Advanced NLP Exercise 1

Github repo - https://github.com/YarinOhayon/ANLP_Ex1.git

Part 1

Question 1

1. **Stanford Question Answering Dataset** - This dataset evaluates reading comprehension by presenting models with questions based on Wikipedia articles. Each question has an answer that is a text span extracted from a relevant paragraph. SQuAD assesses a model's ability to understand text, identify relevant information, and extract precise answers. It requires intrinsic language skills such as syntactic analysis (e.g., identifying subjects and verbs), coreference resolution (tracking entities referred to with pronouns or aliases), and deep contextual understanding.
   Link - https://www.kaggle.com/datasets/stanfordu/stanford-question-answering-dataset

2. **BoolQ** - BoolQ consists of yes/no questions based on short passages derived from naturally occurring texts - primarily Wikipedia. It evaluates a model's comprehension by requiring it to infer the correct answer directly from the passage, relying on semantic interpretation and logical inference, without using external knowledge.
   Link - https://huggingface.co/datasets/google/boolq

3. **ReCoRd** - In this cloze-style dataset, the model is asked to fill in a blank in a sentence by selecting the correct entity (usually one mentioned in the passage). ReCoRD tests a model's ability to handle implicit meaning, resolve coreference, and reason about causal and temporal relationships within the text.
   Link - https://sheng-z.github.io/ReCoRD-explorer/

Question 2

(a) In class we discussed several methods to implement inference-time scaling -

- **Self-Consistency Decoding:**
  - **Description:** Self-consistency decoding extends chain-of-thought prompting by generating multiple diverse reasoning paths for the same input using stochastic sampling. Rather than relying on a single generated response, the model samples a diverse set of reasoning paths and marginalizes over them by aggregating the final answers. This technique aims to identify answers that are robust across varying reasoning trajectories, rather than overcommitting to any one potentially flawed path.
  - **Advantages:**
    * Encourages diversity in reasoning, which reduces bias toward brittle or spurious answers.
    * Aggregation across samples leads to higher factual accuracy and consistency.
  - **Computational Bottlenecks**:
    * Requires generating many responses per input which lead to high inference cost.
    * Sampling and aggregation increase memory and compute usage.
  - **Parallelizable:** Yes - each reasoning path can be sampled and evaluated independently

- **Verifiers:**
  - **Description:** Verifiers are external checks (either rule-based or learned models) used

to evaluate and filter generated outputs at inference time. Rather than selecting an output based on the majority of all generated chains, the model generates multiple candidate completions, and only those that pass the verifier are considered.
  - **Advantages:**
    * Reduces hallucinations by rejecting low-quality completions.
    * Improves factuality and logical correctness, especially in structured domains.
  - **Computational Bottlenecks**:
    * Cost of verifier execution (especially if learned).
    * Generation of multiple candidates.
  - **Parallelizable:** Yes - candidate generation and verification are parallelizable.

- **Increasing Compute Budget:**
  - **Description:** This strategy involves allocating more computational resources at inference time, not by increasing model size but by running the model more times, generating more candidate outputs, or using deeper chains of reasoning. It includes techniques like generating more samples (as in self-consistency), longer reasoning chains, or applying multiple modules. The idea is to trade increased compute for higher output quality, without modifying the model's architecture or training.
  - **Advantages:**
    * Allows small or medium-sized models to perform on par with larger models in some cases.
    * Can combine with other methods.
  - **Computational Bottlenecks**:
    * Directly scales with number of generations and length of outputs.
    * High inference latency and GPU usage for complex tasks.
  - **Parallelizable:** Yes - individual generations or reasoning paths can be run in parallel (e.g., across GPU cores or machines).

- **Planning, Backtracking, and Self-Evaluation:**
  - **Description:** Inspired by models like OpenAI's O1 and DeepSeek R1, this method treats inference as a dynamic problem-solving process. The model begins by generating a high-level plan, executes the reasoning steps, and monitors progress through self-evaluation. If an error is detected, the model will backtrack, revise its approach, and re-plan.
  - **Advantages:**
    * Allows correction of earlier mistakes through backtracking.
    * Enables adaptive, human-like reasoning.
  - **Computational Bottlenecks**:
    * Requires memory to track state and intermediate results.
    * Involves multiple, interdependent generation rounds.
  - **Parallelizable:** Partially. The overall reasoning process is state-dependent and sequential (loop of plan → act → evaluate → revise). But the other components, like generating candidate plans, evaluating multiple partial outputs, or exploring alternative backtracking paths can be parallelized.

(b) Given that the task involves complex scientific reasoning and I have access to a single GPU with large memory capacity, I would choose Self-Consistency Decoding. This method is parallelizable, so the large GPU memory would allow me to run multiple generations in parallel, making self-consistency computationally feasible without overwhelming resources. Self-Consistency Decoding enables the model to explore diverse reasoning paths and aggregate consistent conclusions, which is particularly useful in scientific domains where multiple logical

interpretations may exist.

Part 2

I named each run using the format - *run_#epochs_learningRate_batchSize*. Below are the results:

$$run\_4\_lr5e-05\_bs16: \ val = 0.8652, \ test = 0.8046$$
$$run\_2\_lr5e-05\_bs16: \ val = 0.8627, \ test = 0.7884$$
$$run\_4\_lr3e-05\_bs32: \ val = 0.8235, \ test = 0.7014$$
$$run\_3\_lr2e-05\_bs16: \ val = 0.8456, \ test = 0.4249$$

As we can see, the model with the highest validation accuracy (run_4_lr5e-05_bs16) also achieved the highest test accuracy.

Next, we compare examples where the best-performing model (run_4_lr5e-05_bs16) made the correct prediction while the worst-performing model (run_3_lr2e-05_bs16) failed -

- Example 3
- Sentence 1: The AFL-CIO is waiting until October to decide if it will endorse a candidate .
- Sentence 2: The AFL-CIO announced Wednesday that it will decide in October whether to endorse a candidate before the primaries .
- Ground Truth: 1
- run_4_lr5e-05_bs16 Prediction (correct): 1
- run_3_lr2e-05_bs16 Prediction (wrong): 0
- ---
- Example 24
- Sentence 1: Saddam loyalists have been blamed for sabotaging the nation 's infrastructure , as well as frequent attacks on U.S. soldiers .
- Sentence 2: Hussein loyalists have been blamed for sabotaging the nation 's infrastructure and attacking US soldiers .
- Ground Truth: 1
- run_4_lr5e-05_bs16 Prediction (correct): 1
- run_3_lr2e-05_bs16 Prediction (wrong): 0
- ---

These examples show that the weaker model (run_3_lr2e-05_bs16) struggles to recognize paraphrases and thus predicts the wrong label. for instance:
“decide if it will endorse” vs. “decide whether to endorse.”

- Example 58
- Sentence 1: " This decision is clearly incorrect , " FTC Chairman Timothy Muris said in a written statement .
- Sentence 2: The decision is " clearly incorrect , " FTC Chairman Tim Muris said .
- Ground Truth: 1
- run_4_lr5e-05_bs16 Prediction (correct): 1
- run_3_lr2e-05_bs16 Prediction (wrong): 0
- ---

recognize the same entity in various forms (e.g., “Tim Muris” vs. “Timothy Muris”).

- Example 62
- Sentence 1: " Today , we are trying to convey this problem to Russian President Vladimir Putin and US President George W Bush . "
- Sentence 2: " Today , we are trying to convey this problem to Russian President Vladimir Putin (news - web sites) and President Bush (news - web sites) . "
- Ground Truth: 1
- run_4_lr5e-05_bs16 Prediction (correct): 1
- run_3_lr2e-05_bs16 Prediction (wrong): 0
- --
- Example 79
- Sentence 1: " If we don 't march into Tehran , I think we will be in pretty good shape , " he said .
- Sentence 2: " As long as we don 't march on Tehran , I think we are going to be in pretty good shape , " he said .
- Ground Truth: 1
- run_4_lr5e-05_bs16 Prediction (correct): 1
- run_3_lr2e-05_bs16 Prediction (wrong): 0
- --

In these examples, the weaker model (run_3_lr2e-05_bs16) struggles to correctly handle quotations when the quoted content is phrased slightly different.