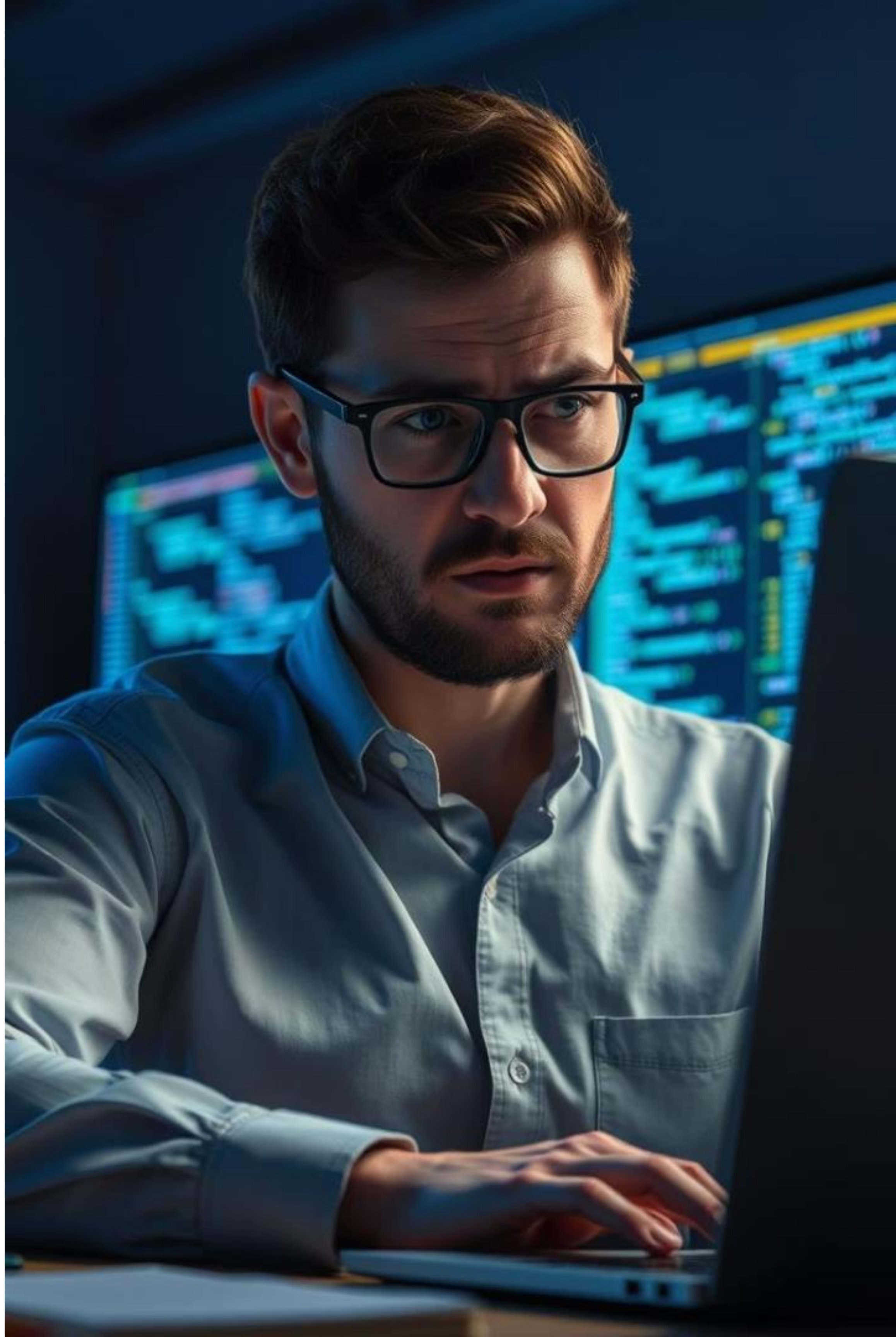
Constrains model size and batch processing

Model Complexity

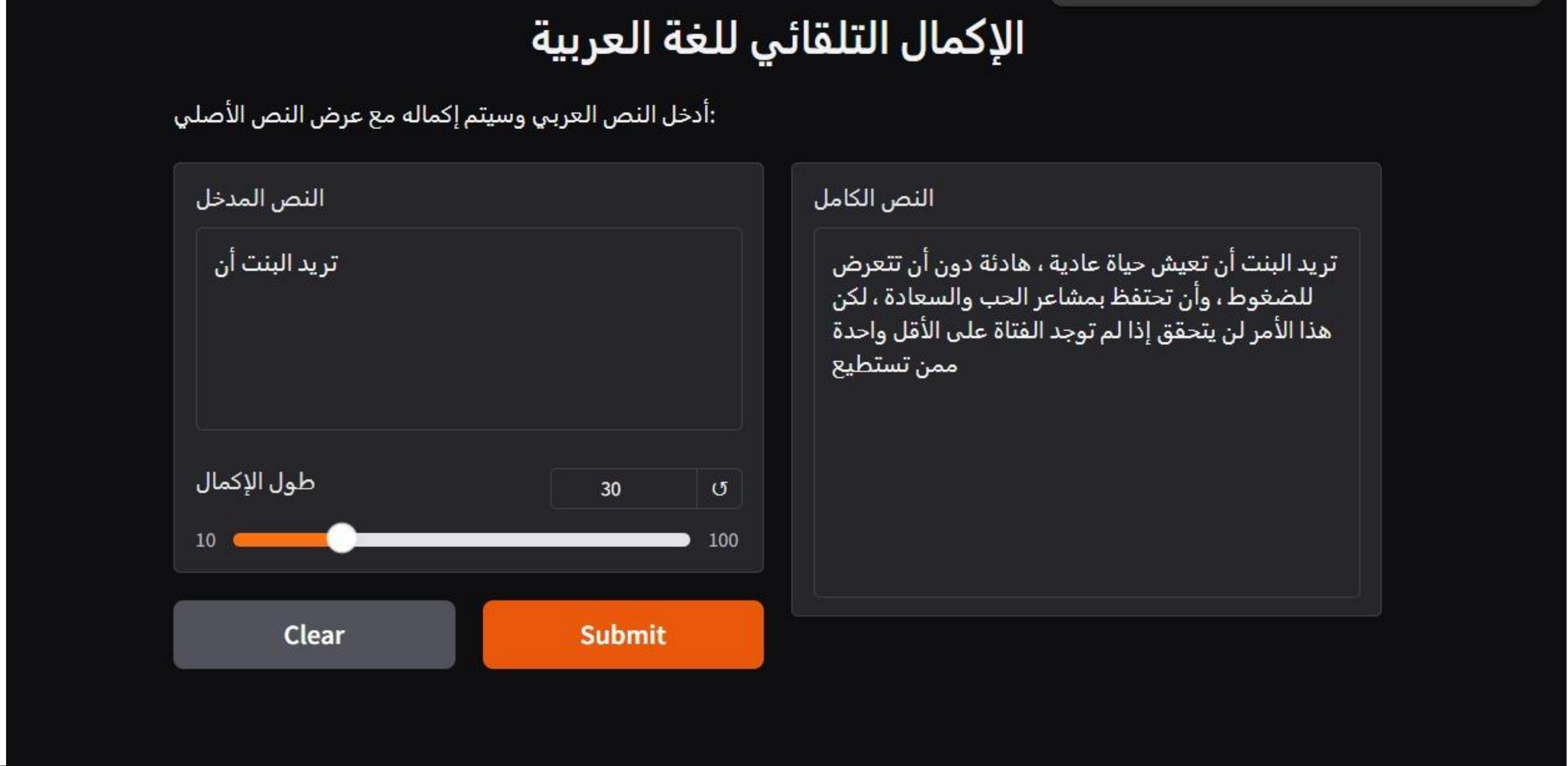
Training and evaluation require significant resources

Evaluation Limitations

Token accuracy insufficient for overall performance



Demo Examples



الإكمال التلقائي للغة العربية

أدخل النص العربي وسيتم إكماله مع عرض النص الأصلي:

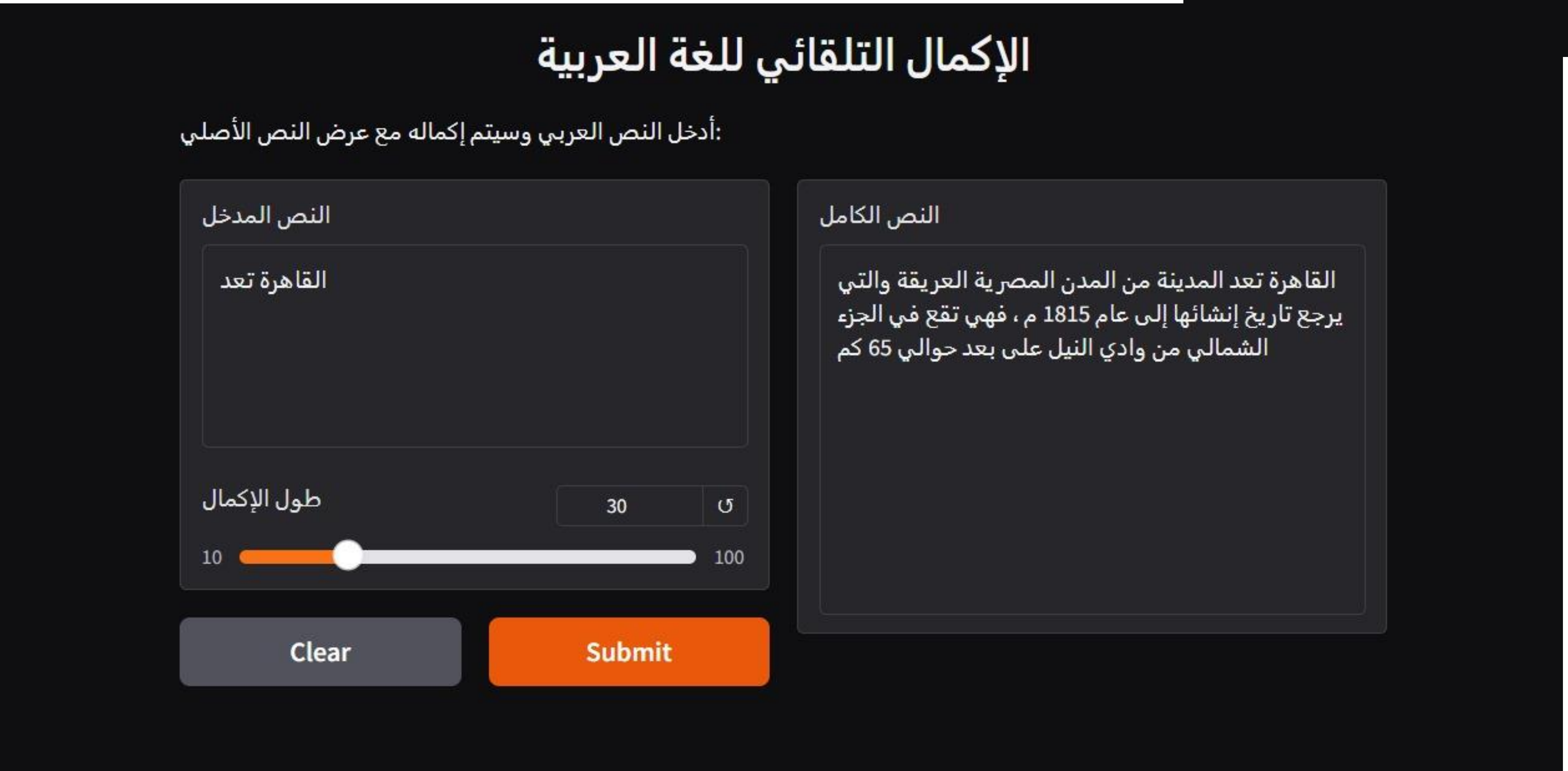
النص المدخل

تريد البنت أن

طول الإكمال: 30

النص الكامل

تريد البنت أن تعيش حياة عادية ، هادئة دون أن تتعرض للضغوط ، وأن تحتفظ بمشاعر الحب والسعادة ، لكن هذا الأمر لن يتحقق إذا لم توجد الفتاة على الأقل واحدة ممن تستطيع



الإكمال التلقائي للغة العربية

أدخل النص العربي وسيتم إكماله مع عرض النص الأصلي:

النص المدخل

القاهرة تعد

طول الإكمال: 30

النص الكامل

القاهرة تعد المدينة من المدن المصرية العريقة والتي يرجع تاريخ إنشائها إلى عام 1815 م ، فهي تقع في الجزء الشمالي من وادي النيل على بعد حوالي 65 كم

Limitations & Future Work

- □ **Hardware Constraint**s

Training and evaluation were limited by available GPU memory.
Larger models, longer sequences, or batch sizes could not be explored due to OOM errors.

- **Limited Training Data**

You used only 1% of the OSCAR Arabic dataset, which limits language coverage and variety.It may underperform on dialects, uncommon phrases, or domain-specific vocabulary.

- **No Semantic Understanding**

The model is based on GPT-2, which predicts the next token based on surface patterns.
It doesn’t truly understand meaning or context, which can result in grammatically correct but nonsensical completions

Achievement

Successfully trained Arabic autocomplete model

Improvement Areas

Larger datasets and enhanced model tuning