NLP Exercise 1 - Kovi Szental 336130588

1. **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**
    a. QUOREF: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning : This dataset tests explicitly coreference resolution which is an intrinsic task. It tests the ability of a model to track entities in text.
    b. QNLI: QNLI reframes the SQUAD dataset as a Natural Language Inference (NLI) problem, where the model must decide if a candidate sentence entails the answer to a question. This setup directly probes intrinsic language understanding by requiring inference at the sentence level.
    c. WebQuestionsSP: WebQuestionsSP pairs natural questions with answers grounded in Freebase entities, requiring the model to link entities in the question to the correct knowledge base entity. This highlights intrinsic concepts of **e**ntity recognition

2. **In class we discussed several methods to implement inference-time scaling.**
   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.

   Self Consistency

   Description - Sample multiple CoT reasoning paths and select the most common or consistent answer
   Advantages- Reduces variance and errors from a single CoT chain, giving more robust answers
   Bottlenecks-Requires multiple forward passes(for multiple generations) which can cause which will cost more
   Parallelization- Yes, can parallelize differenet paths in parallel

C-O-T

Description - Model generates intermediate reasoning steps before producing the final answer.
Advantages- Improves performance on reasoning-intensive tasks and transparent step-by-step logic. Is able to use the thought out logic as context to continue its thinking process
Bottlenecks- Longer output sequences and therefore more tokens to process which causes longer compute and response time
Parallelization- A single reasoning chain is sequential and therefore can not be parallelised, but multiple independent queries can be batched.

RAG

Description - Adds to the model an external retrieval system to fetch relevant documents before answering.
Advantages- Reduces hallucination as well as allowing model to handle knowledge-intensive tasks without retraining.
Bottlenecks- Time in retrieval + search, memory use for large amount of data.
Parallelization- Document retrieval can be fetched in parallel, but the models generation is sequential

Least to most prompting

Description - Model decomposes a hard task into a sequence of simpler subproblems and solves them step by step.
Advantages- Breaks down complex reasoning into easier to compute steps
Bottlenecks- Requires multiple queries to the model
Parallelization- Nope, its sequential

(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?

I would choose Chain-Of-Thought + Self Consistency:
The reason I made this choice is because of the fact that this task requires complex reasoning. We can use CoT to make sure the model reasons step by step in a sequential pattern. Self-Consistency improves reliability by generating multiple reasoning paths and selecting the most consistent outcome.

Since I have a single GPU with large memory, I can batch these reasoning paths in parallel, making efficient use of the available hardware.

3. **Did the configuration that achieved the best validation accuracy also achieve the best test accuracy?** Yes

4.**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: The best model that learned to distinguish between paraphrases and non-paraphrases with 85% accuracy

Worst configuration: The worst model which was trained predicted label 0 for every example, which was not even the majority label in the dataset. The accuracy was around 30%, worse than a randomized guess.

Therefore there was not a specific type of sentence which was harder for the weaker model to predict as it gave everyone the same prediction