Generative Verifiers:
Reward Modeling as Next-Token Prediction

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi,
Aviral Kumar, Rishabh Agarwal

Google DeepMind · University of Toronto · Mila · UCLA · Carnegie Mellon University

ICLR 2025

LLMs are becoming good at reasoning

Write a bash script that takes a matrix represented as a string with format '[1,2],[3,4],[5,6]' and prints the transpose in the same format.

Coding Tasks

Question: For every a, b, $b \neq a$ prove that

$$\frac{a^2 + b^2}{2} > \left(\frac{a+b}{2}\right)^2.$$

Math Tasks

But can **Reward Models** catch the mistakes made by LLMs?

A Motivating Example

Problem: Tim decides to cancel his cable subscription and get streaming services. He gets Netflix for \$10 a month. Hulu and Disney Plus normally cost \$10 a month each but he saves 20% for bundling. How much money does he save by cancelling his \$60 cable package?

Solution: Tim pays \$60 for cable. He gets Netflix for 10 and the bundle of Hulu and Disney Plus costs \$10 * 80% = \$8. So he pays \$10 + \$8 = \$18 for the bundle. Now he saves \$60 - \$18 = \$42. The answer is 42.

Zhang, Lunjun, et al. "Generative Verifiers: Reward Modeling as Next-Token Prediction."

A Motivating Example

Problem: Tim decides to cancel his cable subscription and get streaming services. He gets Netflix for $10 a month. Hulu and Disney Plus normally cost $10 a month **each** but he saves 20% for bundling. How much money does he save by cancelling his $60 cable package?

Solution: Tim pays $60 for cable. He gets Netflix for 10 and the bundle of Hulu and Disney Plus costs $10 * 80% = $8. So he pays $10 + $8 = $18 for the bundle. Now he saves $60 - $18 = $42. The answer is 42.

LLM-generated solutions often sound convincing even when they are wrong

Zhang, Lunjun, et al. "Generative Verifiers: Reward Modeling as Next-Token Prediction."

A Motivating Example

Problem: Tim decides to cancel his cable subscription and get streaming services. He gets Netflix for $10 a month. Hulu and Disney Plus normally cost $10 a month **each** but he saves 20% for bundling. How much money does he save by cancelling his $60 cable package?

Solution: Tim pays $60 for cable. He gets Netflix for 10 and the bundle of Hulu and Disney Plus costs $10 * 80% = $8. So he pays $10 + $8 = $18 for the bundle. Now he saves $60 - $18 = $42. The answer is 42.

Discriminative RM correctness score: 0.999

Zhang, Lunjun, et al. "Generative Verifiers: Reward Modeling as Next-Token Prediction."

A Motivating Example

Problem: Tim decides to cancel his cable subscription and get streaming services. He gets Netflix for \$10 a month. Hulu and Disney Plus normally cost \$10 a month <u>each</u> but he saves 20% for bundling. How much money does he save by cancelling his \$60 cable package?
Solution: Tim pays \$60 for cable. He gets Netflix for 10 and the bundle of Hulu and Disney Plus costs \$10 * 80% = \$8. So he pays \$10 + \$8 = \$18 for the bundle. Now he saves \$60 - \$18 = \$42. The answer is 42.
Discriminative RM correctness score: 0.999

Reward Models (RMs) today are not very good at determining correctness for reasoning tasks.

Zhang, Lunjun, et al. "Generative Verifiers: Reward Modeling as Next-Token Prediction."

Related Work

Reward models for reasoning:

- (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2023)

Prompting the language model to verify a solution:

- **LLM-as-a-Judge** (Bai et al., 2022; Kim et al., 2023; Ling et al., 2024; Zheng et al., 2024)
- "**Large Language Models Cannot Self-Correct Reasoning Yet**", Huang et al, 2023

Training language models to verify self-generated solutions / self-correct:

- **CriticGPT** (McAleese et al, 2024)
- Training Language Models to Self-Correct via Reinforcement Learning (Kumar et al, 2024)

Classical Reward Models

- Finetune a pre-trained LLM to use one of its logits as the **binary classifier**
- Why can't it reliably determine solution correctness?
- Hypothesis: **LLMs need tokens to think, even for verification**

Zhang, Lunjun, et al. "**Generative Verifiers**: Reward Modeling as Next-Token Prediction."

Reward Modeling as Next-Token Prediction

- Idea: output a verification CoT **before** determining the score

Problem: Tim decides to cancel his cable subscription and get streaming services. He gets Netflix for $10 a month. Hulu and Disney Plus normally cost $10 a month **each** but he saves 20% for bundling. How much money does he save by cancelling his $60 cable package?

Solution: Tim pays $60 for cable. He gets Netflix for 10 and the bundle of Hulu and Disney Plus costs $10 * 80% = $8. So he pays $10 + $8 = $18 for the bundle. Now he saves $60 - $18 = $42. The answer is 42.

GenRM-CoT: Let's verify step by step. · · ·
Step 3. **The bundle of Hulu and Disney Plus costs $10 * 80% = $8.** (Incorrect)
- Expected Answer: The bundle of Hulu and Disney Plus costs $10 + $10 = $20. With the 20% discount, the total cost is $20 * 0.8 = $16. · · ·

Zhang, Lunjun, et al. "Generative Verifiers: Reward Modeling as Next-Token Prediction."

Reward Modeling as Next-Token Prediction

- Idea: output a verification CoT **before** determining the score

Problem: Tim decides to cancel his cable subscription and get streaming services. He gets Netflix for $10 a month. Hulu and Disney Plus normally cost $10 a month **each** but he saves 20% for bundling. How much money does he save by cancelling his $60 cable package?

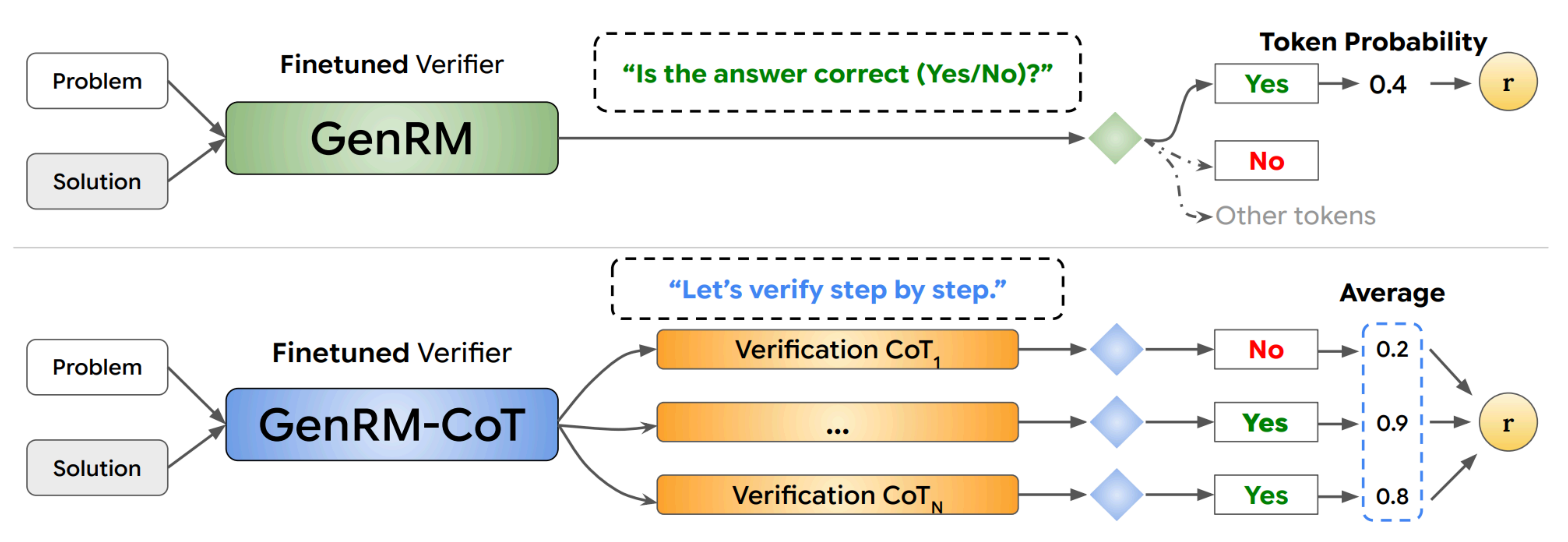
Solution: Tim pays $60 for cable. He gets Netflix for 10 and the bundle of Hulu and Disney Plus costs $10 * 80% = $8. So he pays $10 + $8 = $18 for the bundle. Now he saves $60 - $18 = $42. The answer is 42.

GenRM-CoT: Let's verify step by step. · · ·
Step 3. **The bundle of Hulu and Disney Plus costs $10 * 80% = $8.** (Incorrect)
- Expected Answer: The bundle of Hulu and Disney Plus costs $10 + $10 = $20. With the 20% discount, the total cost is $20 * 0.8 = $16. · · ·
Verification: Is the answer correct (Yes/No)? No

GenRM-CoT (Majority Voting) score: 0.0015

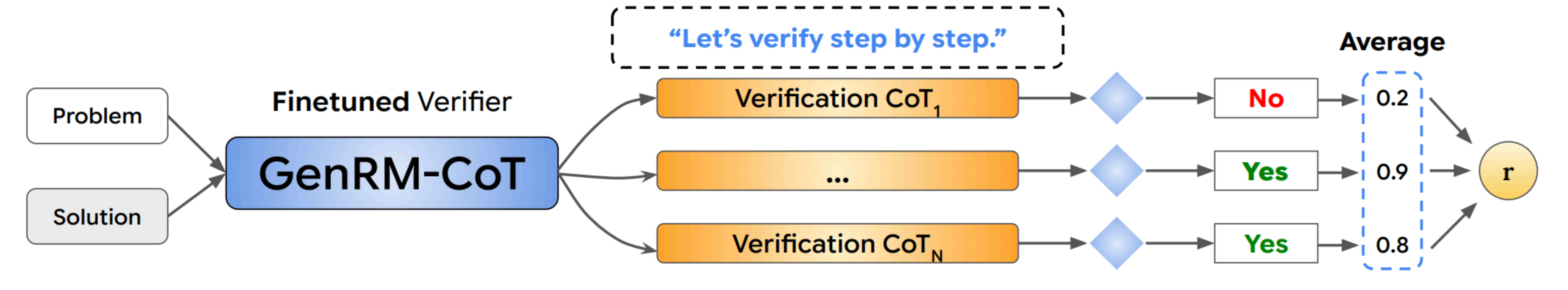
Zhang, Lunjun, et al. "Generative Verifiers: Reward Modeling as Next-Token Prediction."

Reward Modeling as Next-Token Prediction



Zhang, Lunjun, et al. "Generative Verifiers: Reward Modeling as Next-Token Prediction."

Reward Modeling as Next-Token Prediction



At test-time, we sample multiple CoT rationales and use **majority voting** to compute the average probability of 'Yes', enabling GenRM-CoT to utilize **additional inference-compute for better verification**

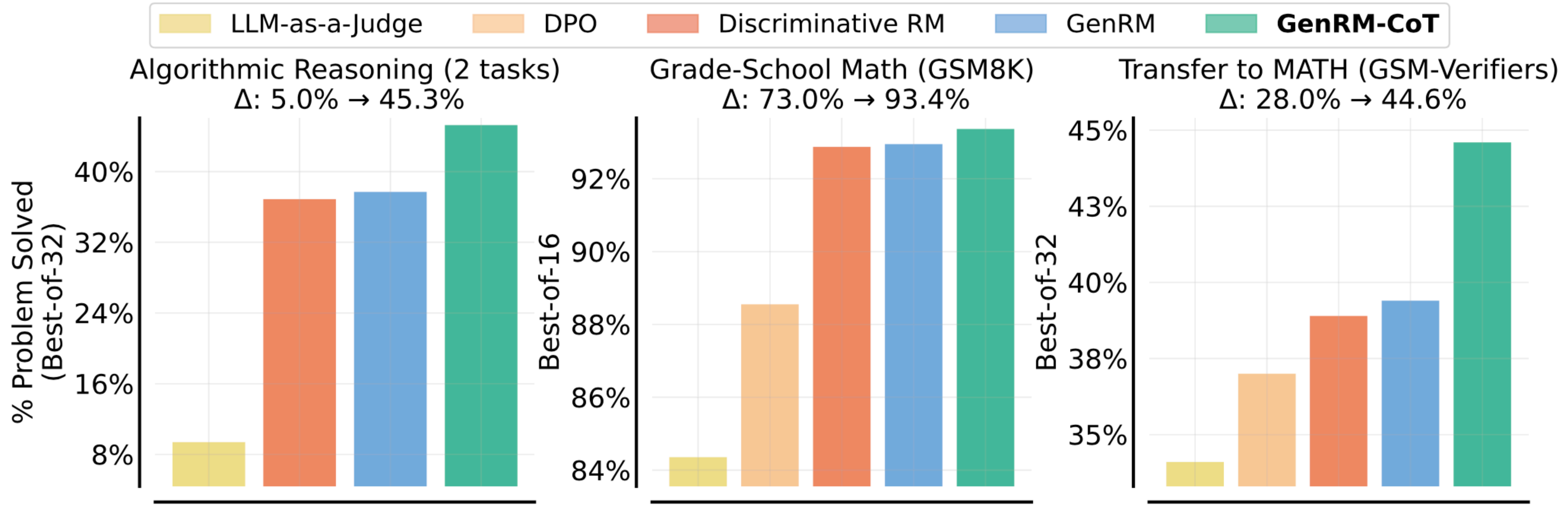
Zhang, Lunjun, et al. "**Generative Verifiers**: Reward Modeling as Next-Token Prediction."

Synthetic Data for Training

- Use model-generated **verification CoT** for training, **filtered** based on correctness
- Provide a **reference solution** during training data generation, making it easier for an LLM to point out any reasoning error
    - Reference solution: any model-generated solution that arrives at the correct final answer
    - Not included during actual finetuning, so **no train/test mismatch**

Zhang, Lunjun, et al. "**Generative Verifiers**: Reward Modeling as Next-Token Prediction."

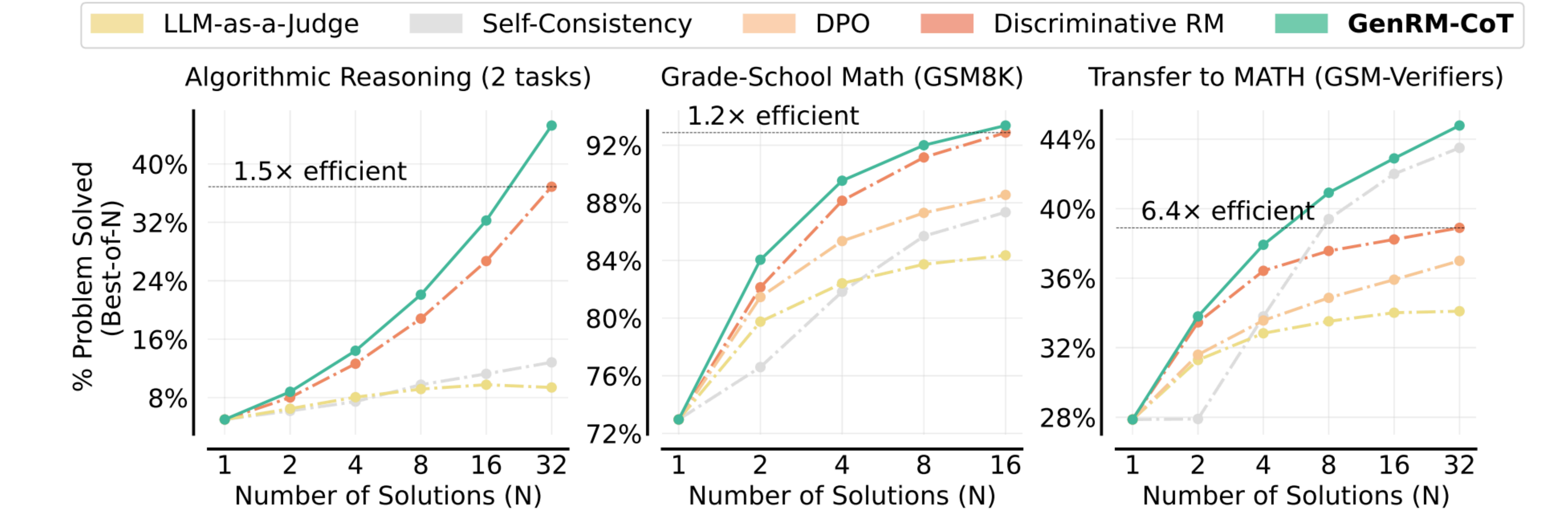
Reward Modeling as Next-Token Prediction



Outperforms LLM-as-a-Judge, DPO, and classical RM on reasoning

Zhang, Lunjun, et al. "**Generative Verifiers**: Reward Modeling as Next-Token Prediction."

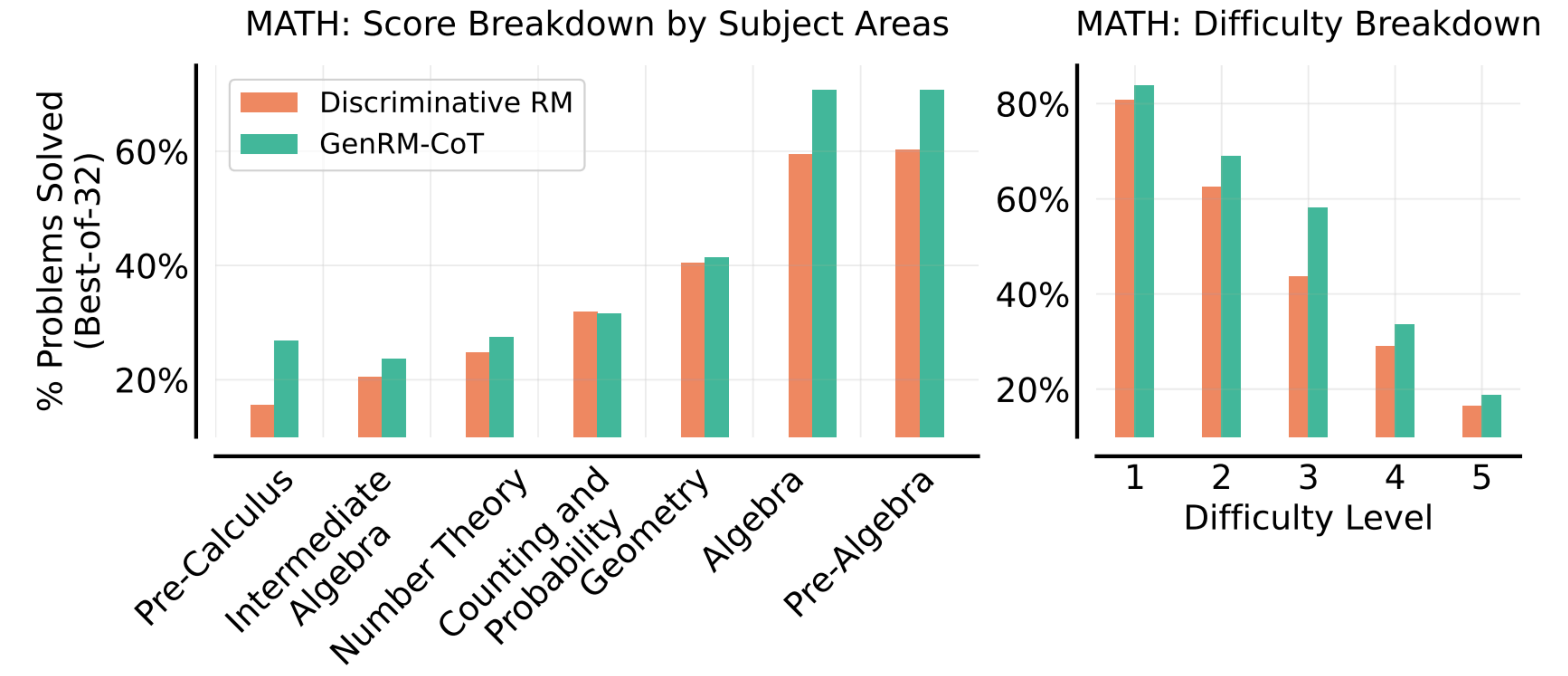
Reward Modeling as Next-Token Prediction



6.4x efficient than Classical RM on MATH

Zhang, Lunjun, et al. "Generative Verifiers: Reward Modeling as Next-Token Prediction."

Reward Modeling as Next-Token Prediction



Easy-to-Hard Generalization

from Grade School Math to high-school math

Zhang, Lunjun, et al. "Generative Verifiers: Reward Modeling as Next-Token Prediction."

Reward Modeling as Next-Token Prediction

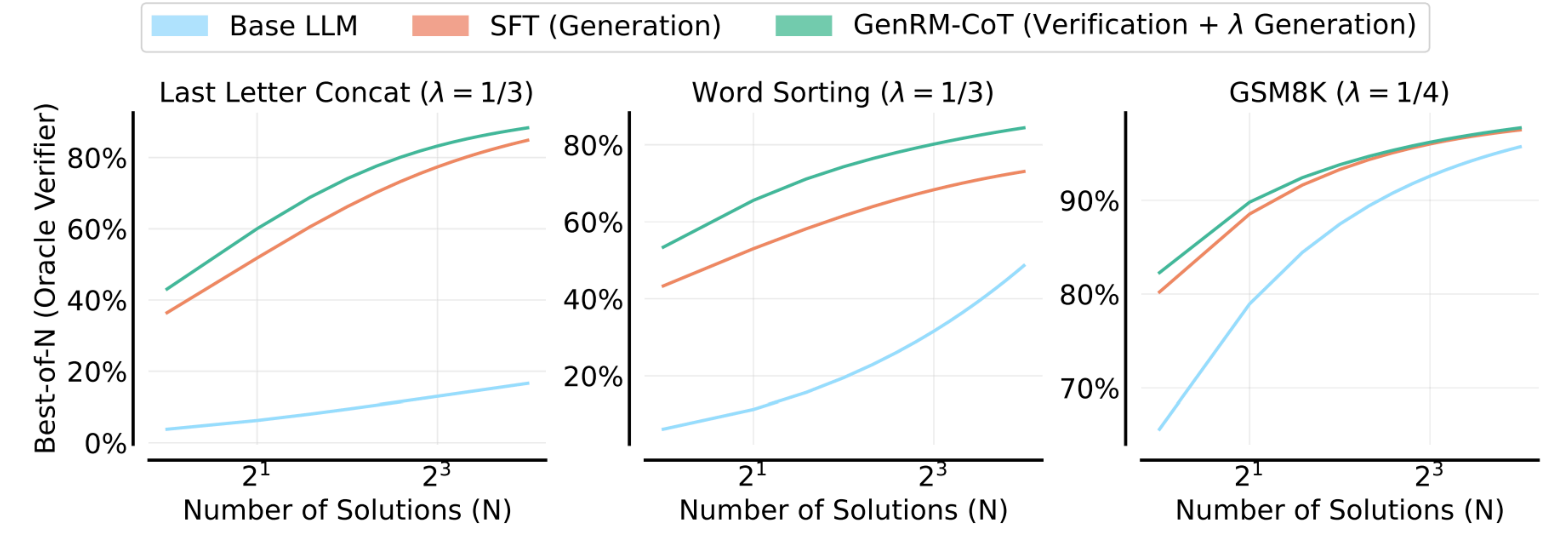
MMLU Dataset	Base Model (Pass@1)	Disc-RM	GenRM-CoT	Improvement
elementary_mathematics	80.1%	90.6%	**91.1%**	+0.5%
high_school_mathematics	52.2%	74.8%	**76.1%**	+1.3%
college_mathematics	47.6%	53%	**56.1%**	+3.1%
abstract_algebra	37.9%	50%	**53.50%**	+3.5%

Easy-to-Hard Generalization

The improvements are more significant on harder tasks

Zhang, Lunjun, et al. "**Generative Verifiers**: Reward Modeling as Next-Token Prediction."

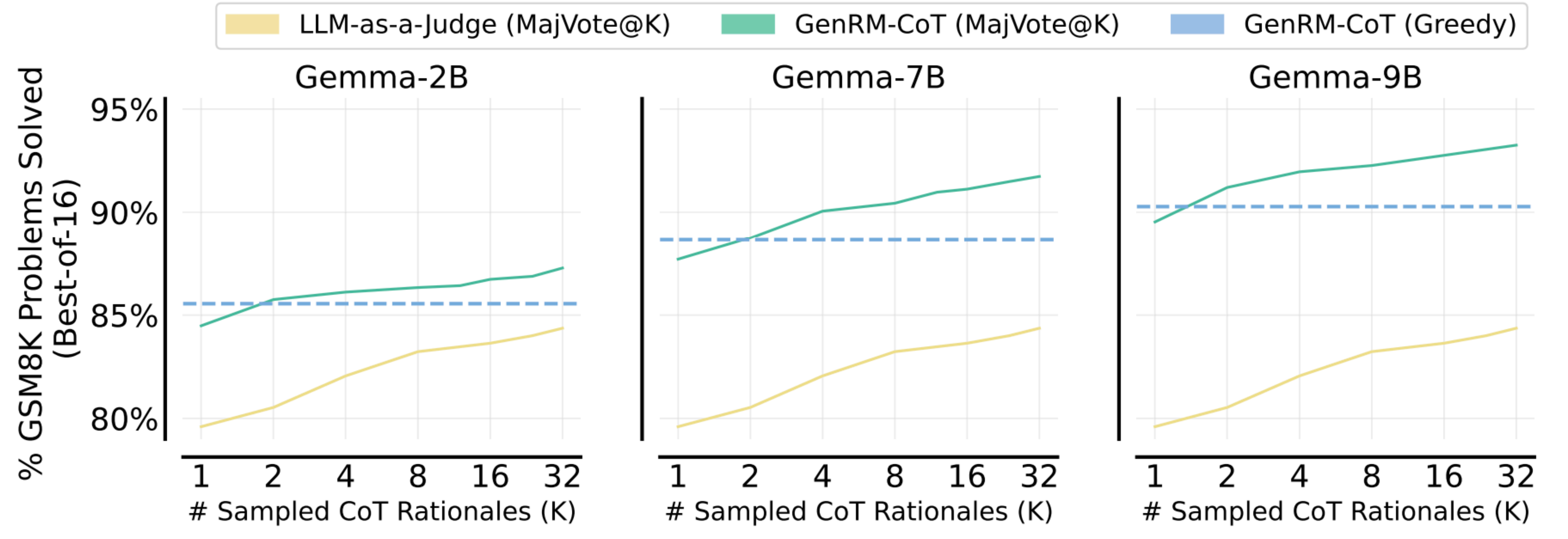
Unifying Generation and Verification



GenRM-CoT allows an LLM policy to also be used as a RM.

Zhang, Lunjun, et al. "**Generative Verifiers**: Reward Modeling as Next-Token Prediction."

Scaling Test-time Compute



GenRM-CoT allows an LLM to think more and perform better

Zhang, Lunjun, et al. "**Generative Verifiers**: Reward Modeling as Next-Token Prediction."

Thank you for listening