David Busbib

1.

- BoolQ - Evaluates *natural language inference* by asking yes/no questions grounded in text. The model must understand the meaning of the passage and question, not just extract facts.
- DROP- Requires models to perform *symbolic reasoning* (arithmetic, counting, comparisons) over text, testing compositional understanding and ability to manipulate information.
- RECLOR- Focuses on *logical reasoning* from standardized reading comprehension exams (GRE). It evaluates the model's grasp of logic, argument structure, and inference.

2.

Verifier

- **Description**: After a model generates an initial answer, a *separate verifier model* (or a second pass through the same model) check whether the answer is okay, correct, or consistent with the input.
- **Advantages**:
    - Catches hallucinations or inconsistencies.
    - Improves reliability of outputs.
    - Can reduce false positives or wrong reasoning chains.
- **Bottlenecks**:
    - Doubles the compute: one pass for generation, one for verification.
    - If using a larger verifier, memory/latency spikes.
- **Parallelizable:**Yes — generation and verification can be batched or pipelined across multiple GPUs or processes.

Increasing Compute Budget

- **Description**: Allocate more compute per query — by using larger models, generating more tokens, sampling more paths, or doing deeper reasoning. This is a generic scaling method.
- **Advantages**:
    - Improves output quality (more reasoning steps, more complex representations).
    - Can yield better answers without changing model weights or training.
- **Bottlenecks**:
    - Significantly increases GPU memory, inference latency, and cost.
    - May hit diminishing returns if not combined with smarter techniques.
- **Parallelizable:** Yes — can distribute samples, model shards, or reasoning paths.

3. Self-Evaluation

- **Description**: The model assesses its own outputs — scoring or reflecting on its response, sometimes using auxiliary prompts like *"Is this correct?"* .

- **Advantages**:
  - No extra model needed — uses the same LLM.
  - Can help select between multiple candidate answers.
  - Encourages introspection and uncertainty awareness.
- **Bottlenecks**:
  - Requires one or more additional forward passes per answer.
  - Sensitive to prompt phrasing and calibration.
- **Parallelizable:**Yes — self-evaluations can be computed independently for each answer.

4. Self-Consistency

- **Description**: Generates multiple reasoning paths (via chain-of-thought + temperature sampling), then selects the most consistent final answer (by majority vote).
- **Advantages**:
  - Greatly improves reasoning reliability.
  - Reduces dependence on any single flawed reasoning path.
- **Bottlenecks**:
  - Requires 10–50× more inference calls.
  - May be memory-bound on single GPU if sampling many in parallel.
- **Parallelizable:**Yes — each sample can be generated independently.

b.

I would choose **self-consistency decoding**, as it improve reasoning quality by generating multiple diverse reasoning paths and selecting the most frequent conclusion. This approach helps reduce random errors and stabilizes the final output, which is especially valuable in complex scientific tasks where multiple logical routes may exist. With a large-memory GPU, it is feasible to generate many reasoning samples in parallel, making self-consistency both effective and computationally practical for this setting.

Part 2

Github_link - https://github.com/david-busbib/Anlp-project_1.git

I name each run the following
epouch_{num of epouch}_{lr}_{batch_size} ,eval_accurcy ,test_accurcy

```
epoch_num: 4, lr: 2e-05, batch_size: 16, eval_acc: 0.8578, test_acc: 0.8470
epoch_num: 2, lr: 3e-05, batch_size: 32, eval_acc: 0.8162, test_acc: 0.8064
epoch_num: 3, lr: 5e-05, batch_size: 32, eval_acc: 0.8529, test_acc: 0.8272
epoch_num: 4, lr: 4e-05, batch_size: 16, eval_acc: 0.8529, test_acc: 0.8458
epoch_num: 1, lr: 6e-05, batch_size: 16, eval_acc: 0.7966, test_acc: 0.7803
```

is the the totatl of run :

as we can see the one with the highest evaluation accurcy has also the highest test accurcy.

Now lets compare between the best and the worst performing configuration, here is some example that we get correct(1) on the best configuration and false in the worst (0)

```
For the first example -
PCCW 's chief operating officer , Mike Butcher , and Alex Arena , the chief
financial officer , will report directly to Mr So .###Current Chief Operating
Officer Mike Butcher and Group Chief Financial Officer Alex Arena will report
to So .

Second example -
" PNC regrets its involvement " in the deals , Chairman and Chief Executive
Officer James Rohr said in a statement .###James Rohr , chairman and chief
executive officer , said PNC regretted the incident .

Third example -
The broad Standard & Poor 's 500 Index .SPX gained 22.25 points , or 2.23
percent , to 1,018.22 , based on the latest available figures .###The broad
Standard & Poor 's 500 Index .SPX gained 11.22 points , or 1.13 percent , at
1,007.19 .

Forth example –
Quinn was assigned to the 2nd Squadron , 3rd Armor Cavalry Regiment .###Quinn
was assigned to the 3rd Armored Cavalry Regiment , based in Fort Carson ,
Colo .

Fifth example -
At this writing , the fate of Alabama Supreme Court Chief Justice Roy Moore
hangs precariously in the hands of the states court of the judiciary
.###Moore , the suspended chief justice of the Alabama Supreme Court , stands
trial before the Alabama Court of the Judiciary .
```

To compare the best and worst configurations, we looked at examples where the best one gave the correct answer (1) and the worst gave a wrong one (0). I think there is some clear patterns. The best model understands the meaning even when the sentences are written in a different way. For example, it knows that “Mr So” and “So” are the same person, or that “PNC regrets its involvement” means the same as “PNC regretted the incident.” It can also deal with small differences in numbers or names. But the worst model gets confused when the words are changed, the sentence structure is different, or when it needs to understand something that’s not said directly. This shows that the best model is better at understanding the real meaning of the sentences, not just matching the exact words.