Part 1: Open Questions

Question 1:

1. Stanford Question Answering Dataset (SQuAD):

This dataset consists of questions on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. By this dataset, we can evaluate the **coreference resolution** ability of the model.

2. Zero-Shot Relation Extraction via Reading Comprehension

The dataset consists of sentence and questions templates about a relation in the text. The dataset was published in an article that showed that **relation extraction** can be reduced to the problem of answering simple reading comprehension questions

3. QA-SRL: Question-Answer Driven Semantic Role Labeling

The dataset uses question-answer pairs to model **semantic role labeling**. The questions start with wh-words (Who, What, Where, When etc.) and contain a verb predicate in the sentence; the answers are phrases in the sentence

Question 2:

a. Inference-time scaling seeks to improve a model's performance not by increasing its size or training data, but by allocating more resources during inference, like running more generation passes, allowing longer chain-of-thought (CoT) sequences, or extending computation time.

In the lecture, we discussed several methods of inference time scaling:

1. Self-Consistency

Description: Instead of producing a single output, the model generates multiple independent reasoning chains (via CoT prompts) and we select the answer that appears most often (the answer is most consistent).

Advantages:

- Leverages the intuition that a complex reasoning problem typically admits multiple different ways of thinking, leading to its unique correct answer.
- Empirical evaluation shows that self-consistency boosts the performance of chain-of-thought prompting.
- This method reduces the impact of any single erroneous chain.

Computational Bottlenecks

- Requires repeated forward passes for each sample, increasing total compute.
- Needs memory to store all candidate outputs before voting

Parallelizability: Fully parallelizable, since each CoT generation is independent and can be executed concurrently..

2. Verifiers

Description: Like self-consistency, this approach generates multiple candidate outputs. Rather than relying on majority voting, each candidate is passed through a set of pre-built verification modules. The final answer is the one that passes the greatest number of these verifiers.

Advantages in addition to self-constitution's advantages:

- Can be performed on unstructured or semi-structured outputs (e.g., code snippets, free-form text).
- Enhances interpretability: you can inspect which verifiers each candidate passed or failed.

Computational Bottlenecks

- Depends on the resource utilization of the verifying unit

Parallelizability:

- CoT generations remain fully parallelizable.
- Verification steps can also run concurrently, provided the modules operate independently and do not form a sequential pipeline.

3. Improved Chain-of-Thought (CoT)

Description: In addition to the regular breakdown of steps the model recognize and correct it mistakes and tries different approch when the correct one does not work.

Advantages: Fewer wasted samples, you don't need to throw away an entire CoT when only one sub-step was wrong.

Computational Bottlenecks – Increase computing time and output length

Parallelizability – No. There is one context and one generation.

b. To solve a complex scientific task on a single GPU with large memory, I would choose Improved Chain-of-Thought (CoT). Self-consistency and Verifiers benefit from multi-GPU parallelism, which we don't have in this setup. Improved CoT focuses computation on fixing errors within a single reasoning path. It uses fewer forward passes overall and better leverages the GPU's memory by avoiding redundant full-chain generations, making it more efficient in this setup.

Part 2: Programming Exercise

GitHub repository link - https://github.com/yehonatan1305/67664-ANLP-EX1

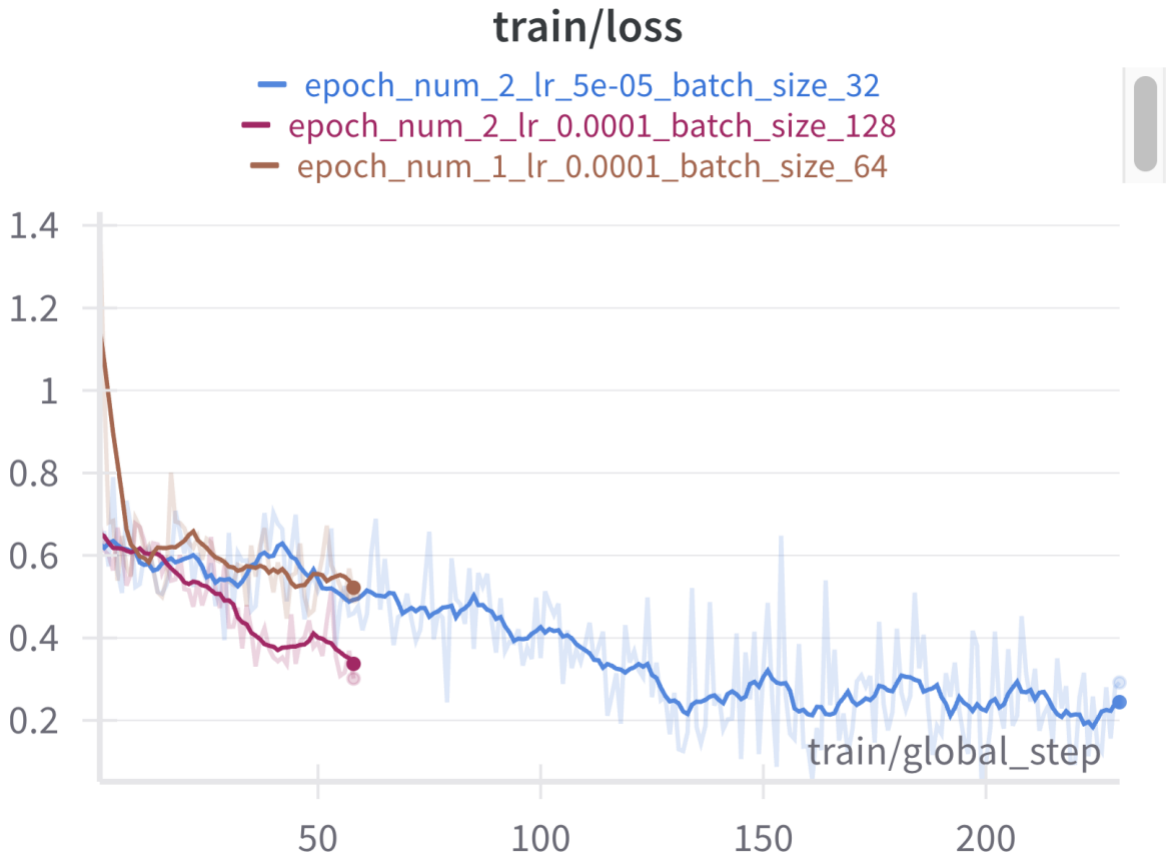
I have trained the model on three sets of hyperparameters. Every training process was performed on the entire training set and validated on the entire validation set.
The prediction stage was done on the whole test set as well.
Below you can see a table that summarizes the results for each model and a ranking of each model by each metric (lower is better).
We can see that using 2 epochs, learning rate 5e-5, and batch size 32 achieved the best score on the entire 3 metrics.

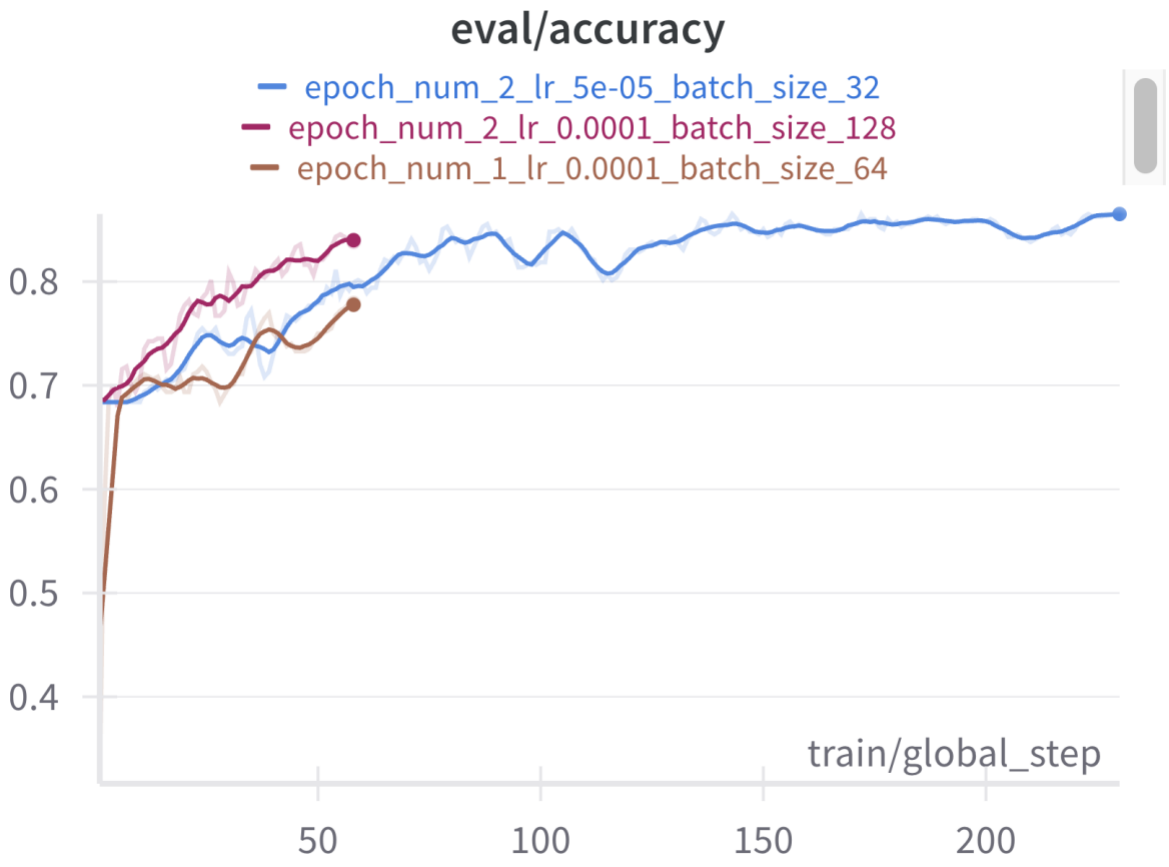
Number of epochs	Learning rate	Batch size	Train loss ranking	Validation accuracy ranking	Test accuracy ranking
2	5e-5	32	1	1	1
2	1e-4	128	2	2	2
1	1e-4	64	3	3	3

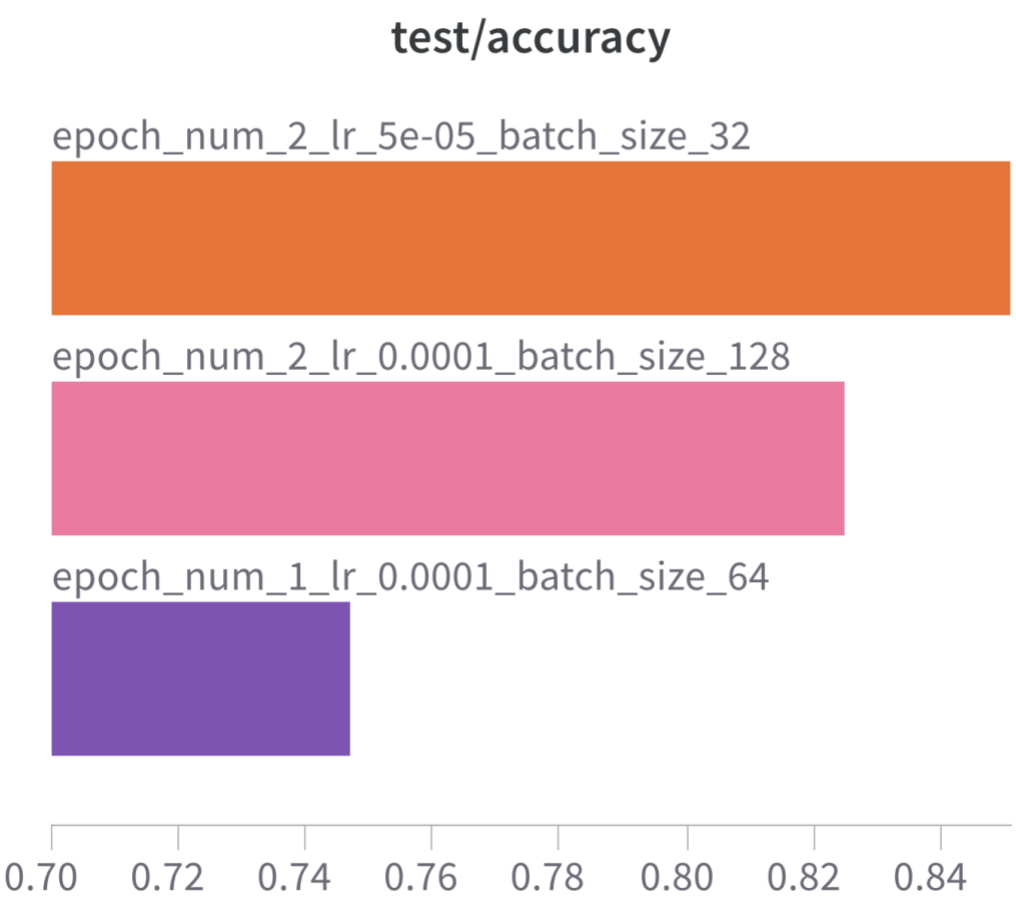
I examined five examples where the best model predicted correctly and the worst model did not. The worst model's false positives often involved high lexical overlap between sentences that were semantically related but not true paraphrases. Its false negatives typically require numerical reasoning or inference, such as understanding that a $300 increase corresponds to a rise from $2,500 to $2,800

```
Label: 0
Best Model Prediction: 0, Worst Model Prediction: 1
Sentence: No dates have been set for the civil or the criminal trial . ||
No dates have been set for the criminal or civil cases , but Shanley has
pleaded not guilty .
--------------------------------------------------
Label: 0
Best Model Prediction: 0, Worst Model Prediction: 1
Sentence: " Sanitation is poor ... there could be typhoid and cholera , "
he said . || " Sanitation is poor , drinking water is generally left
behind . . . there could be typhoid and cholera . "
--------------------------------------------------
Label: 1
Best Model Prediction: 1, Worst Model Prediction: 0
Sentence: Last month Intel raised its revenue guidance for the quarter to
between $ 7.6 billion and $ 7.8 billion . || At the end of the second
quarter , Intel initially predicted sales of between $ 6.9 billion and $
7.5 billion .
--------------------------------------------------
Label: 1
Best Model Prediction: 1, Worst Model Prediction: 0
Sentence: At community colleges , tuition will jump to $ 2,800 from $
2,500 . || Community college students will see their tuition rise by $ 300
to $ 2,800 or 12 percent .
--------------------------------------------------
Label: 0
```

```
Best Model Prediction: 0, Worst Model Prediction: 1
Sentence: The civilian unemployment rate improved marginally last month --
slipping to 6.1 percent -- even as companies slashed payrolls by 93,000 .
|| The civilian unemployment rate improved marginally last month _ sliding
down to 6.1 percent _ as companies slashed payrolls by 93,000 amid
continuing mixed signals about the nation 's economic health .
--------------------------------------------------
```




eval/accuracy
epoch_num_2_lr_5e-05_batch_size_32
epoch_num_2_lr_0.0001_batch_size_128
epoch_num_1_lr_0.0001_batch_size_64
0.8
0.7
0.6
0.5
0.4
train/global_step
50
100
150
200


test/accuracy
epoch_num_2_lr_5e-05_batch_size_32
epoch_num_2_lr_0.0001_batch_size_128
epoch_num_1_lr_0.0001_batch_size_64
0.70
0.72
0.74
0.76
0.78
0.80
0.82
0.84