Bootstrapping models to critique chain-of-thought rationales for n-digit arithmetic problems

Dane Sherburn[1], William Saunders[2], Alex Gray

[1]Machine Learning Alignment Theory Scholar, Stanford Existential Risks Initiative, [2]OpenAI

Stanford Existential Risks Initiative

Problem

Training models to perform tasks that are very time-consuming for humans to evaluate is a major safety challenge.

One solution is to train a model to write critiques of another model as it performs the task [2]. However, training a critique-writing model is itself very time-consuming to create for similar reasons.

We reduce the time to create critique-writing models by 100-1000x on n-digit arithmetic problems by getting the model to learn from its own generated outputs.

Example: generated 2-digit problem

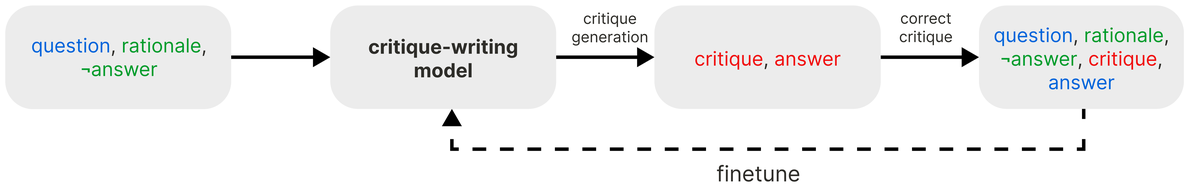
```
Input:
6 5 + 6 2

Target:
<scratch>
6 5 + 6 2 ,  C: 0
6 + 6 , 7 C: 1      <- exactly one error in each rationale
, 3 7 C: 1
1 3 7
</scratch>
1 3 7

Correction:
“C: 1” should be “C: 0” on line 1
<scratch>
6 + 6 , 7 C: 0
, 2 7 C: 1
1 2 7
</scratch>
1 2 7
```

Method

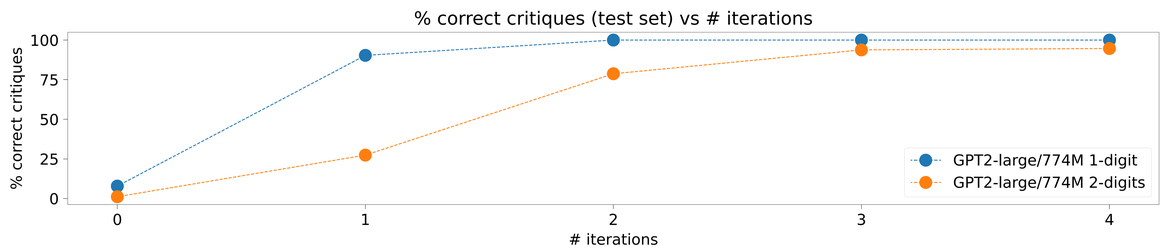
The questions and answers are from an existing dataset [3], the rationales are generated by a model solving the task and the critiques are generated by the critique-writing model.



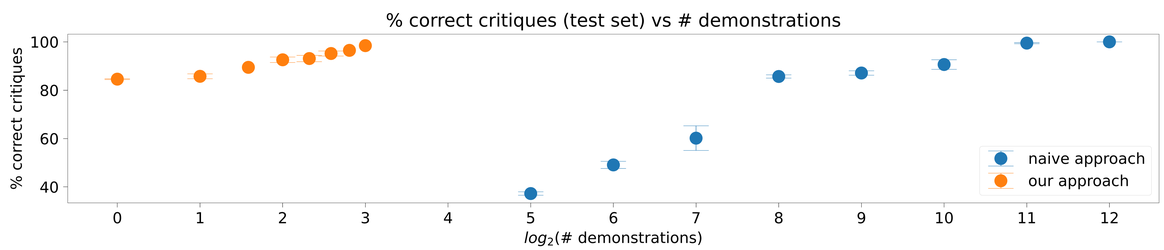
The idea is to generate critiques of many incorrect rationales (prompted with a few examples) and to finetune on all the critiques that arrived at the correct answers, then repeat. [1]

Results

We generate critiques for 10k {1, 2}-digit arithmetic problems, finetune for 5 epochs and test on 1k in-distribution examples. We found that GPT2-large/774M can critique {1, 2}-digit arithmetic workings with just 3 few-shot examples.



We also compare the number of demonstrations needed for our approach vs the naive approach. We found that ours requires 100-1000x less data on 1-digit arithmetic problems.



Conclusions

This work shows that language models can learn to write high-quality critiques from very few human examples by learning from their own generated outputs.

We view this work as an early but promising sign that critiquing models are an economically viable tool which can aid in training models to perform tasks that are time-consuming to evaluate.

References

[1] Eric Zelikman, Yuhuai Wu, Jesse Mu, Noah D. Goodman: “STaR: Bootstrapping Reasoning With Reasoning”, 2022

[2] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, Jan Leike: “Self-critiquing models for assisting human evaluators”, 2022

[3] Maxwell Nye et al.: “Show Your Work: Scratchpads for Intermediate Computation with Language Models”, 2021