DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Zhihong Shao[1,2*†], Peiyi Wang[1,3*†], Qihao Zhu[1,3*†], Runxin Xu[1], Junxiao Song[1]
Xiao Bi[1], Haowei Zhang[1], Mingchuan Zhang[1], Y.K. Li[1], Y. Wu[1], Daya Guo[1*]

[1]DeepSeek-AI, [2]Tsinghua University, [3]Peking University

arXiv preprint 2025

Humane Lab 윤예준

25. 03. 28

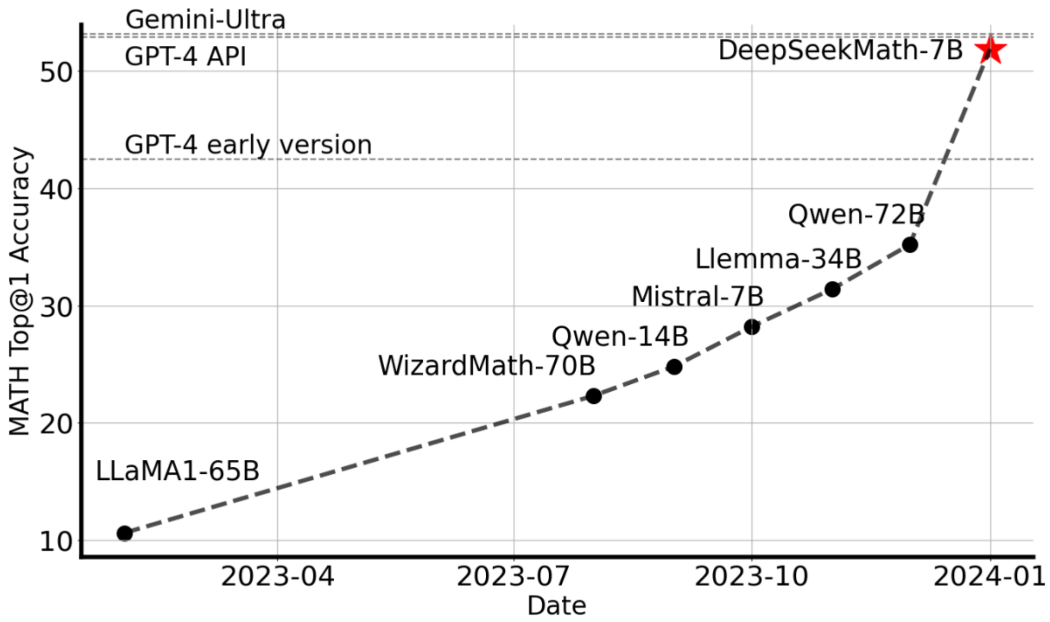
Introduction

- Background
  - Mathematical reasoning is challenging for LLMs due to its complexity and structure
  - Closed-source models (e.g., GPT-4) perform well but are not publicly available
  - Open-source models still significantly lag behind in performance

- Goal
  - Build DeepSeekMath: a powerful open-source model for mathematical reasoning
  - Achieve GPT-4-level performance on academic math benchmarks

Key Contributions

- Constructed a 120B-token high-quality math corpus for pretraining
- Proposed a new RL algorithm: Group Relative Policy Optimization (GRPO)
- Achieved SOTA on the MATH benchmark among open-source models

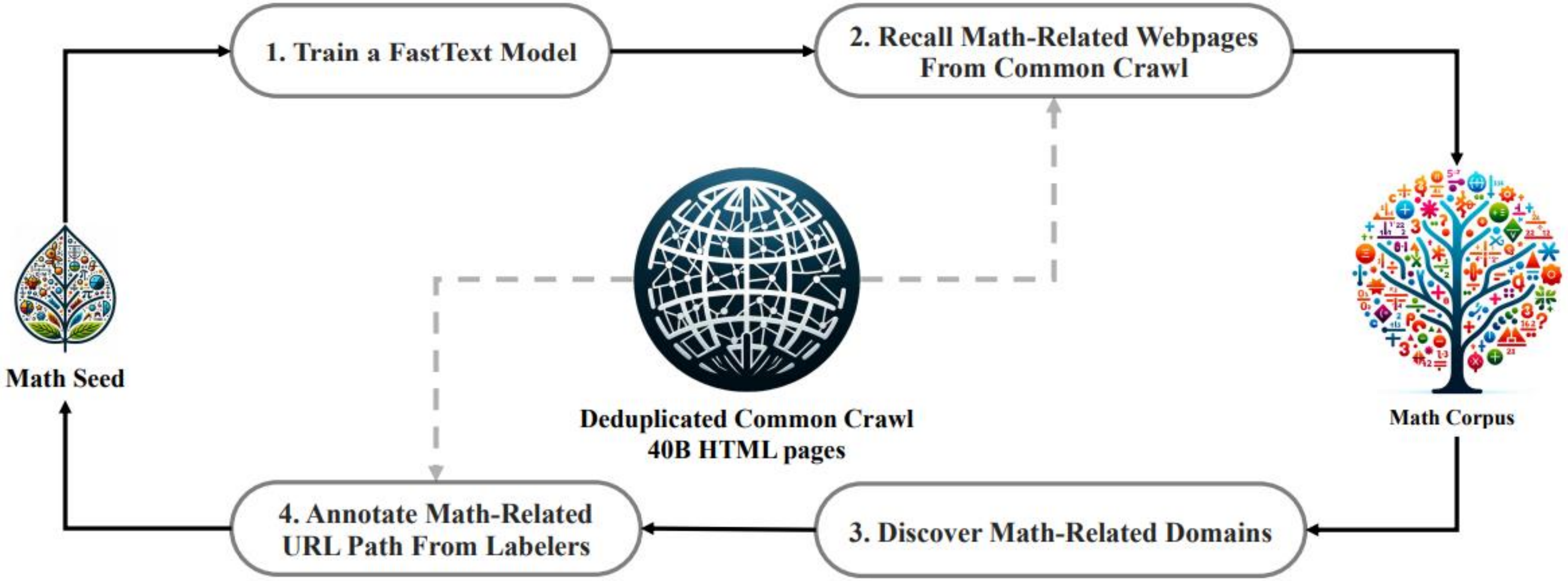


DeepSeekMath Pipeline - Overview

- Build pre-training dataset
- Continued Pre-training
- Supervised Fine-Tuning
- Reinforcement Learning

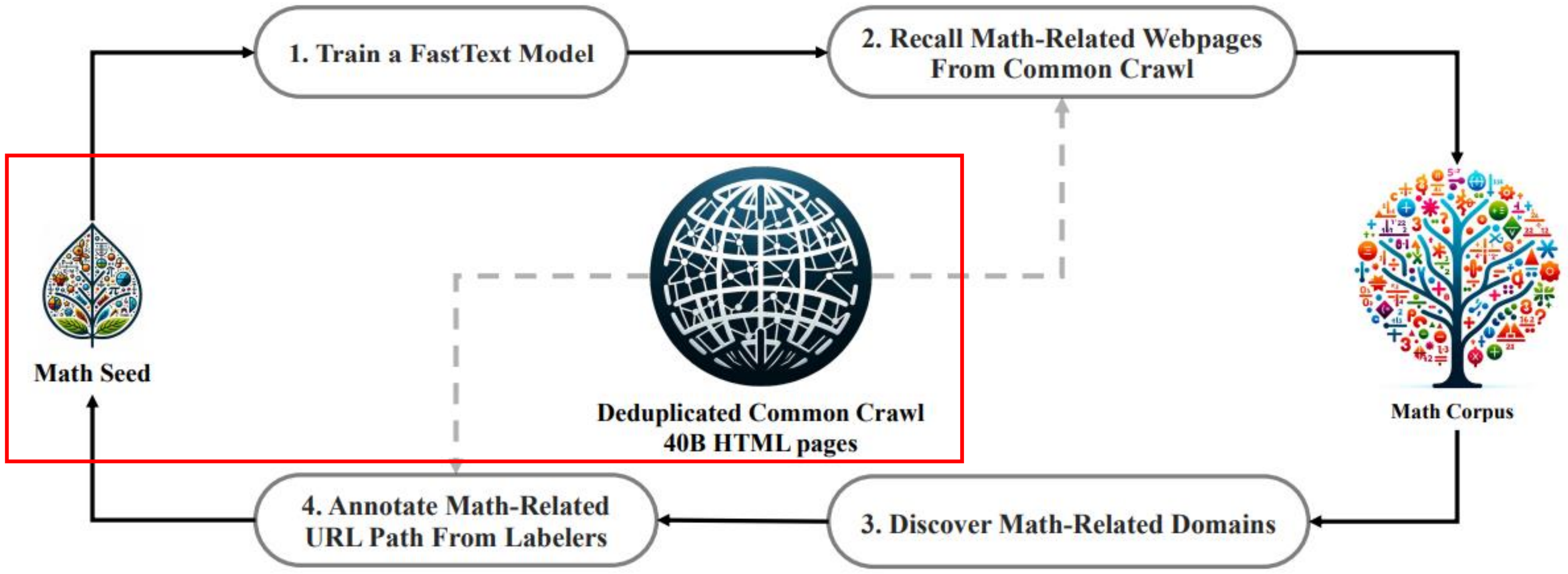
Math Corpus Construction – Overview

- Build a high-quality mathematical corpus (DeepSeekMath Corpus)
- Extracted from Common Crawl (CC) with over 120B tokens



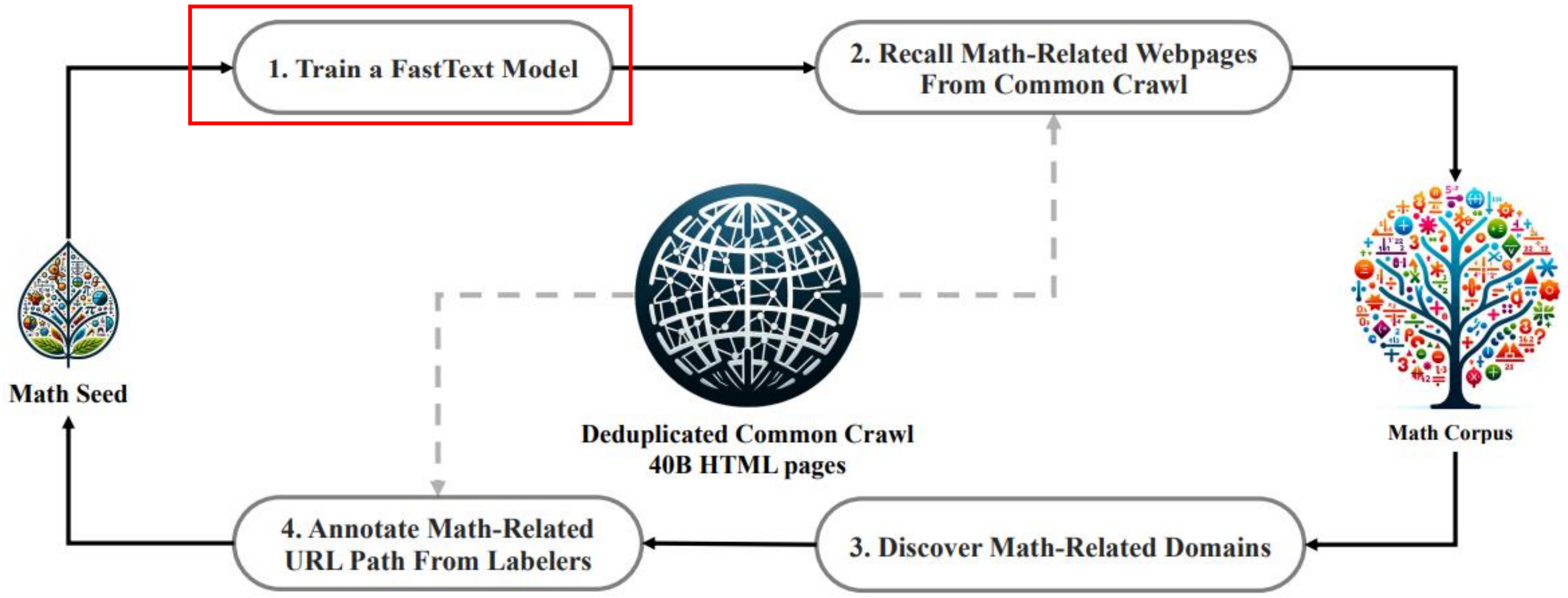
Data Collection

- Initial seed corpus: OpenWebMath
  - OpenWebMath is a high-quality mathematical web texts that is collected from Common Crawl



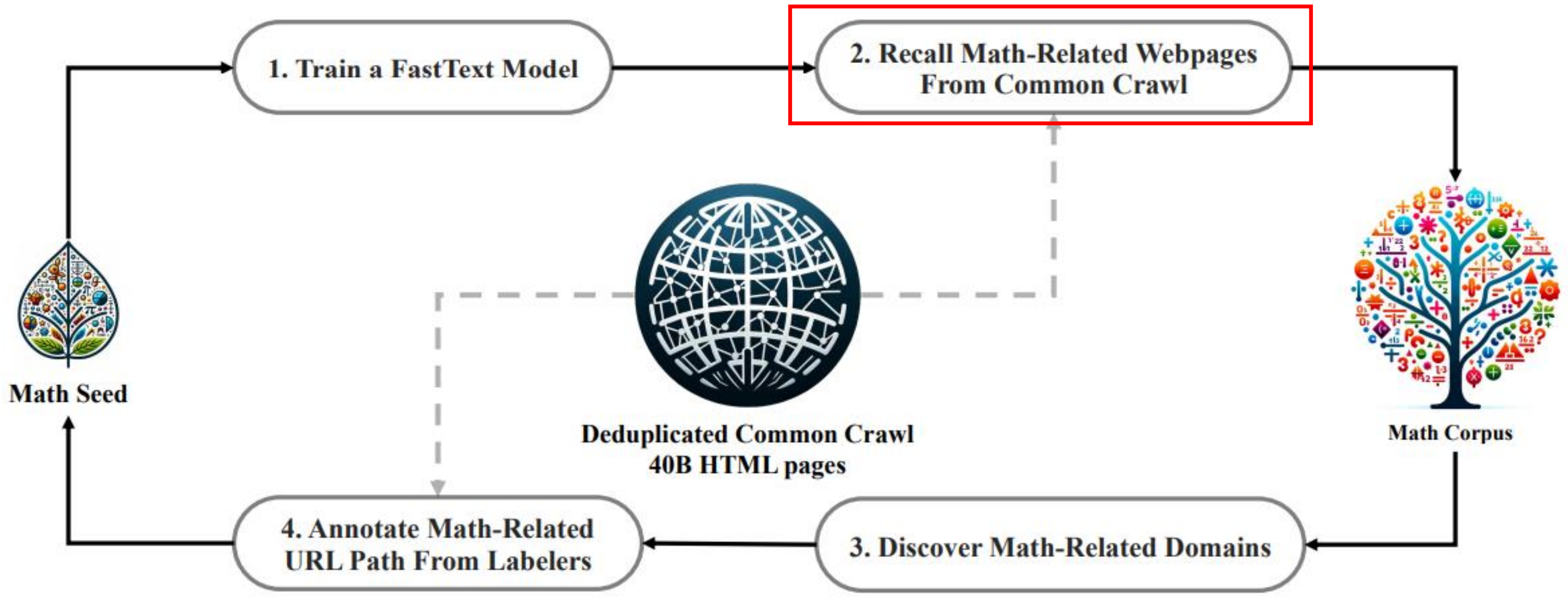
Data Decontamination

- Train a fastText classifier with:
  - 500K positive samples (Math Seed)
  - 500K negative samples (Common Crawl)



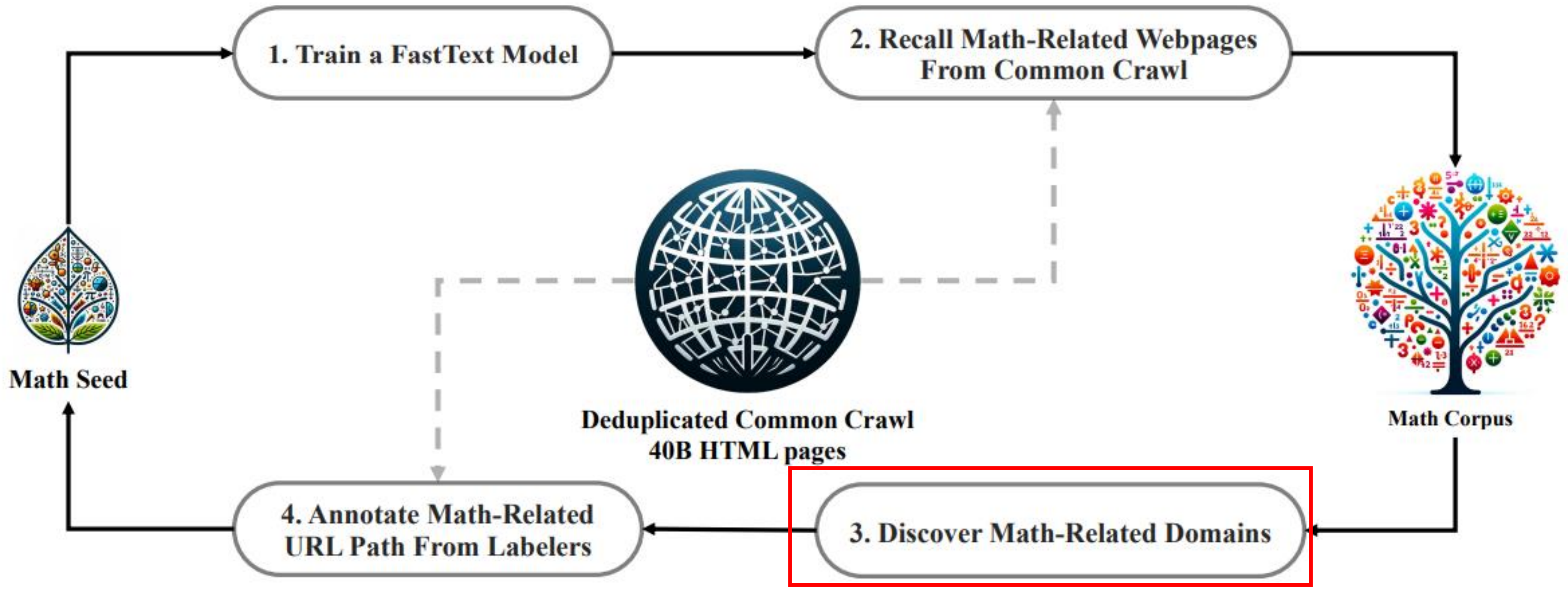
Data Decontamination

- Web Data Recall from Common Crawl
  - Apply trained classifier on deduplicated CC (40B HTML pages)
  - Score and rank each page by math relevance → Select 40B tokens



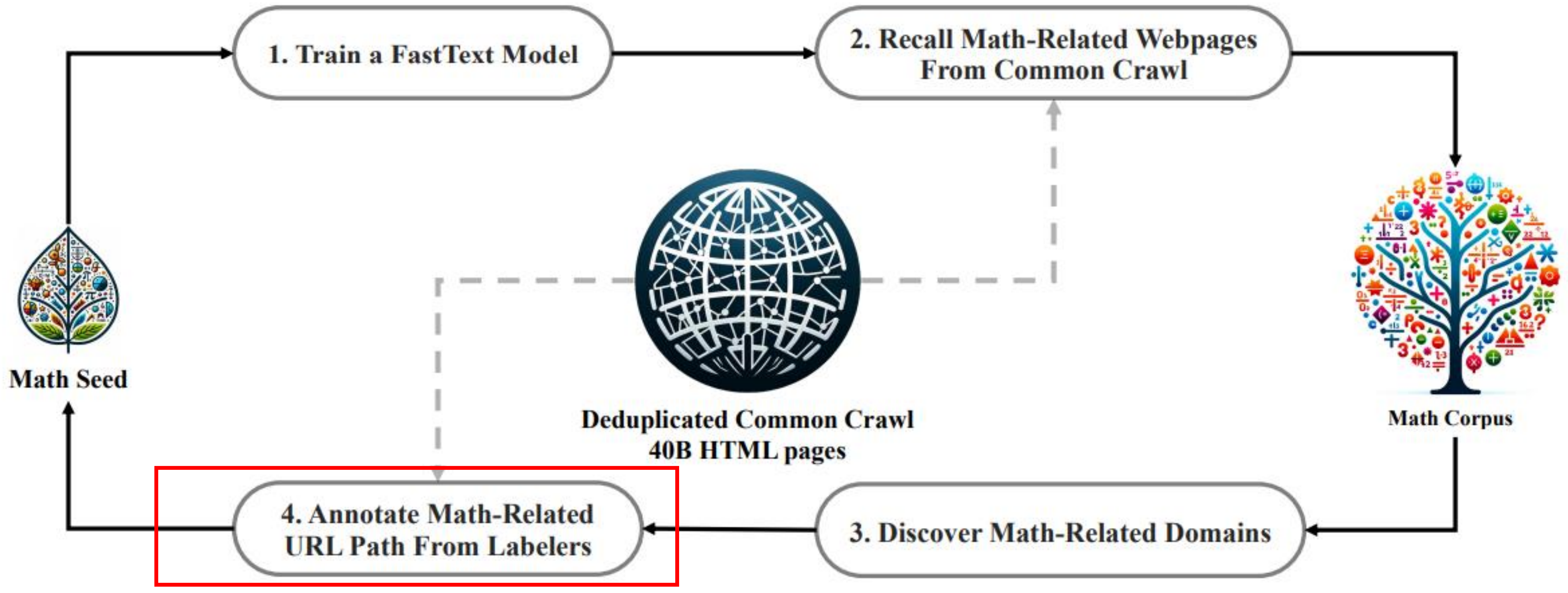
Iterative Data Expansion

- New Domain Discovery
  - Identify math-heavy domains (e.g., mathoverflow.net)
  - If >10% of domain pages are math, classify as math-related



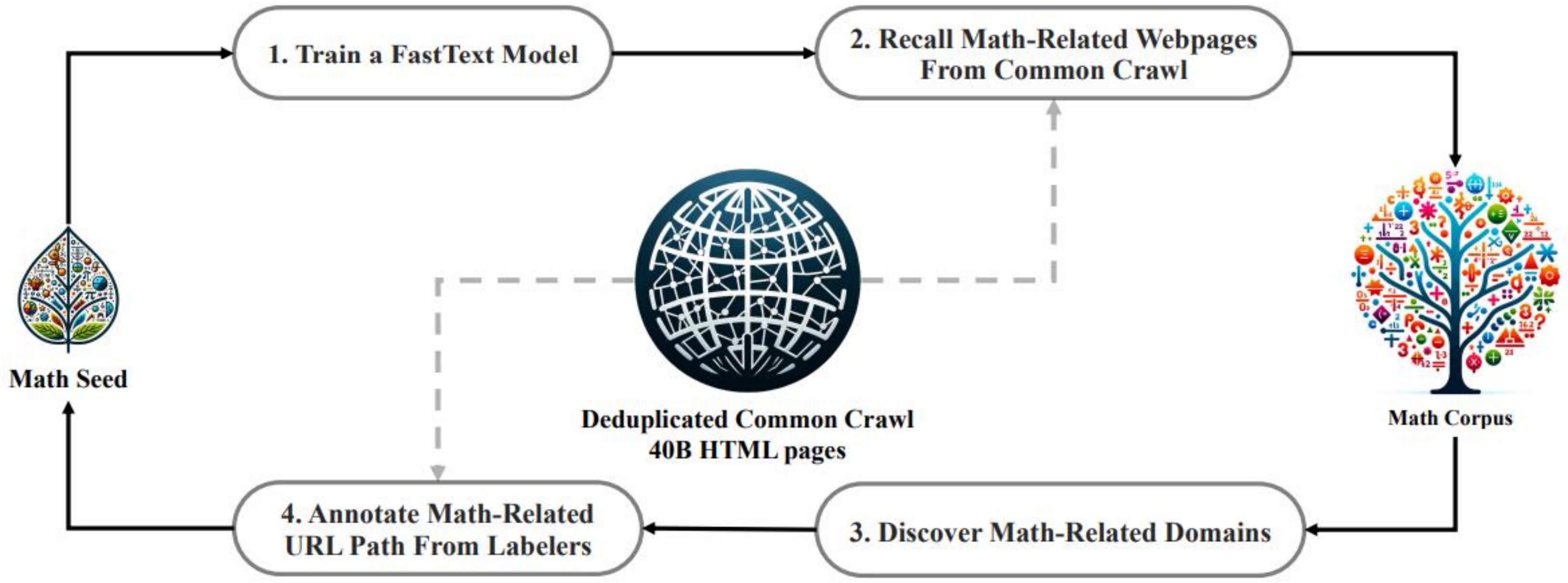
Iterative Data Expansion

- Manual Annotation
  - Annotate math-specific URL paths (e.g., mathoverflow.net/questions)
  - Add missed data to seed corpus → Retrain classifier



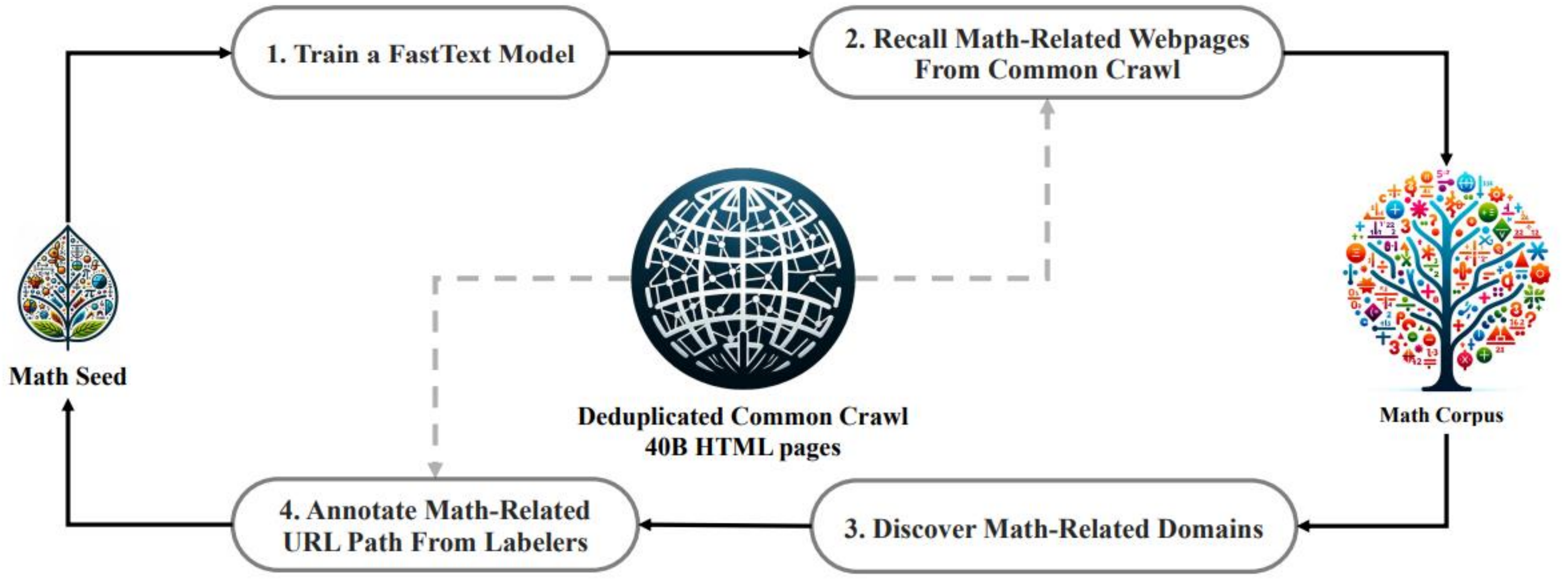
Iterative Data Expansion

- After 4 iterations:
  - Collected 35.5M math pages
  - Total: 120B tokens



Benchmark Decontamination

- Remove overlap with evaluation datasets:
  - GSM8K, MATH, CMATH, AGIEval
- Method: Exact 10-gram matching + rule-based filtering



Superiority of DeepSeekMath Corpus

- DeepSeekMath Corpus is of high quality, covers multilingual mathematical content, and is the largest in size

Math Corpus	Size	English Benchmarks					Chinese Benchmarks		
		GSM8K	MATH	OCW	SAT	MMLU STEM	CMATH	Gaokao MathCloze	Gaokao MathQA
No Math Training	N/A	2.9%	3.0%	2.9%	15.6%	19.5%	12.3%	0.8%	17.9%
MathPile	8.9B	2.7%	3.3%	2.2%	12.5%	15.7%	1.2%	0.0%	2.8%
OpenWebMath	13.6B	11.5%	8.9%	3.7%	31.3%	29.6%	16.8%	0.0%	14.2%
Proof-Pile-2	51.9B	14.3%	11.2%	3.7%	43.8%	29.2%	19.9%	5.1%	11.7%
DeepSeekMath Corpus	**120.2B**	**23.8%**	**13.6%**	**4.8%**	**56.3%**	**33.1%**	**41.5%**	**5.9%**	**23.6%**

Pre-training DeepSeekMath-Base 7B

- Continued Pre-training with DeepSeekMath Corpus
- Pre-training Setup
  - Base model: Initialized from DeepSeek-Coder-Base-v1.5 (7B)
  - Token count: Trained on 500B tokens
    - 56% from DeepSeekMath Corpus
    - 4% AlgebraicStack, 10% arXiv, 20% GitHub code, 10% natural language (CC)
  - Context length: 4K tokens
  - Optimizer: AdamW, LR = 4.2e-4, Batch Size = 10M tokens

DeepSeekMath-Base 7B vs Open/Closed Models

Model	Size	English Benchmarks					Chinese Benchmarks		
		GSM8K	MATH	OCW	SAT	MMLU STEM	CMATH	Gaokao MathCloze	Gaokao MathQA
Closed-Source Base Model									
Minerva	7B	16.2%	14.1%	7.7%	-	35.6%	-	-	-
Minerva	62B	52.4%	27.6%	12.0%	-	53.9%	-	-	-
Minerva	540B	58.8%	33.6%	17.6%	-	63.9%	-	-	-
Open-Source Base Model									
Mistral	7B	40.3%	14.3%	9.2%	71.9%	51.1%	44.9%	5.1%	23.4%
Llemma	7B	37.4%	18.1%	6.3%	59.4%	43.1%	43.4%	11.9%	23.6%
Llemma	34B	54.0%	25.3%	10.3%	71.9%	52.9%	56.1%	11.9%	26.2%
DeepSeekMath-Base	7B	**64.2%**	**36.2%**	**15.4%**	**84.4%**	**56.5%**	**71.7%**	**20.3%**	**35.3%**

- Outperforms much larger models (34B–540B)

Program-of-Thought & Theorem Proving

Model	Size	Problem Solving w/ Tools		Informal-to-Formal Proving	
		GSM8K+Python	MATH+Python	miniF2F-valid	miniF2F-test
Mistral	7B	48.5%	18.2%	18.9%	18.0%
CodeLlama	7B	27.1%	17.2%	16.3%	17.6%
CodeLlama	34B	52.7%	23.5%	18.5%	18.0%
Llemma	7B	41.0%	18.6%	20.6%	22.1%
Llemma	34B	64.6%	26.3%	21.0%	21.3%
DeepSeekMath-Base	7B	**66.9%**	**31.4%**	**25.8%**	**24.6%**

- Strong performance in program-aided math

NLU & Reasoning & Code Tasks

Model	Size	MMLU	BBH	HumanEval (Pass@1)	MBPP (Pass@1)
Mistral	7B	**62.4%**	55.7%	28.0%	41.4%
DeepSeek-Coder-Base-v1.5†	7B	42.9%	42.9%	40.2%	52.6%
DeepSeek-Coder-Base-v1.5	7B	49.1%	55.2%	**43.2%**	**60.4%**
DeepSeekMath-Base	7B	54.9%	**59.5%**	40.9%	52.6%

- Math-specialized pretraining improves general reasoning (MMLU, BBH)
- Maintains competitive performance on coding tasks (HumanEval, MBPP)

Supervised Fine-Tuning

- Fine-tuning performed on DeepSeekMath-Base 7B
- Total of 776K math problems in English and Chinese
- Data includes:
  - Chain-of-Thought (CoT)
  - Program-of-Thought (PoT)
  - Tool-integrated reasoning format
- Data Sources:
  - English: GSM8K, MATH, MathInstruct, Lila-OOD
  - Chinese: K-12 problems across 76 subtopics

Chain-of-Thought Reasoning without Tool Use

- DeepSeekMath-Instruct 7B achieves 46.8% on MATH, SOTA among open-source models

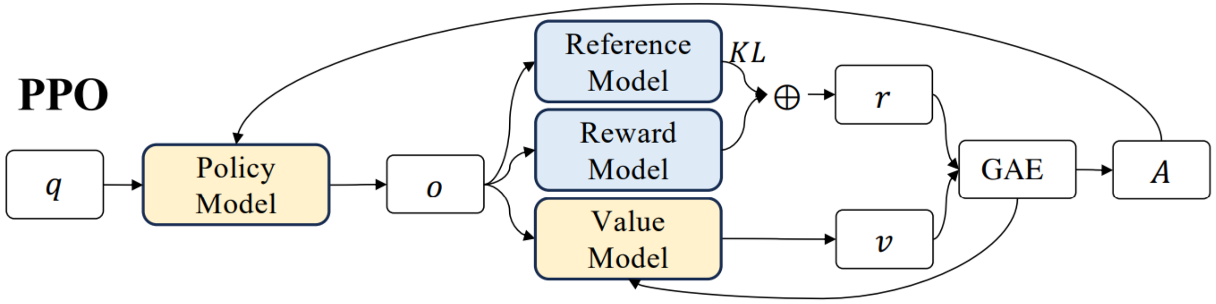
Model	Size	English Benchmarks		Chinese Benchmarks	
		GSM8K	MATH	MGSM-zh	CMATH
Chain-of-Thought Reasoning					
Closed-Source Model					
Gemini Ultra	-	94.4%	53.2%	-	-
GPT-4	-	92.0%	52.9%	-	86.0%
Inflection-2	-	81.4%	34.8%	-	-
GPT-3.5	-	80.8%	34.1%	-	73.8%
Gemini Pro	-	86.5%	32.6%	-	-
Grok-1	-	62.9%	23.9%	-	-
Baichuan-3	-	88.2%	49.2%	-	-
GLM-4	-	87.6%	47.9%	-	-
Open-Source Model					
InternLM2-Math	20B	82.6%	37.7%	-	-
Qwen	72B	78.9%	35.2%	-	-
Math-Shepherd-Mistral	7B	84.1%	33.0%	-	-
WizardMath-v1.1	7B	83.2%	33.0%	-	-
DeepSeek-LLM-Chat	67B	84.1%	32.6%	74.0%	80.3%
MetaMath	70B	82.3%	26.6%	66.4%	70.9%
SeaLLM-v2	7B	78.2%	27.5%	64.8%	-
ChatGLM3	6B	72.3%	25.7%	-	-
WizardMath-v1.0	70B	81.6%	22.7%	64.8%	65.4%
DeepSeekMath-Instruct	7B	82.9%	46.8%	73.2%	84.6%
DeepSeekMath-RL	7B	**88.2%**	**51.7%**	**79.6%**	**88.8%**

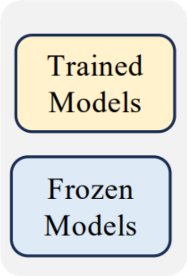
Reasoning with Tool Use: Python Integration

Model	Size	English Benchmarks		Chinese Benchmarks	
		GSM8K	MATH	MGSM-zh	CMATH
Tool-Integrated Reasoning					
Closed-Source Model					
GPT-4 Code Interpreter	-	97.0%	69.7%	-	-
Open-Source Model					
InternLM2-Math	20B	80.7%	54.3%	-	-
DeepSeek-LLM-Chat	67B	86.7%	51.1%	76.4%	85.4%
ToRA	34B	80.7%	50.8%	41.2%	53.4%
MAmmoTH	70B	76.9%	41.8%	-	-
DeepSeekMath-Instruct	7B	83.7%	57.4%	72.0%	84.3%
DeepSeekMath-RL	7B	**86.7%**	**58.8%**	**78.4%**	**87.6%**

- DeepSeekMath-Instruct 7B approaches an accuracy of 60% on MATH, surpassing all existing open-source models

Proximal Policy Optimization (PPO)

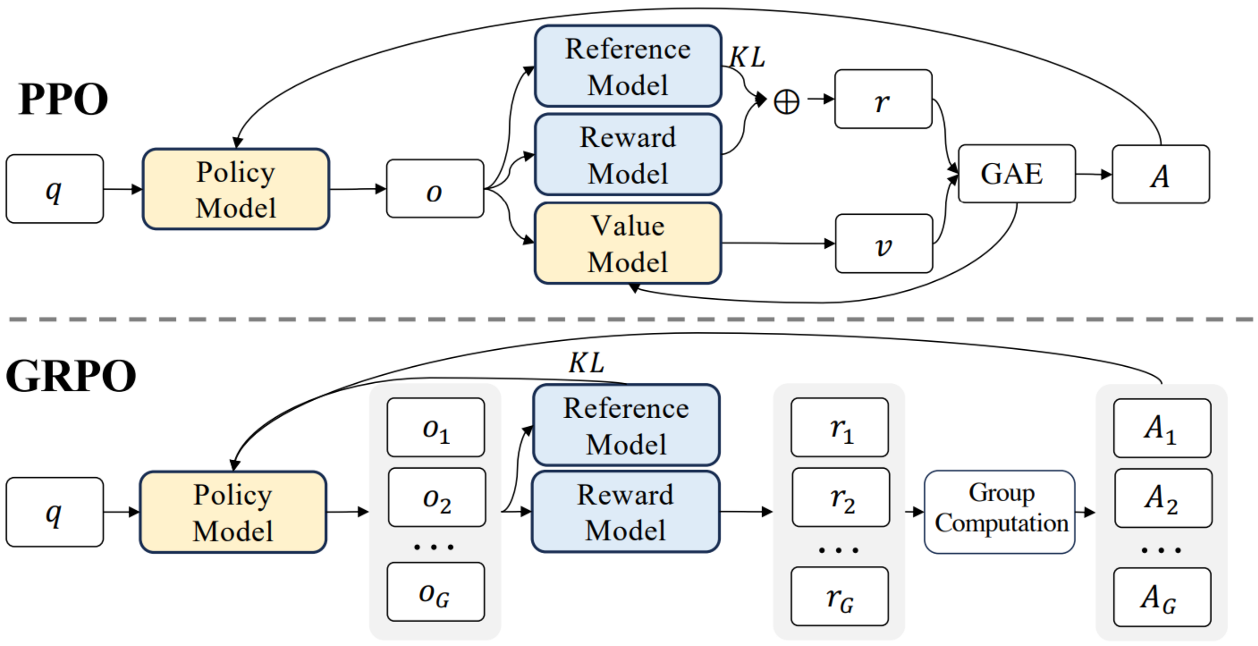


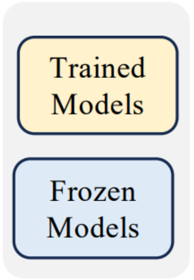


$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min\left[\frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip}\left(\frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1-\varepsilon, 1+\varepsilon \right) A_t \right]$$

$$r_t = r_\varphi(q, o_{\le t}) - \beta \log \frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{ref}(o_t|q, o_{<t})}$$

Group Relative Policy Optimization (GRPO)





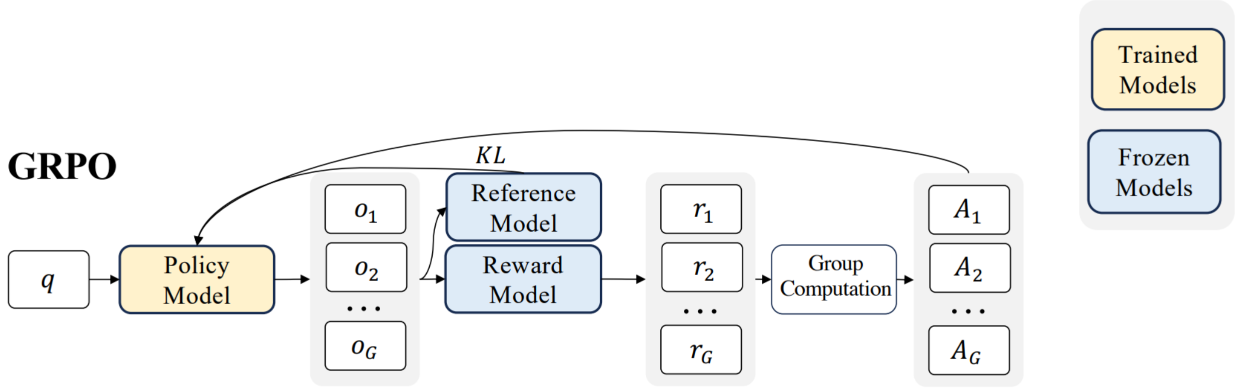
$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1-\varepsilon, 1+\varepsilon \right) A_t \right]$$

$$r_t = r_\varphi(q, o_{\le t}) - \beta \log \frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{ref}(o_t|q, o_{<t})}$$

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left[\frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})}, 1-\varepsilon, 1+\varepsilon \right) \hat{A}_{i,t} \right] - \beta \mathbb{D}_{KL}\left[\pi_\theta || \pi_{ref}\right] \right\}$$

Outcome Supervision RL with GRPO

- Sample multiple outputs $\{o_1, o_2, \ldots, o_G\}$ for a question q
- Use the reward model to score each output $r_1, r_2, \ldots, r_G$
- Normalize rewards: $\hat{A}_{i,t} = \widetilde{r}_i = \frac{r_i - \mathrm{mean}(\mathbf{r})}{\mathrm{std}(\mathbf{r})}$
- Assign $\tilde{r}_i$ to all tokens in output o_i



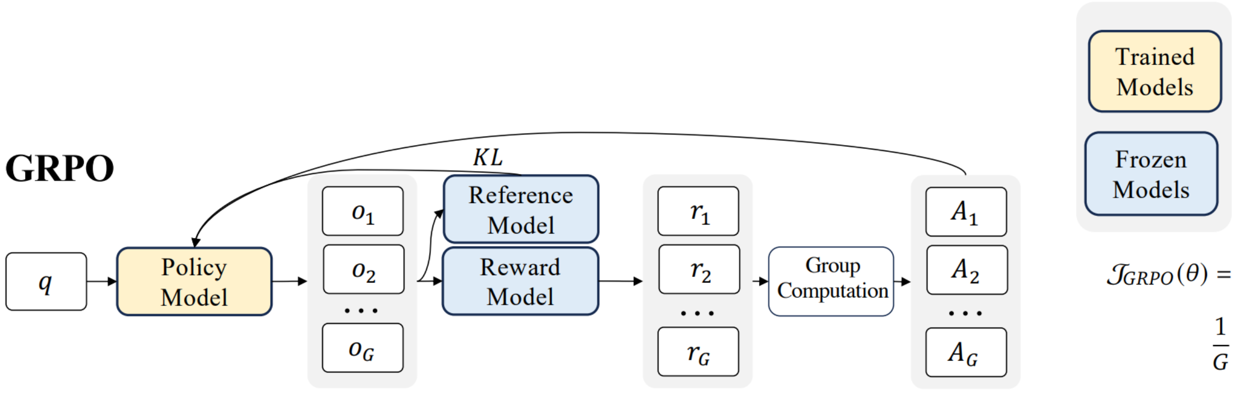
$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \mathrm{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta || \pi_{ref}\right]\right\}$$

Process Supervision RL with GRPO

- Sample multiple outputs $\{o_1, o_2, \dots, o_G\}$ for a question q
- Use a process reward model to score each reasoning step

$$\mathbf{R} = \{\{r_1^{index(1)}, \cdots, r_1^{index(K_1)}\}, \cdots, \{r_G^{index(1)}, \cdots, r_G^{index(K_G)}\}\}$$

- Normalize all step-level rewards: $\widetilde{r}_i^{index(j)} = \frac{r_i^{index(j)} - \mathrm{mean}(\mathbf{R})}{\mathrm{std}(\mathbf{R})}$
- For each token t, compute advantage: $\hat{A}_{i,t} = \sum_{index(j) \geq t} \widetilde{r}_i^{index(j)}$



$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \mathrm{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta || \pi_{ref}\right]\right\}$$

Iterative RL with GRPO

- The old reward model may not be sufficient to supervise the current policy model

Algorithm 1 Iterative Group Relative Policy Optimization

Input initial policy model $\pi_{\theta_{\text{init}}}$; reward models r_φ; task prompts $\mathcal{D}$; hyperparameters ε, β, μ

policy model $\pi_\theta \leftarrow \pi_{\theta_{\text{init}}}$
for iteration = 1, …, I **do**
  reference model $\pi_{ref} \leftarrow \pi_\theta$
  for step = 1, …, M **do**
    Sample a batch $\mathcal{D}_b$ from $\mathcal{D}$
    Update the old policy model $\pi_{\theta_{old}} \leftarrow \pi_\theta$
    Sample G outputs $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(\cdot \mid q)$ for each question $q \in \mathcal{D}_b$
    Compute rewards $\{r_i\}_{i=1}^{G}$ for each sampled output o_i by running r_φ
    Compute $\hat{A}_{i,t}$ for the t-th token of o_i through group relative advantage estimation.
    for GRPO iteration = 1, …, μ **do**
      Update the policy model π_θ by maximizing the GRPO objective (Equation 21)
  Update r_φ through continuous training using a replay mechanism.

Output π_θ

Training Setup

- Starting from: DeepSeekMath-Instruct 7B
- RL Method: Group Relative Policy Optimization (GRPO)
- RL Data: CoT-format GSM8K and MATH (~144K examples)
- For each question:
  - 64 outputs sampled
  - Max token length: 1024
  - Learning rate: 1e-6
  - KL coefficient: 0.04

Evaluation Results

- In-domain performance:
  - GSM8K: 88.2%
  - MATH: 51.7%
- Out-of-domain performance:
  - MGSM-zh: 79.6%
  - CMATH: 88.8%

Model	Size	English Benchmarks		Chinese Benchmarks	
		GSM8K	MATH	MGSM-zh	CMATH
Chain-of-Thought Reasoning					
Closed-Source Model					
Gemini Ultra	-	94.4%	53.2%	-	-
GPT-4	-	92.0%	52.9%	-	86.0%
Inflection-2	-	81.4%	34.8%	-	-
GPT-3.5	-	80.8%	34.1%	-	73.8%
Gemini Pro	-	86.5%	32.6%	-	-
Grok-1	-	62.9%	23.9%	-	-
Baichuan-3	-	88.2%	49.2%	-	-
GLM-4	-	87.6%	47.9%	-	-
Open-Source Model					
InternLM2-Math	20B	82.6%	37.7%	-	-
Qwen	72B	78.9%	35.2%	-	-
Math-Shepherd-Mistral	7B	84.1%	33.0%	-	-
WizardMath-v1.1	7B	83.2%	33.0%	-	-
DeepSeek-LLM-Chat	67B	84.1%	32.6%	74.0%	80.3%
MetaMath	70B	82.3%	26.6%	66.4%	70.9%
SeaLLM-v2	7B	78.2%	27.5%	64.8%	-
ChatGLM3	6B	72.3%	25.7%	-	-
WizardMath-v1.0	70B	81.6%	22.7%	64.8%	65.4%
DeepSeekMath-Instruct	7B	82.9%	46.8%	73.2%	84.6%
DeepSeekMath-RL	7B	**88.2%**	**51.7%**	**79.6%**	**88.8%**

- Outperforms all open-source models from 7B to 70B
- RL improves both in-domain and generalization performance

Discussion

Training Setting	Training Tokens			w/o Tool Use			w/ Tool Use	
	General	Code	Math	GSM8K	MATH	CMATH	GSM8K+Python	MATH+Python
No Continual Training	–	–	–	2.9%	3.0%	12.3%	2.7%	2.3%
Two-Stage Training								
Stage 1: General Training	400B	–	–	2.9%	3.2%	14.8%	3.3%	2.3%
Stage 2: Math Training	–	–	150B	19.1%	14.4%	37.2%	14.3%	6.7%
Stage 1: Code Training	–	400B	–	5.9%	3.6%	19.9%	12.4%	10.0%
Stage 2: Math Training	–	–	150B	**21.9%**	**15.3%**	**39.7%**	17.4%	9.4%
One-Stage Training								
Math Training	–	–	150B	20.5%	13.1%	37.6%	11.4%	6.5%
Code & Math Mixed Training	–	400B	150B	17.6%	12.1%	36.3%	**19.7%**	**13.5%**

- Code training benefits program-aided mathematical reasoning

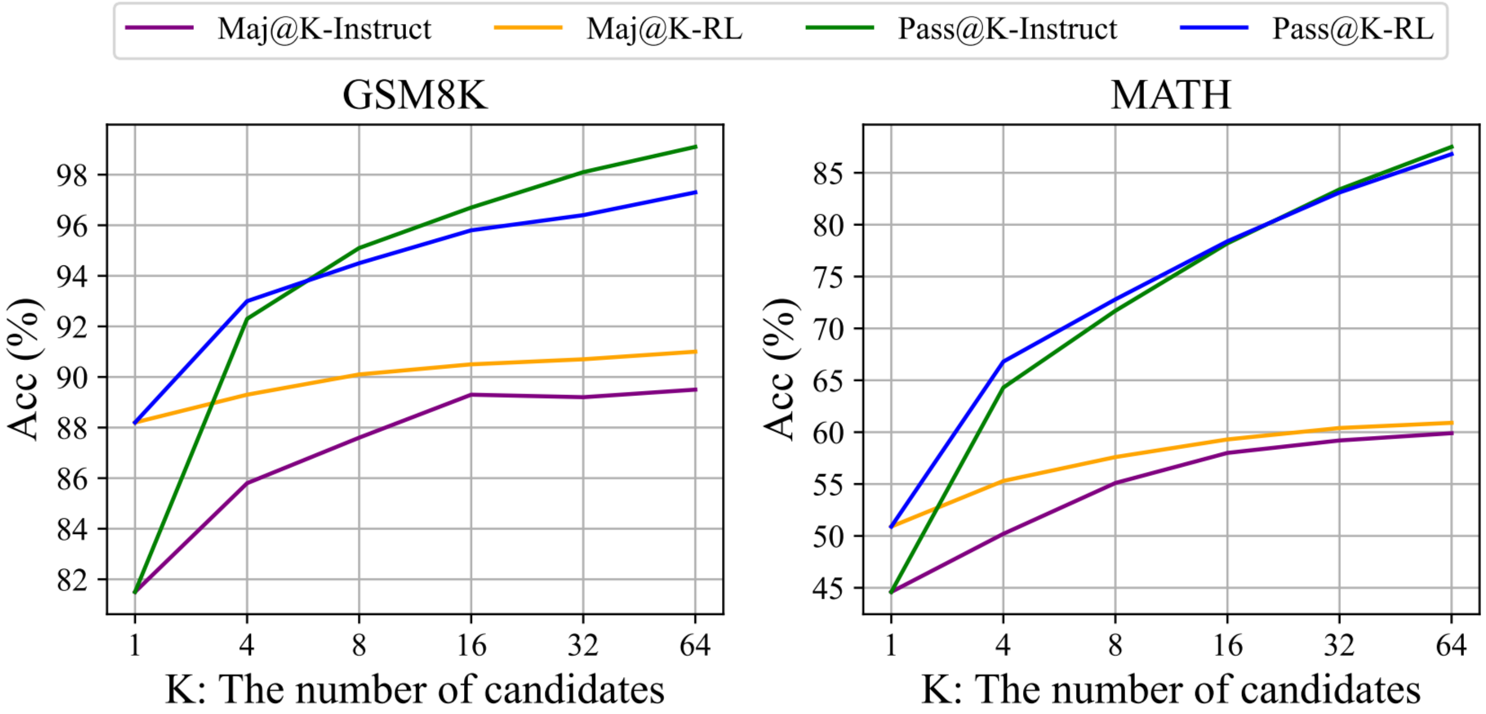
Discussion

Training Setting	Training Tokens			MMLU	BBH	HumanEval (Pass@1)	MBPP (Pass@1)
	General	Code	Math				
No Continual Training	–	–	–	24.5%	28.1%	12.2%	13.0%
Two-Stage Training							
Stage 1: General Training	400B	–	–	25.9%	27.7%	15.2%	13.6%
Stage 2: Math Training	–	–	150B	33.1%	32.7%	12.8%	13.2%
Stage 1: Code Training	–	400B	–	25.0%	31.5%	25.0%	**40.0%**
Stage 2: Math Training	–	–	150B	**36.2%**	35.3%	12.2%	17.0%
One-Stage Training							
Math Training	–	–	150B	32.3%	32.5%	11.6%	13.2%
Code & Math Mixed Training	–	400B	150B	33.5%	**35.6%**	**29.3%**	39.4%

- Under the one-stage training setting, mixing code tokens and math tokens effectively mitigates the issue of catastrophic forgetting that arises from two-stage training

Discussion

- Why reinforcement learning works



  - It seems that the improvement is attributed to boosting the correct response from Top-K rather than the enhancement of fundamental capabilities

Conclusion

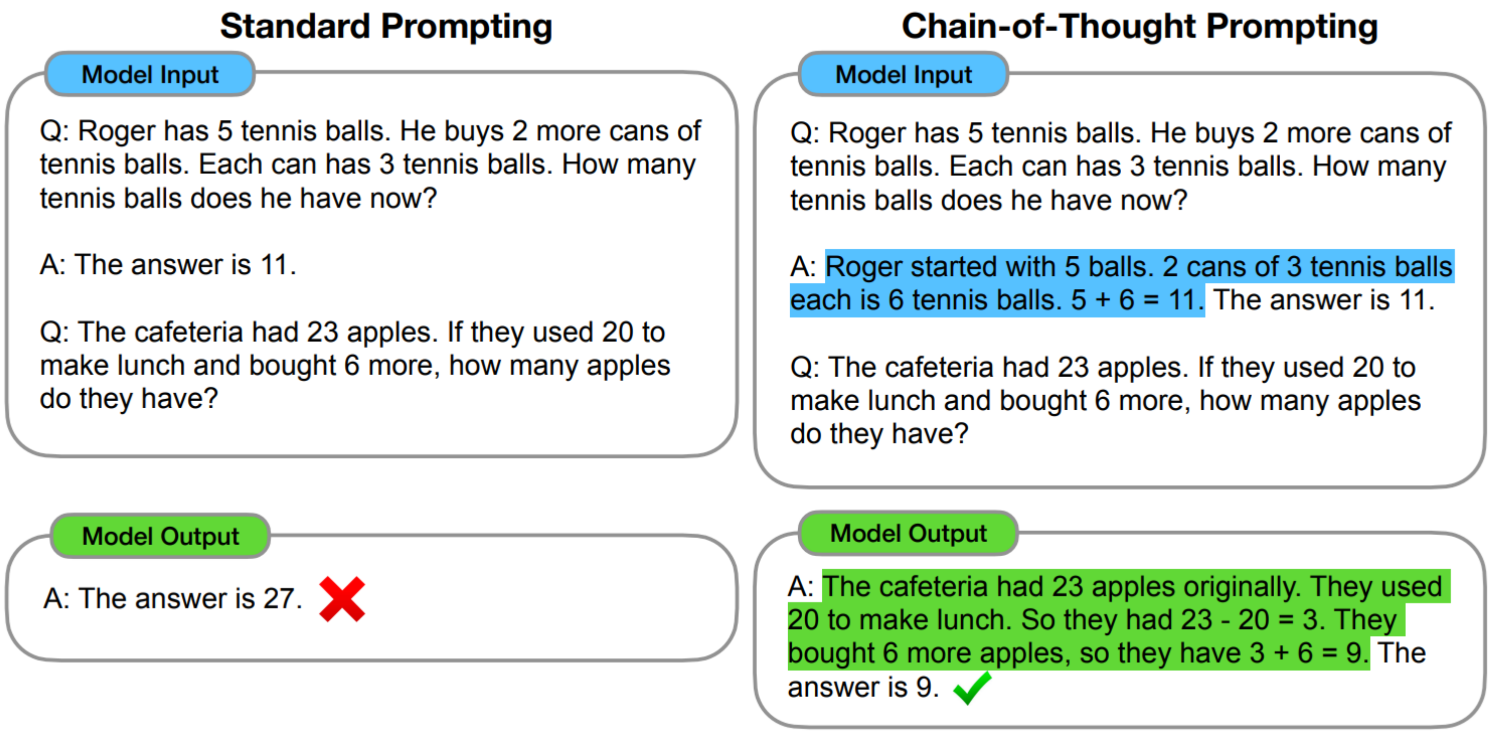
- Introduced DeepSeekMath 7B, a powerful open-source model for mathematical reasoning
- Constructed a 120B-token high-quality math training dataset
- Proposed GRPO, a new RL algorithm that improves performance efficiently

Open Questions

- Although DeepSeekMath performs similarly in zero-shot and few-shot settings, how can we enhance its performance in the few-shot scenario?

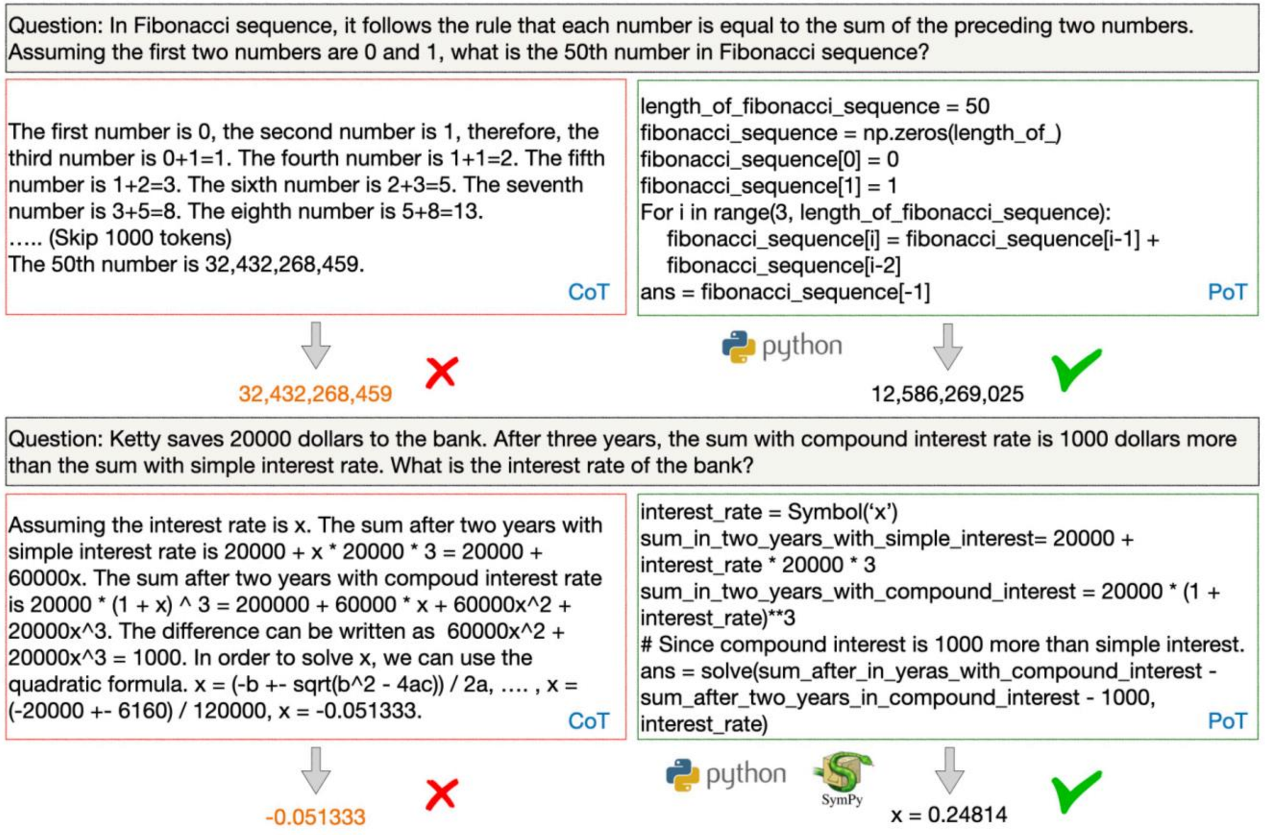
Appendix

- Chain-of-Thought Prompting



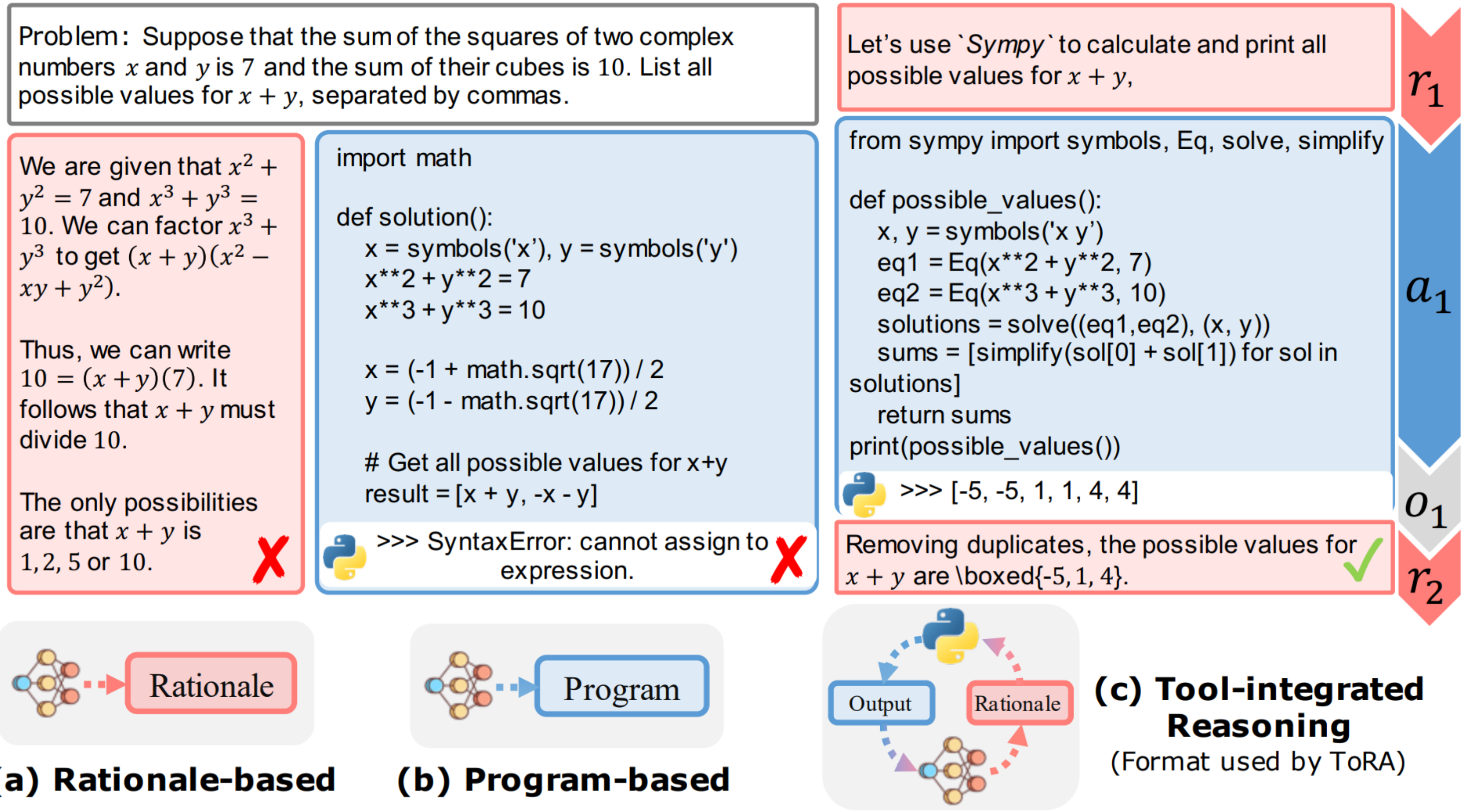
Appendix

- Program-of-thought



Appendix

- Tool-integrated reasoning format



(a) Rationale-based

(b) Program-based

(c) Tool-integrated Reasoning
(Format used by ToRA)