Naftali Grossman
209768688
ANLP Ex1
Public Github Repo
`https://github.com/Naftali-Grossman/ANLP-ex1.git`

1.1
SQuAD – Abstracted by human annotators created question-answer pairs based on Wikipedia paragraphs. Through this we allow it to evaluate core reading comprehension by requiring a model to extract an answer span from a passage given a natural-language question.
SRL - Can target a model's ability to recognize who did what to whom**,** when**,** and why, giving it a great understanding of relationship within text.
NER – Is great for it's ability to recognize and categorize entities

1.2.1
Self-Consistency-
Brief Description:
Instead of a single chain of thought output, we generate multiple CoT outputs and choose the majority

Advantages:
Increased answer reliability
Prevents random reasoning errors

Computational Bottlenecks:
Increased inference time and cost

Parallelizable:
Yes

Verifiers (Best of N/Rejection Sampling)
Brief Description:
Trains an automatic verifier to check whether the generated output was correct. Afterwards we take the best/majority verified answers.

Advantages:
More quality outputs

Computational Bottlenecks:
Increased computation and training on the verifiers

Parallelizable:

Yes

Backtracking/Self-Evaluation:
Brief Description:
At inference time, if needed, the model will dynamically backtrack and attempt a different solution based on self evaluation.

Advantages:
Mistakes can be dynamically corrected
Improved Accuracy

Computational Bottlenecks:
Requires more memory and computing power

Parallelizable:
No

Problem Decomposition:
Brief Description:
Model breaks each task into sub tasks and solves each sub-task in order to solve task

Advantages:
Able to solve more complex tasks

Computational Bottlenecks:
Requires more memory and takes more time

Parallelizable:
No

1.2.2
I would likely choose the self consistency method as it has been proven to be able to solve complex tasks and is parallelizable thus allowing to run with a single GPU

2.1 For me the Configuration for where I got highest validation was on the 3^{rd} epoch for a model in which I had a learning rate of 2e-5 and a train batch size of 16. In total for this model we ran 5 epochs. Validation accuracy was 0.882353. This was not the highest test accuracy as this was achieved by the model which had a learning rate of 5e-5, and a train batch size of 16 on its second epoch

2.2

```
Example 21
Sentence 1: Still , he said , " I 'm absolutely confident we 're going to have a bill . "
Sentence 2: " I 'm absolutely confident we 're going to have a bill , " Frist , R-Tenn . , said Thursday .
Label: 1
Best prediction: 1
Worst prediction: 0

Example 30
Sentence 1: Dynes will get $ 395,000 a year , up from Atkinson 's current salary of $ 361,400 .
Sentence 2: In his new position , Dynes will earn $ 395,000 , a significant increase over Atkinson 's salary of $ 361,400 .
Label: 1
Best prediction: 1
Worst prediction: 0

Example 31
Sentence 1: The daily Hurriyet said the raid aimed to foil a Turkish plot to kill an unnamed senior Iraqi official in Kirkuk .
Sentence 2: The daily Hurriyet said the raid aimed to foil a Turkish plot to kill an unnamed senior Iraqi Kurdish official in Kirkuk , but Gul has denied any Turkish plot .
Label: 0
Best prediction: 0
Worst prediction: 1

Example 32
Sentence 1: " It was a little bit embarrassing the way we played in the first two games , " Thomas said .
Sentence 2: " We 're in the Stanley Cup finals , and it was a little bit embarrassing the way we played in the first two games .
Label: 0
Best prediction: 0
Worst prediction: 1

Example 35
Sentence 1: The women said victims of rape who came forward routinely were punished for minor infractions while their assailants escaped judgment .
Sentence 2: The women said victims of rape who came forward were routinely punished for minor infractions while their attackers escaped judgment , prompting most victims to remain silent .
Label: 0
Best prediction: 0
Worst prediction: 1

Example 40
Sentence 1: Cisco pared spending during the quarter to compensate for sluggish sales .
Sentence 2: In response to sluggish sales , Cisco pared spending .
Label: 1
Best prediction: 1
Worst prediction: 0

Example 44
Sentence 1: On July 22 , Moore announced he would appeal the case directly to the U.S. Supreme Court .
Sentence 2: Moore of Alabama says he will appeal his case to the nation 's highest court .
Label: 1
Best prediction: 1
Worst prediction: 0
```

This is an example of sentence pairs in which our worst configuration (WC) got the prediction wrong but the best configuration (BC) got the prediction right.

Judging by all the examples which were witnessed (where the BC was correct and the WC was incorrect), it seems in the case where the two sentences did in fact have the same meaning, these examples would show up when the two sentences were not similar in terms of text.
Example:

```
Example 40
Sentence 1: Cisco pared spending during the quarter to compensate for
sluggish sales .
Sentence 2: In response to sluggish sales , Cisco pared spending .
Label: 1
Best prediction: 1
Worst prediction: 0
```

But when the two sentences did not in fact have the same meaning, the two sentences were often very similar in text:
Example:

```
Example 31
Sentence 1: The daily Hurriyet said the raid aimed to foil a Turkish plot
to kill an unnamed senior Iraqi official in Kirkuk .
Sentence 2: The daily Hurriyet said the raid aimed to foil a Turkish plot
to kill an unnamed senior Iraqi Kurdish official in Kirkuk , but Gul has
denied any Turkish plot .
Label: 0
Best prediction: 0
Worst prediction: 1
```

This is likely due to the worst configuration relying on similarity of words and not relying enough on similarity of meaning