UNIVERSITY OF HOUSTON

EDS 6352 - Natural Language Processing

Fall - 2025

English QA: Transformer-Based Extractive Question Answering using SQuAD v2

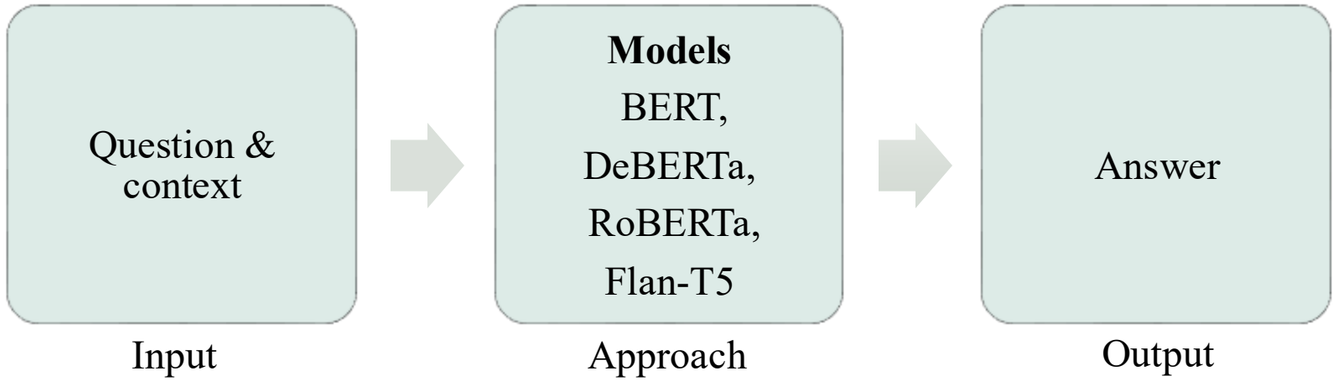
Group - 9

Team Members:

Adi Karthikeya S B – 2417324
Amrutha Vaishnavi Alla - 2378344
Gayatri Chekuri - 2404773

Objective

- To develop and compare transformer-based extractive QA systems on SQuAD 2.0 dataset containing both answerable and unanswerable questions.



Dataset

Stanford Question Answering Dataset – SQuAD v2

- Over 140,000 QA pairs.
- Includes both answerable and unanswerable questions.
- We will use a subset of 10,000 QA pairs for efficient training within Colab GPU limits.

Dataset Link: SQuAD v2

Data Split:

•Training: 10,000 examples (80%)

•Validation: 2,000 examples (20%)

Model Architectures:

- Extractive

BERT

Basic text understanding

DeBERTa

Smarter attention reading

RoBERTa

BERT but stronger

- Generative

Flan–T5

Writes answer itself

BERT - Bidirectional Encoder Representations from Transformers.

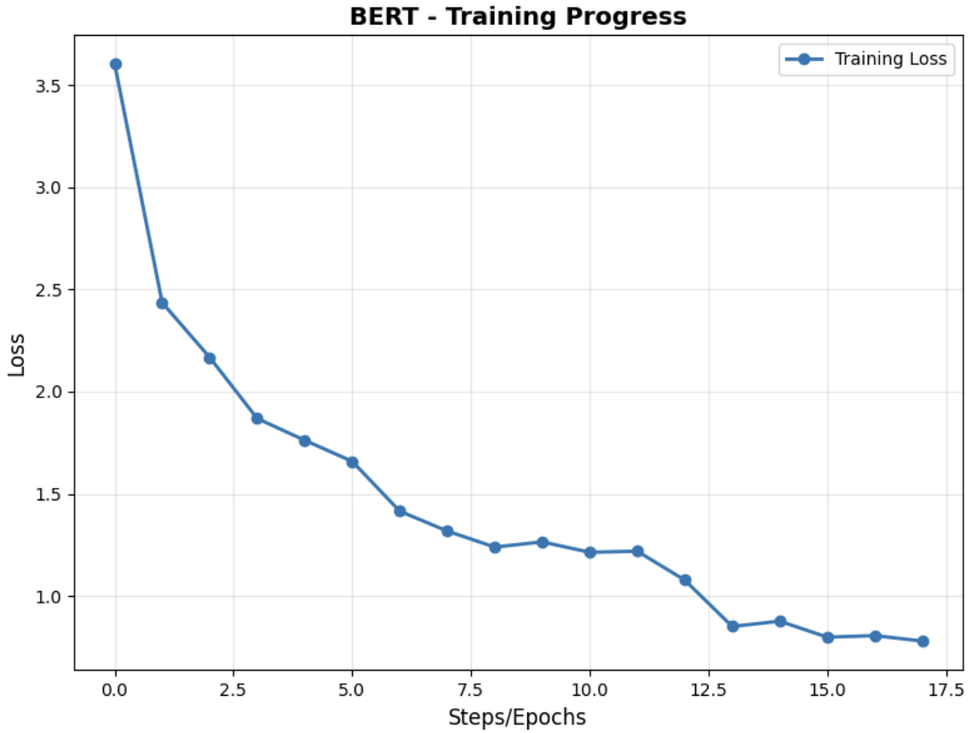
- It is a pre-trained transformer encoder that reads text in both directions simultaneously (left-to-right AND right-to-left) to understand context, like reading a sentence while knowing what comes before AND after each word.

- Designed by Google in 2018.

Note:
The Bert score for the metrics (precision, recall, f1) was mistakenly multiplied by 100, resulting in thousands. We will correct this error by removing the 1000, and the actual percentage will be displayed as the number divided by 100.

```
BERT Results:
exact_match          : 33.12
f1                   : 38.70
bleu                 : 0.00
bertscore_precision  : 9355.71
bertscore_recall     : 9388.98
bertscore_f1         : 9367.39
macro_f1             : 33.30
micro_f1             : 49.93
precision_macro      : 24.96
recall_macro         : 50.00
answerable_accuracy  : 100.00
unanswerable_accuracy: 0.00
```

Training Curve



Example

Question: When was the University of Houston founded?
Context: The University of Houston is a public research university in Houston, Texas. It was founded in 1927.

BERT Answer: 1927.

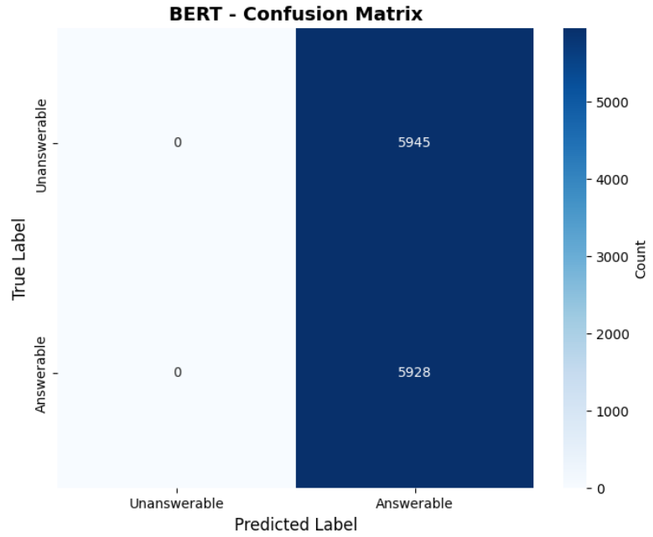
Metrics

```
============================================================
BERT
============================================================
                precision    recall  f1-score   support

Unanswerable        0.000     0.000     0.000      5945
  Answerable        0.499     1.000     0.666      5928

    accuracy                          0.499     11873
   macro avg        0.250     0.500     0.333     11873
weighted avg        0.249     0.499     0.333     11873
```

Confusion Matrix



Overall Accuracy: 49.93%

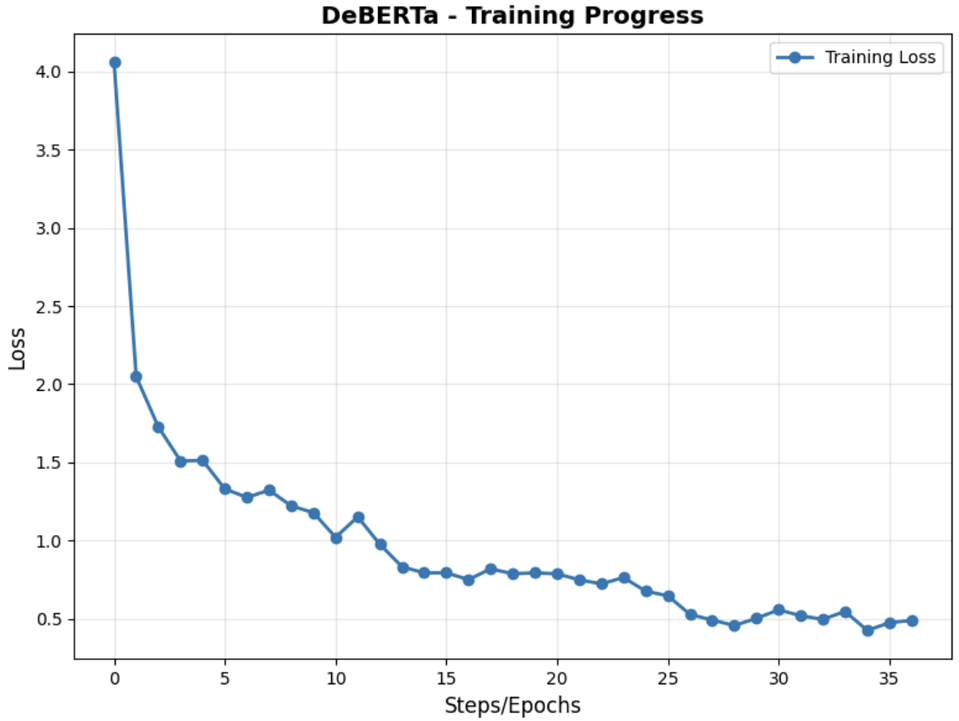
DeBERTa - Decoding-enhanced BERT with disentangled attention.

- It is an enhanced BERT model with disentangled attention which separates word meaning from word position, understanding that "bank" (financial) and "bank" (river) are different based on WHERE they appear in a sentence

- Designed by Microsoft in 2021.

```
DeBERTa Results:
exact_match          : 39.43
f1                   : 44.13
bleu                 : 0.00
bertscore_precision  : 9533.41
bertscore_recall     : 9542.19
bertscore_f1         : 9533.97
macro_f1             : 33.30
micro_f1             : 49.93
precision_macro      : 24.96
recall_macro         : 50.00
answerable_accuracy  : 100.00
unanswerable_accuracy: 0.00
```

Training Curve



Example

Question: Who invented the telephone?
Context: Alexander Graham Bell invented and patented the first practical telephone in 1876.

DeBERTa Answer: Alexander Graham Bell.

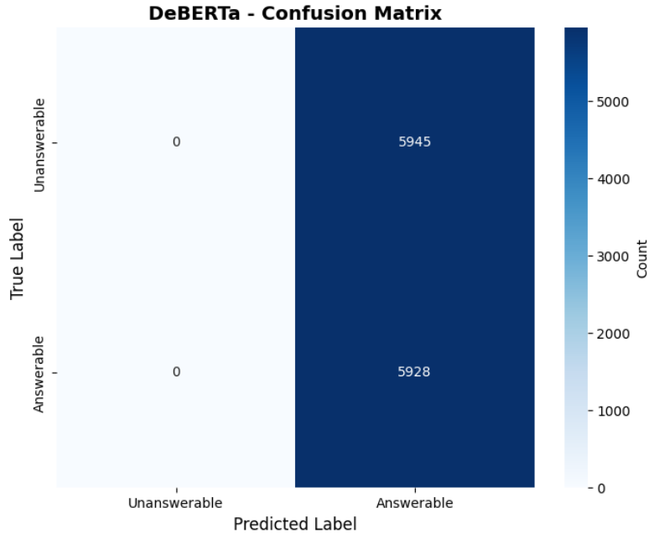
Metrics

```
============================================================
DeBERTa
============================================================
                precision    recall  f1-score   support

Unanswerable        0.000     0.000     0.000      5945
  Answerable        0.499     1.000     0.666      5928

    accuracy                            0.499     11873
   macro avg        0.250     0.500     0.333     11873
weighted avg        0.249     0.499     0.333     11873
```

Confusion Matrix



Overall Accuracy: 49.93%

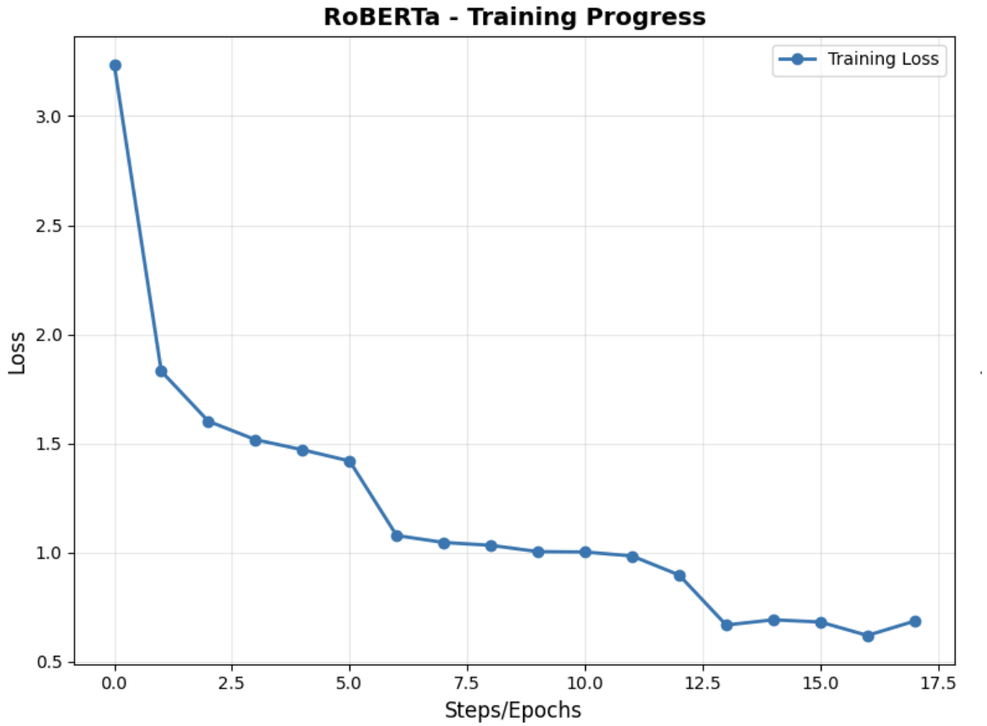
RoBERTa - Robustly optimized BERT approach.

- It is an Optimized BERT model that trains better with more data, longer training, large batch sizes and without the next sentence prediction just like studying for an exam with more practice problems and better study techniques.

- Designed by Meta AI (Facebook) in 2019.

```
RoBERTa Results:
exact_match          : 38.37
f1                   : 42.88
bleu                 : 0.00
bertscore_precision  : 9499.56
bertscore_recall     : 9504.51
bertscore_f1         : 9497.99
macro_f1             : 33.41
micro_f1             : 49.96
precision_macro      : 62.47
recall_macro         : 50.03
answerable_accuracy  : 99.97
unanswerable_accuracy: 0.10
```

Training Curve



Example

Question: What is the capital of France?
Context: Paris is the capital and largest city of France.

RoBERTa Answer: Paris.

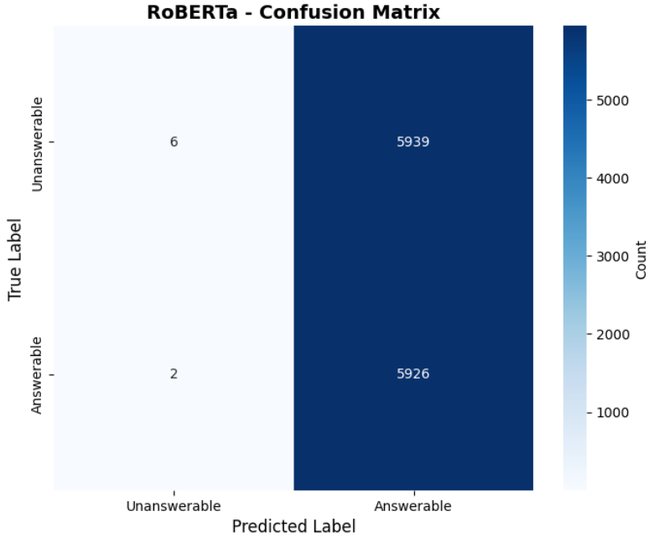
Metrics

```
============================================================
RoBERTa
============================================================
                precision    recall  f1-score   support

Unanswerable        0.750     0.001     0.002      5945
  Answerable        0.499     1.000     0.666      5928

    accuracy                          0.500     11873
   macro avg        0.625     0.500     0.334     11873
weighted avg        0.625     0.500     0.334     11873
```

Confusion Matrix



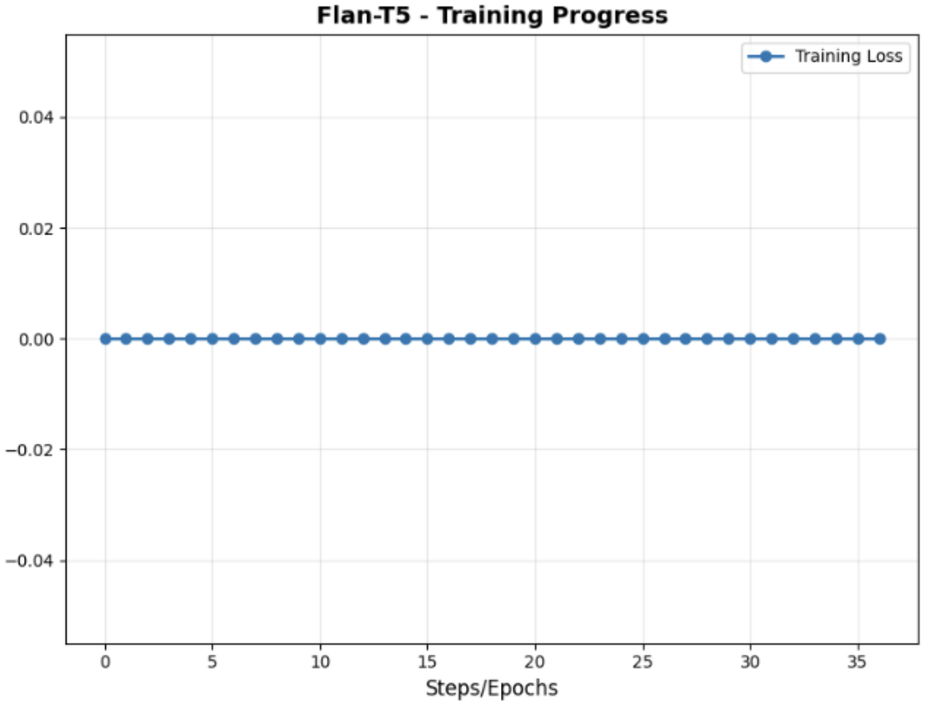
Overall Accuracy: 49.96%

Flan-T5 - Fine-tuned Language Net - Text-to-Text Transfer Transformer

- It is an instruction-tuned T5 model trained on 1,800+ diverse tasks using natural language instructions that treats every task as "read this, write that" - including question answering.
- Instead of pointing to the answer like extractive models (BERT, RoBERTa, DeBERTa), it generates the answer from scratch using encoder- decoder architecture for text generation.
- Designed by Google in 2022.

```
Flan-T5 Results:
exact_match          : 40.28
f1                   : 44.15
bleu                 : 0.00
bertscore_precision  : 9579.97
bertscore_recall     : 9534.13
bertscore_f1         : 9553.45
macro_f1             : 33.32
micro_f1             : 49.93
precision_macro      : 49.96
recall_macro         : 50.00
answerable_accuracy  : 99.98
unanswerable_accuracy: 0.02
```

Training Curve



Example

Question: When was the University of Houston founded?
Context: The University of Houston (UH) was founded in 1927 as Houston Junior College.

Flan-T5 Answer: 1927.

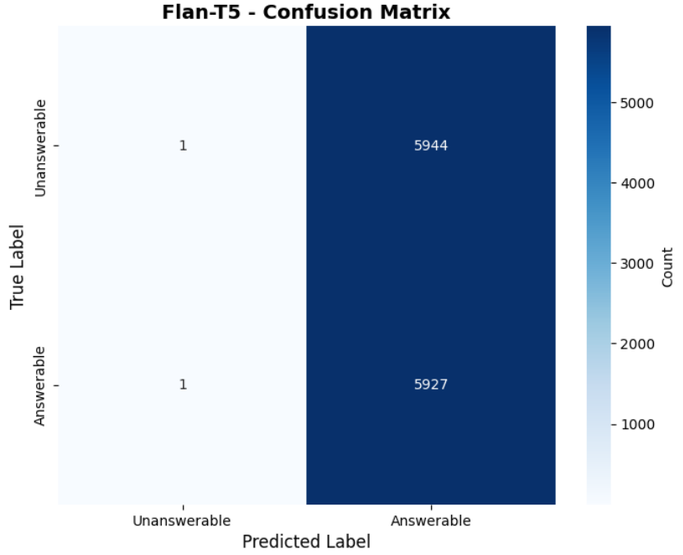
Metrics

```
============================================================
Flan-T5
============================================================
              precision    recall  f1-score   support

Unanswerable      0.500     0.000     0.000      5945
  Answerable      0.499     1.000     0.666      5928

    accuracy                          0.499     11873
   macro avg      0.500     0.500     0.333     11873
weighted avg      0.500     0.499     0.333     11873
```

Confusion Matrix



Overall Accuracy: 49.93%

UNIVERSITY OF HOUSTON

Approach

Data Preprocessing

Structure dataset into (question, context, answer) triples. Apply native tokenizers.

Model Fine-Tuning

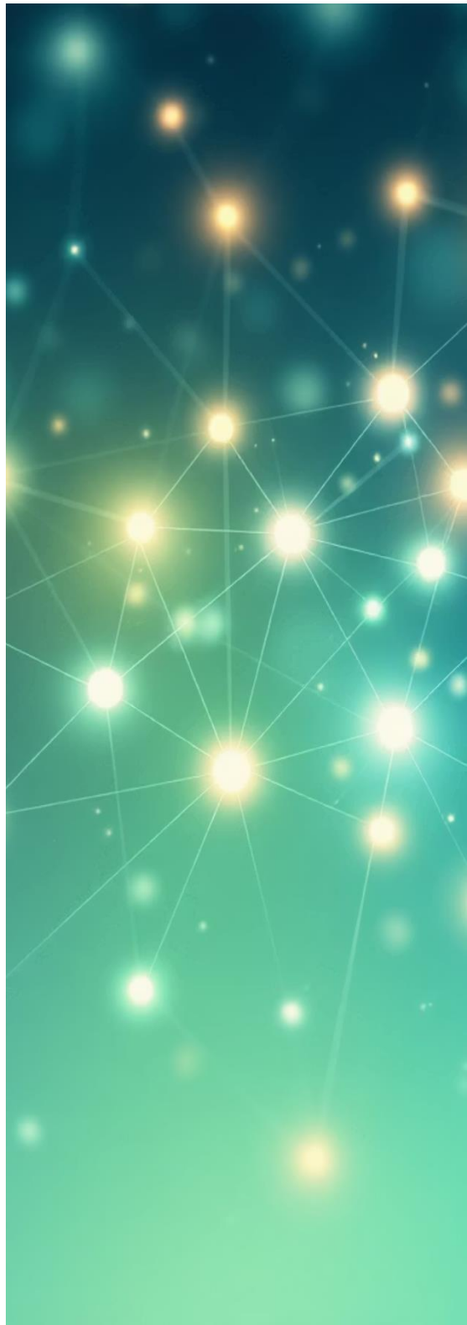
Fine-tune BERT, DeBERTa, and RoBERTa on the SQuAD v2 subset.

Optional: Flan-T5

Explore a generative comparison with Flan-T5-base model.

Evaluation Metrics

Assess performance using Exact Match (EM), F1-score, and BERTScore.



Results

Evaluation Metrics

Exact Match (EM)

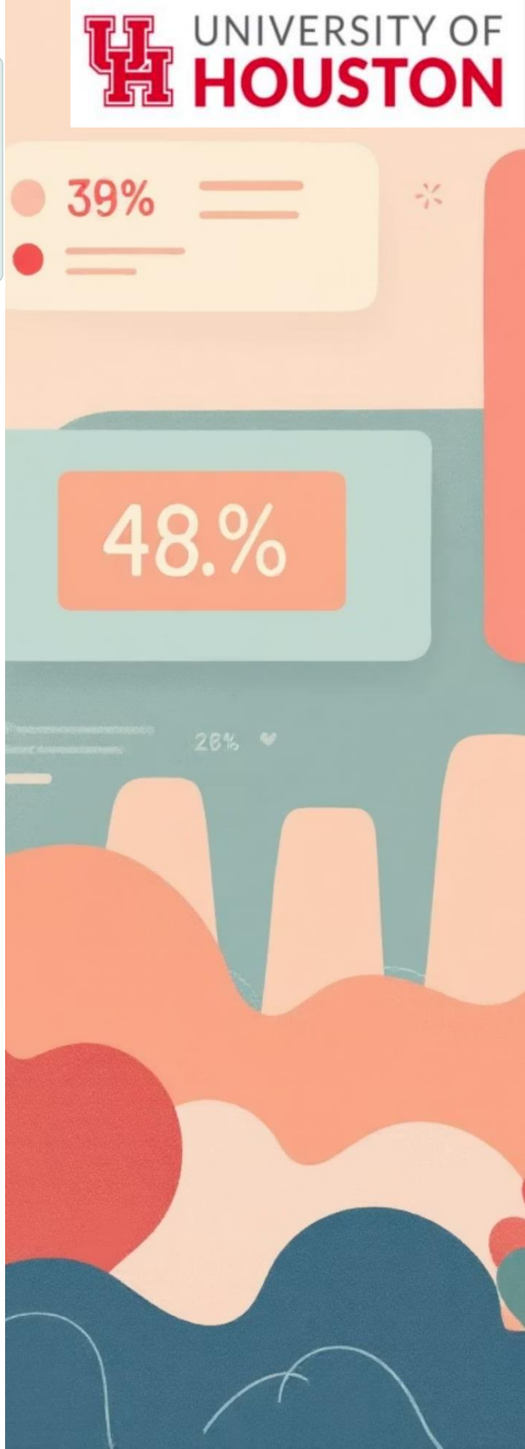
Measures if the predicted answer exactly matches the ground truth.

F1-score

Calculates the overlap between the predicted and true answers, considering partial matches.

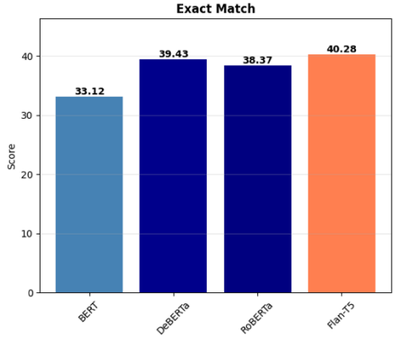
BERTScore

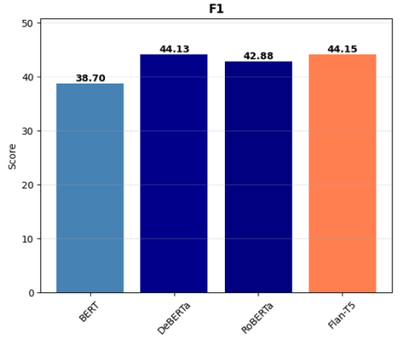
Leverages BERT embeddings to compute a semantic similarity score, capturing lexical and semantic alignment.

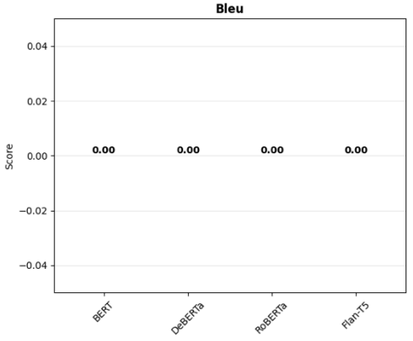


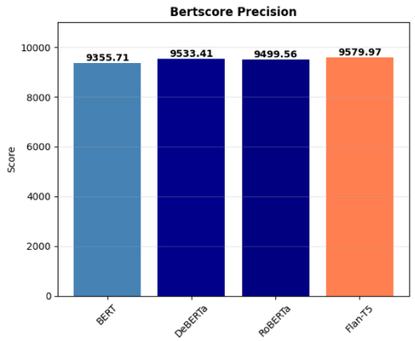
UNIVERSITY OF HOUSTON

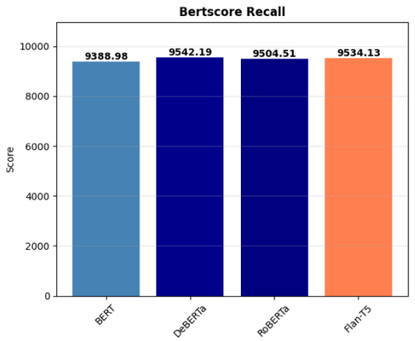
Metric Visualizations

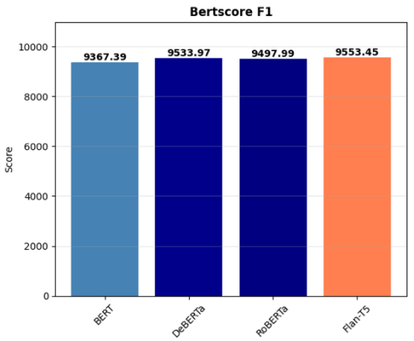


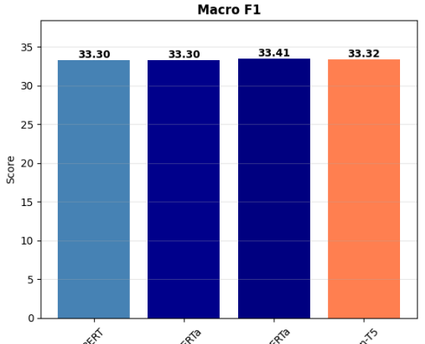


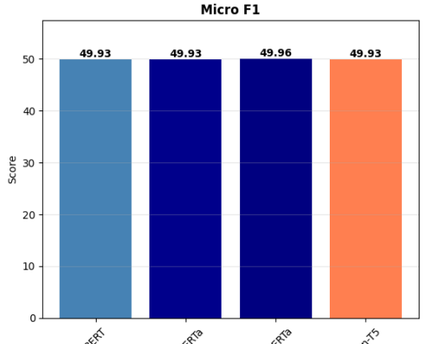


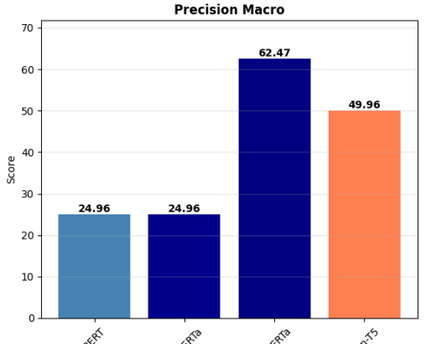


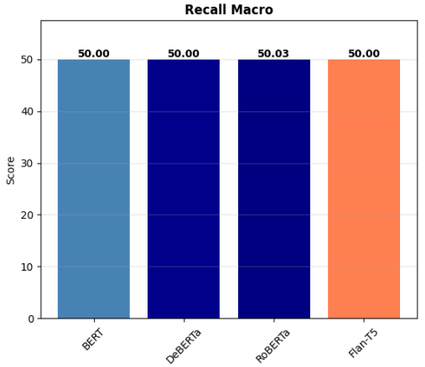


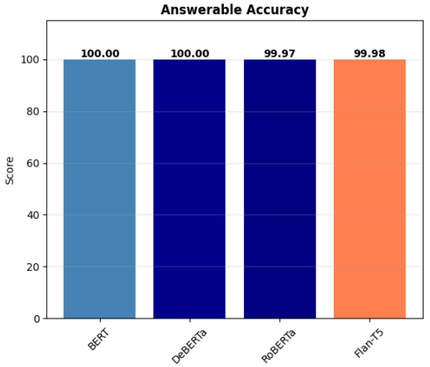


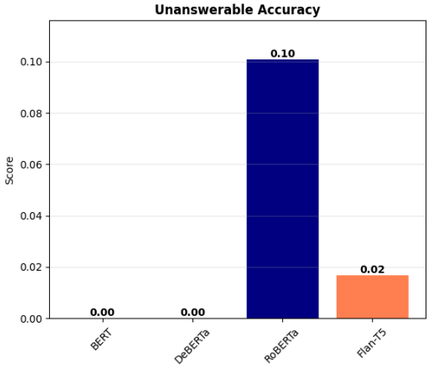




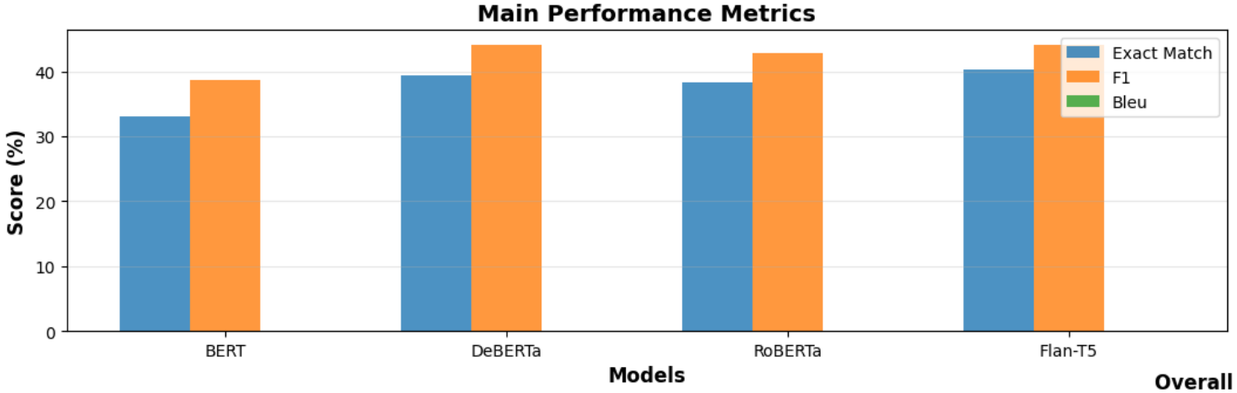


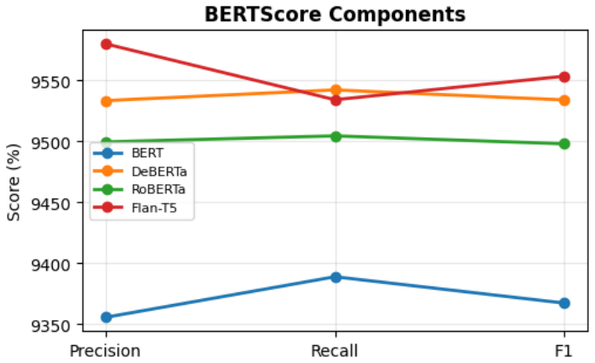


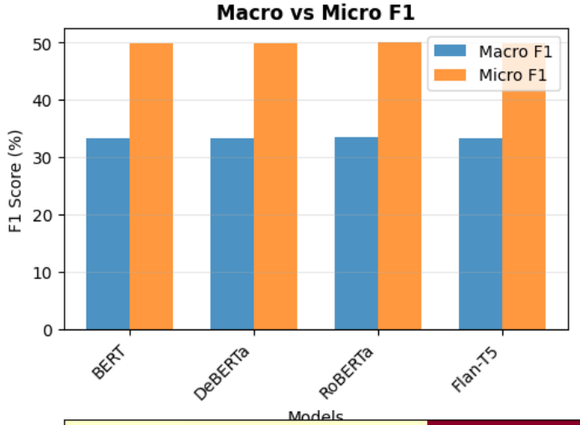


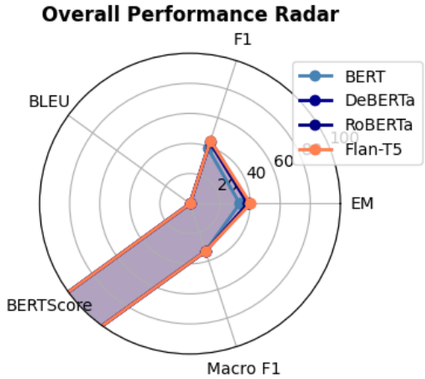


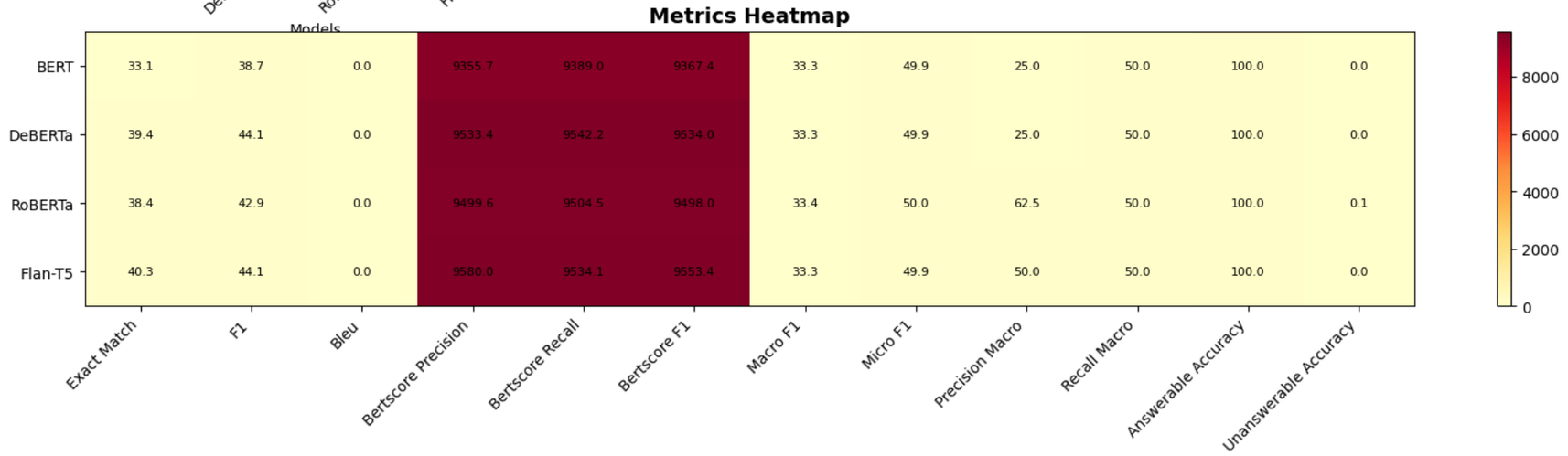
Detailed Comparison Plots











UNIVERSITY OF HOUSTON

Results Summary:

📊 Comprehensive Test Set Results:

Model	Type	exact_match	f1	bleu	bertscore_precision	bertscore_recall	bertscore_f1	macro_f1	micro_f1	precision_macro	recall_macro	answerable_accuracy	unanswerable_accuracy
BERT	Extractive	33.117157	38.700195	0.0	9355.714345	9388.975395	9367.388176	33.301500	49.928409	24.964204	50.000000	100.000000	0.000000
DeBERTa	Extractive	39.434010	44.126860	0.0	9533.409402	9542.187844	9533.973458	33.301500	49.928409	24.964204	50.000000	100.000000	0.000000
RoBERTa	Extractive	38.372779	42.877292	0.0	9499.559281	9504.512891	9497.994121	33.406022	49.962099	62.472609	50.033593	99.966262	0.100925
Flan-T5	Generative	40.284680	44.147216	0.0	9579.966860	9534.129794	9553.445915	33.316439	49.928409	49.964198	49.999976	99.983131	0.016821

📊 COMPREHENSIVE MODEL COMPARISON (3 CORE MODELS)

Model	Type	EM (%)	F1 (%)	BLEU	BERTScore F1 (%)	Answerable Acc (%)	Unanswerable Acc (%)
BERT	Extractive	33.12	38.70	0.00	9367.39	100.00	0.00
DeBERTa	Extractive	39.43	44.13	0.00	9533.97	100.00	0.00
RoBERTa	Extractive	38.37	42.88	0.00	9497.99	99.97	0.10

✅ Comparison table saved to: /content/drive/MyDrive/squad_final_checkpoints/final_model_comparison.csv

🏆 BEST PERFORMERS:
- Best F1 Score: DeBERTa (44.13%)
- Best Answerable: BERT (100.00%)
- Best Unanswerable: RoBERTa (0.10%)

📝 GENERATIVE MODEL (OPTIONAL):

Flan-T5:
- EM: 40.28%
- F1: 44.15%
- BLEU: 0.00
- Answerable Acc: 99.98%
- Unanswerable Acc: 0.02%

Model Performance Rankings (by F1 Score):
1. Flan-T5 – F1: 44.15%, EM: 40.28%
2. DeBERTa – F1: 44.13%, EM: 39.43%
3. RoBERTa – F1: 42.88%, EM: 38.37%
4. BERT – F1: 38.70%, EM: 33.12%

🏆 Best Overall Model:
Flan-T5 with F1=44.15%

UNIVERSITY OF HOUSTON

Demo

User Interface (UI):

🎓 Multi-Model Question Answering System

Compare BERT, DeBERTa, RoBERTa & Flan-T5 on SQuAD 2.0

Final Project: Natural Language Processing | Group 9

📊 System Status

- **Models Loaded:** 4/4 (BERT, DeBERTa, RoBERTa, Flan-T5)
- **Device:** cuda
- **Ready to Answer Questions!** ✅

❓ **Your Question**

what is my career

📄 **Context (paste relevant text here)**

i want to do my masters and become a lecturer in my university

🚀 **Get Answers** | 🔄 **Clear**

🤖 Model Predictions:

🔵 **BERT**

`masters and become a lecturer`

🟢 **DeBERTa**

`lecturer`

🟣 **RoBERTa**

`lecturer`

🟡 **Flan-T5**

`professor`

UNIVERSITY OF HOUSTON

Challenges:

- Leveraging powerful tools for development and training.

Google Colab Pro

- GPU memory constraint - Primary environment for development and training.

GPU Acceleration

- Took a very long time for execution (Used 10k subset as solution).

Python Notebooks

- Managed all code cells, training, and evaluation for reproducibility.

UNIVERSITY OF HOUSTON

QnA

UNIVERSITY OF HOUSTON

Thank You