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* (e.g., 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) checks whether the answer is plausible, correct, or consistent with the input.
* **Advantages**:
  + Catches hallucinations or inconsistencies.
  + Improves trustworthiness 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 — e.g., 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?”* or *“Rate the confidence”*.
* **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 (e.g., 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**, as it significantly improves reasoning accuracy by aggregating diverse thought paths. A large-memory GPU allows efficient parallel sampling to support thi