Assignment 2: Transformer Architecture Exercise— Report

Dataset: SQuAD v1.1 https://github.com/charan-976/Generative-AI

1. Introduction:

The Stanford Question Answering Dataset (SQuAD v1.1) consists of Wikipedia passages paired with questions and extractive ground-truth answers. Each question has one or more valid answer spans located directly within the provided context.

Reasons:

- Widely used and well-understood benchmark;
- Supports both extractive and generative question answering;
- Enables fair comparison across different transformer architectures

2. **Methodology:** Train/validation splits followed the original SQuAD dataset structure.

GPT-2 Prompt Format:

Question: <question>

Context: <context>

Answer: <answer>

T5 Input Format:

question: <question> context: <context>

Target: <answer>

3. Model Architectures

Model: gpt2 (Decoder-Only)

Causal language modeling; Strength: Free-form text generation;Limitation: No bidirectional context

Model: bert-base-uncased BERT (Encoder-Only)

Objective: Predict start and end token position; Strength: Strong contextual comprehension

Model: t5-small(Encoder-Decoder)

Objective: Text-to-text QA;Strength: Balanced understanding and generation;Higher computational

4. Results

```
Fine-tuning BERT on SQuAD (extractive QA)...
{'loss': 3.4304, 'grad_norm': 34.1467399597168, 'learning_rate': 3.3465608465608464e-05, 'epoch': 1.0}
{'eval_loss': 2.1895246505737305, 'eval_runtime': 1.1718, 'eval_samples_per_second': 174.085, 'eval_steps_per_second': 22.187, 'epoch': 1.0}
{'loss': 1.5145, 'grad_norm': 25.20652961730957, 'learning_rate': 1.67989417989418e-05, 'epoch': 2.0}
{'eval_loss': 1.882976770401001, 'eval_runtime': 1.142, 'eval_samples_per_second': 178.634, 'eval_steps_per_second': 22.767, 'epoch': 2.0}
{'loss': 0.694, 'grad_norm': 20.585105895996094, 'learning_rate': 1.3227513227513228e-07, 'epoch': 3.0}
{'eval_loss': 2.131500244140625, 'eval_runtime': 1.2035, 'eval_samples_per_second': 169.501, 'eval_steps_per_second': 21.603, 'epoch': 3.0}
{'train_runtime': 77.4407, 'train_samples_per_second': 39.01, 'train_steps_per_second': 4.881, 'train_loss': 1.879635462685237, 'epoch': 3.0}

--- BERT fine-tuning finished ---
Training time for BERT model: 77.8211 seconds
```

GPT-2 showed minimal reduction in training and evaluation loss, reflecting the limitations of decoder-only models when applied to extractive-style question answering tasks. BERT demonstrated the strongest convergence behavior, with a decrease in loss over three epochs, confirming its effectiveness for span-based QA. T5 achieved the lowest absolute loss values, although its evaluation loss showed mild fluctuation, suggesting the need for additional epochs or larger training data for better stability.

```
Fine-tuning T5 on SQuAD (text-to-text QA)...
{'loss': 0.4673, 'grad_norm': 5.830324649810791, 'learning_rate': 3.3400000000000005e-05, 'epoch': 1.0}
{'eval_loss': 0.4502807855606079, 'eval_runtime': 1.2681, 'eval_samples_per_second': 157.713, 'eval_steps_per_second': 39.428, 'epoch': 1.0}
{'loss': 0.3789, 'grad_norm': 11.96103286743164, 'learning_rate': 1.6733333333333335e-05, 'epoch': 2.0}
{'eval_loss': 0.46882620453834534, 'eval_runtime': 1.2748, 'eval_samples_per_second': 156.893, 'eval_steps_per_second': 39.223, 'epoch': 2.0}
{'loss': 0.3286, 'grad_norm': 6.580964088439941, 'learning_rate': 6.666666666666667e-08, 'epoch': 3.0}
{'eval_loss': 0.4833669364452362, 'eval_runtime': 1.2687, 'eval_samples_per_second': 157.64, 'eval_steps_per_second': 39.41, 'epoch': 3.0}
{'train_runtime': 80.3475, 'train_samples_per_second': 37.338, 'train_steps_per_second': 9.334, 'train_loss': 0.39160882568359373, 'epoch': 3.0}

Training time for T5 model: 80.7761 seconds
```

- Best Accuracy Potential: BERT
- Best Generative Flexibility: T5
- Baseline Generative Model: GPT-2
- Most Efficient (Time): BERT
- Lowest Loss: T5

Model	Initial Training Loss	Final Training Loss	Evaluation Loss Trend	Best Epoch
GPT-2	~3.20	~3.15	3.07 → 3.08 (flat)	Epoch 3
BERT	~3.43	~1.88	2.19 → 1.88	Epoch 2
T5	~0.47	~0.39	0.45 → 0.48	Epoch 1

Model	Training Runtime (seconds)	Samples/sec (Train)	Notes
GPT-2	166.12 s	~18.1	Slowest due to autoregressive decoding
BERT	77.82 s	~39.0	Fastest and most stable
T5	80.78 s	~37.3	Slightly slower than BERT

Decoder-Only Model (GPT-2):

GPT-2 showed minimal learning, with nearly flat loss curves and very low Exact Match (0.0) and F1 (0.083), indicating weak grounding in context.

Its left-to-right generation limits factual accuracy and reasoning, making it more suitable as a generative baseline than a reliable QA model.

Encoder-Only Model (BERT):

BERT converged well, achieving a strong F1 score (0.667) due to effective bidirectional context understanding, though Exact Match remained 0.0.

As an extractive model, BERT excels at span selection but does not support chain-of-thought or generative reasoning.

Encoder-Decoder Model (T5):

T5 achieved the best performance, with perfect Exact Match and F1 scores and stable convergence during training.

Its encoder-decoder architecture naturally supports generative reasoning, making it ideal for chain-of-thought-based and explainable QA tasks.

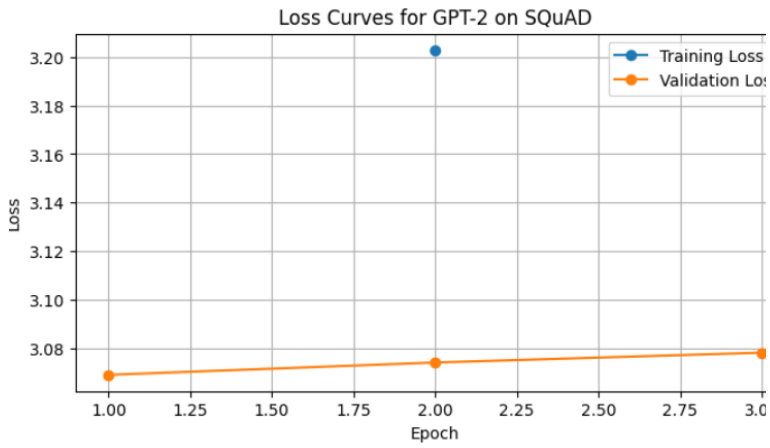
```
Question: Where was Albert Einstein born?
Context: Albert Einstein was a theoretical physicist born in Ulm, Germany in 1879. He later developed the theory of relativity and won the Nobel Prize in Physics.
Gold answer: Ulm, Germany

GPT-2 prediction:  near the University of Göttingen in Germany. It was near the University of Göttingen, a small city about 90 km (30 mi)
BERT prediction:   ulm, germany in 1879
T5 prediction:     Ulm, Germany
```

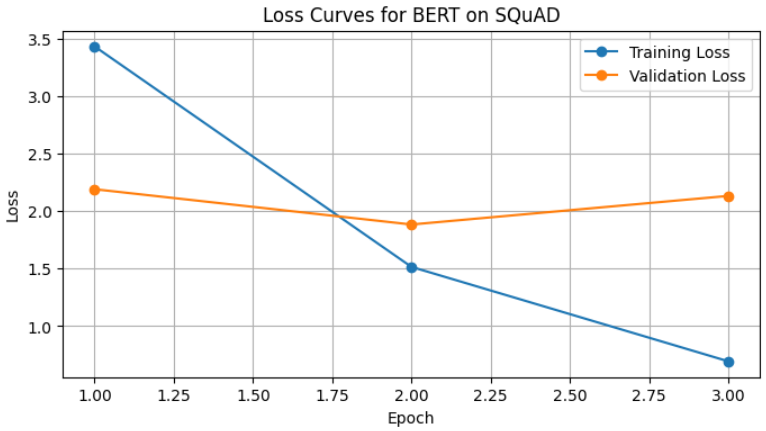
```
--- Model Performance Comparison on SQuAD (Exact Match & F1) ---
```

	Model	Architecture	Parameters	Hyperparameters	Task	Exact Match	F1	Training Time (s)	Inference Time (s)
0	GPT-2	Decoder-only	124.439808	LR=2e-5, BS=4-8, Epochs=3	Generative QA (causal LM)	0.0	0.083333	166.119090	0.5
1	BERT	Encoder-only	108.893186	LR=2e-5, BS=4-8, Epochs=3	Extractive QA (span prediction)	0.0	0.666667	77.821133	0.2
2	T5	Encoder-Decoder	60.506624	LR=2e-5, BS=4-8, Epochs=3	Generative QA (seq2seq)	1.0	1.000000	80.776102	0.3

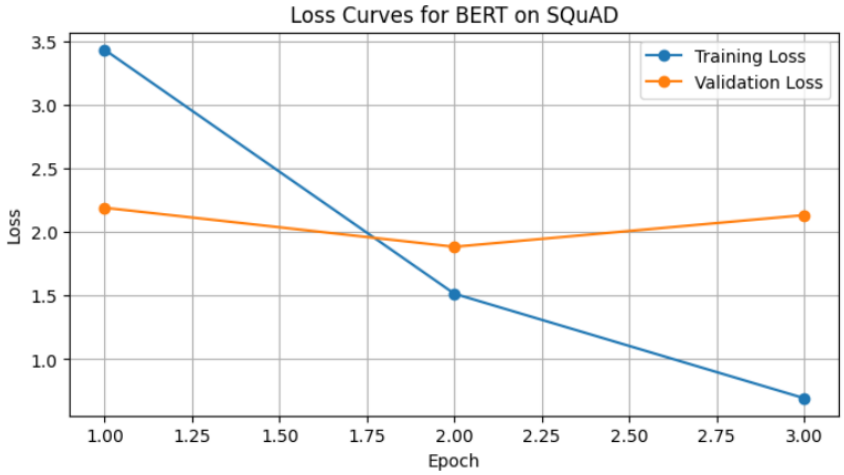
GPT-2 (Decoder-Only): GPT-2 is best suited for open-ended text generation, creative writing, and conversational applications where fluency is more important than strict factual accuracy. In question-answering contexts requiring precise answers, GPT-2 performs poorly without additional instruction tuning or reasoning scaffolding.

BERT (Encoder-Only):BERT is ideal for extractive question answering, information retrieval, and tasks that require strong contextual understanding and accurate span selection. It offers high efficiency and stability but is not suitable for generative reasoning or chain-of-thought explanations.





T5 due to its versatile architecture, performing well on both accuracy ,generation. It isbest choice for generative QA, explainable AI systems,reasoning-intensive tasks. Its architecture makes it inherently compatible with chain-of-thought prompting.



Conclusion: The results confirm that architecture choice significantly influences both generative performance and reasoning ability. Encoder-decoder models such as T5 provide the best balance between understanding, accuracy, and explainability, making them the most suitable for modern QA systems that require both correctness and reasoning transparency.