Question Answering on SQuAD 2.0

Introduction

In the dynamic realm of Natural Language Processing (NLP), the advent of question-answering (QA) systems marks a significant stride in our ability to interact with and process digital information. This project is dedicated to developing a reading comprehension-based QA system inspired by the comprehensive insights in Speech and Language Processing by Daniel Jurafsky & James H. Martin (2023). We aim to create a system that can interpret and respond to questions posed in natural language, drawing answers from provided text passages.

NLP Task - Question Answering

In the Natural Language Processing (NLP) field, extractive Question Answering (QA) is a pivotal task involving locating the answer to a question within a specified text passage. This task is inherently challenging, as it requires the system to comprehend the posed question and accurately extract the specific text portion that contains the answer. As detailed in Hugging Face's documentation and task library (n.d.), extractive QA demands the capability to sift through extensive text and pinpoint information that precisely responds to the query.

Figure 01

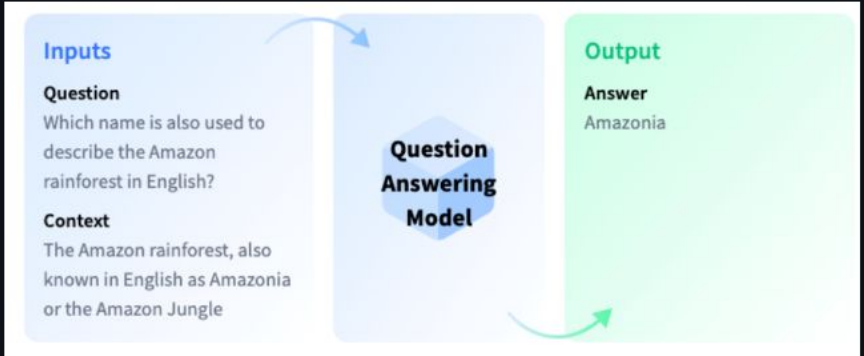


Fig-1: Question Answering Task (Hugging Face. 2001)

Jurafsky and Martin (2023), in their seminal work,Speech and Language Processing,elucidate the complexities of extractive QA, highlighting the necessity for advanced NLP techniques and models.These models are crucial for understanding the context and semantics embedded in both the question and the passage, thus enabling the identification of the exact text span that answers the question.Extractive QA is particularly vital in scenarios necessitating factual answers directly sourced from the provided text, such as in academic research or specific information retrieval tasks.In our project, we embrace the challenges of extractive QA by training our model on the SQuAD 2.0 dataset. This dataset, encompassing diverse questions and passages, provides a comprehensive framework for the system to learn from varying contexts and question types.The model is meticulously trained to parse the subtleties of language in questions and passages, enhancing its ability to discern and extract the relevant answers accurately.This endeavor underscores the significance of sophisticated text processing and comprehension in NLP, laying the groundwork for more intelligent and adept information retrieval systems.

Dataset

- SQuAD 2.0: Stanford Question Answering Dataset 2.0 (SQuAD 2.0) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles.

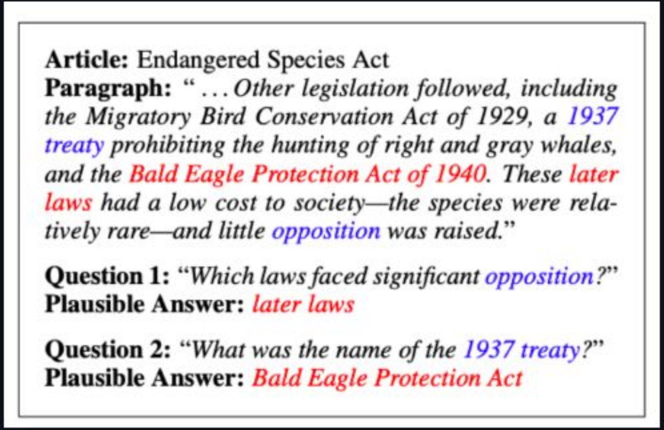
- Introduction of Unanswerable Questions: SQuAD 2.0 is enhanced with 53,775 unanswerable questions, created by crowdworkers, in addition to the answerable ones from SQuAD 1.1.

- Objective of SQuAD 2.0: It aims to challenge machine learning models in determining when a correct answer is not present in the text, elevating their proficiency in reading comprehension.

- Increased Complexity: Unlike SQuAD 1.1, which focused solely on finding the correct text span for a given question, SQuAD 2.0 adds the dimension of identifying unanswerable questions, making it more complex.

- Public Benchmark: SQuAD 2.0 is available publicly and serves as the primary benchmark on the official SQuAD leaderboard.

Figure 02

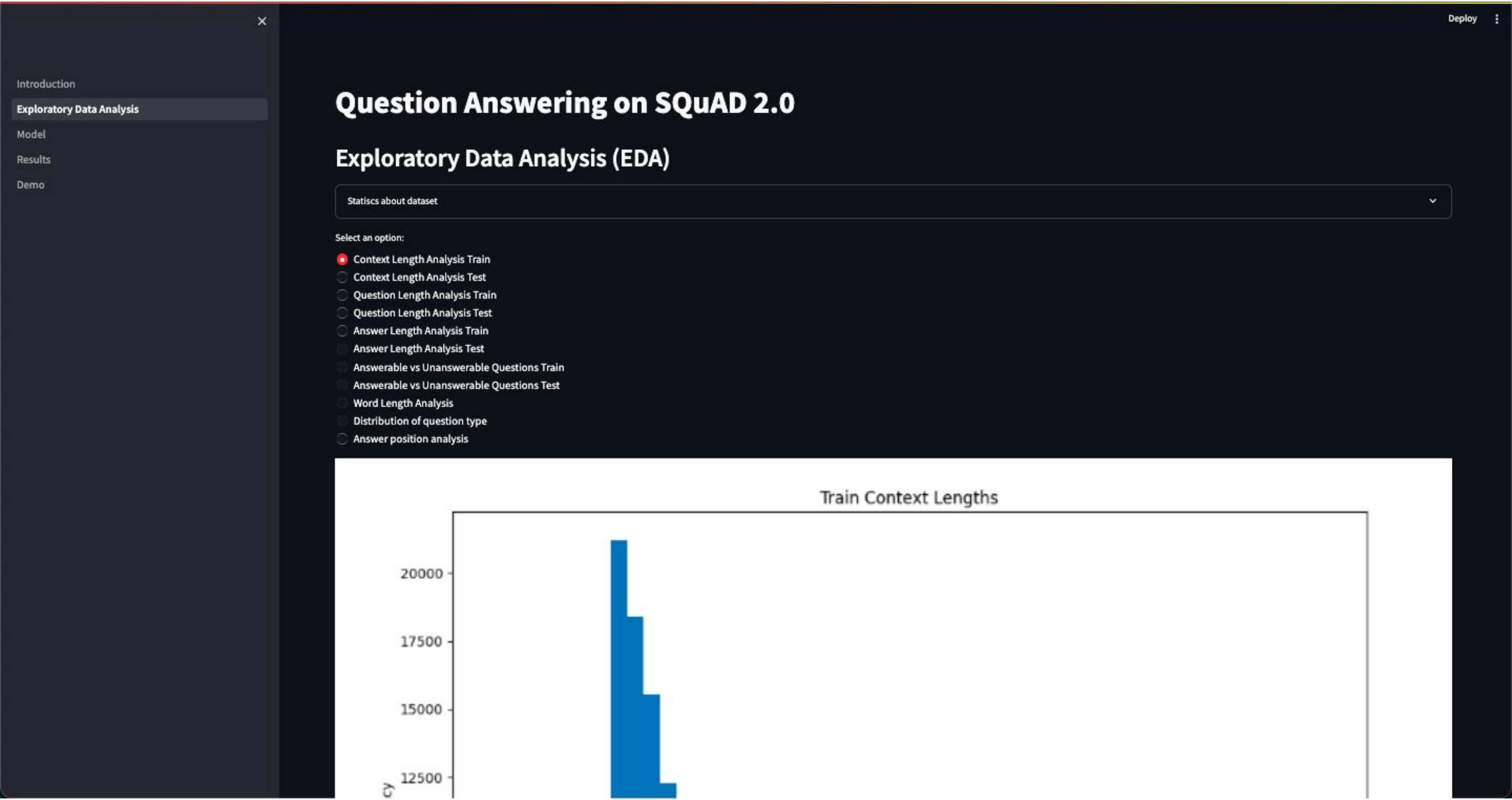


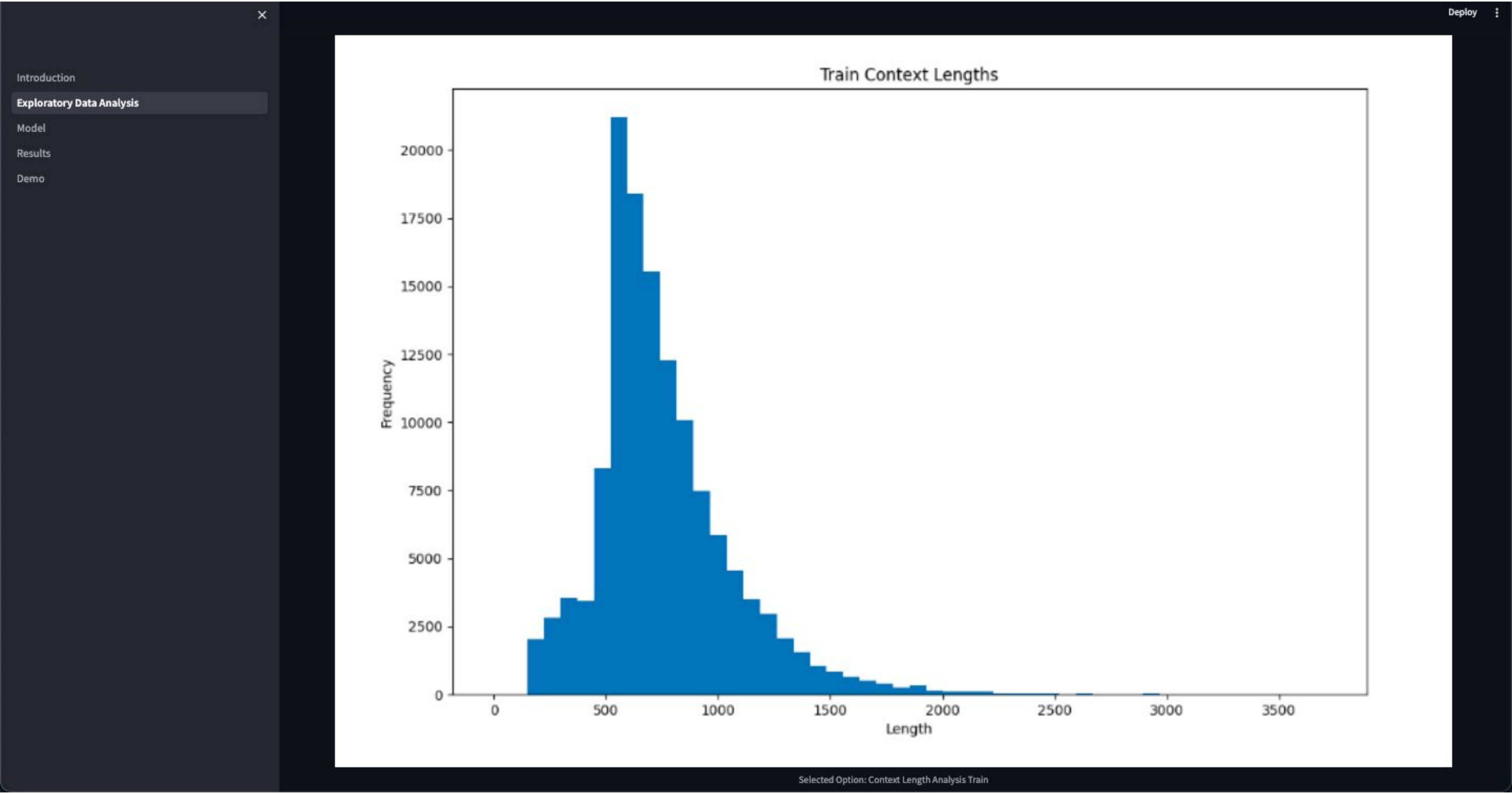
Article: Endangered Species Act
Paragraph: " *...Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

Figure 03


Deploy
Introduction
Exploratory Data Analysis
Model
Results
Demo
Question Answering on SQuAD 2.0
Exploratory Data Analysis (EDA)
Statiscs about dataset
Select an option:
Context Length Analysis Train
Context Length Analysis Test
Question Length Analysis Train
Question Length Analysis Test
Answer Length Analysis Train
Answer Length Analysis Test
Answerable vs Unanswerable Questions Train
Answerable vs Unanswerable Questions Test
Word Length Analysis
Distribution of question type
Answer position analysis
Train Context Lengths
20000
17500
15000
12500


Deploy
Introduction
Exploratory Data Analysis
Model
Results
Demo
Train Context Lengths
Frequency
20000
17500
15000
12500
10000
7500
5000
2500
0
0
500
1000
1500
2000
2500
3000
3500
Length
Selected Option: Context Length Analysis Train

Introduction

Exploratory Data Analysis

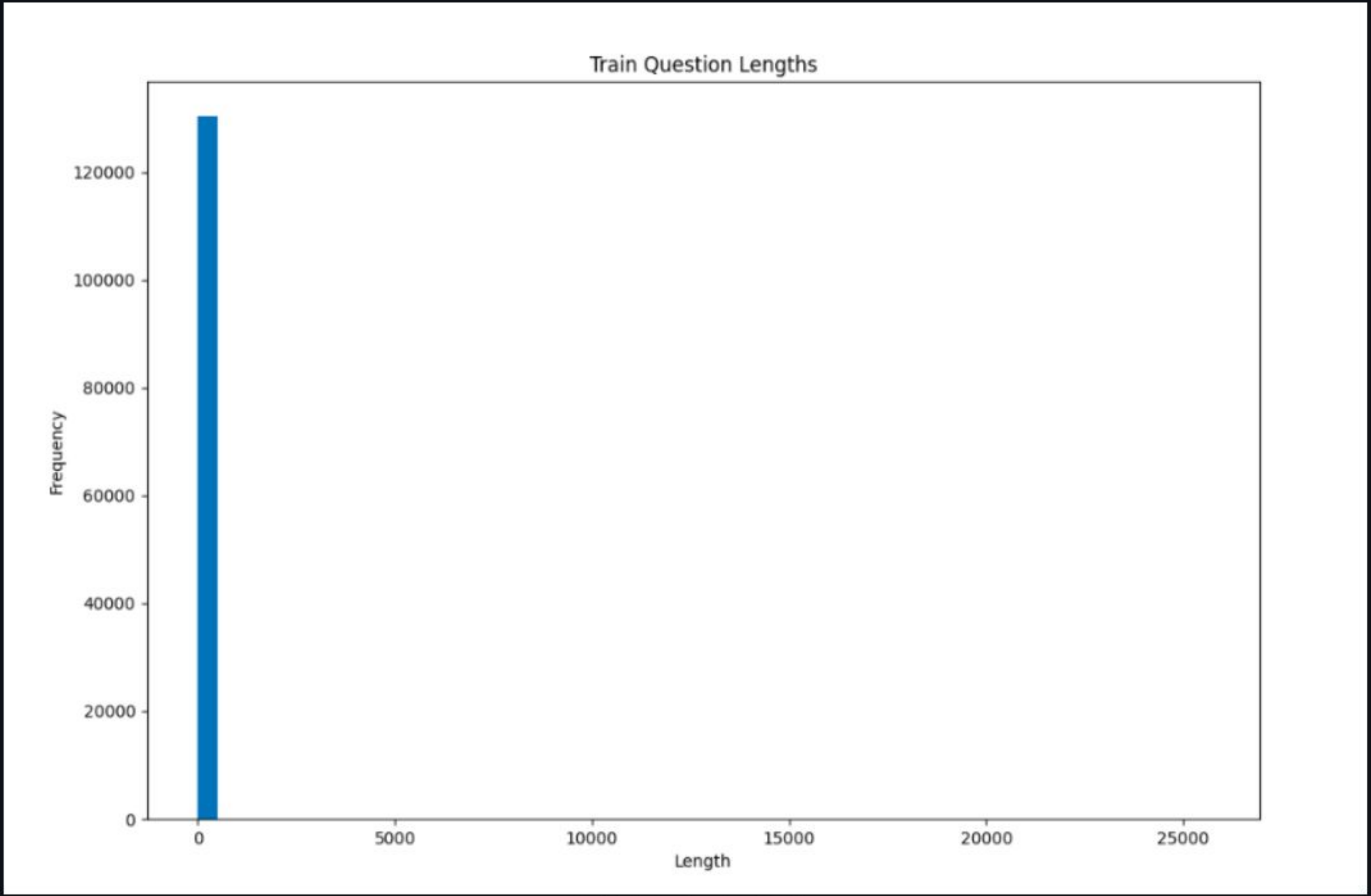
Model

Results

Demo

Deploy

Answer position analysis



Selected Option: Question Length Analysis Train

Deploy

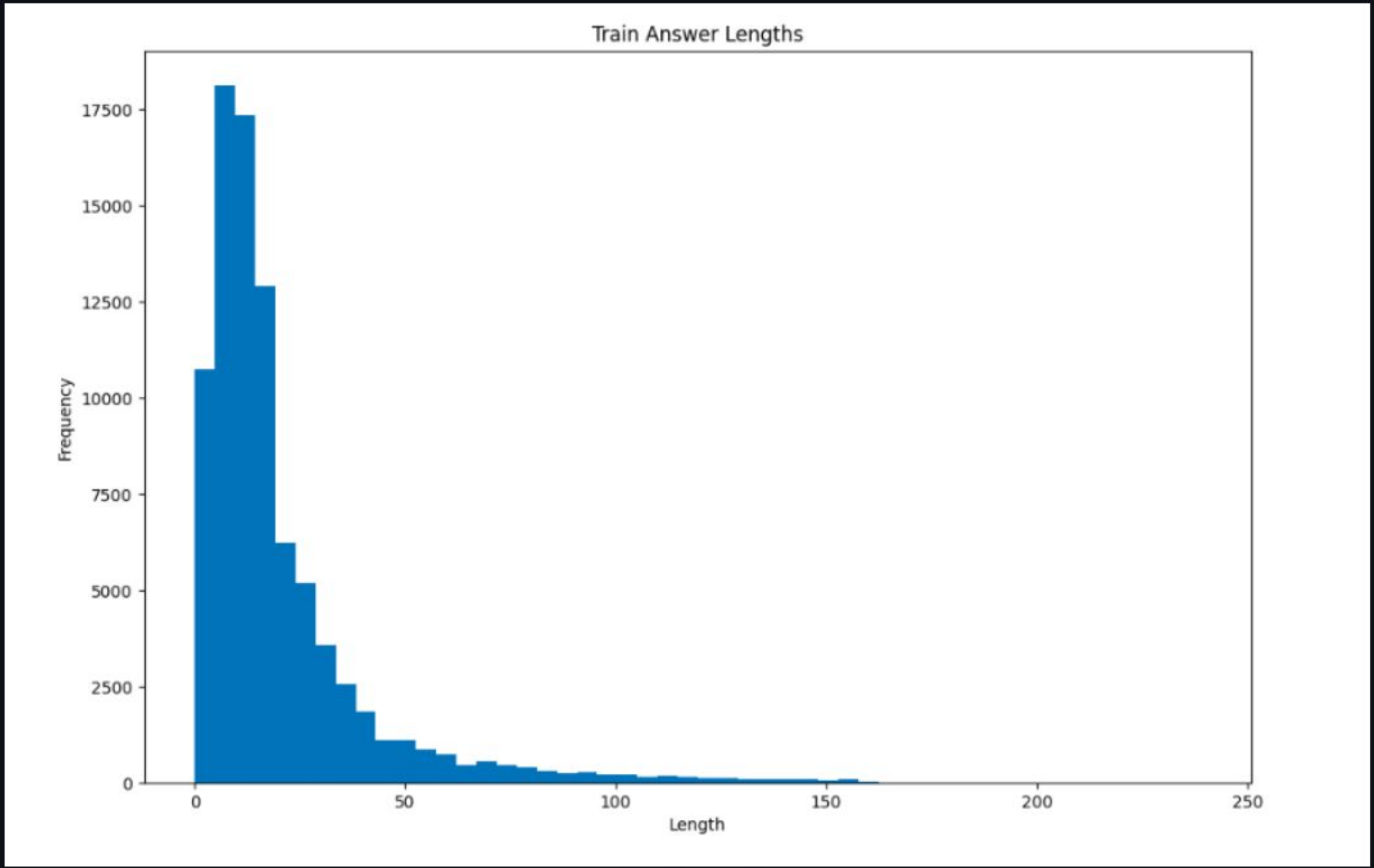
Introduction

Exploratory Data Analysis

Model

Results

Demo



Selected Option: Answer Length Analysis Train

The below figure is the comparison of the SQuAD 1.1 and SQuAD 2.0 representing the train, development and test values of the total examples, negative examples, total articles and articles with negatives

Deploy

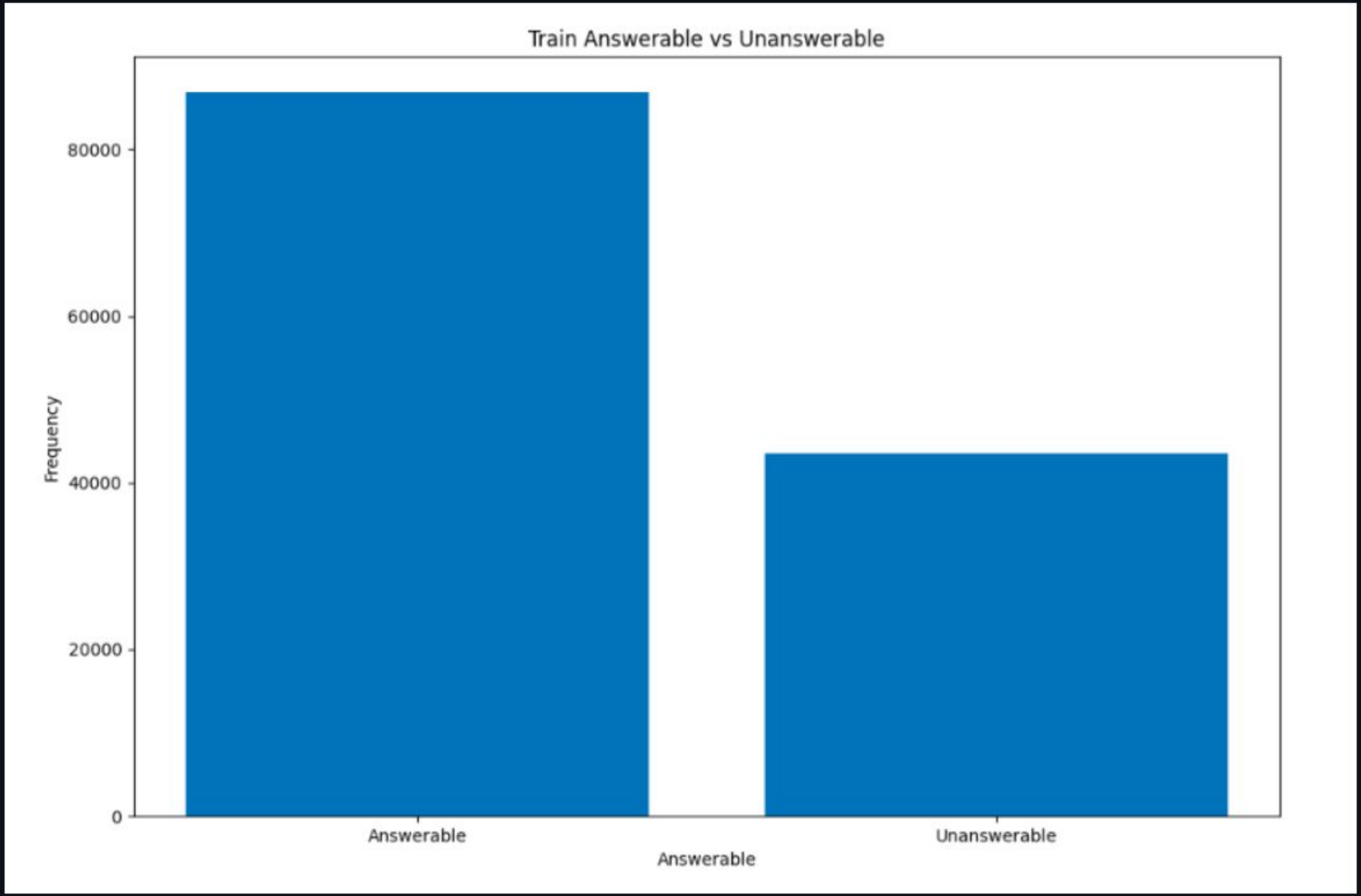
Introduction

Exploratory Data Analysis

Model

Results

Demo



Selected Option: Answerable vs Unanswerable Questions Train

Deploy

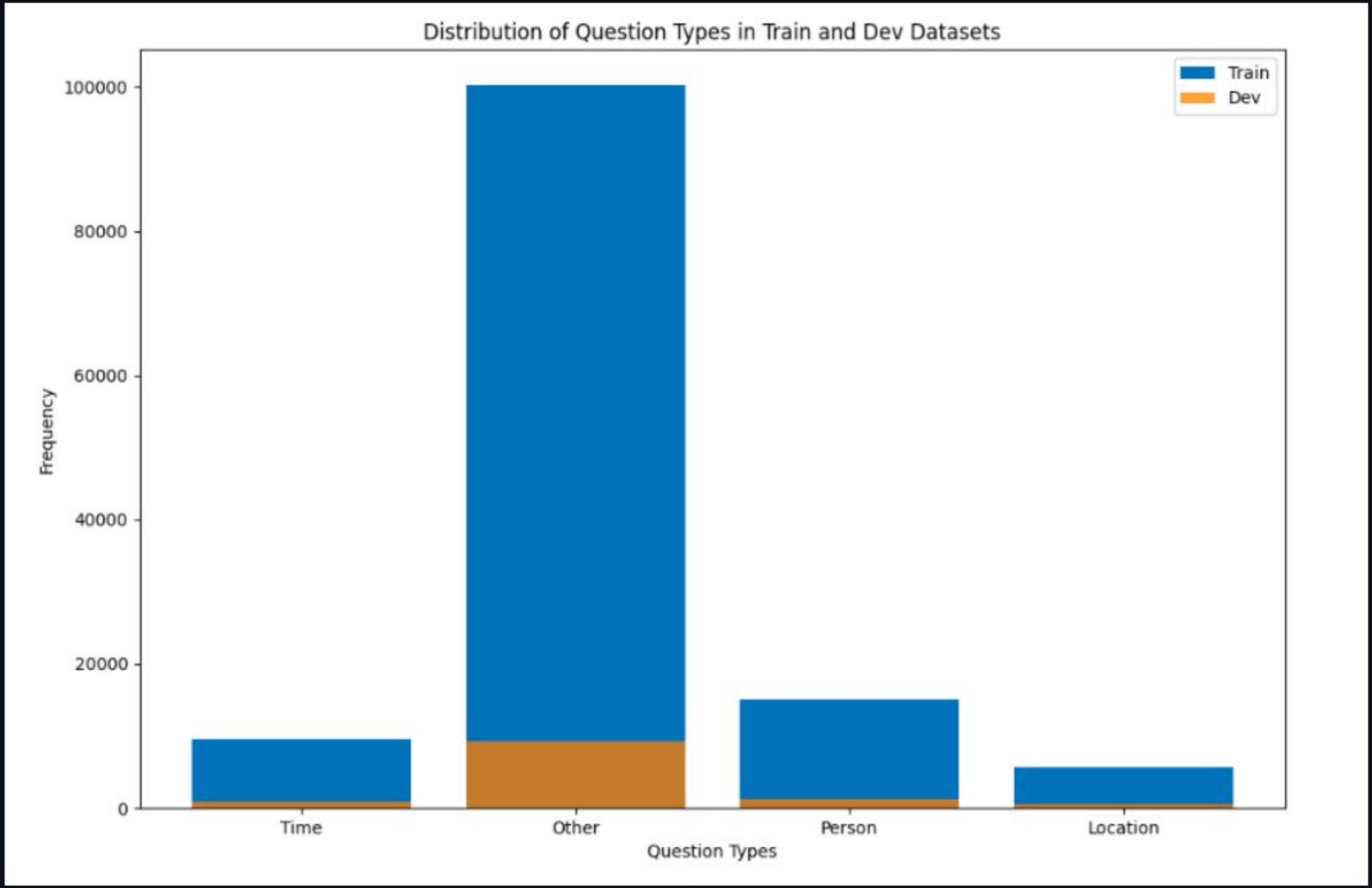
Introduction

Exploratory Data Analysis

Model

Results

Demo



Selected Option: Distribution of question type

- Introduction
- **Exploratory Data Analysis**
- Model
- Results
- Demo

Deploy

Selected Option: Distribution of question type

The below figure is the comparison of the SQuAD 1.1 and SQuAD 2.0 representing the train, development and test values of the total examples, negative examples, total articles and articles with negatives

	SQuAD 1.1	SQuAD 2.0
Train		
Total examples	87,599	130,319
Negative examples	0	43,498
Total articles	442	442
Articles with negatives	0	285
Development		
Total examples	10,570	11,873
Negative examples	0	5,945
Total articles	48	35
Articles with negatives	0	35
Test		
Total examples	9,533	8,862
Negative examples	0	4,332
Total articles	46	28
Articles with negatives	0	28

Deploy

Question Answering on SQuAD 2.0

Model Description

ELECTRA

• **Two-Component Structure**

• **Generator: Its role is to replace some tokens in the input data with a plausible alternatives.**

• **Discriminator: Its role is to determine whether the replaced tokens are real or fake.**

• **Joint Training: Both the generator and the discriminator are trained simultaneously. (It contrasts with GANs.)**

• **Final Model Utilization: After training, only the discriminator is used for downstream tasks. (It contrasts with the traditional approach.)**

• **Efficiency and Scaling: ELECTRA shows that it's more efficient than models like BERT in terms of computational resources needed for training.**

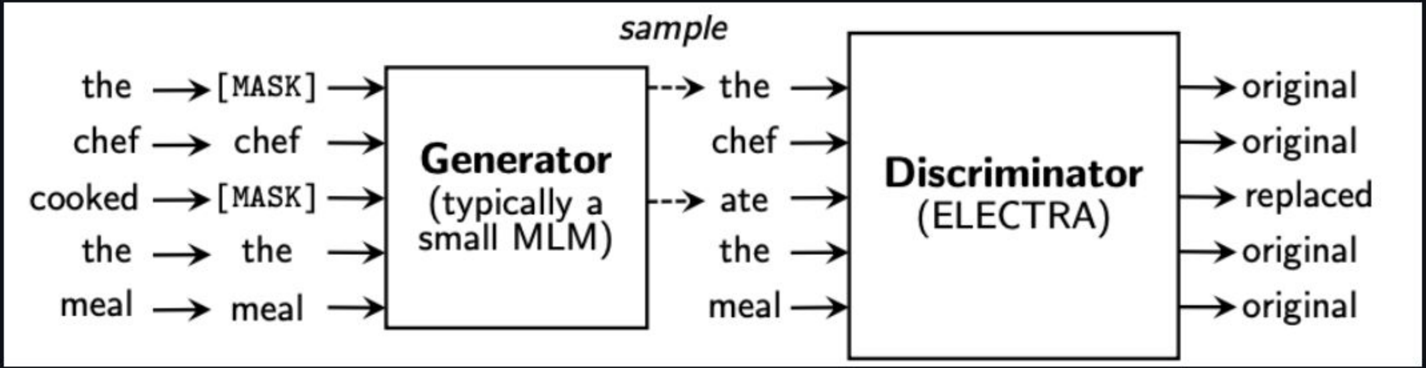


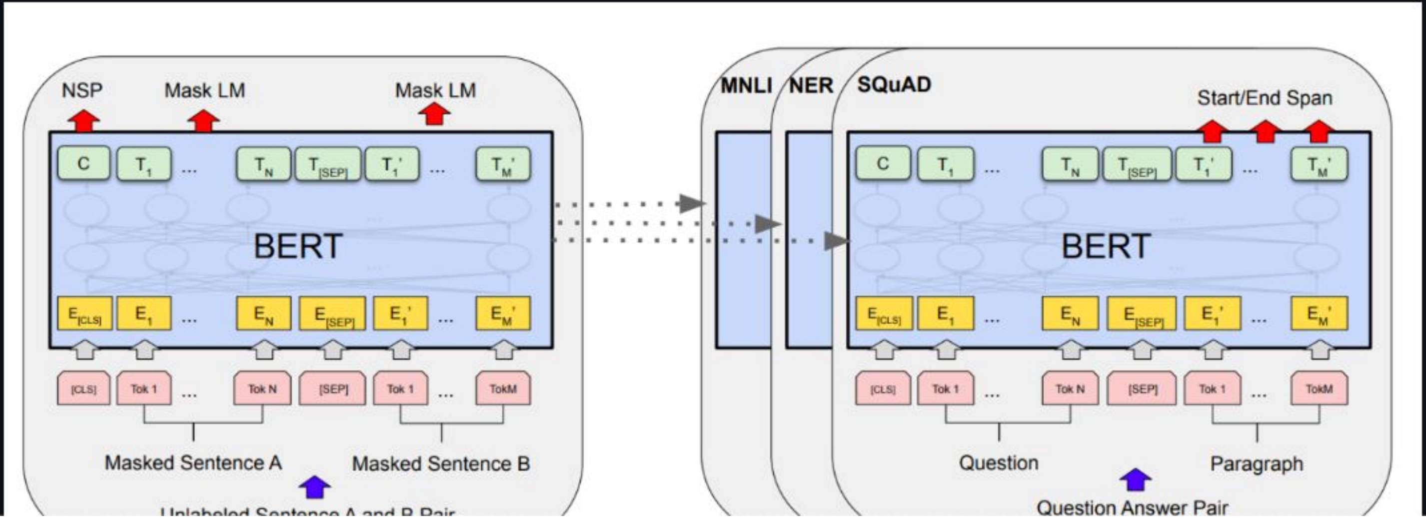
Figure 1: ELECTRA Architecture (Source: Electra Paper)

Introduction
Exploratory Data Analysis
Model
Results
Demo

Deploy

BERT with LORA

- **Transformer, an attention mechanism :**
- **Encoder: Its target is to process the input text and understand the context of each word or token within it.**
- **ATTENTION Mechanism: It gives the model ability to weigh the importance of different words in the sentence relative to each other.**
- **Embedding Layers: Find the Embedding Representation of Each words (Word Embedding, Positional Embedding, Segment Embedding)**
- **Pretraining: Masked Language Models/ Next Sentence Prediction**
- **Fine-Tuning on QA: Start Vector/ End Vector**
- **Why using LORA? Base BERT model: 110 millions parameters**
- **A method to learn a lower-dimensional, task-specific representation of the layer's weights.**



Introduction

Exploratory Data Analysis

Model

Results

Demo

Deploy

- **Pretraining: Masked Language Models/ Next Sentence Prediction**
- **Fine-Tuning on QA: Start Vector/ End Vector**
- **Why using LORA? Base BERT model: 110 millions parameters**
- **A method to learn a lower-dimensional, task-specific representation of the layer's weights.**

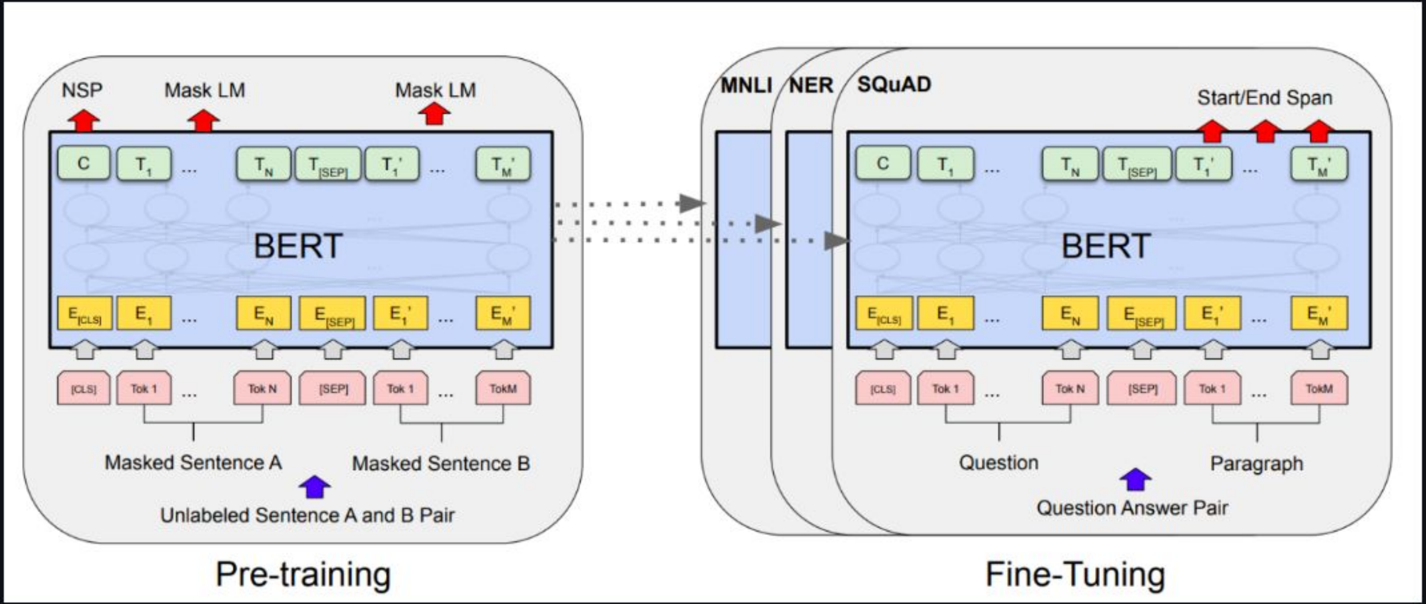
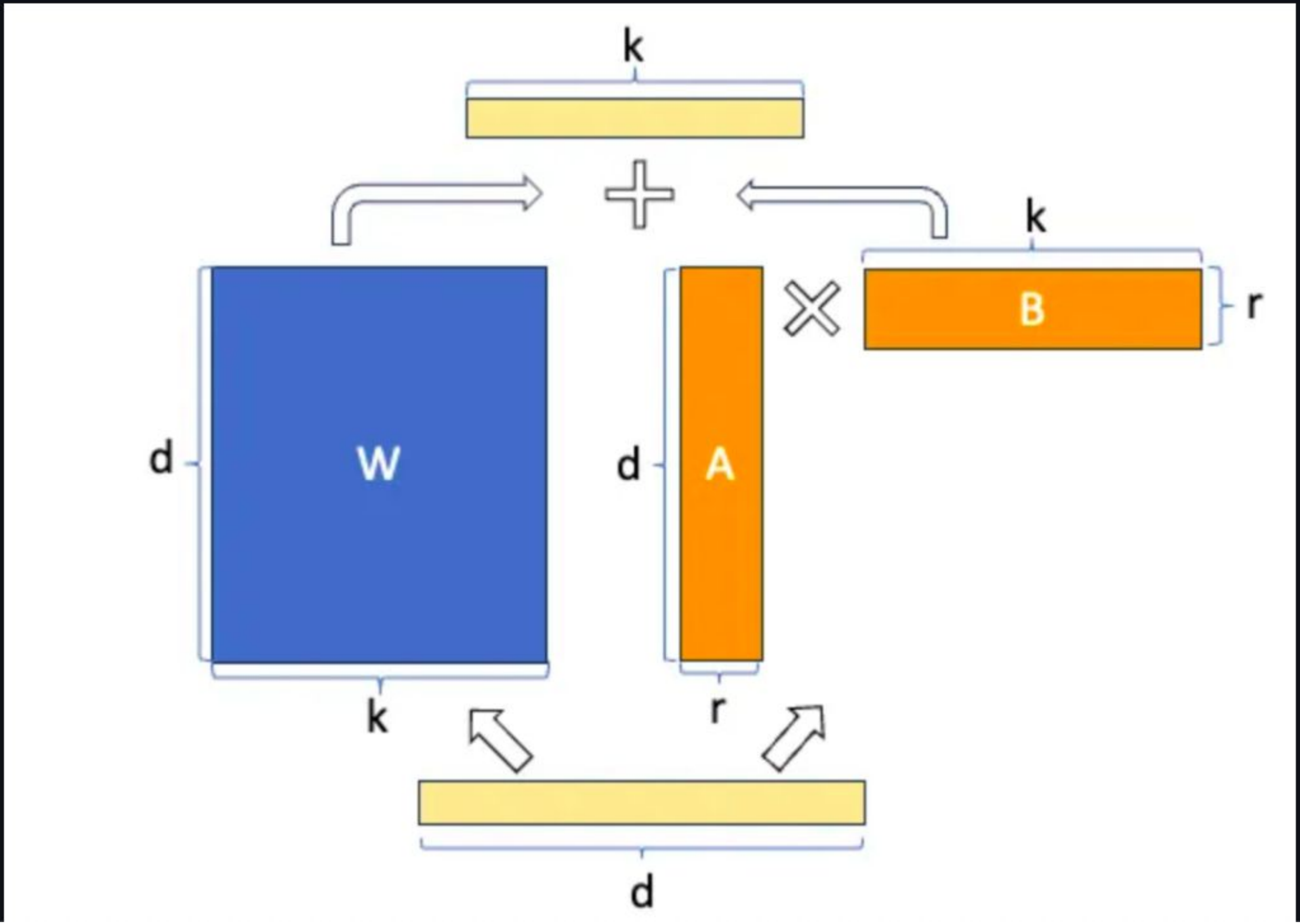


Figure 2: BERT Architecture (Source: Google AI Language)

Introduction
Exploratory Data Analysis
Model
Results
Demo

Deploy



Training Setups

- Dataset: Utilized the SQuAD 2.0 dataset with 130,319 training, 11,873 development examples.

- Preprocessing: Tokenization, sliding windows, padding, and answer localization.

- Evaluation Metrics: Exact Match (EM) and F1 score.

ELECTRA Model Fine-Tuning for SQuAD 2.0

Hyperparameters:

- Maximum Sequence Length: 384

- Stride: 128

- Number of Predictions to Generate: 20

- Maximum Answer Length: 30

Training Parameters:

- Learning Rate: 5e-5

- Batch Size: 32

- Number of Epochs: 3 (Training duration: 1.5 hours)

The model is fine-tuned using HuggingFace's implementation with an AdamW optimizer.

BERT Model Fine-Tuning for SQuAD 2.0

Hyperparameters:

- Model: bert-base-uncased with optional LoRA configuration.

- Maximum Sequence Length: 512

- Learning Rate: 5e-5

Introduction
Exploratory Data Analysis
Model
Results
Demo

Deploy

Training Parameters:

- Learning Rate: 5e-5

- Batch Size: 32

- Number of Epochs: 3 (Training duration: 1.5 hours)

The model is fine-tuned using HuggingFace's implementation with an AdamW optimizer.

BERT Model Fine-Tuning for SQuAD 2.0

Hyperparameters:

- Model: bert-base-uncased with optional LoRA configuration.

- Maximum Sequence Length: 512

- Learning Rate: 5e-5

- Batch Size: 16

- Number of Epochs: 3 (Training duration: 2 hours)

LORA Parameters:

- r: 16

- lora_alpha=32

- lora_dropout=0.05

The model is fine-tuned using HuggingFace's implementation with an AdamW optimizer.

Introduction
Exploratory Data Analysis
Model
Results
Demo

Question Answering on SQuAD 2.0

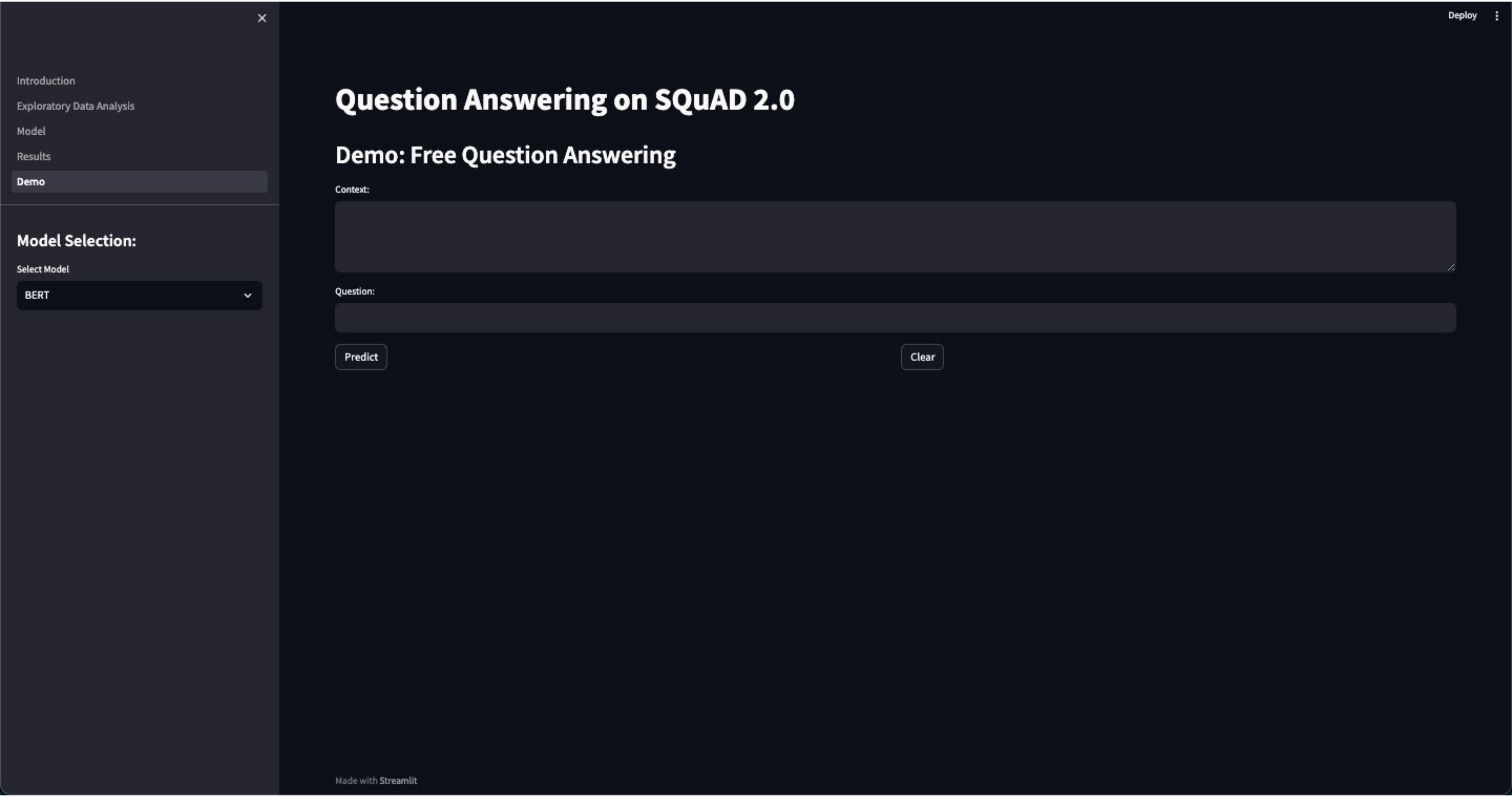
Evaluation / Results

ELECTRA

- F1 Score: 0.50071
- Exact Match: 0.50071

BERT with LORA

- All params : 109,484,548
- Trainable Parameters: 54% 591,362
- Training Loss(epoch 1 -3): 2.6399 1.6731 1.4301
- Valid Loss(epoch 1 -3): 1.7555 1.3669 1.2309
- F1 Score: 0.50989


Deploy
Introduction
Exploratory Data Analysis
Model
Results
Demo
Model Selection:
Select Model
BERT
Question Answering on SQuAD 2.0
Demo: Free Question Answering
Context:
Question:
Predict
Clear
Made with Streamlit