Rationale-Aware Answer Verification by Pairwise Self-Evaluation

Akira Kawabata* — The Asahi Shimbun Company, Japan
Saku Sugawara — National Institute of Informatics, Japan
* kawabata-a@asahi.com

Arxiv

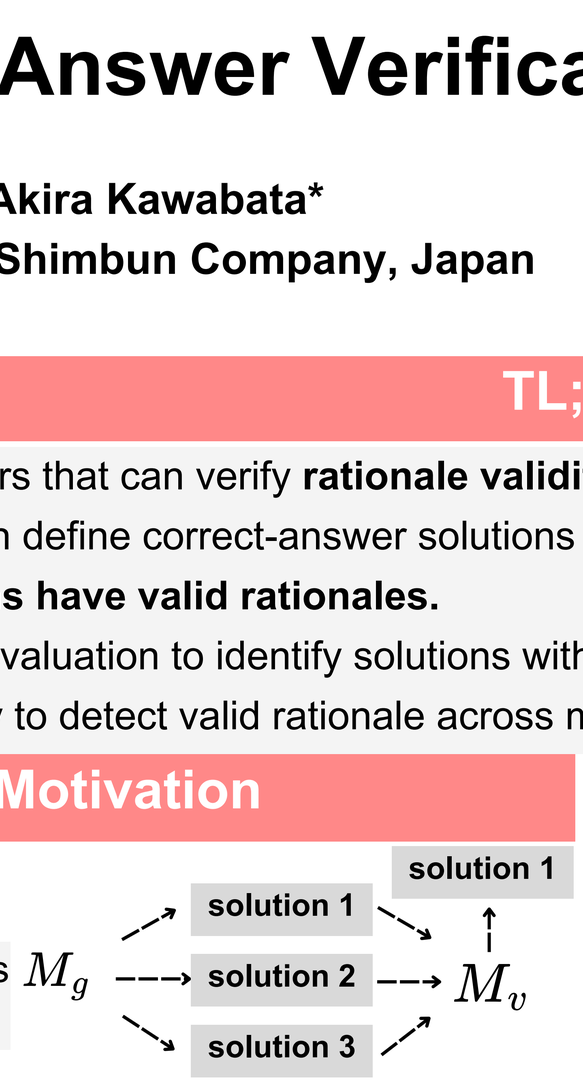
TL;DR

- We propose a method to build verifiers that can verify **rationale validity** as well as answer correctness.
- Conventional training methods, which define correct-answer solutions as positive, fails to train rationale-aware verifiers, as **only 19% of model-generated correct-answer solutions have valid rationales.**
- We introduce iterative pairwise self-evaluation to identify solutions with high-quality rationales among correct-answer solutions.
- Our method improves verifiers' ability to detect valid rationale across multiple reasoning datasets.

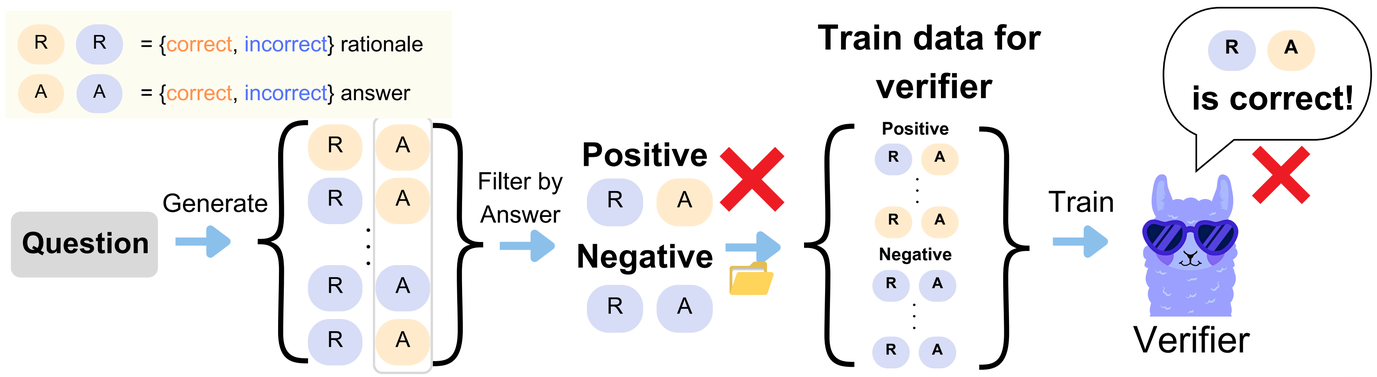
Background & Motivation

Q. What is verifier?

A. A model that evaluates the correctness of model-generated solutions.



Q. How is verifier (usually) trained?



A. Trained to distinguish positive and negative solutions. Conventionally, the labels are defined whether the solutions reach correct answers or not.

Limitations

It is unclear whether the conventional approach can build rationale-aware verifiers.

What We Did

We show that

- the conventional approach is insufficient for rationale-aware verifier due to the unfaithfulness in model-generated solutions.
- refining training data by iterative self-evaluation improves the verifiers' ability to distinguish valid rationales.

Preliminary Experiments

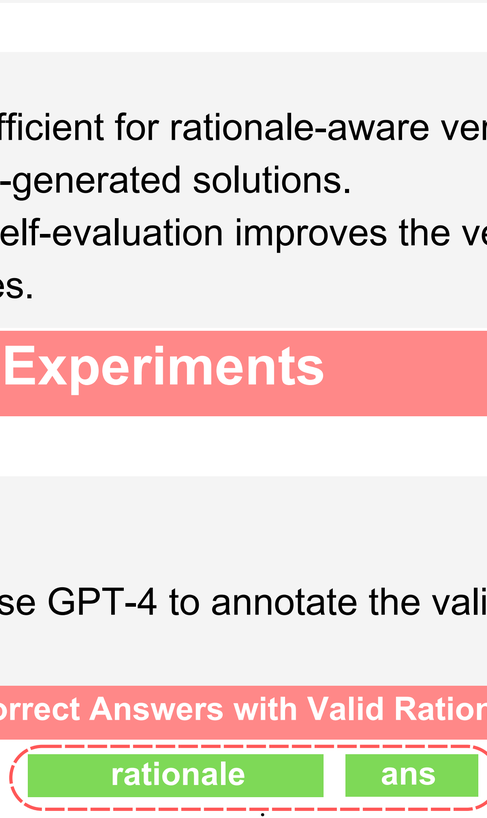
Settings

Model: Llama2-7B
Dataset: StrategyQA
Rationale-quality annotation: We use GPT-4 to annotate the validity of model-generated rationales.

How often Does Model Generate Correct Answers with Valid Rationale?

Result

Only 19% of correct-answer solutions are judged as having valid rationales.



How Does Rationale Quality Affect Verifiers' Selection of Valid Rationales?

Method

Verifiers' Training Setting
Incrementally replace 10% of correct-answer solutions with valid solutions.

Eval Setting
For each Q, verifier selects from 5 solutions (1 valid, 2 correct-but-flawed, 2 incorrect)

× 1 × 2 × 2

Rationale Accuracy

$$\frac{1}{|D|}\sum_{i=1}^{|D|}\mathbb{I}\left[M_v(q_i, cands_i) = \text{[valid rationale]}\right]$$

Answer Accuracy

$$\frac{1}{|D|}\sum_{i=1}^{|D|}\mathbb{I}\left[M_v(q_i, cands_i) = \text{[valid rationale] or [valid answer]}\right]$$

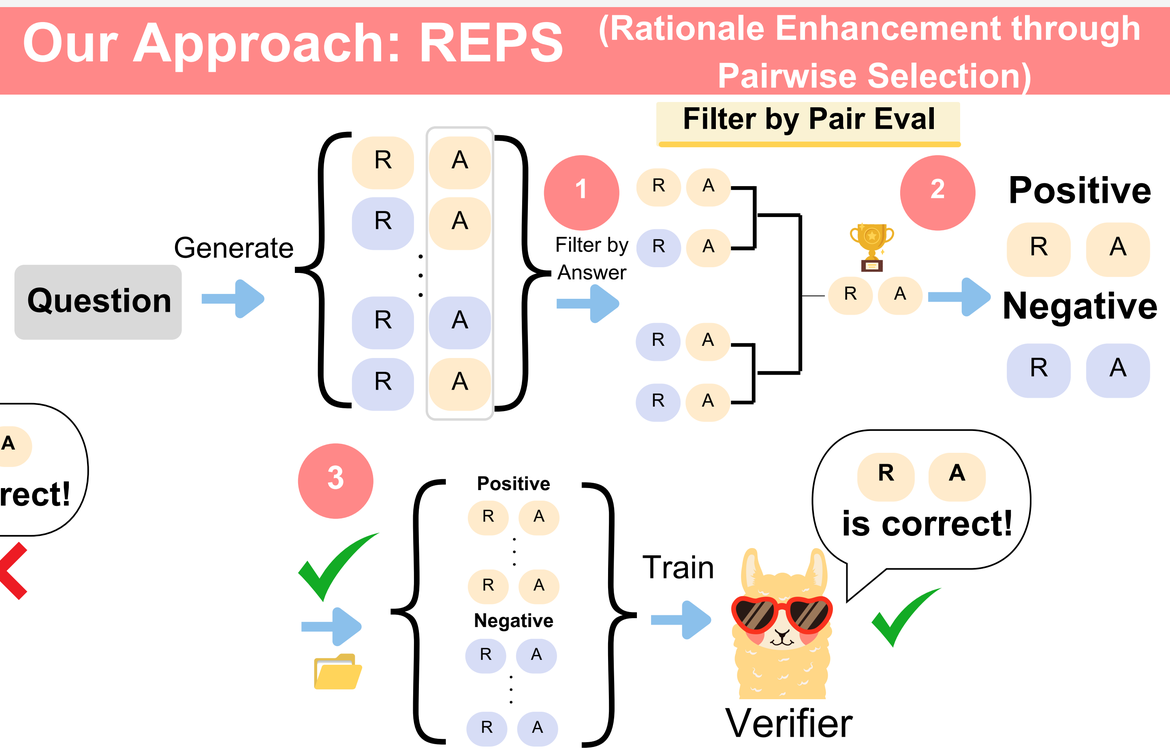
Result

As the proportion of valid solutions increases, Rationale Acc. improves while Answer Acc. remains largely stable.

Implication

Large room for improving rationale accuracy in conventional verifier.
The key lies in how to identify high-quality model-generated rationales.

Our Approach: REPS (Rationale Enhancement through Pairwise Selection)



1. **Selects N (e.g., 4) correct-answer solutions** as candidates from model-generated solutions for each question.
2. **Applies tournament-style pairwise self-evaluation** to these candidates, determining **each round's winner through majority voting of S pairwise evaluations.**
3. Adds the final remaining solution to the positive samples for training the verifier.

Experiments

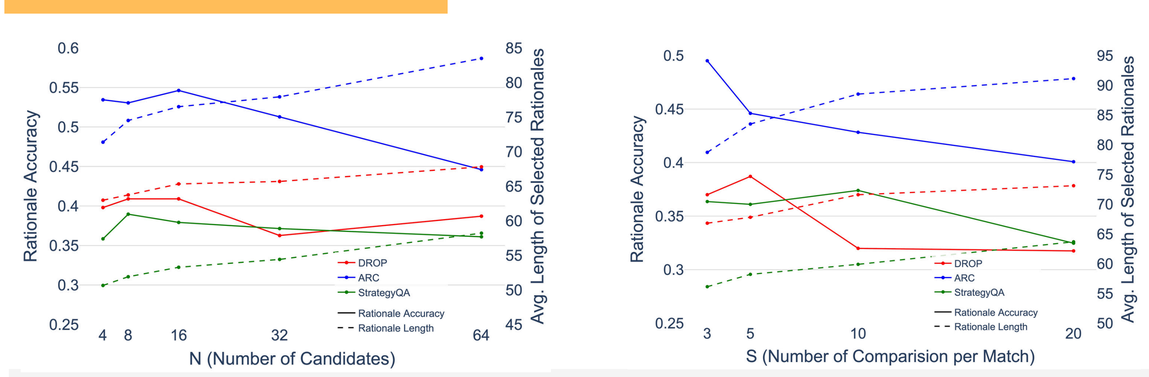
Settings

Model: Llama2-7B **Dataset:** ARC-Challenge, DROP, StrategyQA
REPS Parameters: N=8, S=5
Metrics: Rationale Accuracy, Task Performance (the accuracy of the verifier in selecting the answer from 5 generated solutions)

Main Results

	ARC-Challenge		DROP		StrategyQA	
	Rationale Acc.	Task Perf.	Rationale Acc.	Task Perf.	Rationale Acc.	Task Perf.
Baseline	38.90	52.40	36.02	45.80	30.13	67.10
REPS	**53.05**	**54.75**	**40.90**	**46.90**	**38.96**	**67.25**

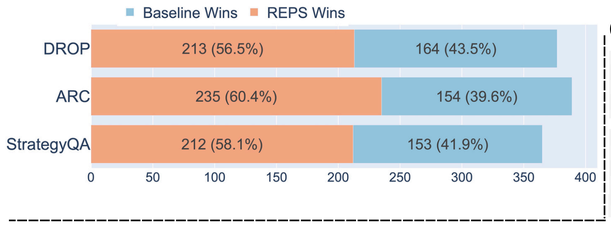
Effects of Parameter Choice



- Unexpectedly, Rationale Accuracy decreases as N and S grow.
- This can be considered as the amplification of self-evaluation biases toward the superficial cue, i.e., the length of rationale.

Which Verifier Can Select Better Rationales? Win Rate by GPT-4

Result



Examples

Q. Is Disney associated with Los Angeles County?

Disney has theme parks in Los Angeles County, California. Thus, Disney is associated with Los Angeles County. REPS

Disney is a large company that has many locations. Disneyland is located in Anaheim, which is 24 miles south of Los Angeles. Thus, Disney does not have a location in Los Angeles County. Baseline

Q. A type of small mammal from the mountain regions of the western United States makes its home out of piles of rock. During summer months, the mammal places grasses and seeds in protected places in the rock piles. Which of the following is the most likely reason for this behavior? (correct answer: to store food that will be eaten over the winter months)

This is the behavior of a prairie dog, which stores grass, seeds, and other food items in rock piles during summer months. Prairie dogs will consume the stored food during winter months. Thus, the prairie dog's behavior is to store food for use during the winter months. REPS

To protect and preserve its food, the mammal buries grasses and seeds in piles of rock. Grasses and seeds decay over time. Thus, the mammal buries its food to protect it from decay. Baseline