Itamar Cohen – 318432382 – ANLP – ex1

Part 1:

A – 3 QA Models:

1) SQuAD
   SQuAD requires models to read a passage and answer questions based on it, testing reading comprehension. This is an intrinsic task because it directly evaluates a model's ability to understand and reason over natural language text.

2) BoolQ
   BoolQ contains yes/no questions paired with a passage, where the model has to decide if the answer is "yes" or "no" based on the given text. It reflects intrinsic language understanding as it challenges the model to assess little detailes and infer truth based on context.

3) ReCoRD
   ReCoRD involves answering questions by filling in masked entities in text, combining reading comprehension with commonsense reasoning. This task measures intrinsic understanding by testing how well the model uses contextual and world knowledge to infer correctly from the text.

B – The main techniques we discussed in class are:

1) Verifiers:
   Using additional model to check or score the output of the main model before accepting it. Can be used for Unitesting.
   Advantages – can improve correctness by rejecting a faulty or low quality output, and allows modular evaluation pipelines
   Bottlenecks – The varifier model adds additional compute time and resources, and can increase time/resources especially if there are several varifiers or the varifier is large.
   Parallelizable – Partially, each varifier can check one output at a time (but we can aggregate the outputs to make it more cost efficient), but each varifier operates on it's own if there aremore than one.
2) Self Consistency:
   instead of relying on single output from a stochastic model, generating multiple outputs and selecting the most consistent using majority vote or aggregation.

Advantages – improves output reliability

Bottlenecks – requires several forward passes, and cost grows linearly with the number of samples

paralliazable – Yes, multiple outputs can be generated in parallal.

3) Computing budgeting:

Efficient use of budget – we discussed using several small models and take the majority decision or the best one.

Advantages- often yields better results than to train one large model, thus improving accuracy.

Bottlenecks – needs to store several answers until the decision time. Also, need to run many small models (increased scale for runtime)

Parallelization – Yes, outputs of each model do not depend on each other.

4) Chain of Thought Methods:

guide the model to explicitly generate reasoning steps in the way to the solution, increasing the output of the model.

Advantage – improved reasoning, makes model more interperatable and debuggable, and an amplify correctness.

Bottlenecks – significant increase in “attention” stage cost, as there are many forward passes. Also, requires multiple generations per input.

Parallelization – possible in term of generations per input in some cases, but sequential in some cases where there’s a need to connect several logic steps.

5) Least-To-Most:

decomposing questions into sub questions and then sequentially solve sub questions.

Advantages – raises model accuracy greatly

bottleneck – generates much more tokens than a regular model. Can run out of space for answers.

Parallelization – No, as each question is solved sequentially.

6) B-

7) Lets say we have complex scientific task – probably logic-intensive with several analysis steps – and we have access to one gpu – meaning no parallelization – and a large memory – enabling long sequences/batches.

This is a classic case for a chain of thought model, and specifically Least-to-Most in that context. We’ve discussed it’s proven success in scientific olympiad-level questions. That is because it offers a structured way to approach a complicated task by breaking it down to several sequential tasks and solving them sequentially.

Our lack of more than one gpus does not interrupt this method in any way, and our access to large memory enables the model to perform many sub questions one by one and provide a complete and well-thought-out question.
if there is no human involved in the task (to break down the prompt) – I would then use self ask instead which imitates this process, but by the model itself. Same advantages apply.

Part 2:

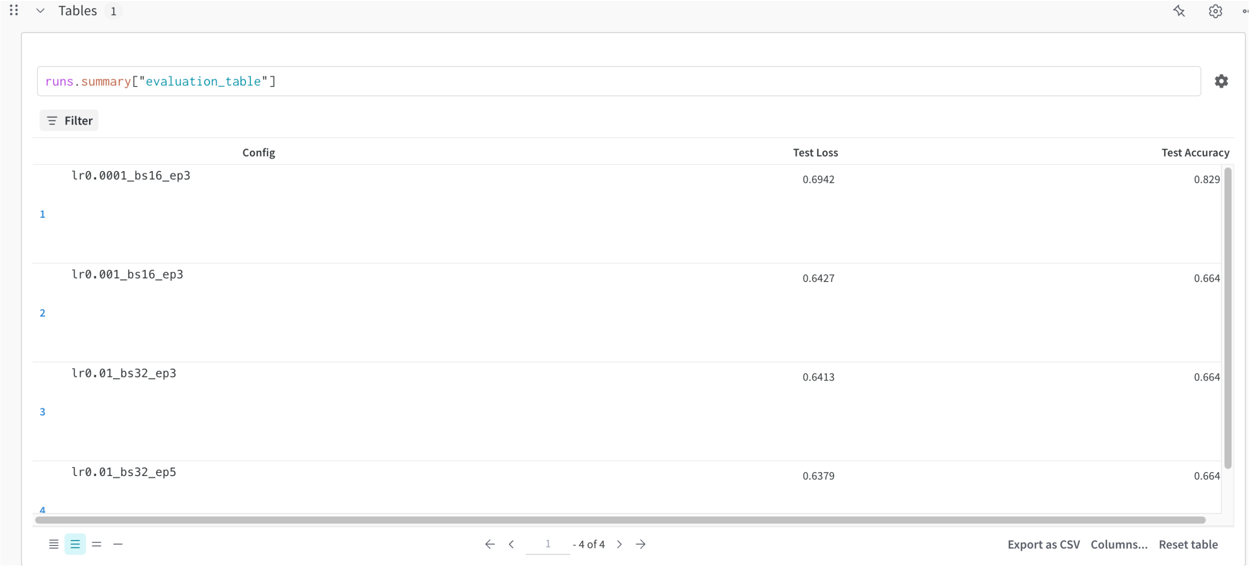
A-

In this table, we can see that the best configuration on the validation:

learning rate 0.01
batch size 16
epochs 3

Did in fact perform the best on the test data out of all other tested configurations, by a significant margin.

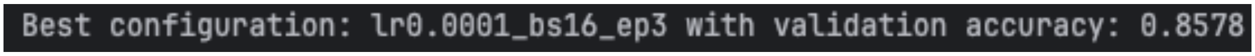
I attempted many configurations. The main conclusion I reached is that under 5 epochs, learning rate of 0.01 or even 0.001 is too aggressive, and all tries with these learning rate ended with the ee=xacr same accuracy, which is the baseline accuracy of predicting everything as 0 or everything as 1 (0 in our case)



```
runs.summary["evaluation_table"]
```

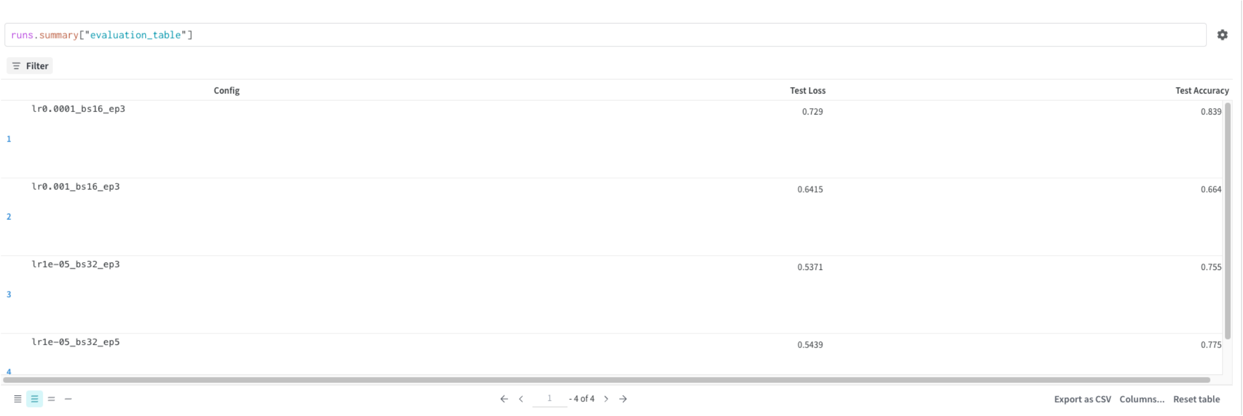
	Config	Test Loss	Test Accuracy
1	lr0.0001_bs16_ep3	0.6942	0.829
2	lr0.001_bs16_ep3	0.6427	0.664
3	lr0.01_bs32_ep3	0.6413	0.664
4	lr0.01_bs32_ep5	0.6379	0.664

After this attemt, I tried running lower learning rates that will be good for testing and interesting in the second part of the exercise.
in training, the best configuration was:



```
Best configuration: lr0.0001_bs16_ep3 with validation accuracy: 0.8578
```

The different configurations performances on the validation set were as follows:

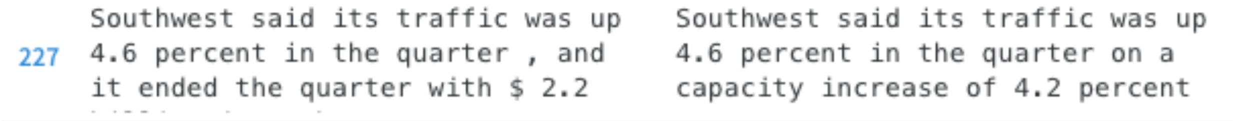


runs.summary["evaluation_table"]

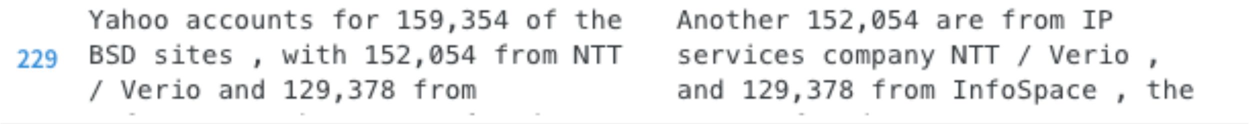
	Config	Test Loss	Test Accuracy
1	lr0.0001_bs16_ep3	0.729	0.839
2	lr0.001_bs16_ep3	0.6415	0.664
3	lr1e-05_bs32_ep3	0.5371	0.755
4	lr1e-05_bs32_ep5	0.5439	0.775

It is still the best configuration, and I want to compare it to the worst configuration that “actually learned something” (better than 0.664) – so I compared it tolr1e-05_bs32_ep3.

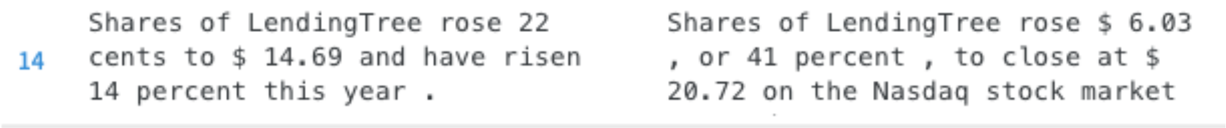
These are some of the examples where the best model succeeded, and the worst failed:



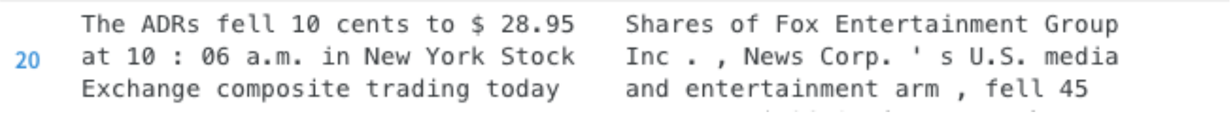
227	Southwest said its traffic was up 4.6 percent in the quarter , and it ended the quarter with $ 2.2	Southwest said its traffic was up 4.6 percent in the quarter on a capacity increase of 4.2 percent



229	Yahoo accounts for 159,354 of the BSD sites , with 152,054 from NTT / Verio and 129,378 from	Another 152,054 are from IP services company NTT / Verio , and 129,378 from InfoSpace , the



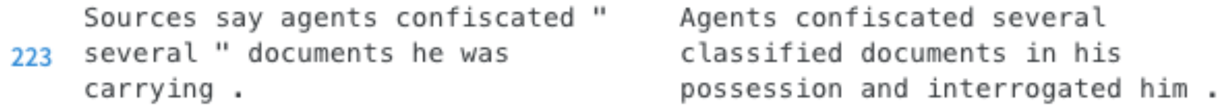
14	Shares of LendingTree rose 22 cents to $ 14.69 and have risen 14 percent this year .	Shares of LendingTree rose $ 6.03 , or 41 percent , to close at $ 20.72 on the Nasdaq stock market



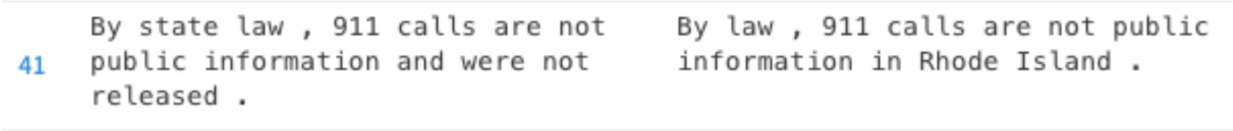
20	The ADRs fell 10 cents to $ 28.95 at 10 : 06 a.m. in New York Stock Exchange composite trading today	Shares of Fox Entertainment Group Inc . , News Corp. ' s U.S. media and entertainment arm , fell 45

In many number-heavy examples, the worst model seemed unsure or confused, often missing subtle rewordings of quantities. The best model, on the other hand, handled these cases well, showing a better grasp of numerical phrasing and meaning.

In addition, in many cases as such:



223	Sources say agents confiscated " several " documents he was carrying .	Agents confiscated several classified documents in his possession and interrogated him .



41	By state law , 911 calls are not public information and were not released .	By law , 911 calls are not public information in Rhode Island .

there are small detailes that are changed, or were added at the end of the sentence, the worst model wasn’t able to grasp that this is not paraphrasing, as completely new information was added in one sentence. The best model was able to comprehand this and label these examples correctly.

So to summarize, the worse model got confused by cases which presented added information, and cases where there were are many numbers and much data.