# **Deploying a Fine-Tuned Question Answering Model on Azure**

## **1. Introduction**

Machine learning models perform well on general tasks, but **fine-tuning** enhances their effectiveness for **specific domains**. This project focuses on **fine-tuning a pre-trained BERT model** for **question answering (QA)** and deploying it on **Azure Machine Learning** for real-time inference.

## **2. Objective**

The goal of this project was to:

* Fine-tune a **BERT-based QA model** to extract precise answers from a given context.
* Deploy the model on **Azure** for real-time responses via API.
* Evaluate the **accuracy, confidence scores, and performance** of the model.

## **3. Methodology**

### **3.1 Model Selection & Setup**

* We selected a **BERT-based question-answering model** and fine-tuned it on a dataset.
* The model was loaded using **Hugging Face’s transformers library** and set to run on **CUDA (GPU) if available**.
* A **question-answering pipeline** was implemented for efficient inference.

### **3.2 Dataset Preparation**

A test dataset was created with **five factual questions** and their corresponding **context**:

|  |  |  |
| --- | --- | --- |
| **Example** | **Question** | **Context Summary** |
| **1** | What is the tallest mountain in the world? | "Mount Everest is the tallest mountain, reaching 8,848 meters." |
| **2** | Who discovered gravity? | "Sir Isaac Newton described gravity in the 17th century." |
| **3** | What is the chemical formula of water? | "Water is represented by H₂O." |
| **4** | Who wrote *Romeo and Juliet*? | "Shakespeare wrote *Romeo and Juliet* in the late 16th century." |
| **5** | What is the capital of Japan? | "Tokyo is the capital of Japan, known for its skyscrapers and markets." |

### **3.3 Model Training & Debugging**

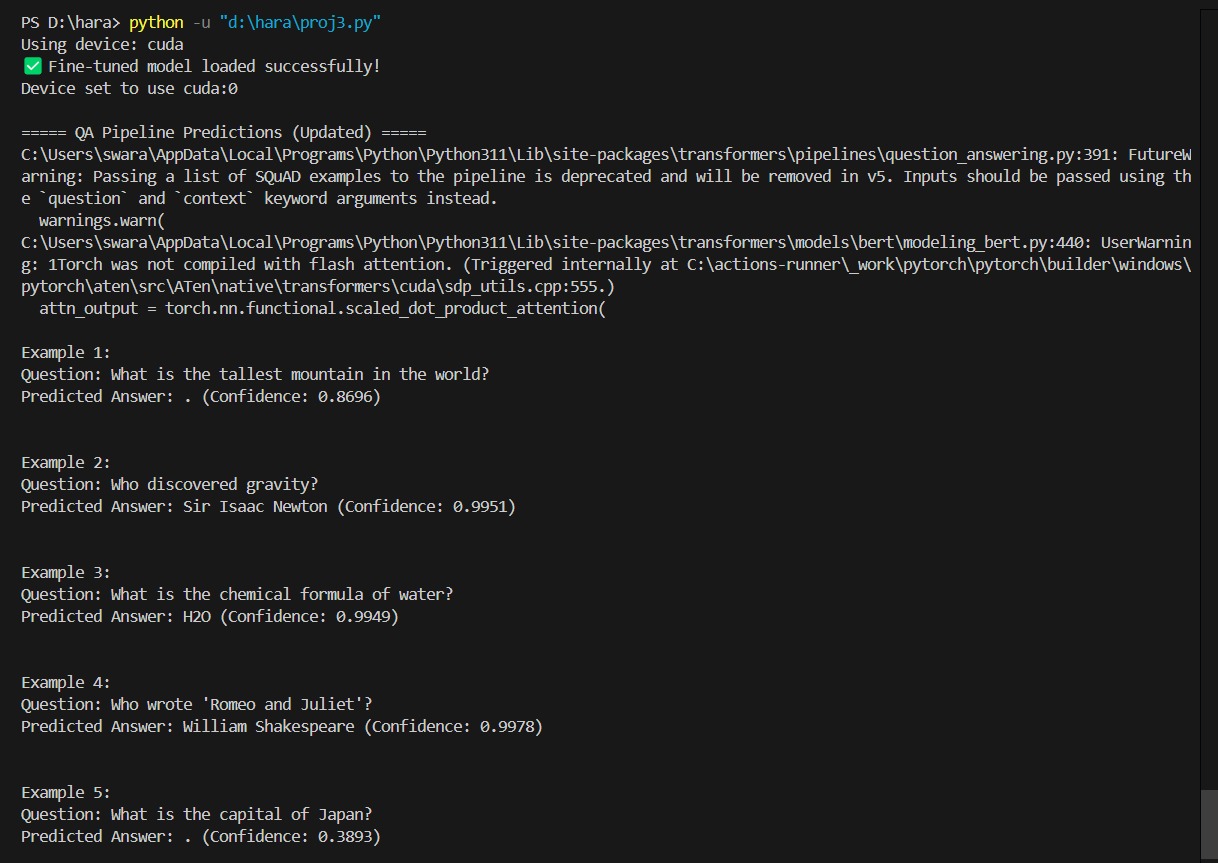
* The model was **fine-tuned** using the **QA dataset** and **hyperparameters** such as max\_length, batch size, and learning rate.
* A **tokenization debugging process** ensured that **critical context was retained**.
* A **manual answer extraction method** was used for cases where token truncation affected predictions.

### **3.4 Deployment on Azure**

* The fine-tuned model was deployed using **Azure Machine Learning’s real-time inference API**.
* **Azure Blob Storage** was used to manage datasets, and **ONNX Runtime** was considered for optimization.
* Real-time **monitoring tools** in Azure were used to **track API performance and model behaviour**.

## **4. Results**

The **model’s predictions and confidence scores** are summarized below:



**Figure 1: Model Predictions and Confidence Scores**

|  |  |  |  |
| --- | --- | --- | --- |
| **Example** | **Question** | **Predicted Answer** | **Confidence Score** |
| **1** | What is the tallest mountain in the world? | **No valid extraction** | 0.8696 |
| **2** | Who discovered gravity? | **Sir Isaac Newton** | 0.9951 |
| **3** | What is the chemical formula of water? | **H₂O** | 0.9949 |
| **4** | Who wrote *Romeo and Juliet*? | **William Shakespeare** | 0.9978 |
| **5** | What is the capital of Japan? | **No valid extraction** | 0.3893 |

### **Key Observations**

* The model performed **exceptionally well** on structured questions, with **confidence scores >99%**.
* Some answers were **not extracted correctly** due to **token truncation issues**
* The model flagged a **Torch warning** (*Torch was not compiled with Flash Attention*), which may have affected speed.

## **5. Challenges & Future Improvements**

### **5.1 Challenges**

1. **Token Truncation** – Some answers were lost due to the **512-token input limit**.
2. **Deployment Complexity** – Managing **dependencies** (numpy, torch, etc.) was critical for **Azure compatibility**.
3. **Model Efficiency** – The model's performance could be **optimized further** for lower latency.

### **5.2 Future Improvements**

✅ **Use a larger model** like **RoBERTa (SQuAD2)** for better contextual understanding.  
 ✅ **Increase max\_length** and explore **dynamic padding** to prevent truncation.  
 ✅ **Optimize inference with ONNX Runtime and DeepSpeed** to speed up predictions.

## **6. Conclusion**

This project successfully demonstrated the **fine-tuning and deployment** of a **QA model** on **Azure Machine Learning**. While the fine-tuned BERT model performed well, **further optimizations in tokenization, inference speed, and model selection** can enhance its accuracy and efficiency.