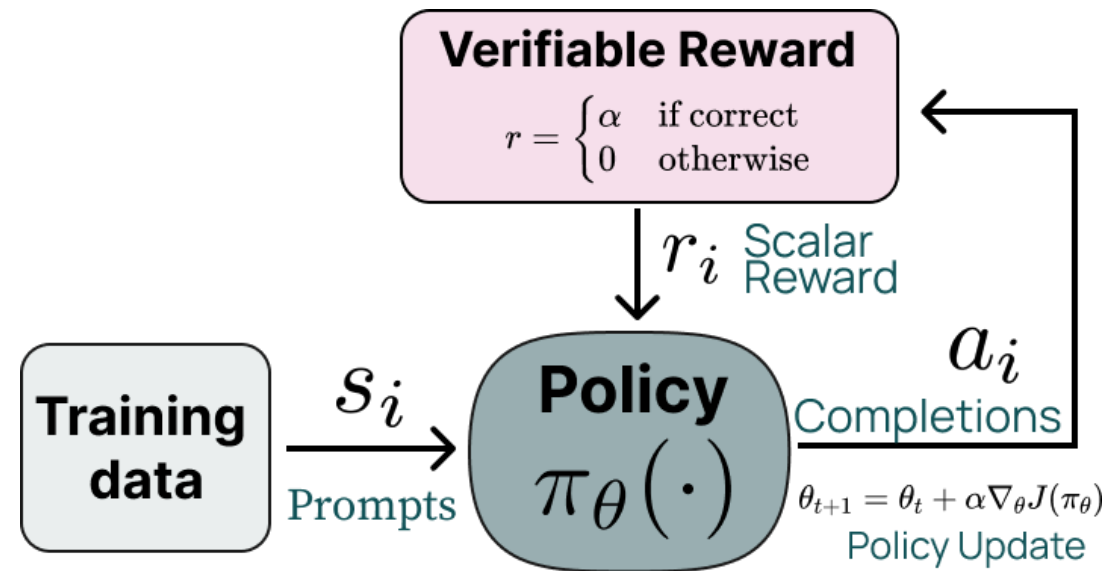


ICLR' 26 Curriculum Reinforcement Learning from Easy to Hard Tasks Improves LLM Reasoning

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, Shuiwang Ji

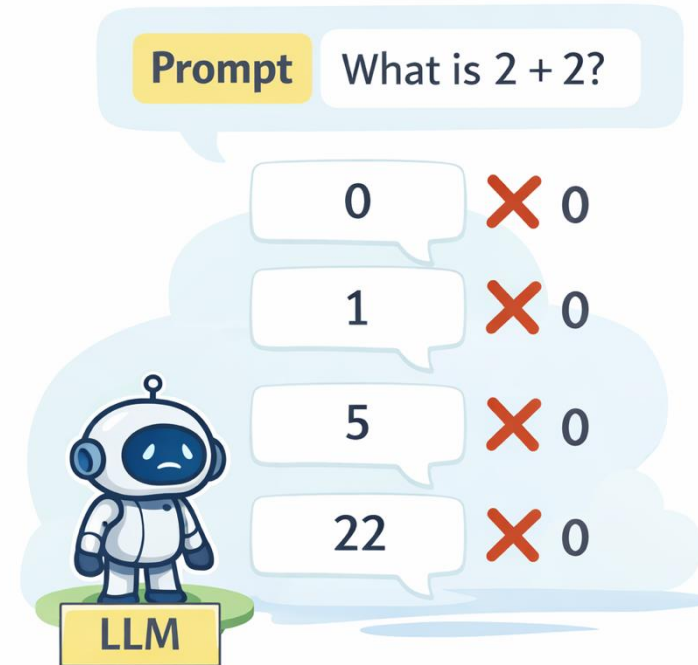
Reinforcement Learning with Verifiable Rewards

- **RLVR (Reinforcement Learning with Verifiable Rewards)** improves LLM reasoning by optimizing on binary rewards.
- Incorrect outputs yield zero reward, resulting in a sparse but high-precision learning signal.
- **Hence**, RLVR requires a base model with non-trivial probability of getting the question right.



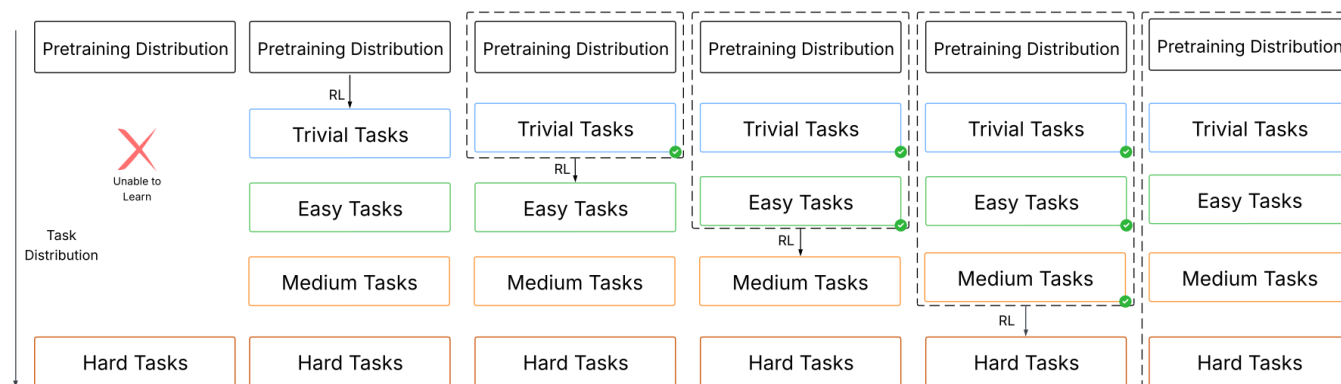
Sparse rewards are a problem

- Binary, verifiable rewards make the signal sparse.
- Small LLMs, or weaker models may never get rewards and learning can stall.



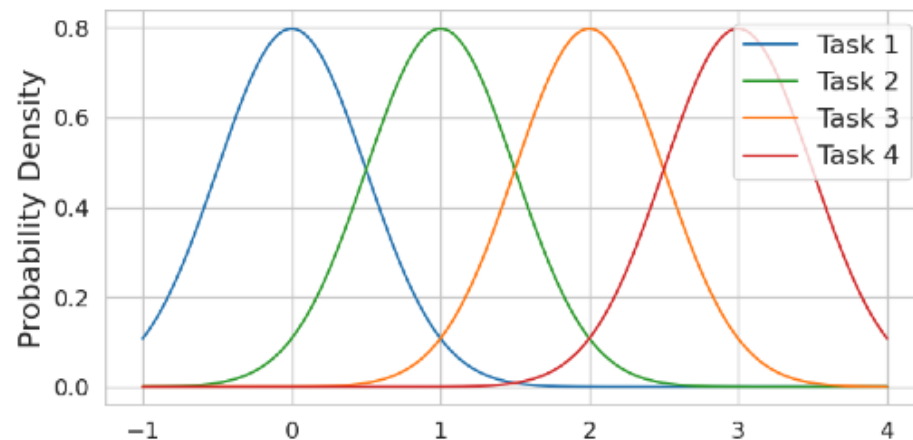
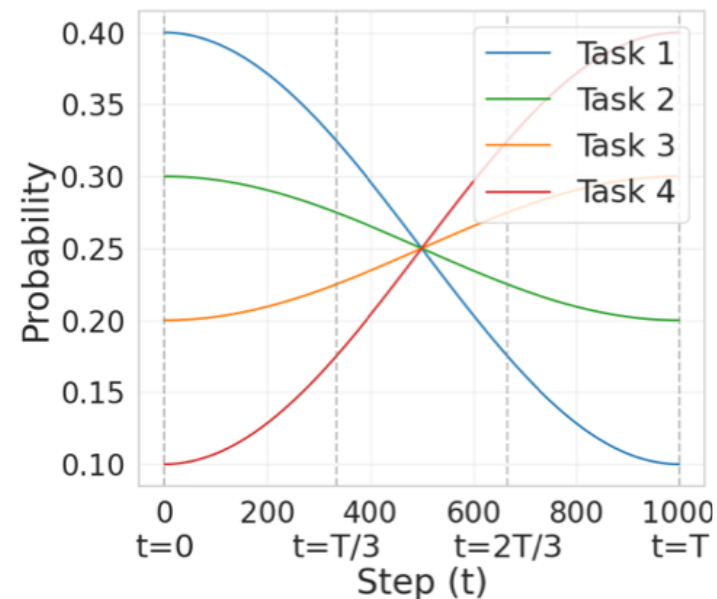
Curriculum RL for Reasoning

- Learn reasoning tasks in increasing order of difficulty using curriculum learning.
- We divide the dataset into subsets of **trivial**, **easy**, **medium** and **hard** problems.
- Goal: By gradually training on harder tasks, mitigate the sparse rewards problem



Easy 2 Hard Reasoner (E2H)

- To enable effective learning, E2H probabilistically samples problems from trivial to hard. E2H has 2 variants.
- **E2H-C (top)**: a parameter free cosine-based curriculum sampler.
- **E2H-G (bottom)**: a tunable Gaussian Mixture Model inspired curriculum sampler.



Theoretical Analysis

- As E2H is a curriculum RL based training method, we establish the theoretical benefits of curriculum RL
- **Convergence:** We show that curriculum RL converges under standard assumptions for policy iteration.
- **Sample efficiency:** We show that curriculum RL improves has better sample complexity than direct training.

Experimental Setup

- Models – Qwen 2.5 1.5B and 3B, LLaMA 3.2 3B
- Datasets – Blocksworld, Countdown, GSM8K, MATH500, AQuA
- Policy Optimization – GRPO, DAPO
- Baselines –
 - CoT – Zero Shot Accuracy
 - GRPO (All) – Vanilla GRPO on the entire dataset.
 - GRPO (**Hard**) – GRPO on the hard task only.
 - GRPO (OOD) – GRPO on the Out of Distribution Difficulty
 - CL – Classical Curriculum Learning
 - Self-Evolve – A Curriculum Learning RL baseline

Results and Insights

- E2H gets the best results across benchmarks.
- (Top) Despite not being trained on OOD difficulties, E2H outperforms GRPO that is explicitly trained on the OOD split.
- (Bottom) E2H results on GSM8K and AQuA.

| | | Blocksworld | | | | | Countdown | | | | | MATH | | | | |
|-----------------------|-------------|-------------|------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | | Trivial | Easy | Med | Hard | OOD | Trivial | Easy | Med | Hard | OOD | Trivial | Easy | Med | Hard | OOD |
| Qwen 1.5B Instruct | CoT | 4.0 | 0.0 | 0.0 | 0.0 | 0.0 | 16.0 | 5.6 | 1.7 | 0.1 | 0.1 | 40.1 | 27.9 | 22.7 | 17.6 | 8.2 |
| | GRPO (All) | <u>98.0</u> | 100 | <u>83.3</u> | <u>21.1</u> | <u>2.6</u> | 96.1 | 64.9 | 28.8 | 18.1 | 9.2 | 87.2 | 72.0 | 61.6 | 46.3 | 25.7 |
| | GRPO (Hard) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 43.9 | 16.4 | 18.1 | 6.5 | 82.3 | 64.7 | 53.4 | 38.2 | 20.6 |
| | GRPO (OOD) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.1 | 23.1 | 18.1 | 11.3 | 5.3 | 37.0 | 21.1 | 15.0 | 8.5 | 3.7 |
| | CL | 46.0 | 100 | 45.2 | 5.8 | 0.7 | 57.7 | <u>85.8</u> | <u>57.2</u> | <u>31.5</u> | <u>12.6</u> | <u>86.2</u> | 71.5 | 62.4 | 46.7 | 25.6 |
| | Self Evolve | 100 | 100 | 70.2 | 13.8 | 2.1 | 96.6 | 65.3 | 34.2 | 17.8 | 9.5 | 84.0 | 70.6 | <u>62.6</u> | <u>48.6</u> | 26.1 |
| | E2H-G | <u>98.0</u> | 100 | 95.3 | 32.9 | 7.3 | 97.9 | 87.2 | 70.4 | 41.0 | 19.2 | 85.3 | <u>71.7</u> | 62.5 | 48.7 | <u>27.6</u> |
| E2H-C | 100 | 100 | 15.5 | 0.0 | 0.0 | <u>96.7</u> | 64.0 | 25.9 | 15.8 | 6.4 | 84.6 | 69.6 | 63.0 | 47.6 | 28.6 | |
| LLaMa 3.2 3B Instruct | CoT | 24.0 | 0.0 | 1.2 | 1.0 | 0.0 | 37.1 | 4.6 | 0.3 | 0.0 | 0.0 | 65.9 | 44.6 | 35.2 | 24.1 | 13.6 |
| | GRPO (All) | 100 | 100 | <u>94.1</u> | <u>38.9</u> | <u>13.3</u> | 99.9 | <u>89.5</u> | 71.6 | 47.9 | 2.7 | 65.9 | 47.0 | 36.0 | 22.0 | 10.2 |
| | GRPO (Hard) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 40.5 | 33.8 | 3.0 | 9.7 | 1.4 | 22.7 | 14.4 | 10.3 | 7.5 | 0.3 |
| | GRPO (OOD) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 8.0 | 0.3 | 0.0 | 0.0 | 0.0 | 63.6 | 39.0 | 31.3 | 19.0 | 7.6 |
| | CL | 100 | 0.0 | 0.0 | 0.0 | 0.0 | 17.2 | 36.0 | 22.7 | 11.2 | 4.1 | 74.1 | 54.1 | 43.9 | 28.0 | 12.5 |
| | Self-Evolve | 100 | 100 | 91.1 | 35.8 | 16.6 | 96.7 | 66.6 | 37.9 | 27.5 | 18.5 | 79.1 | 61.4 | 48.9 | <u>33.1</u> | 14.1 |
| | E2H-G | 100 | 100 | 98.8 | 44.1 | 17.6 | 95.0 | 89.9 | 73.3 | <u>46.5</u> | 24.3 | <u>78.7</u> | 58.4 | 46.4 | 32.3 | <u>14.5</u> |
| E2H-C | 100 | 0.0 | 0.0 | 0.0 | 0.0 | 100 | 55.3 | 0.0 | 0.0 | 0.0 | 74.8 | <u>60.6</u> | <u>48.3</u> | 34.3 | 15.8 | |

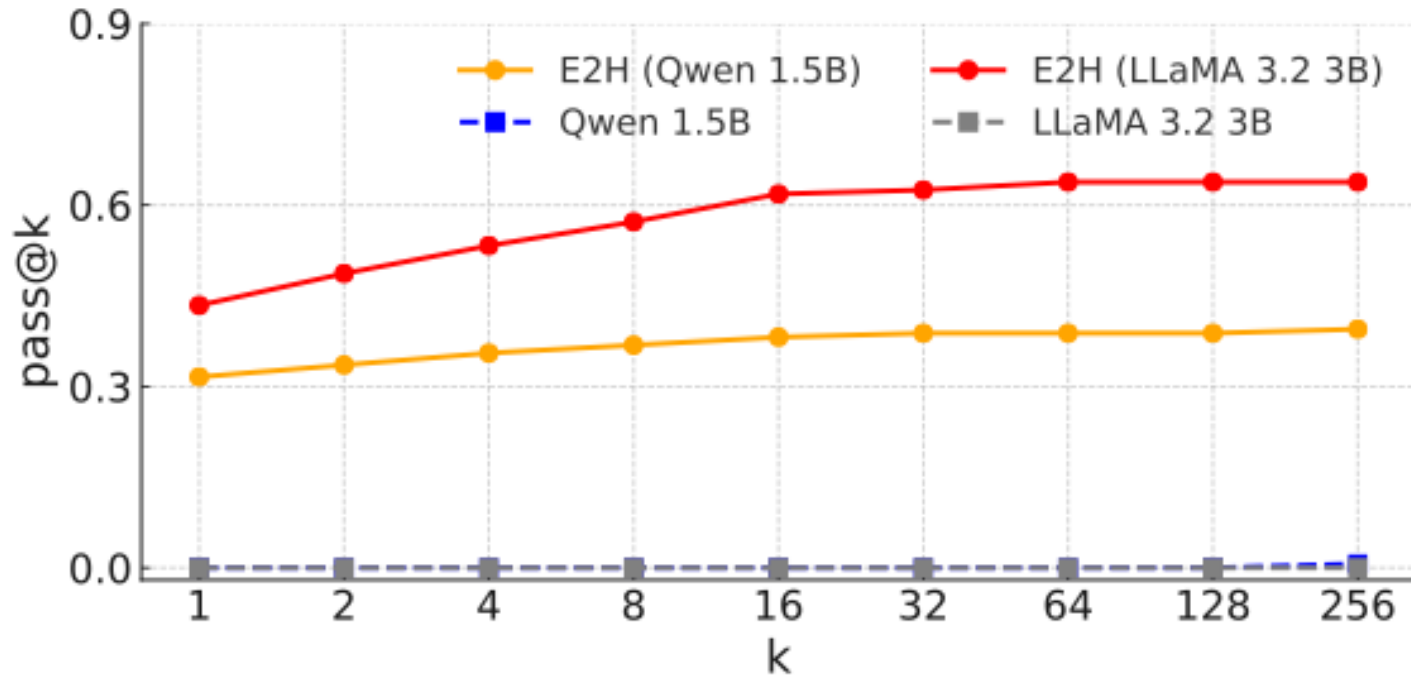
| Qwen 1.5B Instruct | | | | | | | | | | |
|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | GSM8K | | | | | AQuA | | | | |
| | Trivial | Easy | Med | Hard | Avg | Trivial | Easy | Med | Hard | Avg |
| CoT | 90.2 | 87.3 | 76.5 | 38.1 | 67.7 | 70.8 | 51.3 | 20.8 | 2.6 | 40.9 |
| GRPO | 99.0 | <u>95.3</u> | 84.1 | 49.9 | 77.1 | 95.8 | 68.0 | 48.6 | 21.0 | 63.3 |
| CL | 98.0 | 97.2 | 85.8 | 52.2 | 78.6 | 88.8 | 72.2 | 36.1 | 18.4 | 58.6 |
| Self-Evolve | <u>98.1</u> | <u>95.3</u> | <u>87.0</u> | 50.3 | 77.8 | <u>94.4</u> | <u>75.0</u> | 40.3 | 31.6 | <u>64.2</u> |
| E2H-G | 97.6 | 94.7 | 89.0 | <u>51.8</u> | 78.7 | 90.2 | 81.9 | <u>43.0</u> | 34.2 | 66.1 |
| E2H-C | 98.0 | <u>95.3</u> | 83.9 | 46.6 | 75.7 | 86.1 | 72.2 | 48.6 | <u>26.3</u> | 62.5 |

| Method | Variant | Blocksworld | | | | | Countdown | | | | | MATH | | | | |
|--------|----------|-------------|------|------|------|-----|-----------|------|------|------|------|-------|------|------|------|------|
| | | Triv. | Easy | Med. | Hard | OOD | Triv. | Easy | Med. | Hard | OOD | Triv. | Easy | Med. | Hard | OOD |
| GRPO | Baseline | 98 | 100 | 83.3 | 21.1 | 2.6 | 96.1 | 64.9 | 28.8 | 18.1 | 9.2 | 87.2 | 72.0 | 61.6 | 46.3 | 25.7 |
| | E2H-G | 98 | 100 | 95.3 | 32.9 | 7.3 | 97.9 | 87.2 | 70.4 | 41.0 | 19.2 | 85.3 | 71.7 | 62.5 | 48.7 | 27.6 |
| | E2H-C | 100 | 100 | 15.5 | 0.0 | 0.0 | 96.7 | 64.0 | 25.9 | 15.8 | 6.4 | 84.6 | 69.6 | 63.0 | 47.6 | 28.6 |
| DAPO | Baseline | 100 | 100 | 89.3 | 24.3 | 3.4 | 96.5 | 67.3 | 35.7 | 22.0 | 10.9 | 85.6 | 73.1 | 63.3 | 47.9 | 28.3 |
| | E2H-G | 100 | 100 | 98.8 | 46.7 | 9.3 | 99.9 | 90.6 | 82.4 | 59.4 | 30.1 | 86.3 | 72.7 | 64.4 | 49.1 | 28.5 |
| | E2H-C | 100 | 100 | 96.4 | 44.1 | 8.7 | 96.5 | 64.1 | 26.0 | 16.7 | 8.7 | 86.3 | 72.3 | 64.5 | 50.2 | 29.2 |

E2H is
complimentary
to DAPO

- DAPO (a GRPO variant) adaptively focuses training on problems with non zero advantage, where the batch contains a mix of correct and incorrect samples.
- Combining DAPO with E2H is consistently strong and yields additional gains.
- E2H calibrates DAPO's adaptive sampling to problems that are learnable but not yet saturated.

E2H enables Small Language Model Reasoning



E2H enables LLMs to solve problems they fail on even after up to 256 attempts.

Q: Using the numbers [89, 48, 99, 28, 5, 39], create an equation that equals 14

Base LLaMA 3.2 3B
<answer> I am not able to find an equation </answer>

E2H Reasoner (LLaMA 3.2 3B)
<think>First, we can add 5 and 99 to get 104, then subtract 48 from that result to get 56, then subtract 28 from that to get 28, then subtract 39 from that to get -11, then subtract 5 from that to get -16, then subtract 89 from that to get -105, which is not equal to 14. Next, we can subtract 99 from 48 to get -51, then add 89 to get 38, then subtract 5 from that to get 33, then subtract 28 from that to get 5, then add 39 to get 44, which is not equal to 14. We can also try subtracting 48 from 89 to get 41, then subtracting 99 from that to get -58, then adding 5 to get -53, then adding 28 to get -25, then adding 39 to get 14.</think>
<answer>89 - 48 - 99 + 28 + 5 + 39</answer>

LLaMA 3.2 3B Reasoning trace