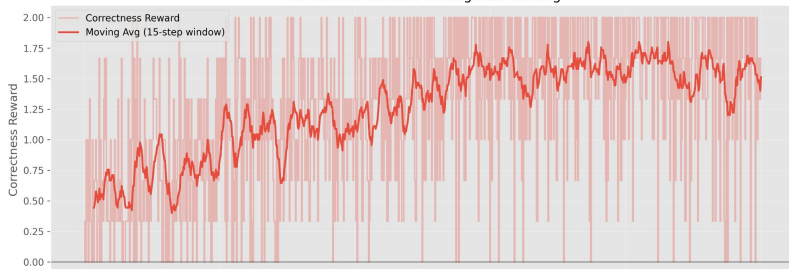


Diverse Parallel Responses with Entropic Reasoners

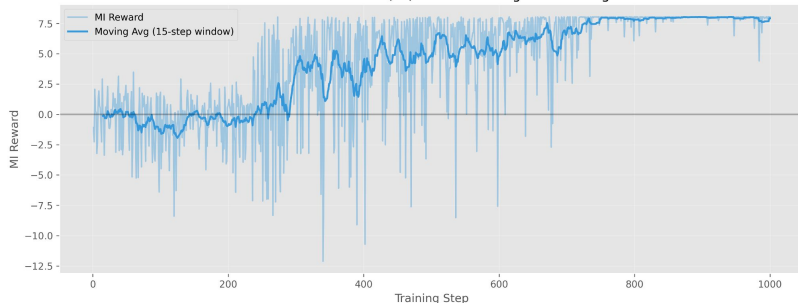
Devan Shah, COS '26; Daniel Yang, COS '26; Owen Yang, MAT '26

(a) GRPO Training Graph

Correctness Reward Throughout Training



Mutual Information (MI) Reward Throughout Training



(b) Specific example + test results

Question (GSM8K): Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did he make?

Correct Answer: \$70,000

Baseline (LLaMA 3.1 8B with 1000 GRPO Steps):

Explanation:

- Josh bought the house for \$80,000
- Then, he added \$50,000 in repairs
- The value of the house increased by 150% after repairs
- Let's find the original cost of house post repairs
 $= \$80,000 + \$50,000 = \$130,000$
 $= \$130,000$ increased value
 $= 1.5 \times \$87,000$

Now subtract price he has spent \$130,000
 $(\$80,000 + \$50,000) = \$87,000$
 $= \$43,000$ profit.

(0/5 attempts correct)

Ours (Above + Mutual Information Reward):

Strategy 1:

Josh buys a house for \$80,000 and then puts in \$50,000 in repairs. This means the total cost for Josh is \$80,000 + \$50,000 = \$130,000.

The value of the house increased by 150% after the repairs. This means the new value is 150% of \$80,000 plus the original \$80,000.

150% of \$80,000 is $1.5 \times \$80,000 = \$120,000$.

The new value is \$120,000 + \$80,000 = \$200,000.

The profit Josh made is the difference between the new value and the total cost, which is \$200,000 - \$130,000 = \$70,000.

[...]

(2/5 attempts correct)

(n=100)	Pass@1	Pass@5	Plu@5	Cons@5
Base	0.61	0.93	0.71	0.60
Ours	0.80	0.95	0.87	0.79

Thanks to Prof. Eysenbach and Chongyi Zheng,
Date: 4/30/25