ANLP ex1 – Gilad Shenhar 208278325

1.Open Questions:

1.1. **SQuAD (Stanford Question Answering Dataset)**

Why intrinsic**:** It tests reading comprehension and semantic understanding by requiring models to extract answers from context passages. Success relies on parsing syntax, resolving coreference, and grasping question semantics.

BoolQ

Why intrinsic**:** Questions are yes/no, based on short paragraphs from Wikipedia. The task evaluates a model's ability to integrate factual knowledge and make logical inferences based on context, rather than perform task-specific classification.

DROP (Discrete Reasoning Over Paragraphs)

Why intrinsic**:** DROP pushes models to perform discrete reasoning like counting, comparison, and arithmetic based on text. It explicitly measures a model's capacity for multi-step inference and numerical reasoning, beyond surface-level understanding.

1.2. (a).

CoT – Chain of Thoughts:

- Description: prompts are extended with step-by-step intermediate reasoning to guide the model toward correct conclusions.
- Advantages: improving performance on reasoning-intensive tasks such as math or logic problems.
- Computational Bottlenecks: longer outputs lead to slower generation and higher memory use.
- Can be parallelized: no.

Self-Consistency:

- Description: sample multiple reasoning paths (with CoT) and choose the most frequent final answer.
- Advantages: reduces variance and hallucinations, leading to more reliable outputs.
- Computational Bottlenecks: require multiple forward passes and sampling chains, which increases runtime.
- Can be parallelized: yes, samples can be generated independently.

Verifiers:

- Description: use rule-based or learned verifiers to filter or score generated outputs and keep only high-confidence ones.
- Advantages: better accuracy by discarding incorrect or inconsistent generations.
- Computational Bottlenecks: adds a verification step per generation, or if verifier is learned, it adds model inference time.
- Can be parallelized: yes, multiple candidate generations and verifications can run in parallel.

(b).

I would choose self-consistency.

While this method increases inference cost, a single GPU with large memory can handle multiple forward passes for sampling (multiple CoT generations) and slightly longer outputs due to reasoning steps.

It is better than the Verifier method because we can't use parallel computation, and thus it will be more expensive than the self-consistency method. Also, we depend on the Verifier size.

2. Programming Exercise:

2.1.2.

- The model with the configurations: lr – 5e-05, batch_size – 32, num_epoch – 5 and the model with the configurations: lr – 2.5e-05, batch_size – 15, num_epoch – 5, got the same best of all models eval accuracy – 0.81863.
  The second model got better test accuracy – 0.81159.

- **Qualitative analysis:**
  Examples where the best configuration succeeded but the worst failed:
  Sentence1: Prof Sally Baldwin, 63, from York, fell into a cavity which opened up when the structure collapsed at Tiburtina station, Italian railway officials said.
  Sentence2: Sally Baldwin, from York, was killed instantly when a walkway collapsed and she fell into the machinery at Tiburtina station.
  Label: 0

  Sentence1: Anything less is unacceptable, " said Gordon, the ranking Democrat on the House Space and Aeronautics subcommittee.
  Sentence2: Gordon is the senior Democrat on the House Subcommittee on Space and Aeronautics.
  Label: 0

Sentence1: The Justice Department and the Federal Communications Commission have the final say.
Sentence2: But the deal must get approval from the Federal Communications Commission and the Justice Department 's antitrust division.
Label: 1

It seems like the worst configuration is wrong on examples where the two sentences are the same but a slight change, when most of the words in the sentences are the same but the meaning of the sentences may or may not be the same.

Link to github repository: https://github.com/shenhars/ANLP_ex1