Question 1

- 1. Question answering (QA) can be an expressive format for annotating both intrinsic as well as extrinsic tasks. List three QA datasets that use QA to annotate intrinsic concepts. For each, write a short explanation (1-2 sentences) for why it measures an intrinsic property of language Understanding.

IWhoQA: It is a QA dataset which contains several context-answer pairs for each question. It is designed to evaluate how language models balance intrinsic knowledge reliance. It measures intrinsic knowledge reliance by testing whether models prioritize their parametric memory over conflicting information in provided documents, particularly in long-context scenarios.

SPIQA: it is a scientific QA dataset which focuses on a specific domain. It evaluates the models ability to synthesize visual elements with latent scientific concepts learned while pre-training rather than relying on textual context only.

SCQA: it is a multimodal QA on scientific papers dataset. Unlike context-dependent QA, it shows that models achieve higher accuracy when leveraging domain-specific internal knowledge rather than external retrieval, even without additional pre-training.

- 2. In class we discussed several methods to implement inference-time scaling.
  (a) For each method we covered, answer the following:
  • Provide a brief description of the method.
  • Outline its advantages.
  • Identify its computational bottlenecks (i.e., the resources heavily consumed during its execution).
  • Indicate whether the method can be parallelized.
  (b) Suppose you must solve a complex scientific task requiring reasoning, and you have access to a single GPU with large memory capacity. Which method would you choose, and why?

Self-consistency:
a method that improves Large Language Model (LLM) reasoning by generating multiple, diverse reasoning paths for a given problem and then selecting the most consistent final answer. Self-consistency replaces greedy decoding with multiple, diverse reasoning paths. It uses techniques like sampling to generate multiple chains of thought for the same problem.
Advantage: it does not require additional training or architectural changes.
Bottleneck: Memory overhead for storing all candidate solutions.
Parallelizable: Yes - candidate generations are independent.

Chain-of-thoughts:
CoT prompting encourages LLMs to generate a chain of thought or reasoning steps to arrive at an answer, rather than directly answering a question. This helps improve performance on multi-step reasoning tasks.
Advantage: as in self-consistency it does not require additional training or architectural changes.
Bottleneck: Quadratic memory scaling with sequence length (from longer outputs).
Parallelizable: No - sequential token generation.

Verifiers:
Verifiers are used to evaluate the quality of generated samples and guide the search for better candidates. Verifiers are essentially functions that assess the goodness of generated outputs, often using pre-trained models, and provide a score or feedback. They involve pre-trained models or functions that take the generated sample as input and output a scalar value representing a score or feedback.
Advantage: Enables error correction through iterative filtering.
Bottleneck: model loading (reasoner + verifier) increases memory usage.
Parallelizable: Partial - parallel candidate scoring but sequential refinement.

I would choose the verifier method. The high memory will be valuable for the reasoner and the verifier, and this method is not parallel anyway. It is most suitable for reasoning tasks.

Question 2.1

- Did the configuration that achieved the best validation accuracy also achieve the best test accuracy?

No. In the evaluation accuracy the best configuration was num_train_epochs_5_lr 2e-5_batch_size_16, and in the test accuracy the best configuration was num_train_epochs_5_lr 5e-5_batch_size_32.

- Qualitative analysis: Compare the best and worst performing configurations. Examine validation examples where the best configuration succeeded but the worst failed. Can you characterize the types of examples that were harder for the lower-performing model?

Best configuration was : num_train_epochs_5_lr 5e-5_batch_size_32 with test accuracy of 0.8480.
Worst configuration was: num_train_epochs_5_lr 2e-5_batch_size_16 with test accuracy of 0.82957.

An examination of the best and worst configuration by their predictions:

```
Sentence 1: A Washington County man may have the countys first human case of West Nile virus , the health department said Friday .
Sentence 2: The countys first and only human case of West Nile this year was confirmed by health officials on Sept . 8 .
True Label: 1 | Best Pred: 1 | Worst Pred: 0
--------------------------------------------------------------------------------
Sentence 1: During a screaming match in 1999 , Carolyn told John she was still sleeping with Bergin .
Sentence 2: She , in turn , occasionally told John that she was still sleeping with an ex-boyfriend , " Baywatch " hunk Michael Bergin .
True Label: 0 | Best Pred: 0 | Worst Pred: 1
--------------------------------------------------------------------------------
Sentence 1: Licensing revenue slid 21 percent , however , to $ 107.6 million .
Sentence 2: License sales , a key measure of demand , fell 21 percent to $ 107.6 million .
True Label: 1 | Best Pred: 1 | Worst Pred: 0
--------------------------------------------------------------------------------
Sentence 1: Stock futures were mixed in early Thursday trading , but trading below fair value , pointing to a lower open for the major market indexes .
Sentence 2: Stock futures were trading lower early on Thursday , below fair value , pointing to a lower open .
True Label: 1 | Best Pred: 1 | Worst Pred: 0
--------------------------------------------------------------------------------
Sentence 1: Gartner 's report said global WLAN equipment shipments reached 19.5 million last year , a 120 percent increase over 2001 's 8.9 million units .
Sentence 2: Total shipments reached 19.5 million units last year , compared with 8.9 million units in 2001 .
True Label: 1 | Best Pred: 1 | Worst Pred: 0
--------------------------------------------------------------------------------
Sentence 1: The service is deploying Cisco 's BTS 10200 Softswitch cable modem termination system and MGXR 8850 voice gateway products .
Sentence 2: This solution includes the BTS 10200 soft switch , uBR7246VXR cable modem termination system and MGX 8850 voice gateway products .
True Label: 1 | Best Pred: 1 | Worst Pred: 0
--------------------------------------------------------------------------------
Sentence 1: By state law , 911 calls are not public information and were not released .
Sentence 2: By law , 911 calls are not public information in Rhode Island .
True Label: 0 | Best Pred: 0 | Worst Pred: 1
--------------------------------------------------------------------------------
Sentence 1: Police then called a bomb squad , but the device exploded , killing Wells , before bomb technicians arrived .
Sentence 2: While waiting for a bomb squad to arrive , the bomb exploded , killing Wells .
True Label: 1 | Best Pred: 1 | Worst Pred: 0
--------------------------------------------------------------------------------
Sentence 1: Its shares jumped to $ 54.50 in pre-open trading from $ 50.90 at Wednesday 's close .
Sentence 2: Shares jumped almost 7 percent in pre-open trading , rising to $ 18.26 from $ 17.05 at Tuesday 's close .
True Label: 1 | Best Pred: 1 | Worst Pred: 0
--------------------------------------------------------------------------------
Sentence 1: Blair 's government included the charge that Saddam sought uranium from Niger in a September 2002 dossier setting out the case for military action .
Sentence 2: Britain included the accusation in a September 2002 dossier setting out the case for war in Iraq .
True Label: 1 | Best Pred: 1 | Worst Pred: 0
```

I could not find a unique pattern to which the best and worst configurations were divided. In all sentences I sampled it seems that if the sentences were related and if they were not, the worst configuration mis-labeled them (was equally wrong in both types). I did notice that when more text was added to one of the sentences but they were both equal (0 label), the worse configuration tended to be wrong and label them as unrelated.