DART-Math

Difficulty-Aware Rejection Tuning

for Mathematical Problem-Solving

Yuxuan Tong[1], Xiwen Zhang[2], Rui Wang[2], Ruidong Wu[2], Junxian He[3]

[1]Tsinghua University [2]Helixon Research [3]HKUST

Slides & Presented by Yuxuan Tong

Background: Task + Method + SotA

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving
  - a critical cornerstone of human cognition

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving
  - a critical cornerstone of human cognition
  - a significant challenge for SotA large language models (LLMs)

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving
  - a critical cornerstone of human cognition
  - a significant challenge for SotA large language models (LLMs)

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving
  - a critical cornerstone of human cognition
  - a significant challenge for SotA large language models (LLMs)

- **Instruction tuning** (on diverse query-response pairs)

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving
  - a critical cornerstone of human cognition
  - a significant challenge for SotA large language models (LLMs)

- **Instruction tuning** (on diverse query-response pairs)
  - cost-effective

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving
  - a critical cornerstone of human cognition
  - a significant challenge for SotA large language models (LLMs)

- **Instruction tuning** (on diverse query-response pairs)
  - cost-effective
  - achieves SotA across many benchmarks

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving
  - a critical cornerstone of human cognition
  - a significant challenge for SotA large language models (LLMs)

- **Instruction tuning** (on diverse query-response pairs)
  - cost-effective
  - achieves SotA across many benchmarks

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving
  - a critical cornerstone of human cognition
  - a significant challenge for SotA large language models (LLMs)

- **Instruction tuning** (on diverse query-response pairs)
  - cost-effective
  - achieves SotA across many benchmarks

- Current **SotA** instruction tuning datasets for mathematical problem-solving are typically constructed by:

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving
    - a critical cornerstone of human cognition
    - a significant challenge for SotA large language models (LLMs)

- **Instruction tuning** (on diverse query-response pairs)
    - cost-effective
    - achieves SotA across many benchmarks

- Current **SotA** instruction tuning datasets for mathematical problem-solving are typically constructed by:
    - augmenting existing training datasets with **synthetic data**

Background: Task + Method + SotA

- **Complex reasoning** like mathematical problem-solving
  - a critical cornerstone of human cognition
  - a significant challenge for SotA large language models (LLMs)

- **Instruction tuning** (on diverse query-response pairs)
  - cost-effective
  - achieves SotA across many benchmarks

- Current **SotA** instruction tuning datasets for mathematical problem-solving are typically constructed by:
  - augmenting existing training datasets with **synthetic data**
  - from **proprietary models like *ChatGPT***.

Observation: Synthetic Datasets are Biased towards Easier Queries

Observation: Synthetic Datasets are Biased towards Easier Queries

Observation: Synthetic Datasets are Biased towards Easier Queries

We checked most of the existing math instruction tuning datasets,

Observation: Synthetic Datasets are Biased towards Easier Queries

We checked most of the existing math instruction tuning datasets,

Observation: Synthetic Datasets are Biased towards Easier Queries

We checked most of the existing math instruction tuning datasets,

most of which are **synthetic**,

Observation: Synthetic Datasets are Biased towards Easier Queries

We checked most of the existing math instruction tuning datasets,

most of which are **synthetic**,

Observation: Synthetic Datasets are Biased towards Easier Queries

We checked most of the existing math instruction tuning datasets,

most of which are **synthetic**,

and noticed **severe biases towards easier queries**.

E.g., *MetaMathQA-MATH-AnsAug*

E.g., *MetaMathQA-MATH-AnsAug*

MetaMathQA: a popular **synthetic** mathematical instruction tuning dataset

E.g., *MetaMathQA-MATH-AnsAug*

MetaMathQA: a popular **synthetic** mathematical instruction tuning dataset

+

E.g., *MetaMathQA-MATH-AnsAug*

MetaMathQA: a popular **synthetic** mathematical instruction tuning dataset

+

MATH: a prominent mathematical competition problem-solving dataset

E.g., *MetaMathQA-MATH-AnsAug*

MetaMathQA: a popular **synthetic** mathematical instruction tuning dataset

+

MATH: a prominent mathematical competition problem-solving dataset

- 7.5k problem-solution (query-response) pairs (in the training set)

E.g., *MetaMathQA-MATH-AnsAug*

MetaMathQA: a popular **synthetic** mathematical instruction tuning dataset

+

MATH: a prominent mathematical competition problem-solving dataset

- 7.5k problem-solution (query-response) pairs (in the training set)
- with **human-annotated difficulty levels (1-5 from easy to hard)**

E.g., *MetaMathQA-MATH-AnsAug*

MetaMathQA: a popular **synthetic** mathematical instruction tuning dataset

+

MATH: a prominent mathematical competition problem-solving dataset

- 7.5k problem-solution (query-response) pairs (in the training set)
- with **human-annotated difficulty levels (1-5 from easy to hard)**

↓

E.g., *MetaMathQA-MATH-AnsAug*

MetaMathQA: a popular **synthetic** mathematical instruction tuning dataset

+

MATH: a prominent mathematical competition problem-solving dataset

- 7.5k problem-solution (query-response) pairs (in the training set)
- with **human-annotated difficulty levels (1-5 from easy to hard)**

↓

MetaMathQA-MATH-AnsAug: a *MetaMathQA* subset constructed by **augmenting** the *MATH* training set as follows:

E.g., *MetaMathQA-MATH-AnsAug*

MetaMathQA: a popular **synthetic** mathematical instruction tuning dataset

+

MATH: a prominent mathematical competition problem-solving dataset

- 7.5k problem-solution (query-response) pairs (in the training set)
- with **human-annotated difficulty levels (1-5 from easy to hard)**

↓

MetaMathQA-MATH-AnsAug: a *MetaMathQA* subset constructed by **augmenting** the *MATH* training set as follows:

- sampling candidate solutions with *ChatGPT*

E.g., *MetaMathQA-MATH-AnsAug*

MetaMathQA: a popular **synthetic** mathematical instruction tuning dataset

+

MATH: a prominent mathematical competition problem-solving dataset

- 7.5k problem-solution (query-response) pairs (in the training set)
- with **human-annotated difficulty levels (1-5 from easy to hard)**

↓

MetaMathQA-MATH-AnsAug: a *MetaMathQA* subset constructed by **augmenting** the *MATH* training set as follows:

- sampling candidate solutions with *ChatGPT*
- and filtering with correct answers

E.g., *MetaMathQA-MATH-AnsAug*

MetaMathQA: a popular **synthetic** mathematical instruction tuning dataset

+

MATH: a prominent mathematical competition problem-solving dataset

- 7.5k problem-solution (query-response) pairs (in the training set)
- with **human-annotated difficulty levels (1-5 from easy to hard)**

↓

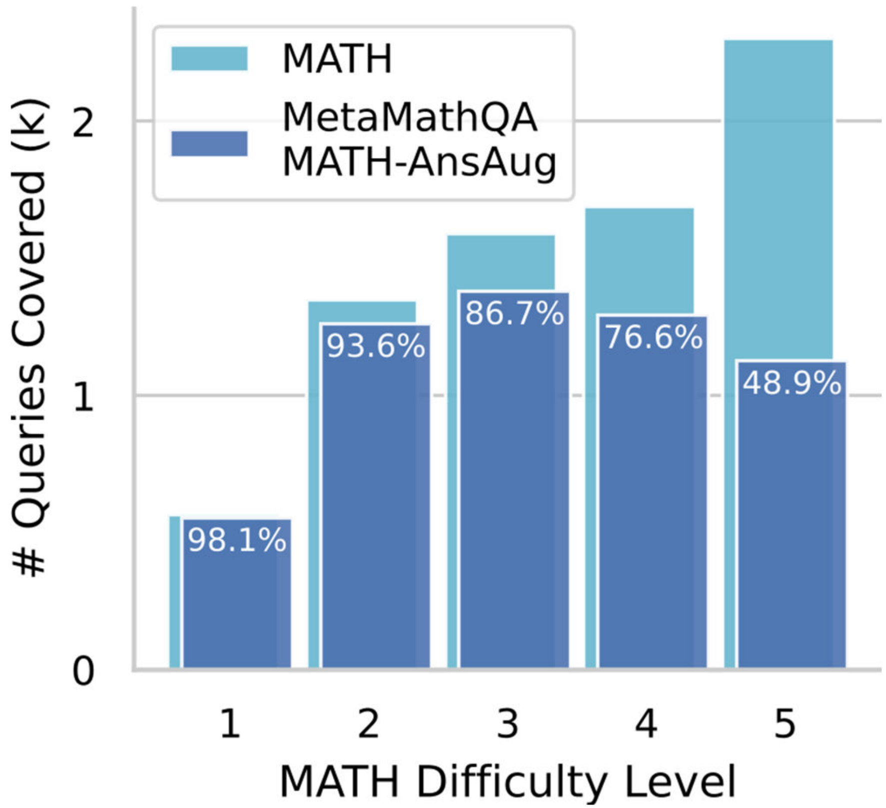
MetaMathQA-MATH-AnsAug: a *MetaMathQA* subset constructed by **augmenting** the *MATH* training set as follows:

- sampling candidate solutions with *ChatGPT*
- and filtering with correct answers
- Finally, there might be **zero to multiple** solutions to one problem in the augmented dataset

E.g., the original *MATH* query distribution

E.g., the original *MATH* query distribution

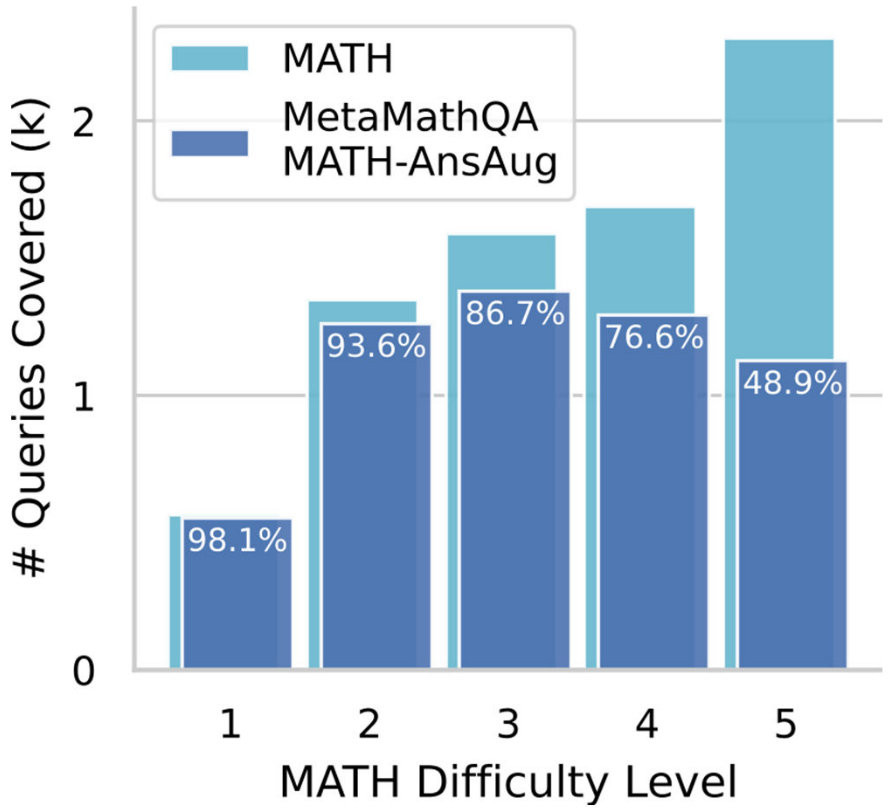
(light blue bars)



E.g., the original *MATH* query distribution

(light blue bars)

The **most difficult (level 4-5)** queries

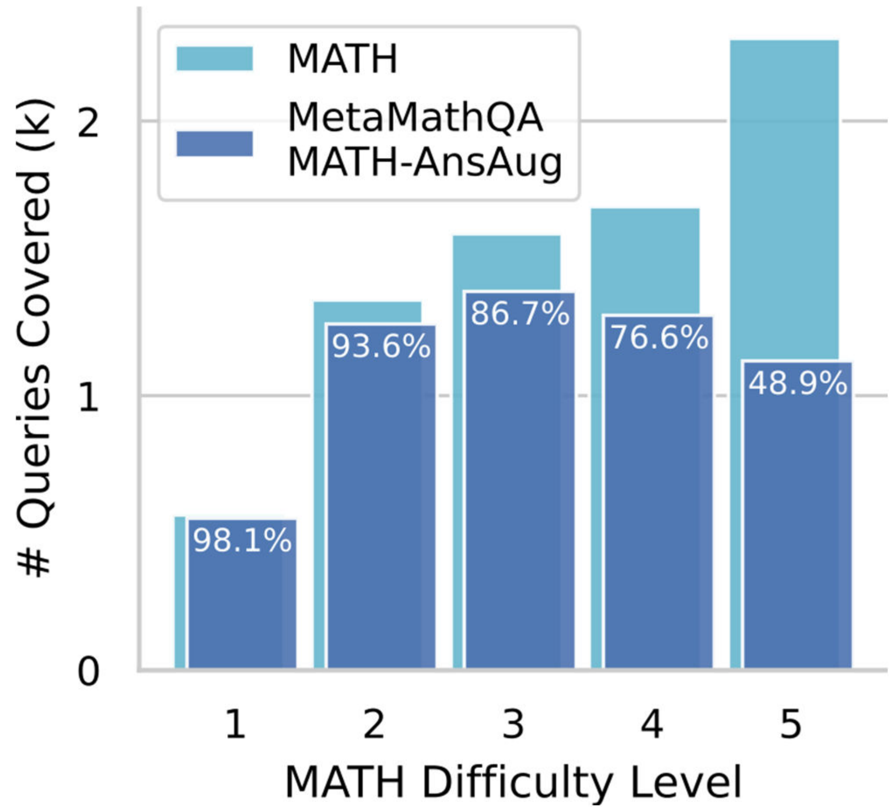


E.g., the original *MATH* query distribution

(light blue bars)

The **most difficult (level 4-5)** queries

↓



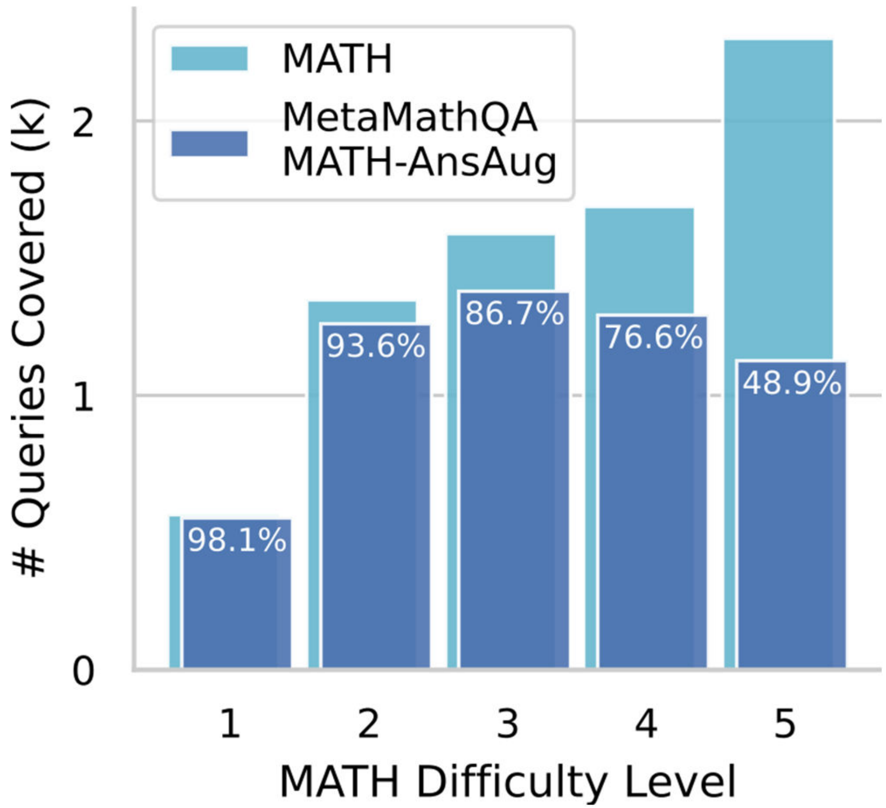
E.g., the original *MATH* query distribution

(light blue bars)

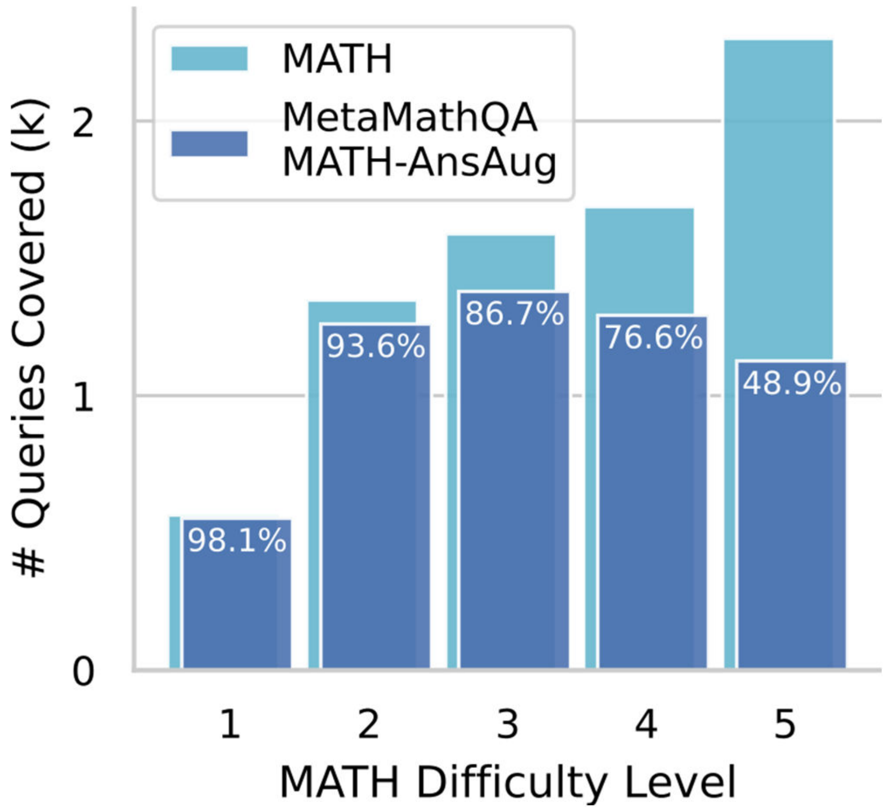
The **most difficult (level 4-5)** queries

↓

take the **largest proportion (> 50%)**.

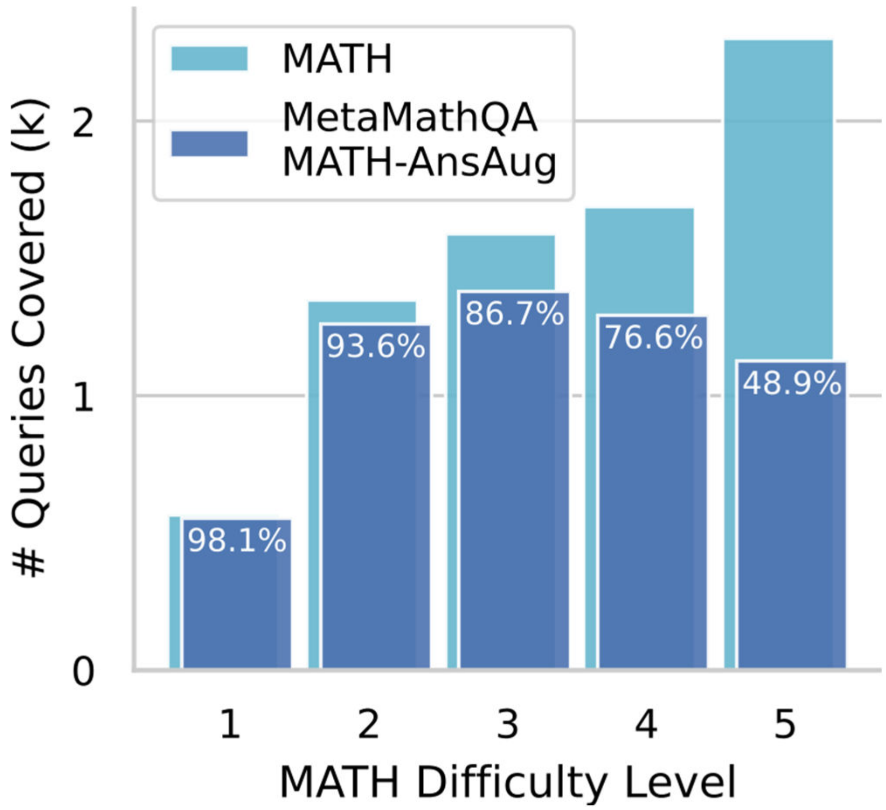


E.g., *MetaMathQA* biases the distribution towards easier queries



E.g., *MetaMathQA* biases the distribution towards easier queries

Query-level:

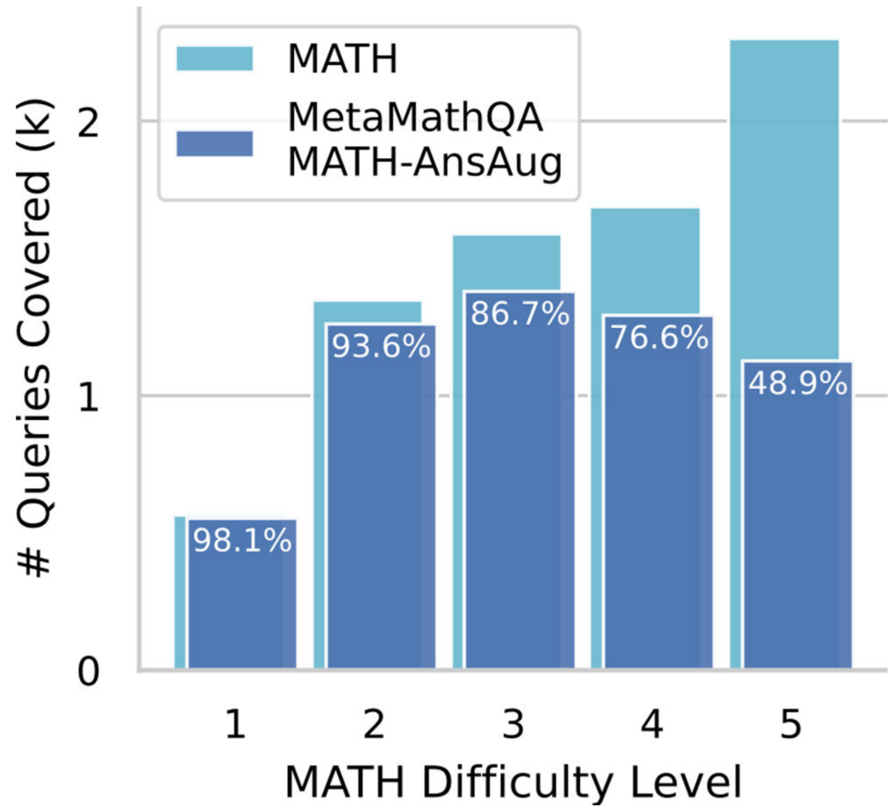


E.g., *MetaMathQA* biases the distribution towards easier queries

Query-level:

Dropping many hard queries

(light blue → dark blue bars)



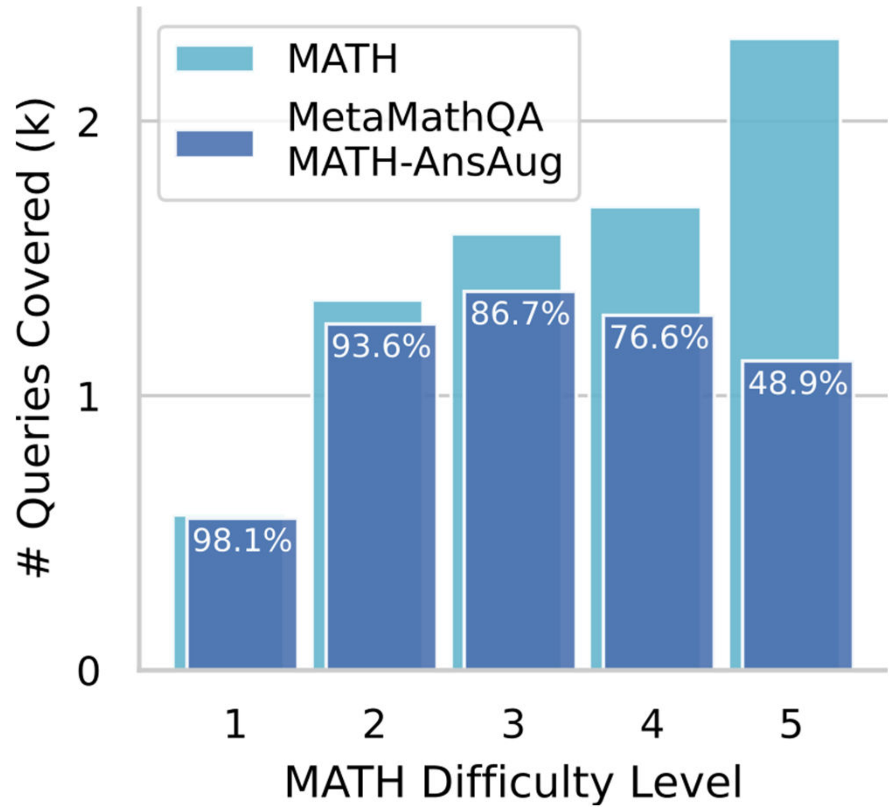
E.g., *MetaMathQA* biases the distribution towards easier queries

Query-level:

Dropping many hard queries

(light blue → dark blue bars)

E.g.,



E.g., *MetaMathQA* biases the distribution towards easier queries

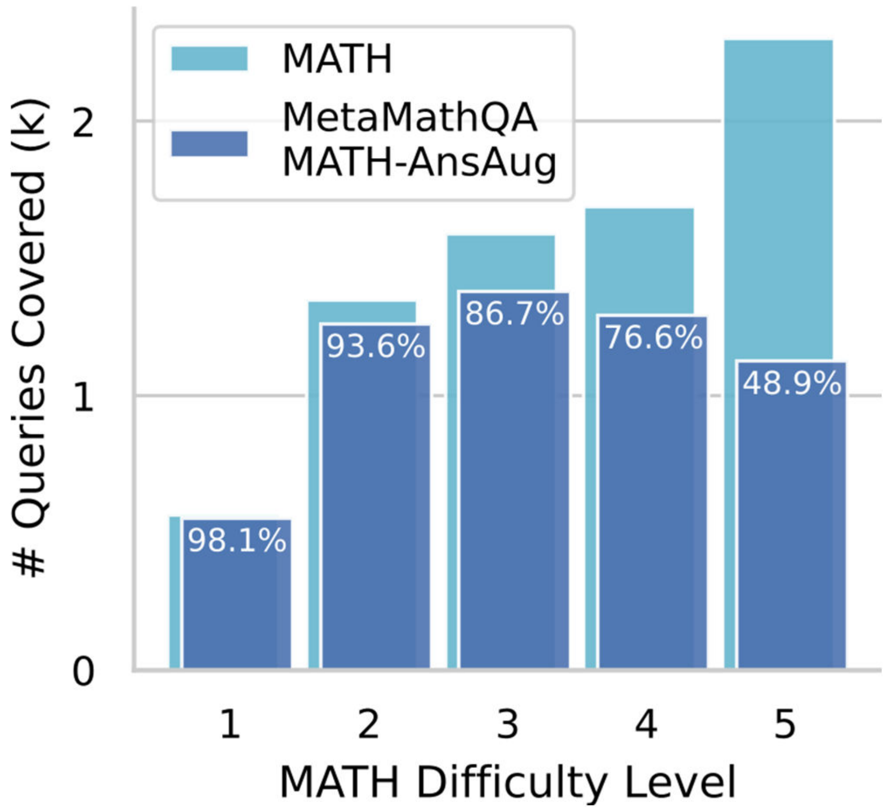
Query-level:

Dropping many hard queries

(light blue → dark blue bars)

E.g.,

The **most difficult (level 5)** queries



E.g., *MetaMathQA* biases the distribution towards easier queries

Query-level:

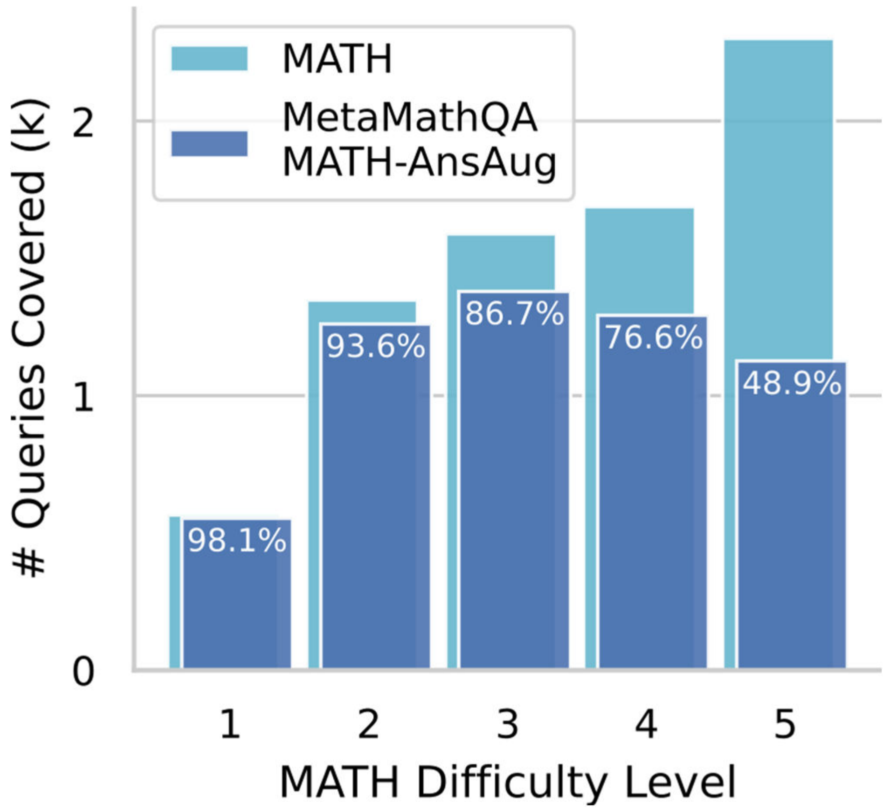
Dropping many hard queries

(light blue → dark blue bars)

E.g.,

The **most difficult (level 5)** queries

↓



E.g., *MetaMathQA* biases the distribution towards easier queries

Query-level:

Dropping many hard queries

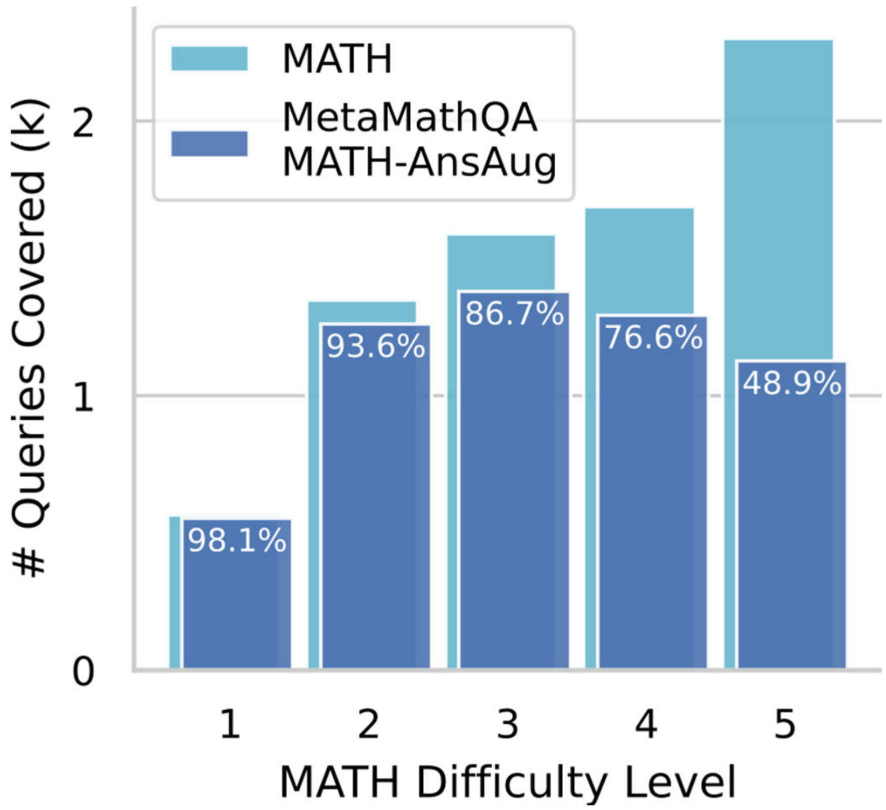
(light blue → dark blue bars)

E.g.,

The **most difficult (level 5)** queries

↓

notably **decreases by >50%** relatively.



E.g., *MetaMathQA* biases the distribution towards easier queries

E.g., *MetaMathQA* biases the distribution towards easier queries

Response-level:

E.g., *MetaMathQA* biases the distribution towards easier queries

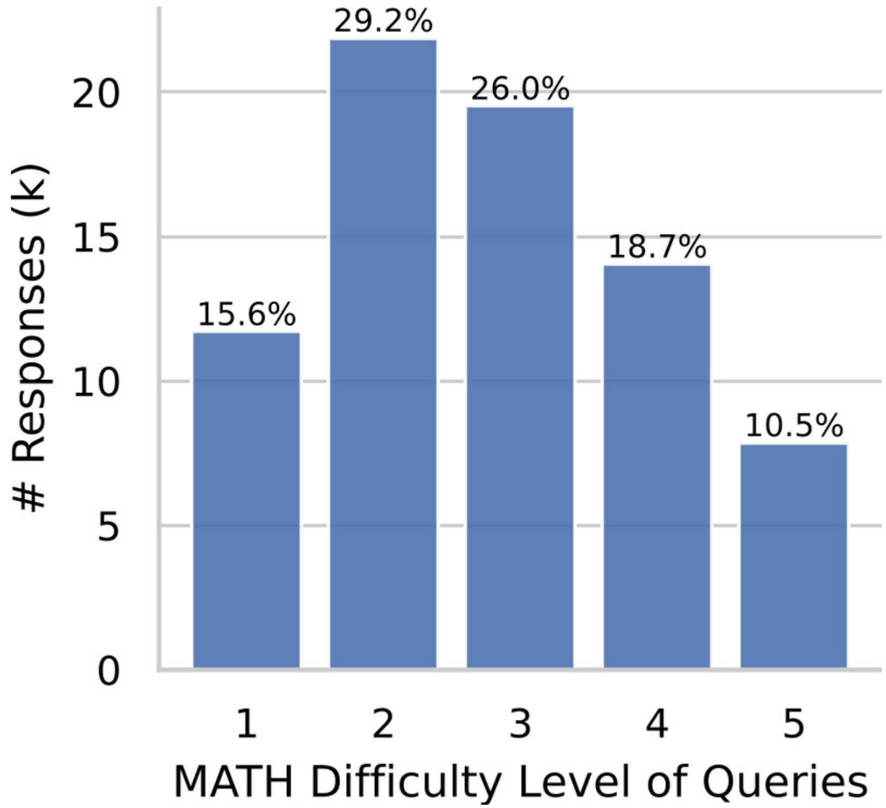
Response-level:

Fewer responses for **harder** queries

E.g., *MetaMathQA* biases the distribution towards easier queries

Response-level:

Fewer responses for **harder** queries

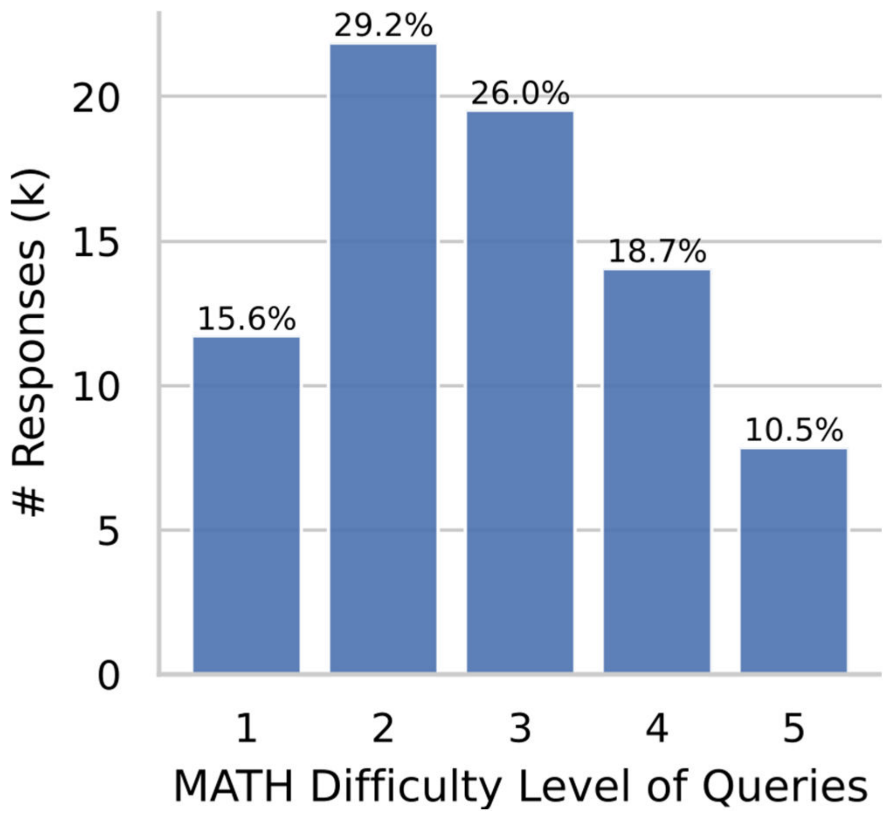


E.g., *MetaMathQA* biases the distribution towards easier queries

Response-level:

Fewer responses for **harder** queries

E.g.,



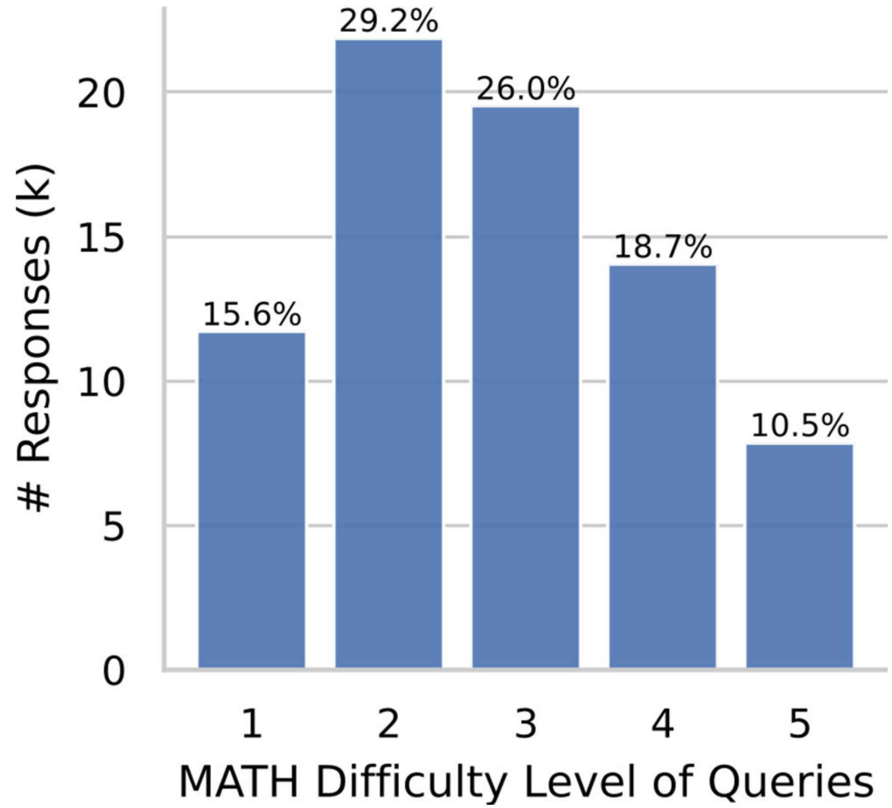
E.g., *MetaMathQA* biases the distribution towards easier queries

Response-level:

Fewer responses for **harder** queries

E.g.,

The responses for **the most difficult (Level 5)** queries



E.g., *MetaMathQA* biases the distribution towards easier queries

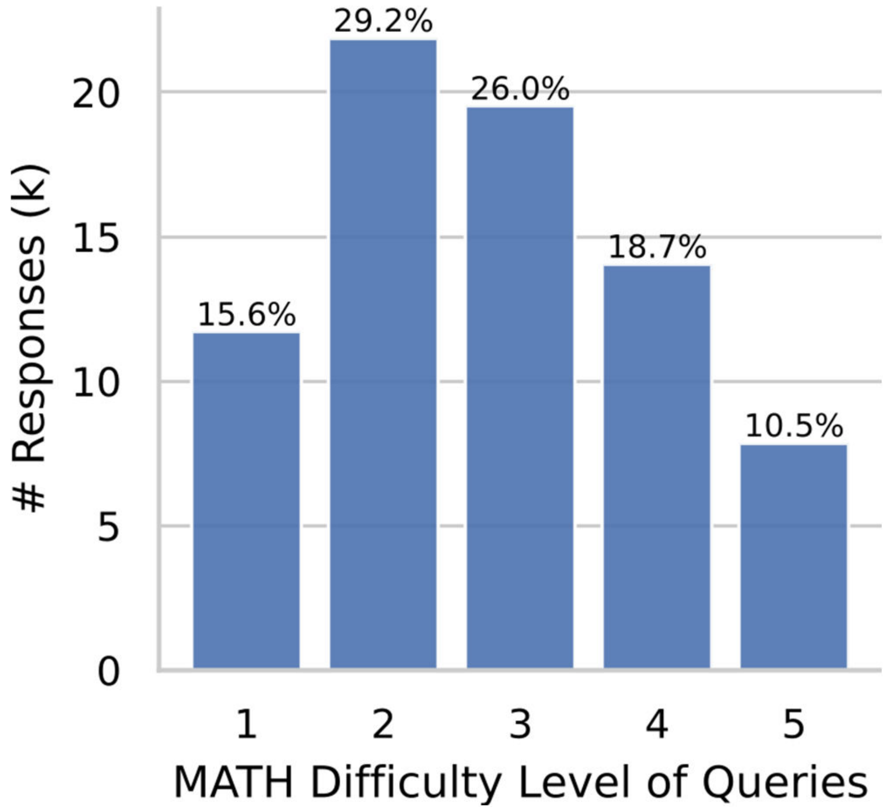
Response-level:

Fewer responses for **harder** queries

E.g.,

The responses for **the most difficult (Level 5)** queries

↓



E.g., *MetaMathQA* biases the distribution towards easier queries

Response-level:

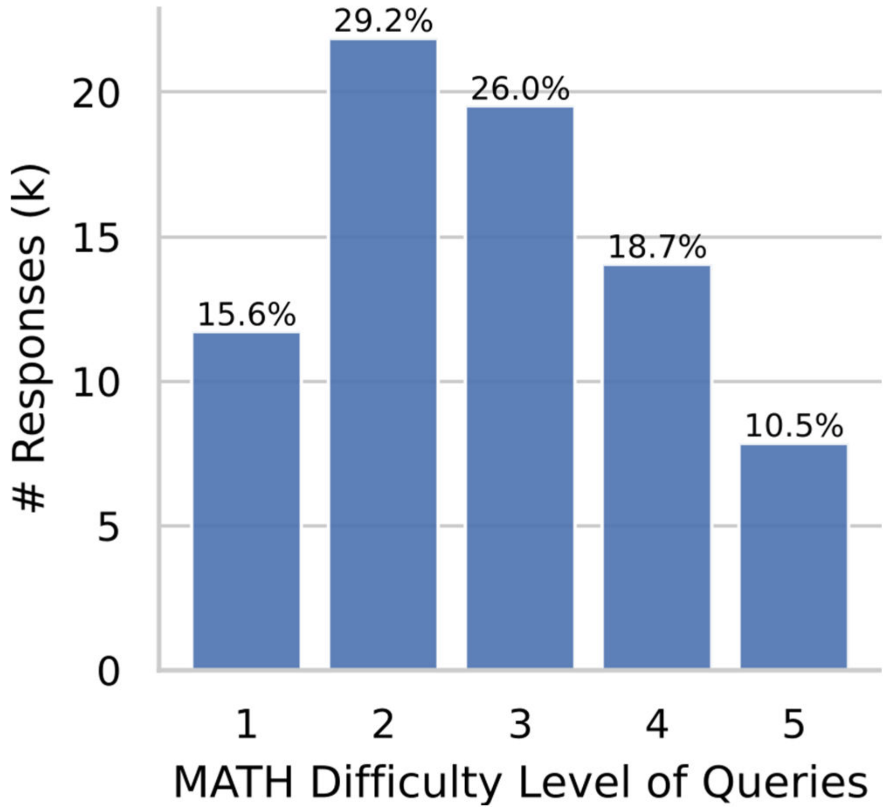
Fewer responses for **harder** queries

E.g.,

The responses for **the most difficult (Level 5)** queries

↓

account for **only ~10%** of all the samples.



E.g., *MetaMathQA* biases the distribution towards easier queries

Response-level:

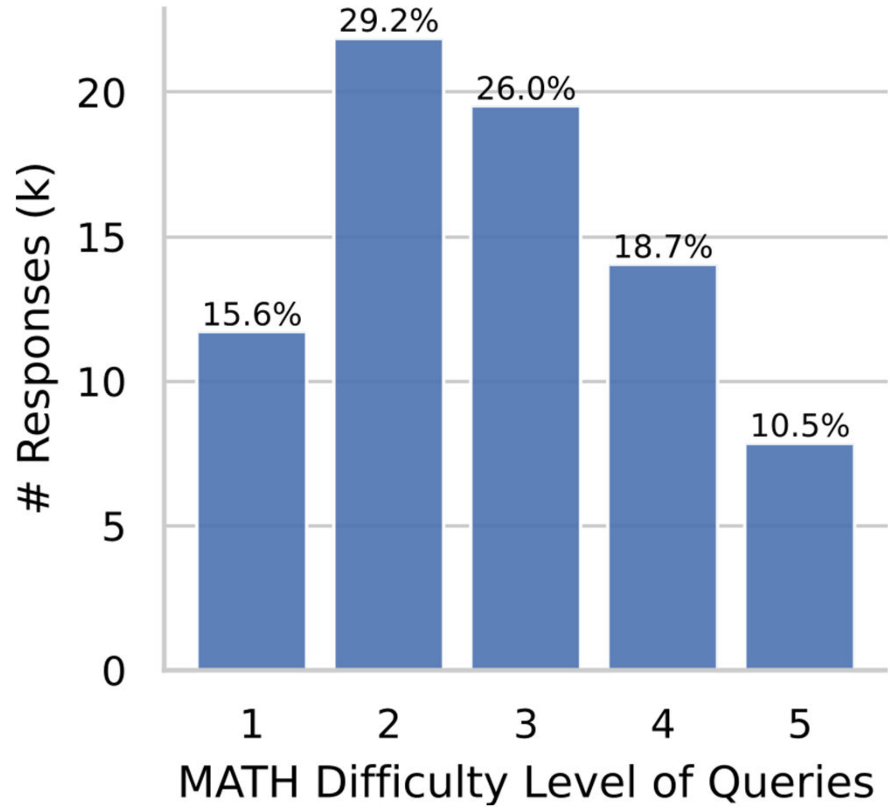
Fewer responses for **harder** queries

E.g.,

The responses for **the most difficult (Level 5)** queries

↓

account for **only ~10%** of all the samples.



MetaMathQA provides an example of

MetaMathQA provides an example of **bias towards easier queries**

MetaMathQA provides an example of
bias towards easier queries

MetaMathQA provides an example of
bias towards easier queries

↓

MetaMathQA provides an example of
bias towards easier queries

↓

MetaMathQA provides an example of
bias towards easier queries

↓

This **doesn’t seem good**.

Motivation: Enough Difficult Data are Critical for Learning (Complex Reasoning)

Motivation: Enough Difficult Data are Critical for Learning (Complex Reasoning)

Analogy: Humans need enough practice on difficult tasks.

Motivation: Enough Difficult Data are Critical for Learning (Complex Reasoning)

Analogy: Humans need enough practice on difficult tasks.

+

Motivation: Enough Difficult Data are Critical for Learning (Complex Reasoning)

Analogy: Humans need enough practice on difficult tasks.

+

Literature: Difficult data are often considered critical, e.g.,

Motivation: Enough Difficult Data are Critical for Learning (Complex Reasoning)

Analogy: Humans need enough practice on difficult tasks.

+

Literature: Difficult data are often considered critical, e.g.,

- Sorscher B, Geirhos R, Shekhar S, et al. *Beyond neural scaling laws: beating power law scaling via data pruning*. *NeurIPS 2022*.

Motivation: Enough Difficult Data are Critical for Learning (Complex Reasoning)

Analogy: Humans need enough practice on difficult tasks.

+

Literature: Difficult data are often considered critical, e.g.,

- Sorscher B, Geirhos R, Shekhar S, et al. *Beyond neural scaling laws: beating power law scaling via data pruning*. *NeurIPS 2022*.
- Liu W, Zeng W, He K, et al. *What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning*. *ICLR 2023*.

Motivation: Enough Difficult Data are Critical for Learning (Complex Reasoning)

Analogy: Humans need enough practice on difficult tasks.

+

Literature: Difficult data are often considered critical, e.g.,

- Sorscher B, Geirhos R, Shekhar S, et al. *Beyond neural scaling laws: beating power law scaling via data pruning*. *NeurIPS 2022*.
- Liu W, Zeng W, He K, et al. *What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning*. *ICLR 2023*.

↓

Motivation: Enough Difficult Data are Critical for Learning (Complex Reasoning)

Analogy: Humans need enough practice on difficult tasks.

\+

Literature: Difficult data are often considered critical, e.g.,

- Sorscher B, Geirhos R, Shekhar S, et al. *Beyond neural scaling laws: beating power law scaling via data pruning*. *NeurIPS 2022*.
- Liu W, Zeng W, He K, et al. *What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning*. *ICLR 2023*.

↓

Hypothesis:

Motivation: Enough Difficult Data are Critical for Learning (Complex Reasoning)

Analogy: Humans need enough practice on difficult tasks.

+

Literature: Difficult data are often considered critical, e.g.,

- Sorscher B, Geirhos R, Shekhar S, et al. *Beyond neural scaling laws: beating power law scaling via data pruning*. *NeurIPS 2022*.
- Liu W, Zeng W, He K, et al. *What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning*. *ICLR 2023*.

↓

Hypothesis:

bias towards **easier** queries

Motivation: Enough Difficult Data are Critical for Learning (Complex Reasoning)

Analogy: Humans need enough practice on difficult tasks.

+

Literature: Difficult data are often considered critical, e.g.,

- Sorscher B, Geirhos R, Shekhar S, et al. *Beyond neural scaling laws: beating power law scaling via data pruning*. *NeurIPS 2022*.
- Liu W, Zeng W, He K, et al. *What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning*. *ICLR 2023*.

↓

Hypothesis:

bias towards **easier** queries

→ **hinder** learning (complex reasoning)

How to eliminate the biases

How to eliminate the biases towards easier queries

How to eliminate the biases towards easier queries in synthetic datasets?

How to eliminate the biases
towards easier queries
in synthetic datasets?

How to eliminate the biases
towards easier queries
in synthetic datasets?

↓

How to eliminate the biases towards easier queries in synthetic datasets?

↓

How to eliminate the biases
towards easier queries
in synthetic datasets?

↓

What **causes** such biases?

How to eliminate the biases
towards easier queries
in synthetic datasets?

↓

What **causes** such biases?

How to eliminate the biases
towards easier queries
in synthetic datasets?

↓

What **causes** such biases?

↑

How to eliminate the biases
towards easier queries
in synthetic datasets?

↓

What **causes** such biases?

↑

How to eliminate the biases
towards easier queries
in synthetic datasets?

↓

What **causes** such biases?

↑

“Vanilla Rejection Sampling” (*VRS*)

What is Vanilla Rejection Sampling?

What is Vanilla Rejection Sampling?

Rejection sampling

What is Vanilla Rejection Sampling?

Rejection sampling
→ sample from a target distribution that is **not directly accessible**

What is Vanilla Rejection Sampling?

Rejection sampling
→ sample from a target distribution that is **not directly accessible**

What is Vanilla Rejection Sampling?

Rejection sampling
→ sample from a target distribution that is **not directly accessible**

- e.g., the distribution of correct solutions conditioned on problems

What is Vanilla Rejection Sampling?

Rejection sampling
→ sample from a target distribution that is **not directly accessible**

- e.g., the distribution of correct solutions conditioned on problems

What is Vanilla Rejection Sampling?

Rejection sampling
→ sample from a target distribution that is **not directly accessible**

- e.g., the distribution of correct solutions conditioned on problems

Rejection sampling in data synthesis usually refers to a two-stage pipeline:

What is Vanilla Rejection Sampling?

Rejection sampling
→ sample from a target distribution that is **not directly accessible**

- e.g., the distribution of correct solutions conditioned on problems

Rejection sampling in data synthesis usually refers to a two-stage pipeline:

What is Vanilla Rejection Sampling?

Rejection sampling
→ sample from a target distribution that is **not directly accessible**

- e.g., the distribution of correct solutions conditioned on problems

Rejection sampling in data synthesis usually refers to a two-stage pipeline:

1. **sampling** from some model to get candidate samples

What is Vanilla Rejection Sampling?

Rejection sampling
→ sample from a target distribution that is **not directly accessible**

- e.g., the distribution of correct solutions conditioned on problems

Rejection sampling in data synthesis usually refers to a two-stage pipeline:

1. **sampling** from some model to get candidate samples

What is Vanilla Rejection Sampling?

Rejection sampling
→ sample from a target distribution that is **not directly accessible**

- e.g., the distribution of correct solutions conditioned on problems

Rejection sampling in data synthesis usually refers to a two-stage pipeline:

1. **sampling** from some model to get candidate samples

1. **filtering** by some criteria to get final samples

What is Vanilla Rejection Sampling?

Rejection sampling
→ sample from a target distribution that is **not directly accessible**

- e.g., the distribution of correct solutions conditioned on problems

Rejection sampling in data synthesis usually refers to a two-stage pipeline:

1. **sampling** from some model to get candidate samples

1. **filtering** by some criteria to get final samples

What is Vanilla Rejection Sampling?

What is Vanilla Rejection Sampling?

The specific method for previous synthetic datasets:

What is Vanilla Rejection Sampling?

The specific method for previous synthetic datasets:

What is Vanilla Rejection Sampling?

The specific method for previous synthetic datasets:

1. Sampling from the model to get **the same number of** candidate responses **for each query**

What is Vanilla Rejection Sampling?

The specific method for previous synthetic datasets:

1. Sampling from the model to get **the same number of** candidate responses **for each query**

What is Vanilla Rejection Sampling?

The specific method for previous synthetic datasets:

1. Sampling from the model to get **the same number of** candidate responses **for each query**

1. Filtering by whether the answer is correct

What is Vanilla Rejection Sampling?

What is Vanilla Rejection Sampling?

We call this **vanilla** because

What is Vanilla Rejection Sampling?

We call this **vanilla** because

1. It ignores the fact that the likelihood for correct responses to hard queries

What is Vanilla Rejection Sampling?

We call this **vanilla** because

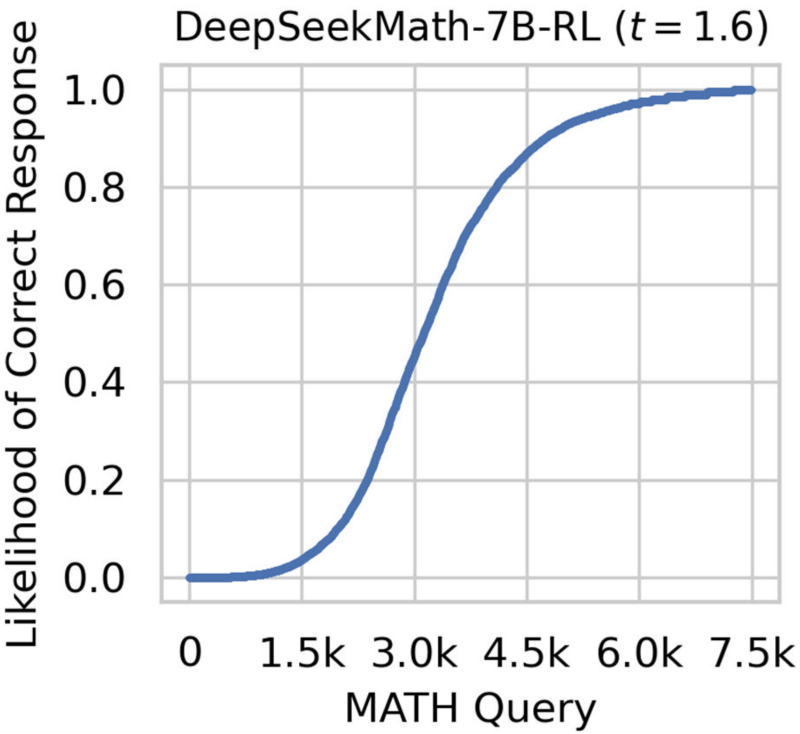
1. It ignores the fact that the likelihood for correct responses to hard queries

can be **low, sometimes even zero**

What is Vanilla Rejection Sampling?

We call this **vanilla** because

1. It ignores the fact that the likelihood for correct responses to hard queries can be **low, sometimes even zero**



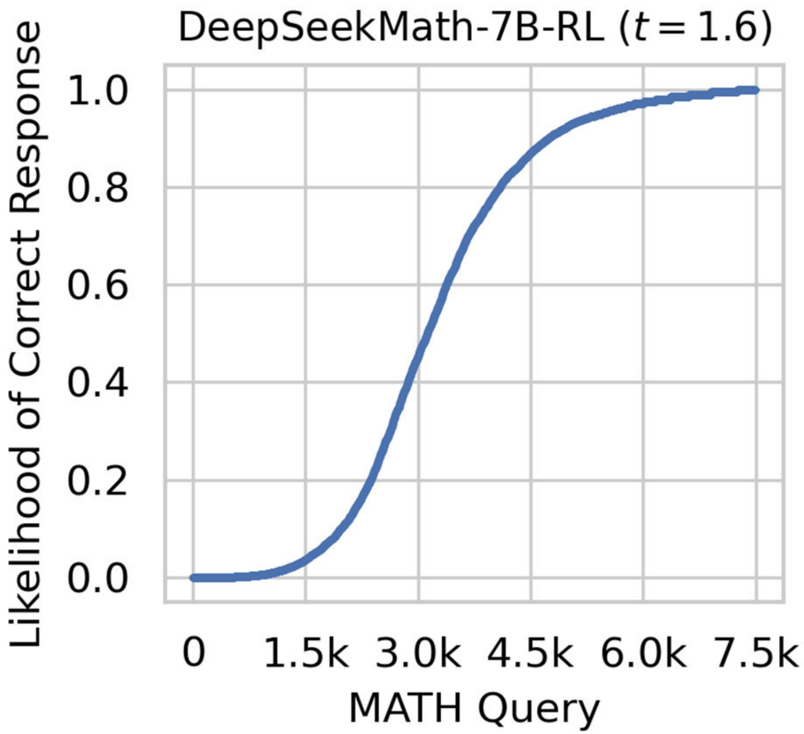
What is Vanilla Rejection Sampling?

We call this **vanilla** because

1. It ignores the fact that the likelihood for correct responses to hard queries

can be **low, sometimes even zero**

VRS →

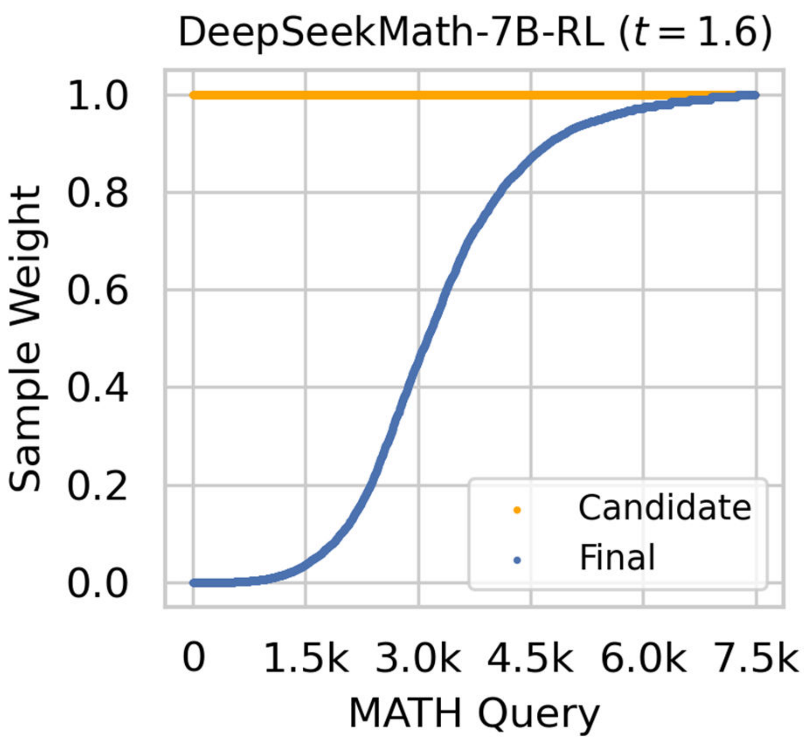


What is Vanilla Rejection Sampling?

We call this **vanilla** because

1. It ignores the fact that the likelihood for correct responses to hard queries can be **low, sometimes even zero**

VRS →

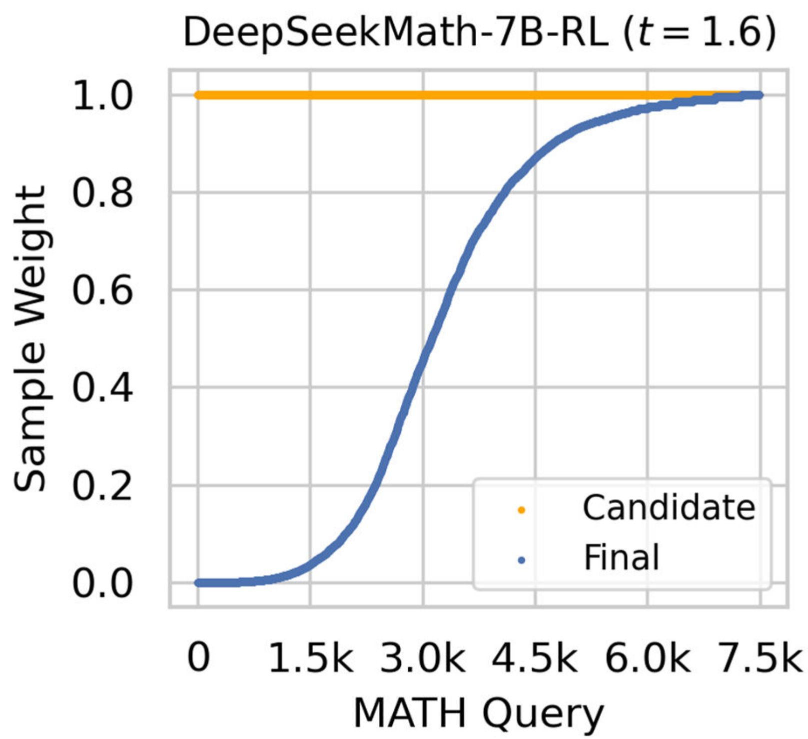


What is Vanilla Rejection Sampling?

We call this **vanilla** because

1. It ignores the fact that the likelihood for correct responses to hard queries can be **low, sometimes even zero**

VRS →



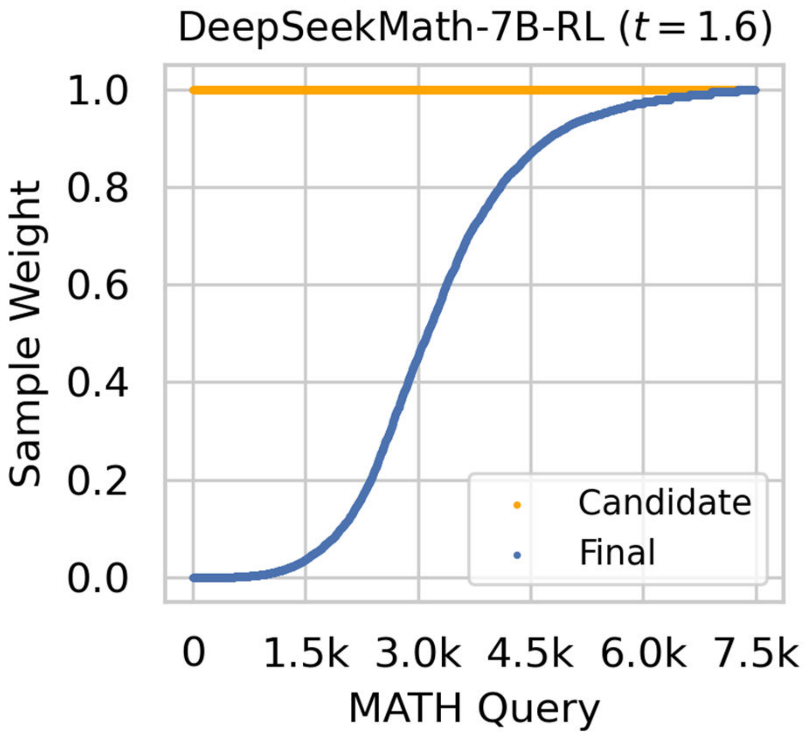
What is Vanilla Rejection Sampling?

We call this **vanilla** because

1. It ignores the fact that the likelihood for correct responses to hard queries can be **low, sometimes even zero**

VRS →

2. It only controls the **candidate** distribution



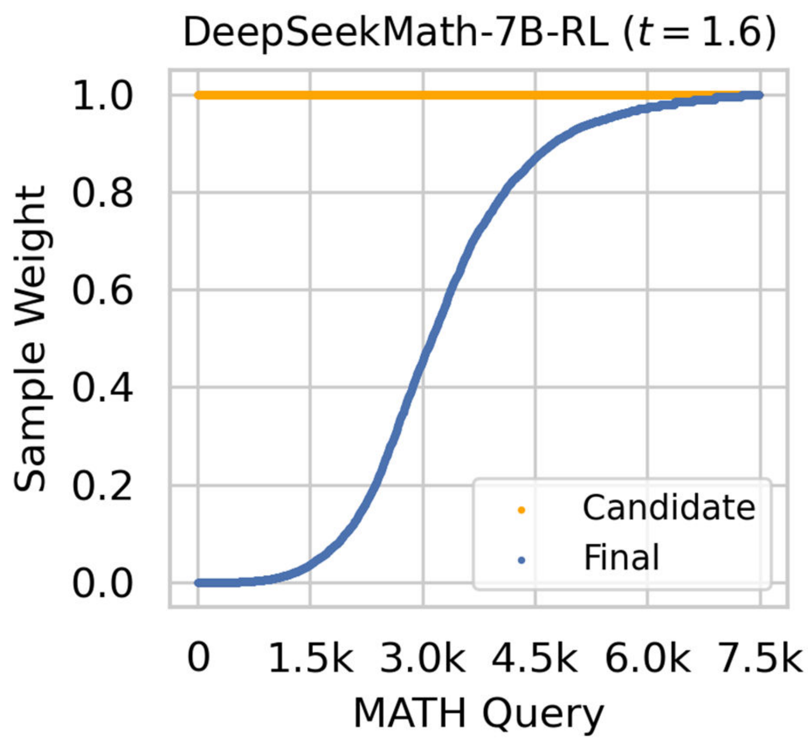
What is Vanilla Rejection Sampling?

We call this **vanilla** because

1. It ignores the fact that the likelihood for correct responses to hard queries
can be **low, sometimes even zero**

VRS →

2. It only controls the **candidate** distribution
instead of the **final** distribution

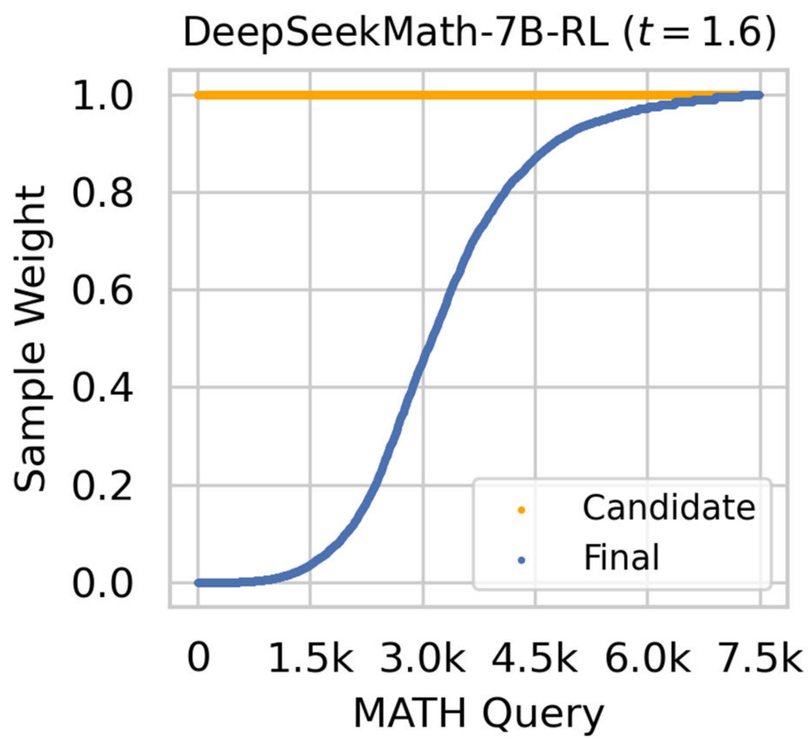


What is Vanilla Rejection Sampling?

We call this **vanilla** because

1. It ignores the fact that the likelihood for correct responses to hard queries can be **low, sometimes even zero**

VRS →



2. It only controls the candidate distribution instead of the final distribution

↓

What is Vanilla Rejection Sampling?

We call this **vanilla** because

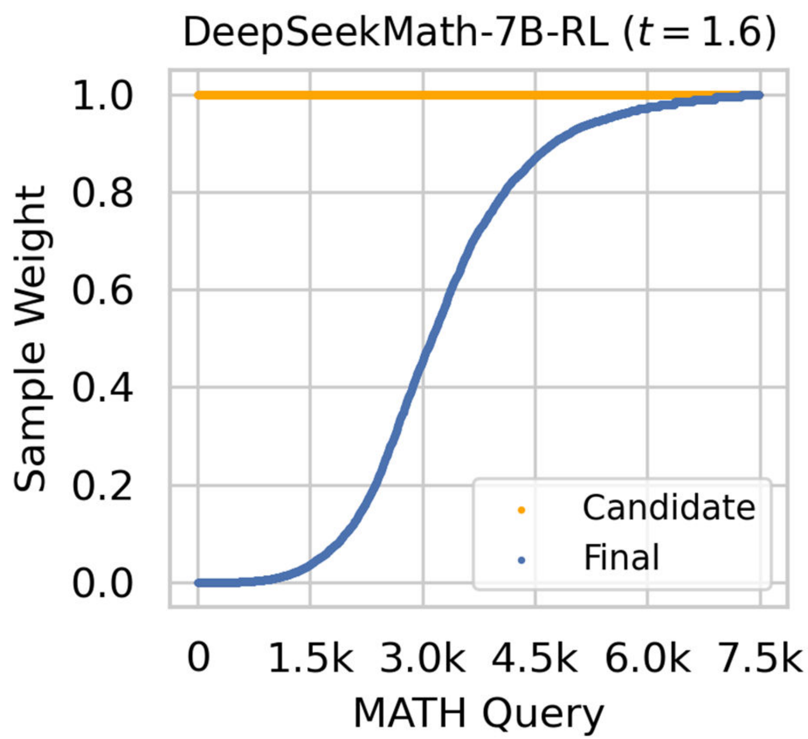
1. It ignores the fact that the likelihood for correct responses to hard queries
can be **low, sometimes even zero**

VRS →

2. It only controls the **candidate** distribution
instead of the **final** distribution

↓

Bias towards easier queries



What is Vanilla Rejection Sampling?

We call this **vanilla** because

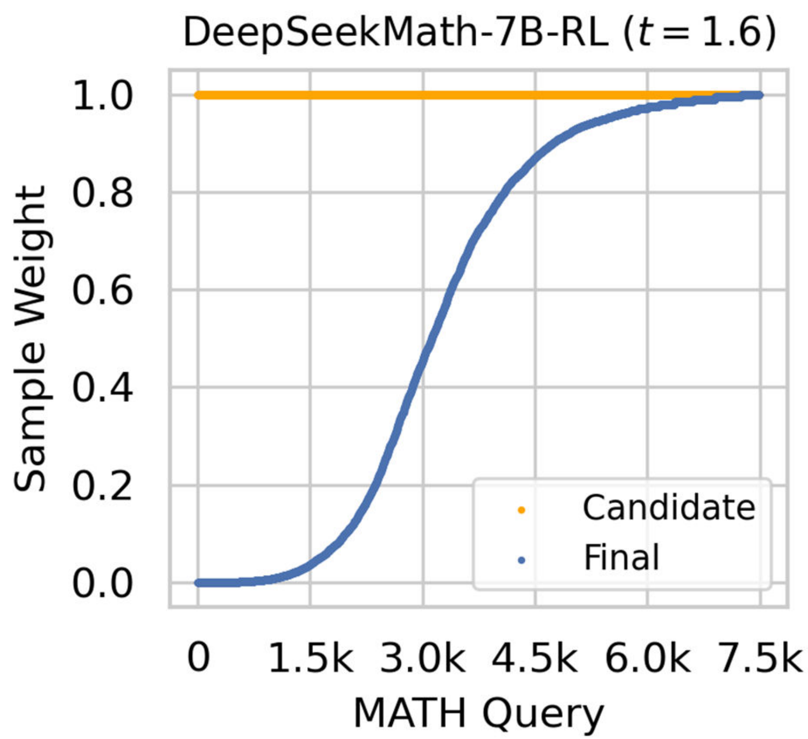
1. It ignores the fact that the likelihood for correct responses to hard queries
can be **low, sometimes even zero**

VRS →

2. It only controls the **candidate** distribution
instead of the **final** distribution

↓

Bias towards easier queries
in the final training distribution



What is Vanilla Rejection Sampling?

We call this **vanilla** because

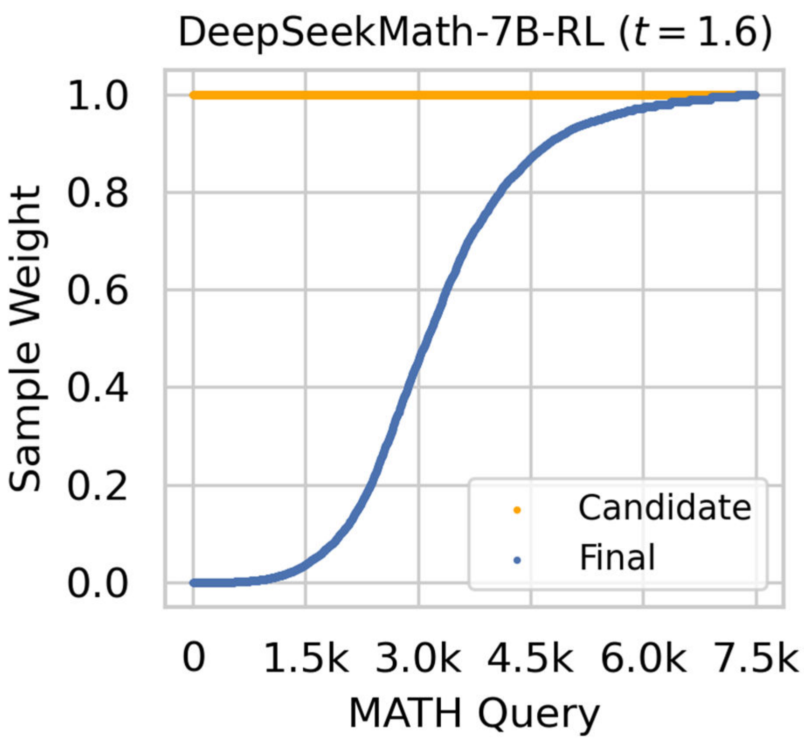
1. It ignores the fact that the likelihood for correct responses to hard queries can be **low, sometimes even zero**

VRS →

2. It only controls the **candidate** distribution instead of the **final** distribution

↓

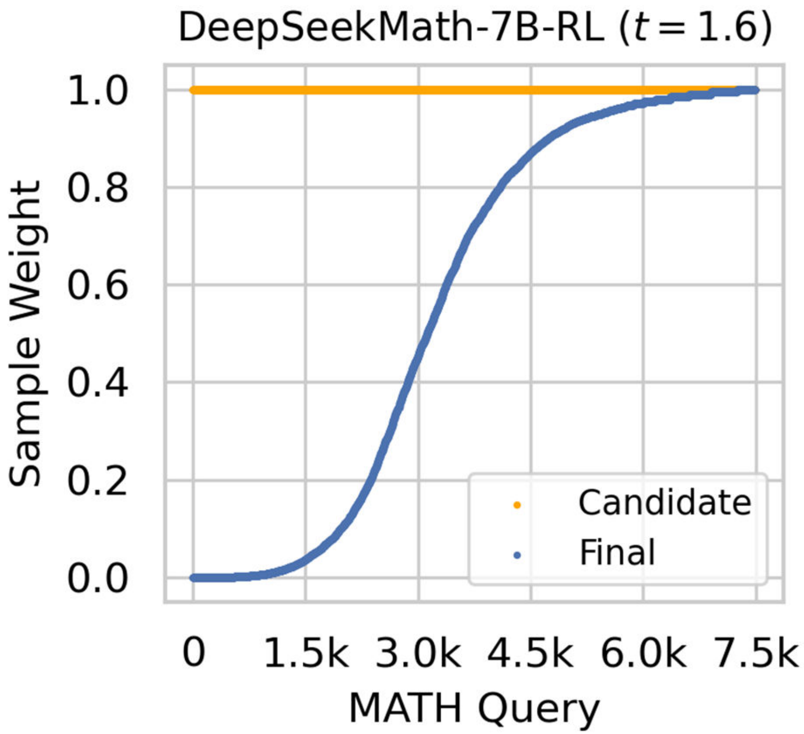
Bias towards easier queries in the final training distribution



How to Avoid the Bias towards Easier Queries by VRS?

How to Avoid the Bias towards Easier Queries by VRS?

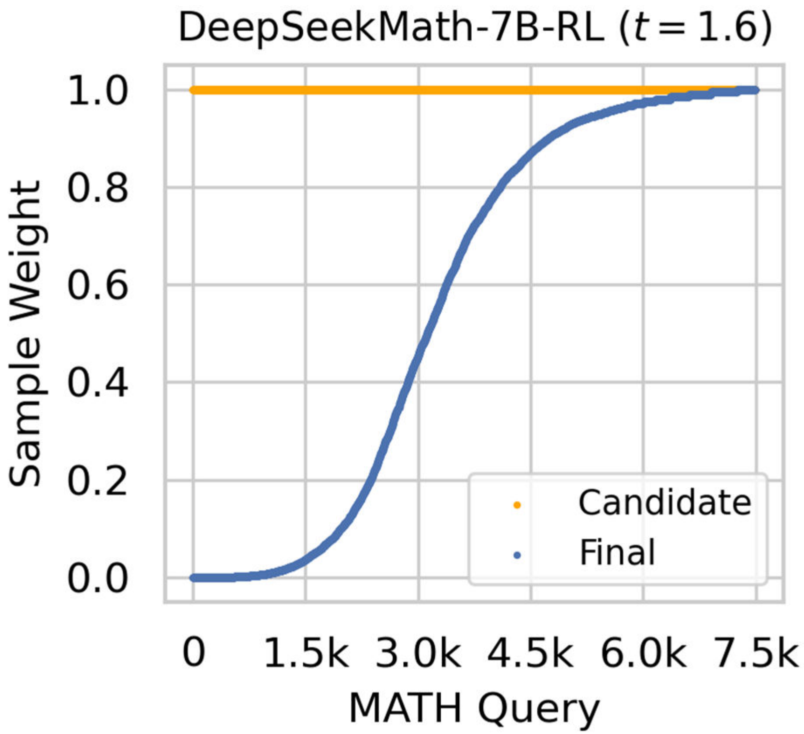
VRS controls the **candidate** distribution



How to Avoid the Bias towards Easier Queries by VRS?

VRS controls the **candidate** distribution

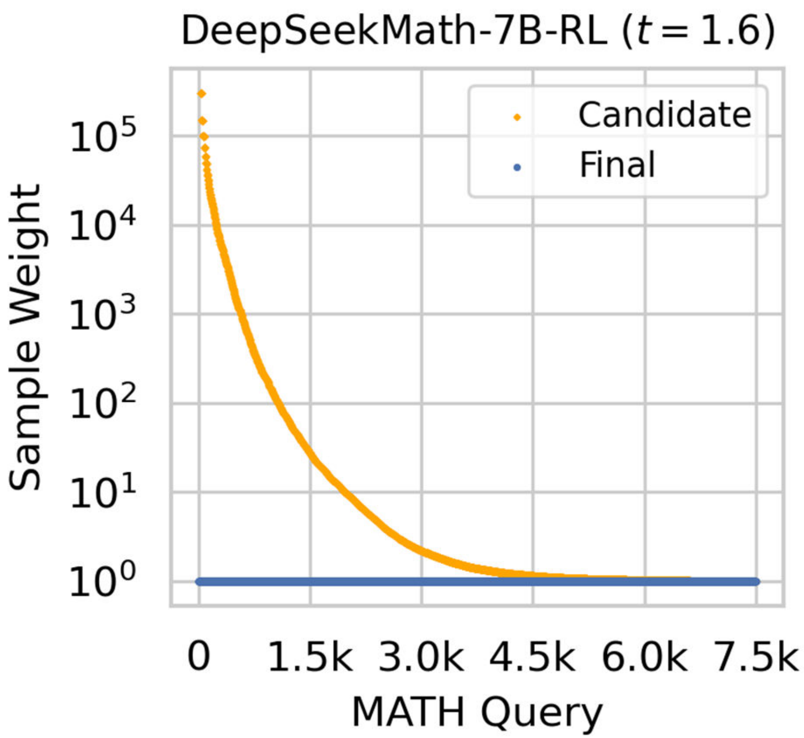
explicitly control the **final** distribution →



How to Avoid the Bias towards Easier Queries by VRS?

VRS controls the **candidate** distribution

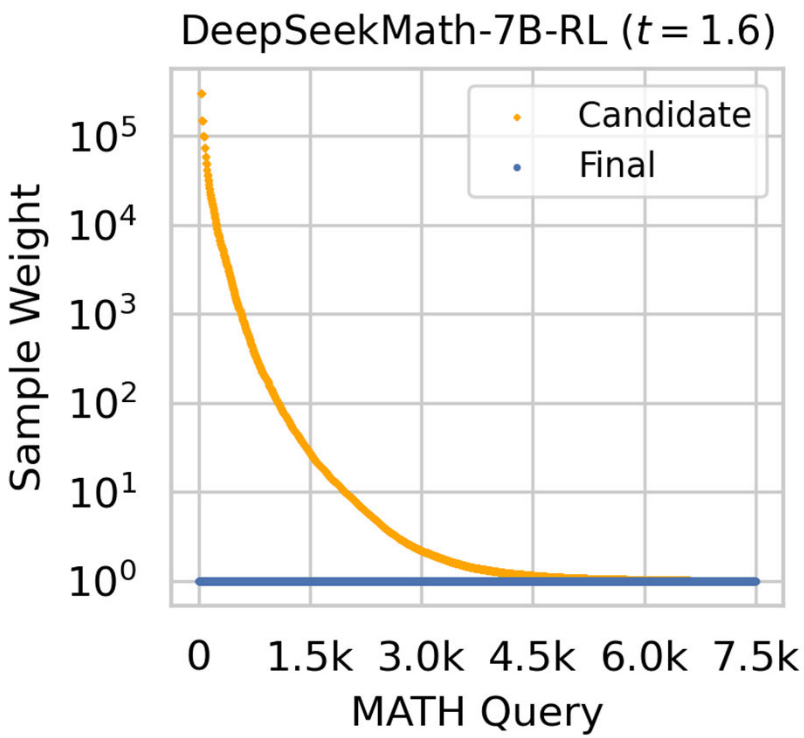
explicitly control the **final** distribution →



How to Avoid the Bias towards Easier Queries by VRS?

VRS controls the **candidate** distribution

explicitly control the **final** distribution →

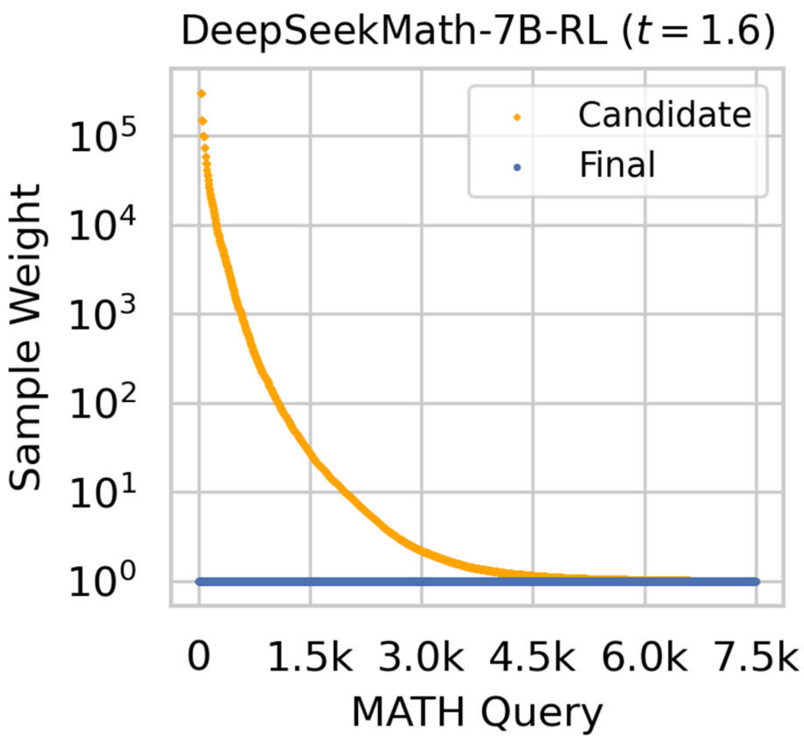


How to Avoid the Bias towards Easier Queries by VRS?

VRS controls the **candidate** distribution

explicitly control the **final** distribution →

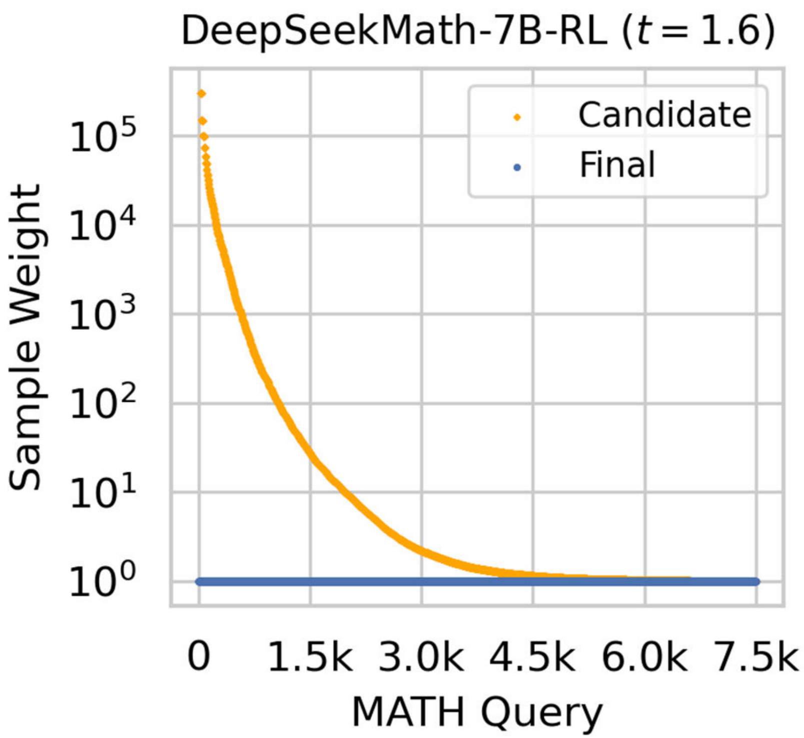
Balanced across difficulties



How to Avoid the Bias towards Easier Queries by VRS?

VRS controls the **candidate** distribution

explicitly control the **final** distribution →



Balanced across difficulties

(Uniform one-sample per query here)

How to Avoid the Bias towards Easier Queries by VRS?

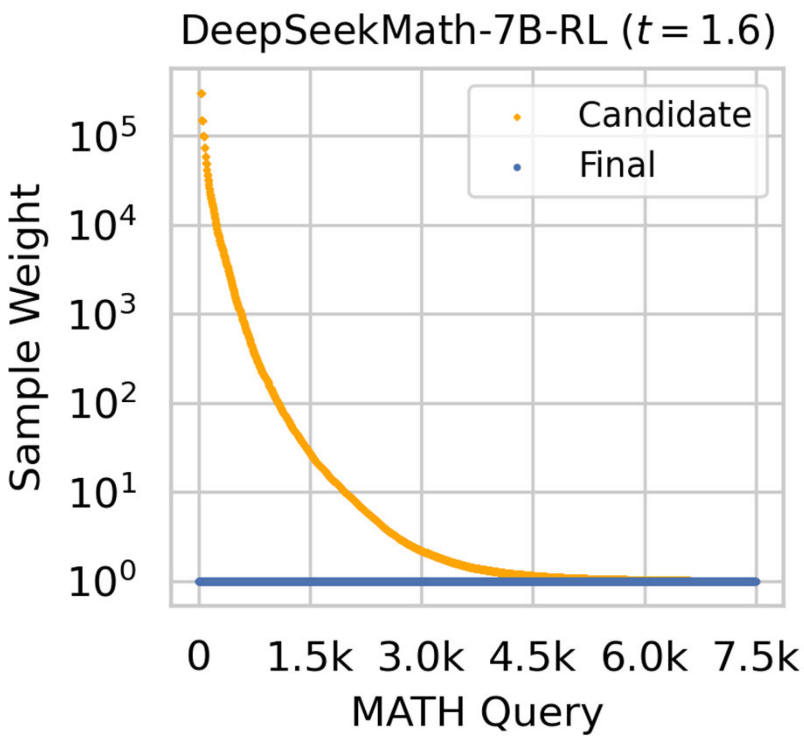
VRS controls the **candidate** distribution

explicitly control the **final** distribution →

Balanced across difficulties

(Uniform one-sample per query here)

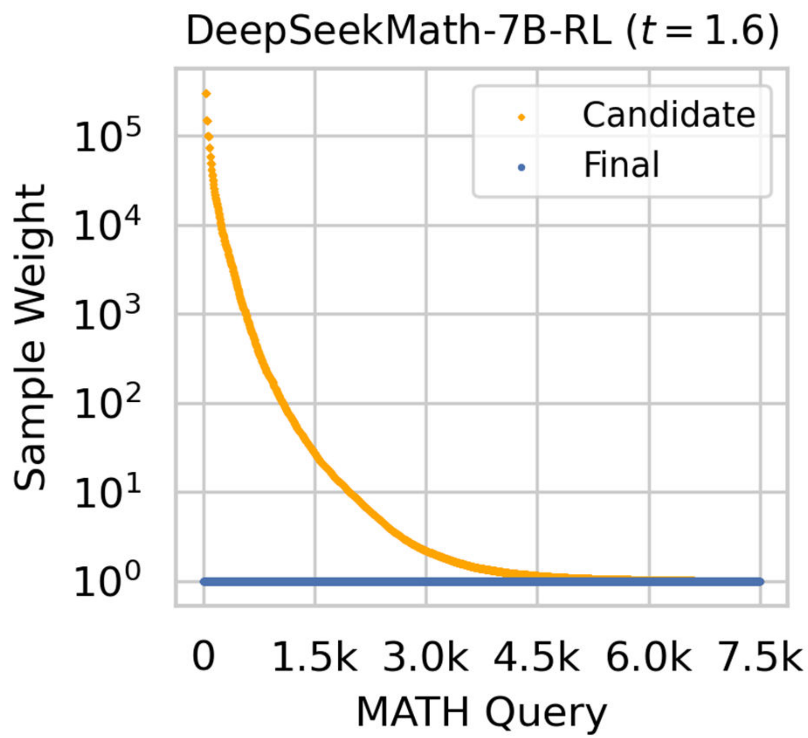
↑



How to Avoid the Bias towards Easier Queries by VRS?

VRS controls the **candidate** distribution

explicitly control the **final** distribution →



Balanced across difficulties

(Uniform one-sample per query here)

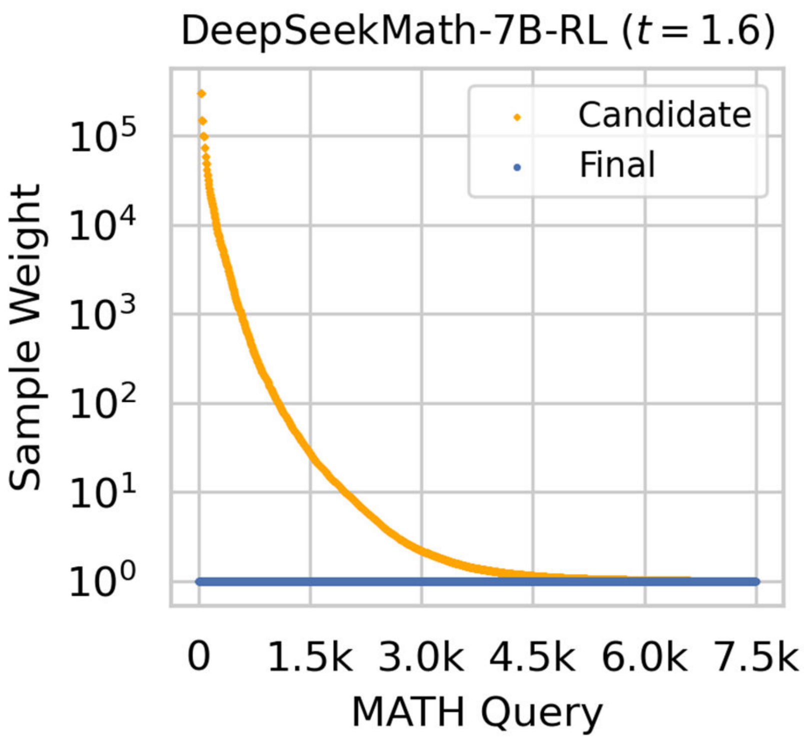
↑

Difficulty-Aware Rejection Sampling

How to Avoid the Bias towards Easier Queries by VRS?

VRS controls the **candidate** distribution

explicitly control the **final** distribution →



Balanced across difficulties

(Uniform one-sample per query here)

↑

Difficulty-Aware Rejection Sampling

(DARS)

Enough Difficult Data are Critical?

Enough Difficult Data are Critical?
→ Deliberate Bias towards Harder Queries?

Enough Difficult Data are Critical?
→ Deliberate Bias towards Harder Queries?

1. ***Uniform***: sampling responses until each query accumulates correct responses to a **fixed** number per query.

Enough Difficult Data are Critical?
→ Deliberate Bias towards Harder Queries?

1. ***Uniform***: sampling responses until each query accumulates correct responses to a **fixed** number per query.

↓

Enough Difficult Data are Critical?
→ Deliberate Bias towards Harder Queries?

1. ***Uniform***: sampling responses until each query accumulates correct responses to a **fixed** number per query.

↓

2. ***Prop2Diff***: sampling responses until each query accumulates correct responses to the number **proportional to** some **difficulty** metric per query.

Enough Difficult Data are Critical?
→ Deliberate Bias towards Harder Queries?

1. ***Uniform***: sampling responses until each query accumulates correct responses to a **fixed** number per query.

↓

2. ***Prop2Diff***: sampling responses until each query accumulates correct responses to the number **proportional to** some **difficulty** metric per query.

Enough Difficult Data are Critical?
→ Deliberate Bias towards Harder Queries?

1. ***Uniform***: sampling responses until each query accumulates correct responses to a **fixed** number per query.

↓

2. ***Prop2Diff***: sampling responses until each query accumulates correct responses to the number **proportional to** some **difficulty** metric per query.

E.g., in this work we use: *fail rate* = likelihood of wrong responses to the query

Results: *DART-Math* Datasets

Results: *DART-Math* Datasets

[(*DARS-Uniform* + *DARS-Prop2Diff*) + *VRS*]

Results: *DART-Math* Datasets

[(*DARS-Uniform* + *DARS-Prop2Diff*) + *VRS*]
*

Results: *DART-Math* Datasets

[(*DARS-Uniform* + *DARS-Prop2Diff*) + *VRS*]

*

(*MATH* + *GSM8K*)

Results: *DART-Math* Datasets

[(*DARS-Uniform* + *DARS-Prop2Diff*) + *VRS*]

*****

(*MATH* + *GSM8K*)

↓

Results: *DART-Math* Datasets

[(*DARS-Uniform* + *DARS-Prop2Diff*) + *VRS*]

*****

(*MATH* + *GSM8K*)

↓

***DART-Math* datasets** + VRT baseline dataset

Results: *DART-Math* Datasets

[(*DARS-Uniform* + *DARS-Prop2Diff*) + *VRS*]

*****

(*MATH* + *GSM8K*)

↓

***DART-Math* datasets** + VRT baseline dataset

Results: *DART-Math* Datasets

[(*DARS-Uniform* + *DARS-Prop2Diff*) + *VRS*]

*

(*MATH* + *GSM8K*)

↓

DART-Math **datasets** + VRT baseline dataset

(RT = Rejection Tuning)

Results: *DART-Math* Datasets

[(*DARS-Uniform* + *DARS-Prop2Diff*) + *VRS*]

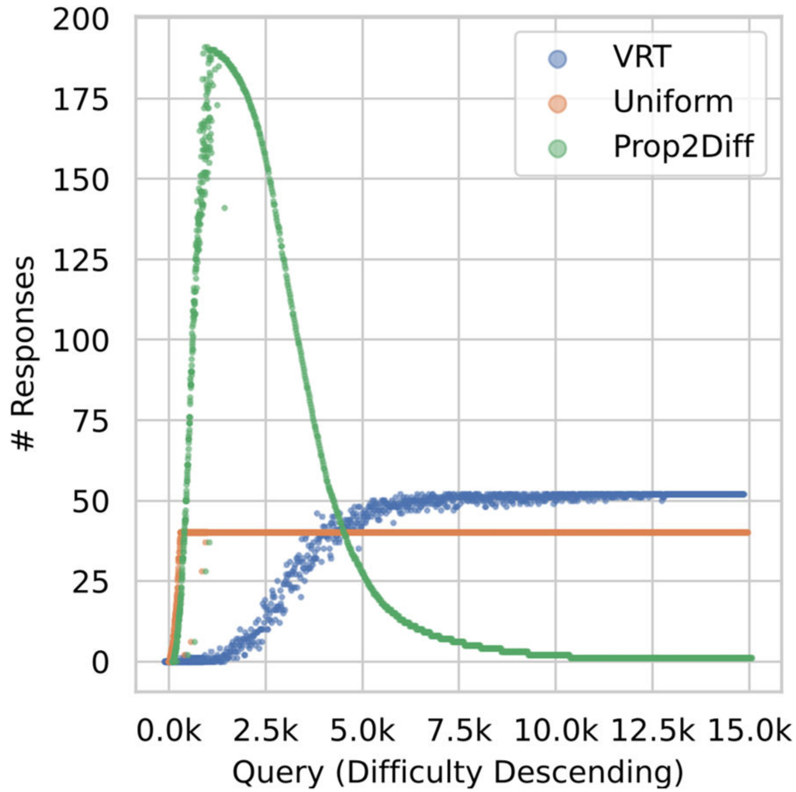
*

(*MATH* + *GSM8K*)

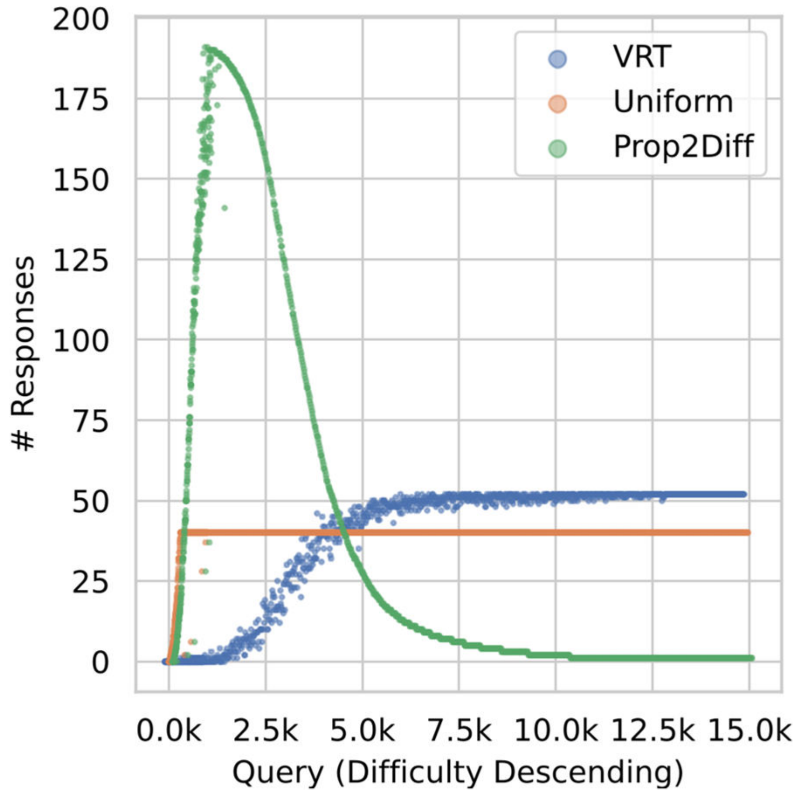
↓

***DART-Math* datasets** + VRT baseline dataset

(RT = Rejection Tuning)

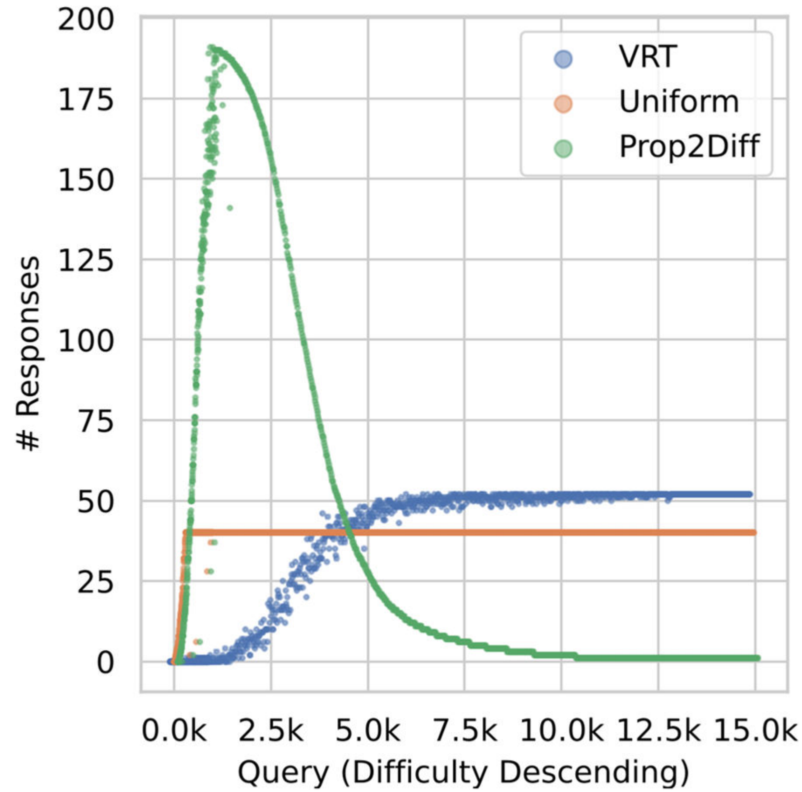


Results: *DART-Math* Datasets



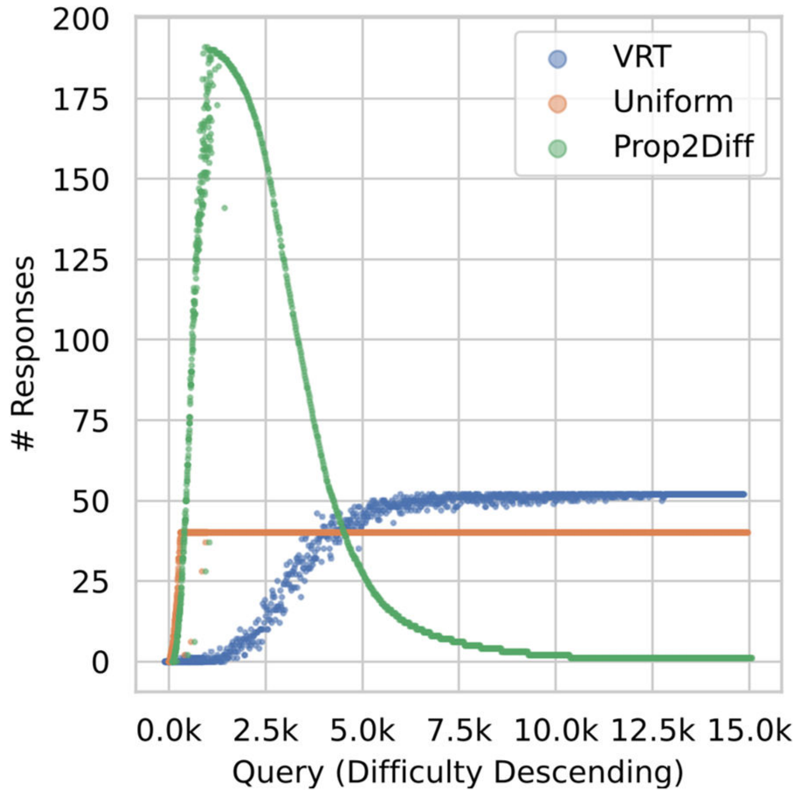
Results: *DART-Math* Datasets

As expected:



Results: *DART-Math* Datasets

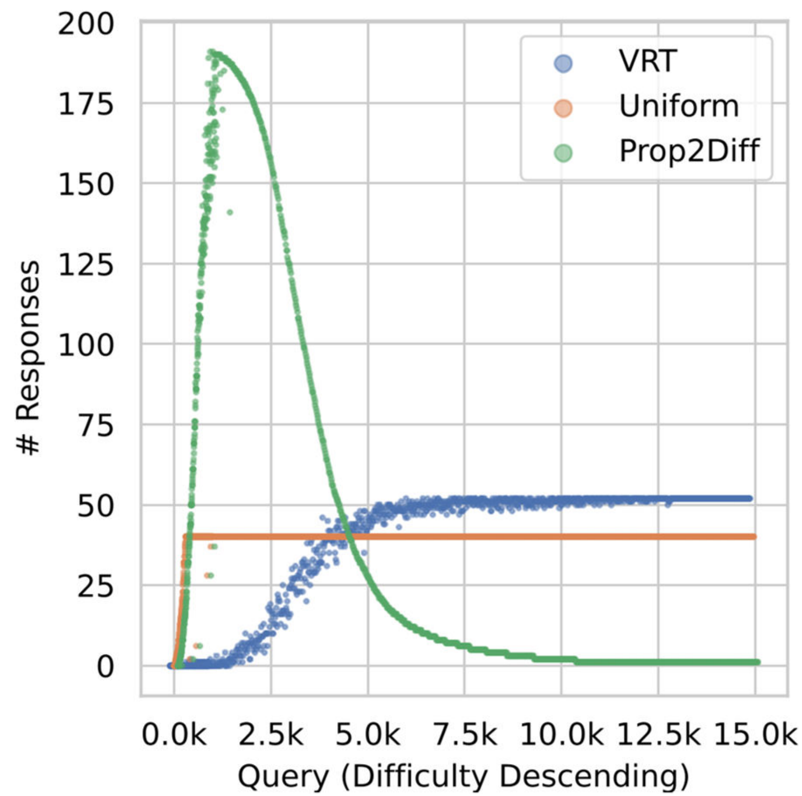
As expected:



Results: *DART-Math* Datasets

As expected:

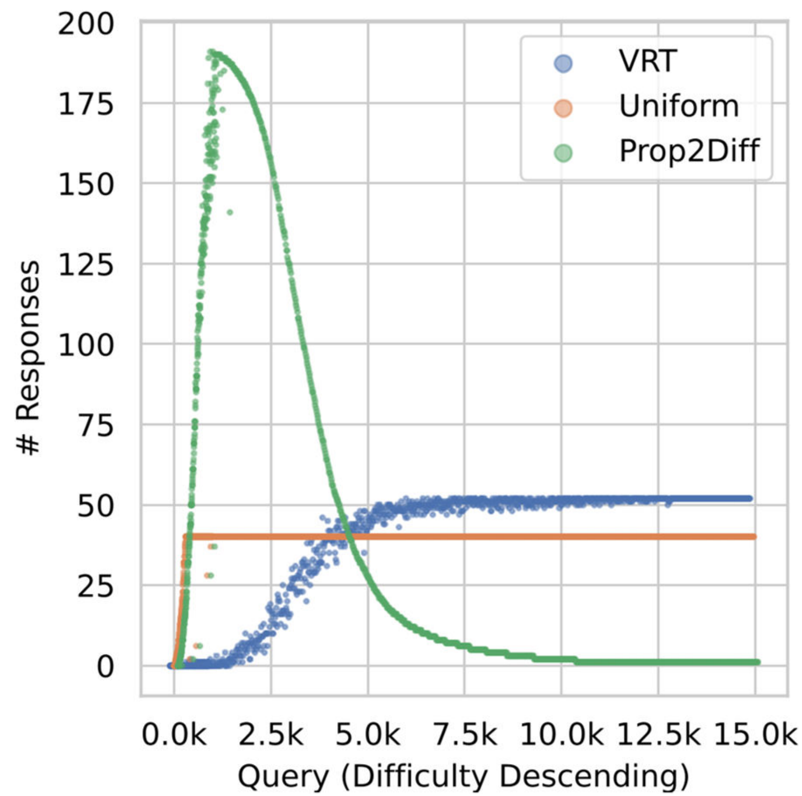
- ***VRT*** baseline is higher for **easier** queries (right)



Results: *DART-Math* Datasets

As expected:

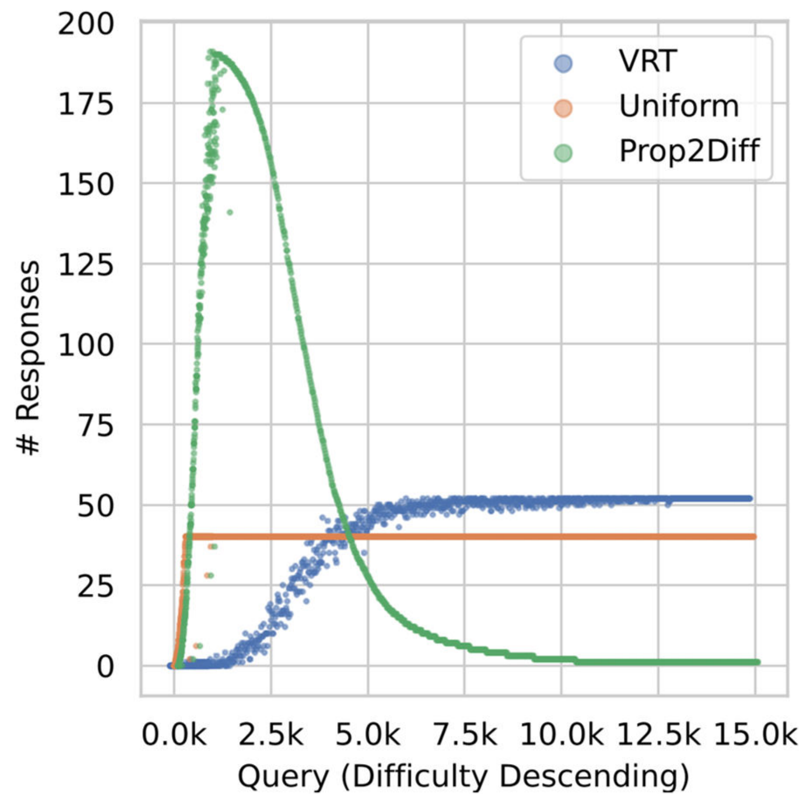
- ***VRT*** baseline is higher for **easier** queries (right)
- ***Uniform*** is almost **horizontal**



Results: *DART-Math* Datasets

As expected:

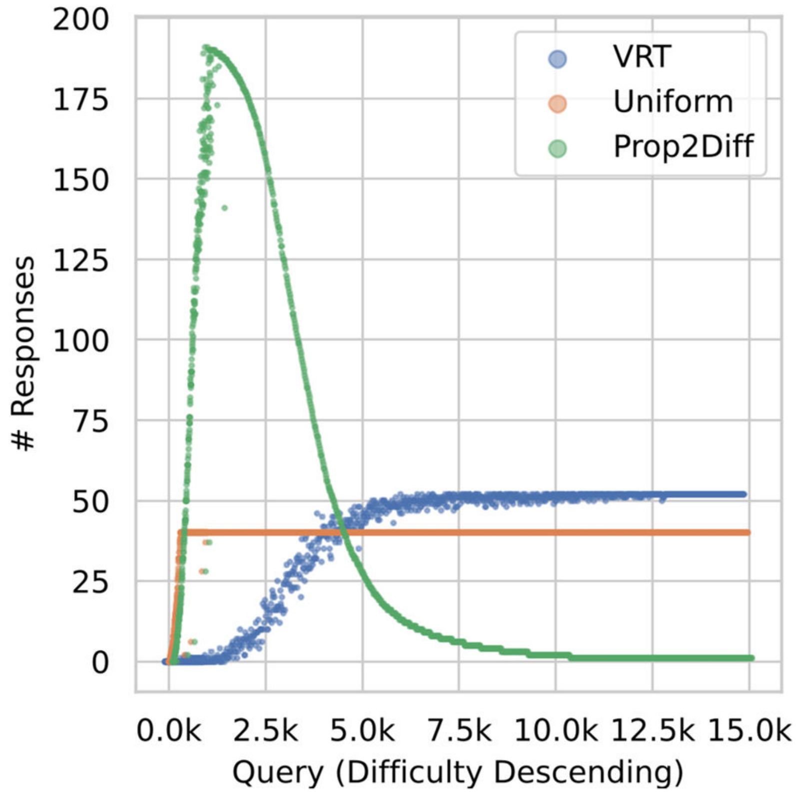
- ***VRT*** baseline is higher for **easier** queries (right)
- ***Uniform*** is almost **horizontal**
- ***Prop2Diff*** is higher for **harder** queries (left)



Results: *DART-Math* Datasets

As expected:

- ***VRT*** baseline is higher for **easier** queries (right)
- ***Uniform*** is almost **horizontal**
- ***Prop2Diff*** is higher for **harder** queries (left)

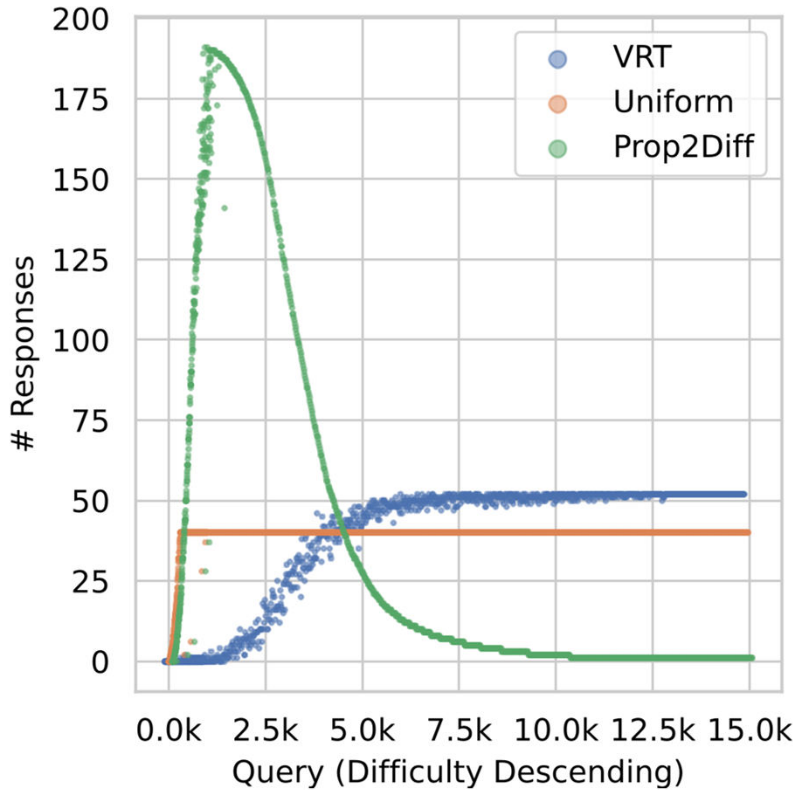


Results: *DART-Math* Datasets

As expected:

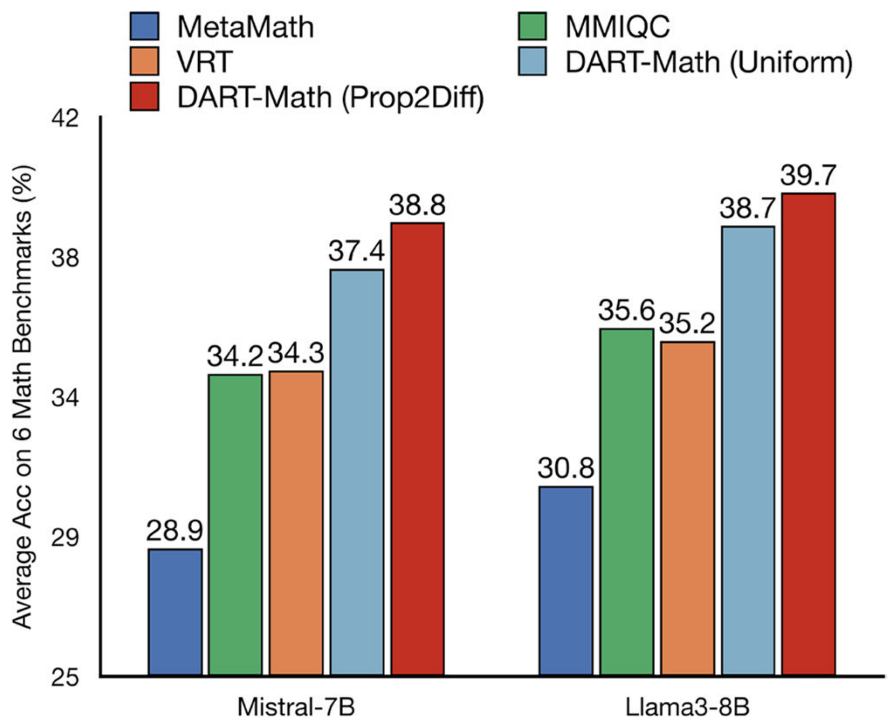
- ***VRT*** baseline is higher for **easier** queries (right)
- ***Uniform*** is almost **horizontal**
- ***Prop2Diff*** is higher for **harder** queries (left)

- **Area** under a line = **Dataset size** (all ~590k here)



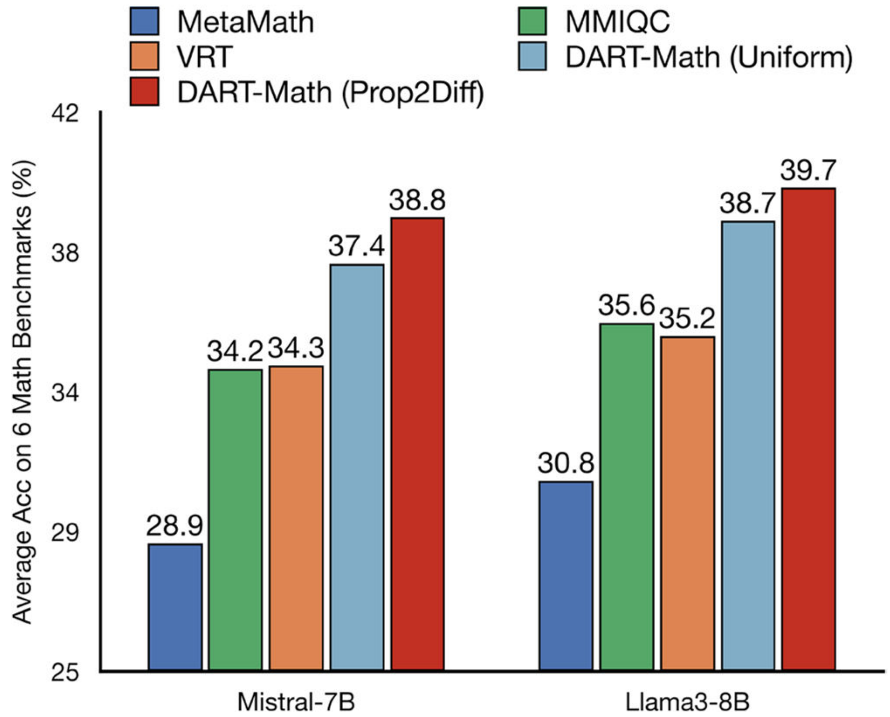
Results: *DART-Math* Models

Results: *DART-Math* Models



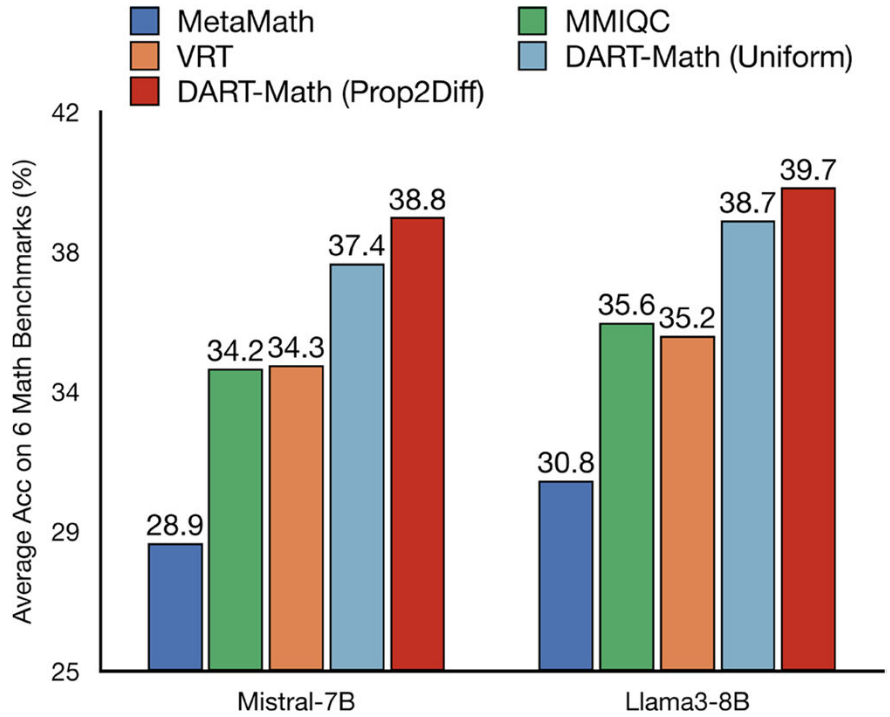
Results: *DART-Math* Models

DART-Math significantly outperforms



Results: *DART-Math* Models

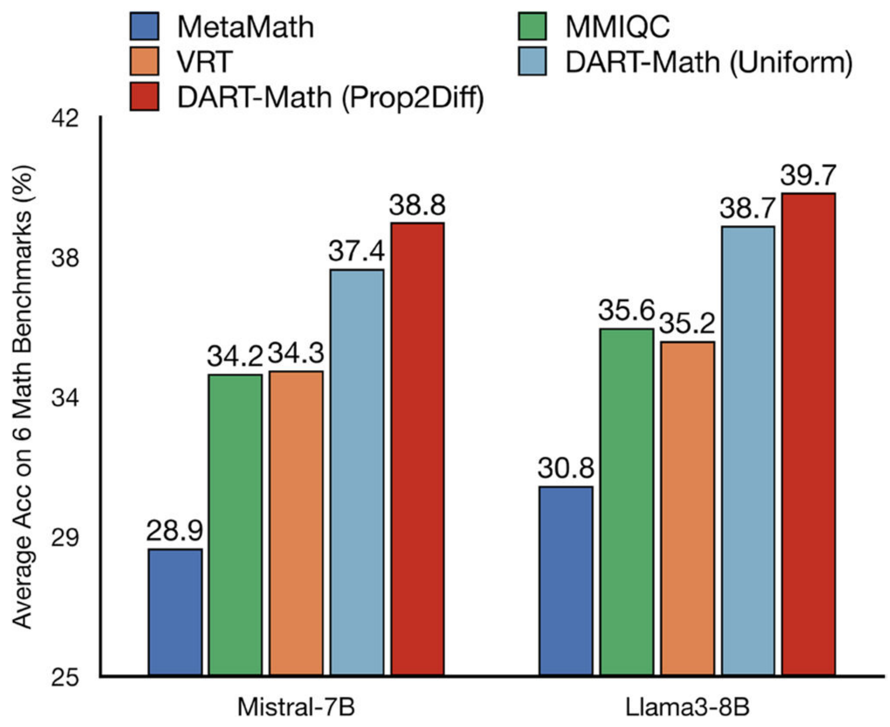
DART-Math significantly outperforms



Results: *DART-Math* Models

DART-Math significantly outperforms

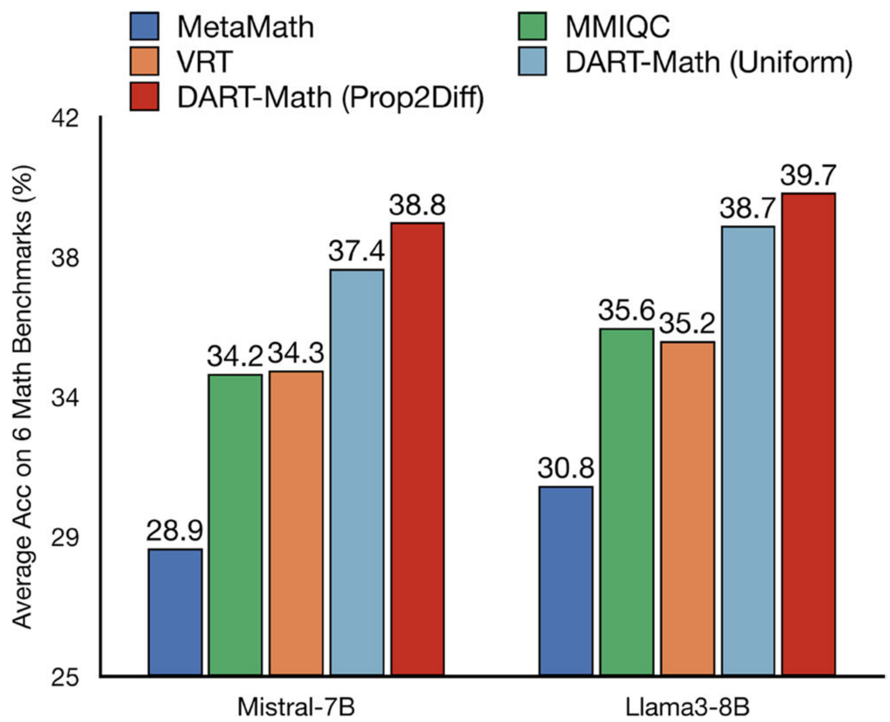
1. *VRT* (identical synthesis model)



Results: *DART-Math* Models

DART-Math significantly outperforms

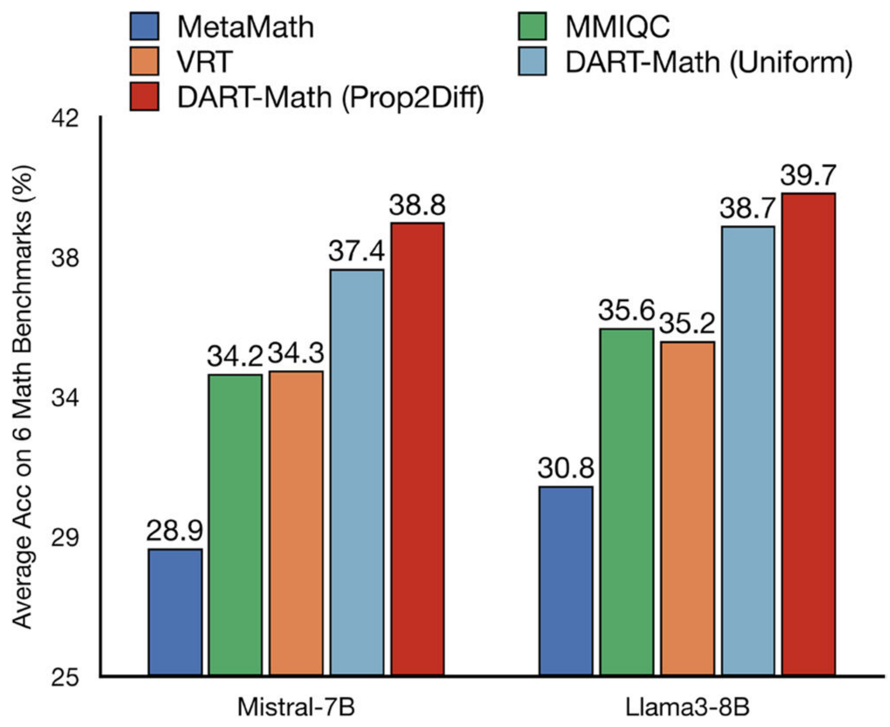
1. *VRT* (identical synthesis model)
2. baselines trained on previous top public datasets, e.g.,



Results: *DART-Math* Models

DART-Math significantly outperforms

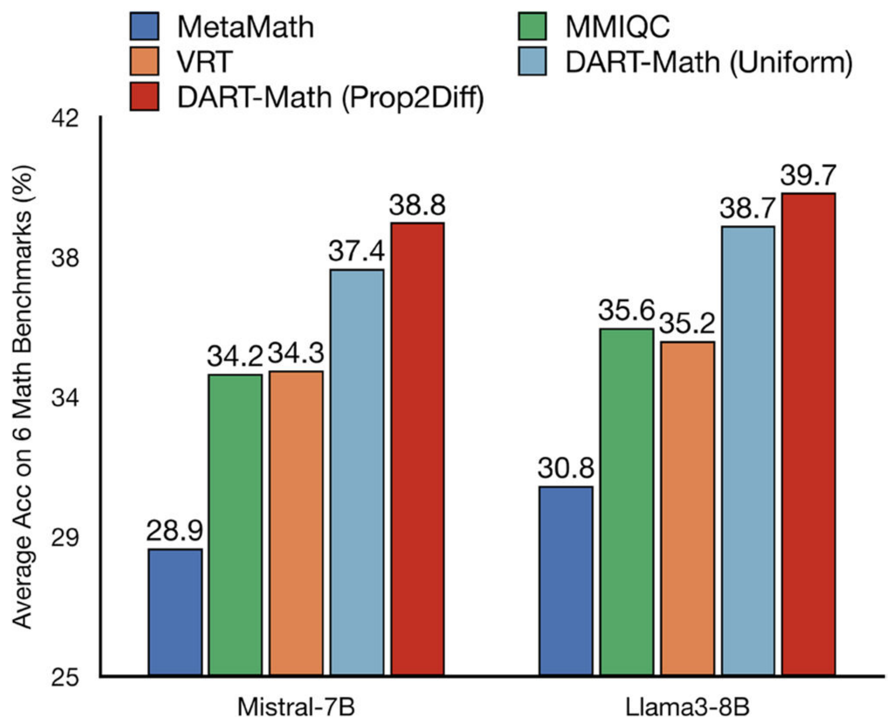
1. *VRT* (identical synthesis model)
2. baselines trained on previous top public datasets, e.g.,
   a. *MMIQC*



Results: *DART-Math* Models

DART-Math significantly outperforms

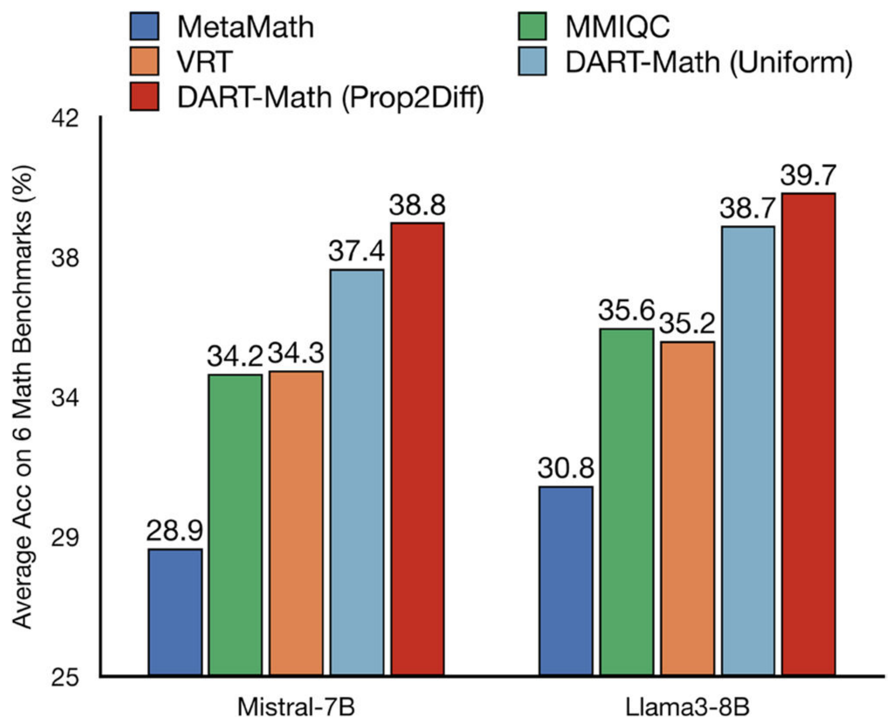
1. *VRT* (identical synthesis model)
2. baselines trained on previous top public datasets, e.g.,
   a. *MMIQC*
   b. *MetaMath*



Results: *DART-Math* Models

DART-Math significantly outperforms

1. *VRT* (identical synthesis model)
2. baselines trained on previous top public datasets, e.g.,
   a. *MMIQC*
   b. *MetaMath*

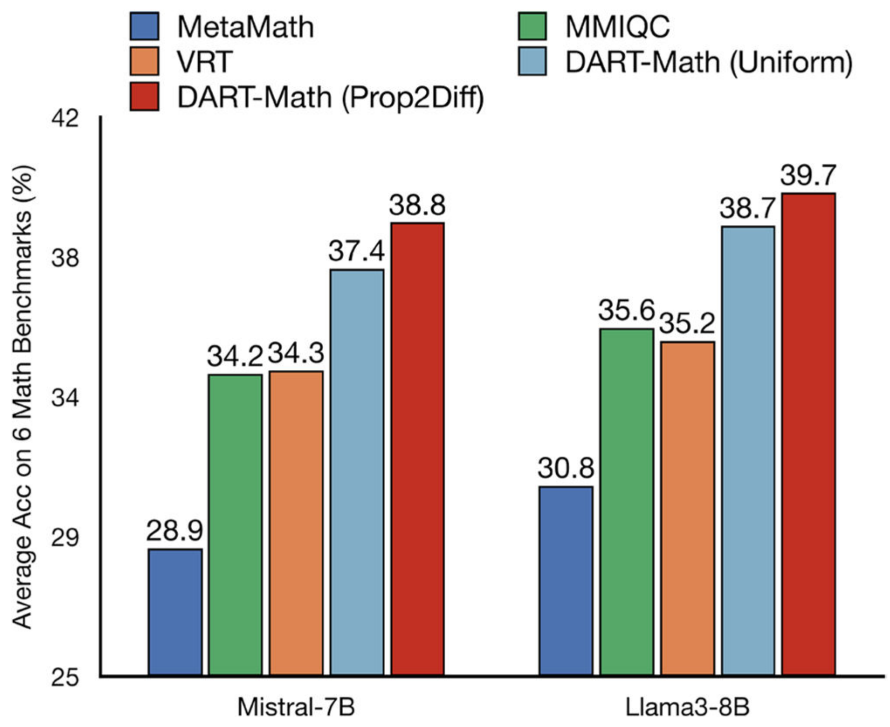


Results: *DART-Math* Models

DART-Math significantly outperforms

1. *VRT* (identical synthesis model)
2. baselines trained on previous top public datasets, e.g.,
    a. *MMIQC*
    b. *MetaMath*

often with smaller training dataset size,



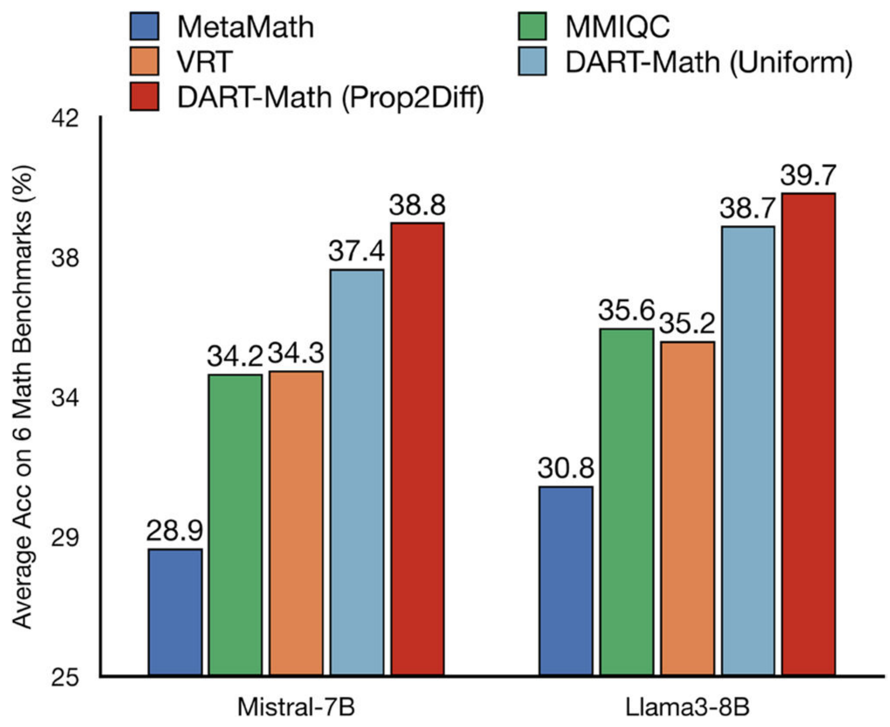
Results: *DART-Math* Models

DART-Math significantly outperforms

1. *VRT* (identical synthesis model)
2. baselines trained on previous top public datasets, e.g.,
   a. *MMIQC*
   b. *MetaMath*

often with smaller training dataset size,

- e.g., 0.59M << 2.2M for *MMIQC*



All Main Results

Model	# Samples	In-Domain MATH	GSM8K	College	Out-of-Domain DM	Olympiad	Theorem	AVG
GPT-4-Turbo (24-04-09)	–	73.4	94.5	–	–	–	48.4	–
GPT-4 (0314)	–	52.6	94.7	24.4	–	–	–	–
Claude-3-Opus	–	60.1	95.0	–	–	–	–	–
Gemini 1.5 Pro	–	67.7	–	–	–	–	–	–
70B General Base Model								
Llama2-70B-Xwin-Math-V1.1†	1.4M	52.5	90.2	33.1	58.0	16.3	14.9	44.2
Llama3-70B-ICL	–	44.0	80.1	33.5	51.7	10.8	27.0	41.2
Llama3-70B-MetaMath	0.40M	44.9	88.0	31.9	53.2	11.6	21.9	41.9
Llama3-70B-MMIQC	2.3M	49.4	89.3	37.6	60.4	15.3	23.5	45.9
Llama3-70B-VRT	0.59M	53.1	90.3	36.8	62.8	19.3	**28.6**	48.5
`DART-Math-Llama3-70B` (Uniform)	0.59M	54.9 ↑1.8	**90.4** ↑0.1	**38.5** ↑1.7	**64.1** ↑1.3	19.1 ↓0.2	27.4 ↓1.2	49.1 ↑0.6
`DART-Math-Llama3-70B` (Prop2Diff)	0.59M	**56.1** ↑3.0	89.6 ↓0.7	37.9 ↑1.1	**64.1** ↑1.3	**20.0** ↑0.7	28.2 ↓0.4	**49.3** ↑0.8
7B Math-Specialized Base Model								
DeepSeekMath-7B-ICL	–	35.5	64.2	34.7	45.2	9.3	23.5	35.4
DeepSeekMath-7B-Instruct	0.78M	46.9	82.7	37.1	52.2	14.2	28.1	43.5
DeepSeekMath-7B-MMIQC	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DeepSeekMath-7B-KPMath-Plus	1.6M	48.8	83.9	–	–	–	–	–
DeepSeekMath-7B-VRT	0.59M	53.0	**88.2**	**41.9**	60.2	19.1	27.2	48.3
`DART-Math-DSMath-7B` (Uniform)	0.59M	52.9 ↓0.1	**88.2**	40.1 ↓1.8	60.2	21.3 ↑2.2	**32.5** ↑5.3	49.2 ↑0.9
`DART-Math-DSMath-7B` (Prop2Diff)	0.59M	**53.6** ↑0.6	86.8 ↓1.4	40.7 ↓1.2	**61.6** ↑1.4	**21.7** ↑2.6	32.2 ↑5.0	**49.4** ↑1.1
7-8B General Base Model								
Llama2-7B-Xwin-Math-V1.1†	1.4M	45.5	84.9	27.6	43.0	10.5	15.0	37.8
Mistral-7B-ICL	–	16.5	45.9	17.9	23.5	3.7	14.2	20.3
Mistral-7B-WizardMath-V1.1 (RL)	–	32.3	80.4	23.1	38.4	7.7	16.6	33.1
Mistral-7B-MetaMath	0.40M	29.8	76.5	19.3	28.0	5.9	14.0	28.9
Mistral-7B-MMIQC	2.3M	37.4	75.4	28.5	38.0	9.4	16.2	34.2
Mistral-7B-MathScale	2.0M	35.2	74.8	21.8	–	–	–	–
Mistral-7B-KPMath-Plus	1.6M	**46.8**	82.1	–	–	–	–	–
Mistral-7B-VRT	0.59M	38.7	82.3	24.2	35.6	8.7	16.2	34.3
`DART-Math-Mistral-7B` (Uniform)	0.59M	43.5 ↑4.8	**82.6** ↑0.3	26.9 ↑2.7	42.0 ↑6.4	13.2 ↑4.5	16.4 ↑0.2	37.4 ↑3.1
`DART-Math-Mistral-7B` (Prop2Diff)	0.59M	45.5 ↑6.8	81.1 ↓1.2	**29.4** ↑5.2	**45.1** ↑9.5	**14.7** ↑6.0	**17.0** ↑0.8	**38.8** ↑4.5
Llama3-8B-ICL	–	21.2	51.0	19.9	27.4	4.2	19.8	23.9
Llama3-8B-MetaMath	0.40M	32.5	77.3	20.6	35.0	5.5	13.8	30.8
Llama3-8B-MMIQC	2.3M	39.5	77.6	**29.5**	41.0	9.6	16.2	35.6
Llama3-8B-VRT	0.59M	39.7	81.7	23.9	41.7	9.3	14.9	35.2
`DART-Math-Llama3-8B` (Uniform)	0.59M	45.3 ↑5.6	**82.5** ↑0.8	27.1 ↑3.2	**48.2** ↑6.5	13.6 ↑4.3	15.4 ↑0.5	38.7 ↑3.5
`DART-Math-Llama3-8B` (Prop2Diff)	0.59M	**46.6** ↑6.9	81.1 ↓0.6	28.8 ↑4.9	48.0 ↑6.3	**14.5** ↑5.2	**19.4** ↑4.5	**39.7** ↑4.5

Table 2: Main results on mathematical benchmarks. College, DM, Olympiad, Theorem denote the CollegeMath, DeepMind-Mathematics, OlympiadBench-Math, TheoremQA benchmarks respectively. We annotate the absolute accuracy change compared to the VRT baseline within the same base model. Bold means the best score within the respective base model. ICL, MetaMath, MMIQC, and VRT baselines are from our own runs, while other numbers are copied from the respective papers or reports. For WizardMath and Xwin-Math, we take the public model checkpoints and evaluate ourselves using their official CoT prompt. †: For Xwin-Math, we take the best public models that are based on Llama2 (Touvron et al., 2023), which is not a very fair comparison with others.

All Main Results

2 in-domain benchmarks

Model	# Samples	In-Domain		Out-of-Domain				
		MATH	GSM8K	College	DM	Olympiad	Theorem	AVG
GPT-4-Turbo (24-04-09)	–	73.4	94.5	–	–	–	48.4	–
GPT-4 (0314)	–	52.6	94.7	24.4	–	–	–	–
Claude-3-Opus	–	60.1	95.0	–	–	–	–	–
Gemini 1.5 Pro	–	67.7	–	–	–	–	–	–
70B General Base Model								
Llama2-70B-Xwin-Math-V1.1†	1.4M	52.5	90.2	33.1	58.0	16.3	14.9	44.2
Llama3-70B-ICL	–	44.0	80.1	33.5	51.7	10.8	27.0	41.2
Llama3-70B-MetaMath	0.40M	44.9	88.0	31.9	53.2	11.6	21.9	41.9
Llama3-70B-MMIQC	2.3M	49.4	89.3	37.6	60.4	15.3	23.5	45.9
Llama3-70B-VRT	0.59M	53.1	90.3	36.8	62.8	19.3	**28.6**	48.5
DART-Math-Llama3-70B (Uniform)	0.59M	54.9 ↑1.8	**90.4** ↑0.1	**38.5** ↑1.7	**64.1** ↑1.3	19.1 ↓0.2	27.4 ↓1.2	49.1 ↑0.6
DART-Math-Llama3-70B (Prop2Diff)	0.59M	**56.1** ↑3.0	89.6 ↓0.7	37.9 ↑1.1	**64.1** ↑1.3	**20.0** ↑0.7	28.2 ↓0.4	**49.3** ↑0.8
7B Math-Specialized Base Model								
DeepSeekMath-7B-ICL	–	35.5	64.2	34.7	45.2	9.3	23.5	35.4
DeepSeekMath-7B-Instruct	0.78M	46.9	82.7	37.1	52.2	14.2	28.1	43.5
DeepSeekMath-7B-MMIQC	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DeepSeekMath-7B-KPMath-Plus	1.6M	48.8	83.9	–	–	–	–	–
DeepSeekMath-7B-VRT	0.59M	53.0	**88.2**	**41.9**	60.2	19.1	27.2	48.3
DART-Math-DSMath-7B (Uniform)	0.59M	52.9 ↓0.1	**88.2**	40.1 ↓1.8	60.2	21.3 ↑2.2	**32.5** ↑5.3	49.2 ↑0.9
DART-Math-DSMath-7B (Prop2Diff)	0.59M	**53.6** ↑0.6	86.8 ↓1.4	40.7 ↓1.2	**61.6** ↑1.4	**21.7** ↑2.6	32.2 ↑5.0	**49.4** ↑1.1
7-8B General Base Model								
Llama2-7B-Xwin-Math-V1.1†	1.4M	45.5	84.9	27.6	43.0	10.5	15.0	37.8
Mistral-7B-ICL	–	16.5	45.9	17.9	23.5	3.7	14.2	20.3
Mistral-7B-WizardMath-V1.1 (RL)	–	32.3	80.4	23.1	38.4	7.7	16.6	33.1
Mistral-7B-MetaMath	0.40M	29.8	76.5	19.3	28.0	5.9	14.0	28.9
Mistral-7B-MMIQC	2.3M	37.4	75.4	28.5	38.0	9.4	16.2	34.2
Mistral-7B-MathScale	2.0M	35.2	74.8	21.8	–	–	–	–
Mistral-7B-KPMath-Plus	1.6M	**46.8**	82.1	–	–	–	–	–
Mistral-7B-VRT	0.59M	38.7	82.3	24.2	35.6	8.7	16.2	34.3
DART-Math-Mistral-7B (Uniform)	0.59M	43.5 ↑4.8	**82.6** ↑0.3	26.9 ↑2.7	42.0 ↑6.4	13.2 ↑4.5	16.4 ↑0.2	37.4 ↑3.1
DART-Math-Mistral-7B (Prop2Diff)	0.59M	45.5 ↑6.8	81.1 ↓1.2	**29.4** ↑5.2	**45.1** ↑9.5	**14.7** ↑6.0	**17.0** ↑0.8	**38.8** ↑4.5
Llama3-8B-ICL	–	21.2	51.0	19.9	27.4	4.2	19.8	23.9
Llama3-8B-MetaMath	0.40M	32.5	77.3	20.6	35.0	5.5	13.8	30.8
Llama3-8B-MMIQC	2.3M	39.5	77.6	**29.5**	41.0	9.6	16.2	35.6
Llama3-8B-VRT	0.59M	39.7	81.7	23.9	41.7	9.3	14.9	35.2
DART-Math-Llama3-8B (Uniform)	0.59M	45.3 ↑5.6	**82.5** ↑0.8	27.1 ↑3.2	**48.2** ↑6.5	13.6 ↑4.3	15.4 ↑0.5	38.7 ↑3.5
DART-Math-Llama3-8B (Prop2Diff)	0.59M	**46.6** ↑6.9	81.1 ↓0.6	28.8 ↑4.9	48.0 ↑6.3	**14.5** ↑5.2	**19.4** ↑4.5	**39.7** ↑4.5

Table 2: Main results on mathematical benchmarks. College, DM, Olympiad, Theorem denote the CollegeMath, DeepMind-Mathematics, OlympiadBench-Math, TheoremQA benchmarks respectively. We annotate the absolute accuracy change compared to the VRT baseline within the same base model. Bold means the best score within the respective base model. ICL, MetaMath, MMIQC, and VRT baselines are from our own runs, while other numbers are copied from the respective papers or reports. For WizardMath and Xwin-Math, we take the public model checkpoints and evaluate ourselves using their official CoT prompt. †: For Xwin-Math, we take the best public models that are based on Llama2 (Touvron et al., 2023), which is not a very fair comparison with others.

All Main Results

2 in-domain benchmarks

+

Model	# Samples	In-Domain MATH	GSM8K	College	Out-of-Domain DM	Olympiad	Theorem	AVG
GPT-4-Turbo (24-04-09)	–	73.4	94.5	–	–	–	48.4	–
GPT-4 (0314)	–	52.6	94.7	24.4	–	–	–	–
Claude-3-Opus	–	60.1	95.0	–	–	–	–	–
Gemini 1.5 Pro	–	67.7	–	–	–	–	–	–
70B General Base Model								
Llama2-70B-Xwin-Math-V1.1†	1.4M	52.5	90.2	33.1	58.0	16.3	14.9	44.2
Llama3-70B-ICL	–	44.0	80.1	33.5	51.7	10.8	27.0	41.2
Llama3-70B-MetaMath	0.40M	44.9	88.0	31.9	53.2	11.6	21.9	41.9
Llama3-70B-MMIQC	2.3M	49.4	89.3	37.6	60.4	15.3	23.5	45.9
Llama3-70B-VRT	0.59M	53.1	90.3	36.8	62.8	19.3	**28.6**	48.5
`DART-Math-Llama3-70B` (Uniform)	0.59M	54.9 ↑1.8	**90.4** ↑0.1	**38.5** ↑1.7	**64.1** ↑1.3	19.1 ↓0.2	27.4 ↓1.2	49.1 ↑0.6
`DART-Math-Llama3-70B` (Prop2Diff)	0.59M	**56.1** ↑3.0	89.6 ↓0.7	37.9 ↑1.1	**64.1** ↑1.3	**20.0** ↑0.7	28.2 ↓0.4	**49.3** ↑0.8
7B Math-Specialized Base Model								
DeepSeekMath-7B-ICL	–	35.5	64.2	34.7	45.2	9.3	23.5	35.4
DeepSeekMath-7B-Instruct	0.78M	46.9	82.7	37.1	52.2	14.2	28.1	43.5
DeepSeekMath-7B-MMIQC	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DeepSeekMath-7B-KPMath-Plus	1.6M	48.8	83.9	–	–	–	–	–
DeepSeekMath-7B-VRT	0.59M	53.0	**88.2**	**41.9**	60.2	19.1	27.2	48.3
`DART-Math-DSMath-7B` (Uniform)	0.59M	52.9 ↓0.1	**88.2**	40.1 ↓1.8	60.2	21.3 ↑2.2	**32.5** ↑5.3	49.2 ↑0.9
`DART-Math-DSMath-7B` (Prop2Diff)	0.59M	**53.6** ↑0.6	86.8 ↓1.4	40.7 ↓1.2	**61.6** ↑1.4	**21.7** ↑2.6	32.2 ↑5.0	**49.4** ↑1.1
7-8B General Base Model								
Llama2-7B-Xwin-Math-V1.1†	1.4M	45.5	84.9	27.6	43.0	10.5	15.0	37.8
Mistral-7B-ICL	–	16.5	45.9	17.9	23.5	3.7	14.2	20.3
Mistral-7B-WizardMath-V1.1 (RL)	–	32.3	80.4	23.1	38.4	7.7	16.6	33.1
Mistral-7B-MetaMath	0.40M	29.8	76.5	19.3	28.0	5.9	14.0	28.9
Mistral-7B-MMIQC	2.3M	37.4	75.4	28.5	38.0	9.4	16.2	34.2
Mistral-7B-MathScale	2.0M	35.2	74.8	21.8	–	–	–	–
Mistral-7B-KPMath-Plus	1.6M	**46.8**	82.1	–	–	–	–	–
Mistral-7B-VRT	0.59M	38.7	82.3	24.2	35.6	8.7	16.2	34.3
`DART-Math-Mistral-7B` (Uniform)	0.59M	43.5 ↑4.8	**82.6** ↑0.3	26.9 ↑2.7	42.0 ↑6.4	13.2 ↑4.5	16.4 ↑0.2	37.4 ↑3.1
`DART-Math-Mistral-7B` (Prop2Diff)	0.59M	45.5 ↑6.8	81.1 ↓1.2	**29.4** ↑5.2	**45.1** ↑9.5	**14.7** ↑6.0	**17.0** ↑0.8	**38.8** ↑4.5
Llama3-8B-ICL	–	21.2	51.0	19.9	27.4	4.2	19.8	23.9
Llama3-8B-MetaMath	0.40M	32.5	77.3	20.6	35.0	5.5	13.8	30.8
Llama3-8B-MMIQC	2.3M	39.5	77.6	**29.5**	41.0	9.6	16.2	35.6
Llama3-8B-VRT	0.59M	39.7	81.7	23.9	41.7	9.3	14.9	35.2
`DART-Math-Llama3-8B` (Uniform)	0.59M	45.3 ↑5.6	**82.5** ↑0.8	27.1 ↑3.2	**48.2** ↑6.5	13.6 ↑4.3	15.4 ↑0.5	38.7 ↑3.5
`DART-Math-Llama3-8B` (Prop2Diff)	0.59M	**46.6** ↑6.9	81.1 ↓0.6	28.8 ↑4.9	48.0 ↑6.3	**14.5** ↑5.2	**19.4** ↑4.5	**39.7** ↑4.5

Table 2: Main results on mathematical benchmarks. College, DM, Olympiad, Theorem denote the CollegeMath, DeepMind-Mathematics, OlympiadBench-Math, TheoremQA benchmarks respectively. We annotate the absolute accuracy change compared to the VRT baseline within the same base model. Bold means the best score within the respective base model. ICL, MetaMath, MMIQC, and VRT baselines are from our own runs, while other numbers are copied from the respective papers or reports. For WizardMath and Xwin-Math, we take the public model checkpoints and evaluate ourselves using their official CoT prompt. †: For Xwin-Math, we take the best public models that are based on Llama2 (Touvron et al., 2023), which is not a very fair comparison with others.

All Main Results

2 in-domain benchmarks

+

4 challenging

Model	# Samples	In-Domain MATH	GSM8K	Out-of-Domain College	DM	Olympiad	Theorem	AVG
GPT-4-Turbo (24-04-09)	–	73.4	94.5	–	–	–	48.4	–
GPT-4 (0314)	–	52.6	94.7	24.4	–	–	–	–
Claude-3-Opus	–	60.1	95.0	–	–	–	–	–
Gemini 1.5 Pro	–	67.7	–	–	–	–	–	–
70B General Base Model								
Llama2-70B-Xwin-Math-V1.1†	1.4M	52.5	90.2	33.1	58.0	16.3	14.9	44.2
Llama3-70B-ICL	–	44.0	80.1	33.5	51.7	10.8	27.0	41.2
Llama3-70B-MetaMath	0.40M	44.9	88.0	31.9	53.2	11.6	21.9	41.9
Llama3-70B-MMIQC	2.3M	49.4	89.3	37.6	60.4	15.3	23.5	45.9
Llama3-70B-VRT	0.59M	53.1	90.3	36.8	62.8	19.3	**28.6**	48.5
DART-Math-Llama3-70B (Uniform)	0.59M	54.9 ↑1.8	**90.4** ↑0.1	**38.5** ↑1.7	**64.1** ↑1.3	19.1 ↓0.2	27.4 ↓1.2	49.1 ↑0.6
DART-Math-Llama3-70B (Prop2Diff)	0.59M	**56.1** ↑3.0	89.6 ↓0.7	37.9 ↑1.1	**64.1** ↑1.3	**20.0** ↑0.7	28.2 ↓0.4	**49.3** ↑0.8
7B Math-Specialized Base Model								
DeepSeekMath-7B-ICL	–	35.5	64.2	34.7	45.2	9.3	23.5	35.4
DeepSeekMath-7B-Instruct	0.78M	46.9	82.7	37.1	52.2	14.2	28.1	43.5
DeepSeekMath-7B-MMIQC	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DeepSeekMath-7B-KPMath-Plus	1.6M	48.8	83.9	–	–	–	–	–
DeepSeekMath-7B-VRT	0.59M	53.0	**88.2**	**41.9**	60.2	19.1	27.2	48.3
DART-Math-DSMath-7B (Uniform)	0.59M	52.9 ↓0.1	**88.2**	40.1 ↓1.8	60.2	21.3 ↑2.2	**32.5** ↑5.3	49.2 ↑0.9
DART-Math-DSMath-7B (Prop2Diff)	0.59M	**53.6** ↑0.6	86.8 ↓1.4	40.7 ↓1.2	**61.6** ↑1.4	**21.7** ↑2.6	32.2 ↑5.0	**49.4** ↑1.1
7-8B General Base Model								
Llama2-7B-Xwin-Math-V1.1†	1.4M	45.5	84.9	27.6	43.0	10.5	15.0	37.8
Mistral-7B-ICL	–	16.5	45.9	17.9	23.5	3.7	14.2	20.3
Mistral-7B-WizardMath-V1.1 (RL)	–	32.3	80.4	23.1	38.4	7.7	16.6	33.1
Mistral-7B-MetaMath	0.40M	29.8	76.5	19.3	28.0	5.9	14.0	28.9
Mistral-7B-MMIQC	2.3M	37.4	75.4	28.5	38.0	9.4	16.2	34.2
Mistral-7B-MathScale	2.0M	35.2	74.8	21.8	–	–	–	–
Mistral-7B-KPMath-Plus	1.6M	**46.8**	82.1	–	–	–	–	–
Mistral-7B-VRT	0.59M	38.7	82.3	24.2	35.6	8.7	16.2	34.3
DART-Math-Mistral-7B (Uniform)	0.59M	43.5 ↑4.8	**82.6** ↑0.3	26.9 ↑2.7	42.0 ↑6.4	13.2 ↑4.5	16.4 ↑0.2	37.4 ↑3.1
DART-Math-Mistral-7B (Prop2Diff)	0.59M	45.5 ↑6.8	81.1 ↓1.2	**29.4** ↑5.2	**45.1** ↑9.5	**14.7** ↑6.0	**17.0** ↑0.8	**38.8** ↑4.5
Llama3-8B-ICL	–	21.2	51.0	19.9	27.4	4.2	19.8	23.9
Llama3-8B-MetaMath	0.40M	32.5	77.3	20.6	35.0	5.5	13.8	30.8
Llama3-8B-MMIQC	2.3M	39.5	77.6	**29.5**	41.0	9.6	16.2	35.6
Llama3-8B-VRT	0.59M	39.7	81.7	23.9	41.7	9.3	14.9	35.2
DART-Math-Llama3-8B (Uniform)	0.59M	45.3 ↑5.6	**82.5** ↑0.8	27.1 ↑3.2	**48.2** ↑6.5	13.6 ↑4.3	15.4 ↑0.5	38.7 ↑3.5
DART-Math-Llama3-8B (Prop2Diff)	0.59M	**46.6** ↑6.9	81.1 ↓0.6	28.8 ↑4.9	48.0 ↑6.3	**14.5** ↑5.2	**19.4** ↑4.5	**39.7** ↑4.5

Table 2: Main results on mathematical benchmarks. College, DM, Olympiad, Theorem denote the CollegeMath, DeepMind-Mathematics, OlympiadBench-Math, TheoremQA benchmarks respectively. We annotate the absolute accuracy change compared to the VRT baseline within the same base model. Bold means the best score within the respective base model. ICL, MetaMath, MMIQC, and VRT baselines are from our own runs, while other numbers are copied from the respective papers or reports. For WizardMath and Xwin-Math, we take the public model checkpoints and evaluate ourselves using their official CoT prompt. †: For Xwin-Math, we take the best public models that are based on Llama2 (Touvron et al., 2023), which is not a very fair comparison with others.

All Main Results

2 in-domain benchmarks

\+

4 challenging

out-of-domain benchmarks

Model	# Samples	In-Domain MATH	GSM8K	Out-of-Domain College	DM	Olympiad	Theorem	AVG
GPT-4-Turbo (24-04-09)	–	73.4	94.5	–	–	–	48.4	–
GPT-4 (0314)	–	52.6	94.7	24.4	–	–	–	–
Claude-3-Opus	–	60.1	95.0	–	–	–	–	–
Gemini 1.5 Pro	–	67.7	–	–	–	–	–	–
70B General Base Model								
Llama2-70B-Xwin-Math-V1.1†	1.4M	52.5	90.2	33.1	58.0	16.3	14.9	44.2
Llama3-70B-ICL	–	44.0	80.1	33.5	51.7	10.8	27.0	41.2
Llama3-70B-MetaMath	0.40M	44.9	88.0	31.9	53.2	11.6	21.9	41.9
Llama3-70B-MMIQC	2.3M	49.4	89.3	37.6	60.4	15.3	23.5	45.9
Llama3-70B-VRT	0.59M	53.1	90.3	36.8	62.8	19.3	**28.6**	48.5
DART-Math-Llama3-70B (Uniform)	0.59M	54.9 ↑1.8	**90.4** ↑0.1	**38.5** ↑1.7	**64.1** ↑1.3	19.1 ↓0.2	27.4 ↓1.2	49.1 ↑0.6
DART-Math-Llama3-70B (Prop2Diff)	0.59M	**56.1** ↑3.0	89.6 ↓0.7	37.9 ↑1.1	**64.1** ↑1.3	**20.0** ↑0.7	28.2 ↓0.4	**49.3** ↑0.8
7B Math-Specialized Base Model								
DeepSeekMath-7B-ICL	–	35.5	64.2	34.7	45.2	9.3	23.5	35.4
DeepSeekMath-7B-Instruct	0.78M	46.9	82.7	37.1	52.2	14.2	28.1	43.5
DeepSeekMath-7B-MMIQC	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DeepSeekMath-7B-KPMath-Plus	1.6M	48.8	83.9	–	–	–	–	–
DeepSeekMath-7B-VRT	0.59M	53.0	**88.2**	**41.9**	60.2	19.1	27.2	48.3
DART-Math-DSMath-7B (Uniform)	0.59M	52.9 ↓0.1	**88.2**	40.1 ↓1.8	60.2	21.3 ↑2.2	**32.5** ↑5.3	49.2 ↑0.9
DART-Math-DSMath-7B (Prop2Diff)	0.59M	**53.6** ↑0.6	86.8 ↓1.4	40.7 ↓1.2	**61.6** ↑1.4	**21.7** ↑2.6	32.2 ↑5.0	**49.4** ↑1.1
7-8B General Base Model								
Llama2-7B-Xwin-Math-V1.1†	1.4M	45.5	84.9	27.6	43.0	10.5	15.0	37.8
Mistral-7B-ICL	–	16.5	45.9	17.9	23.5	3.7	14.2	20.3
Mistral-7B-WizardMath-V1.1 (RL)	–	32.3	80.4	23.1	38.4	7.7	16.6	33.1
Mistral-7B-MetaMath	0.40M	29.8	76.5	19.3	28.0	5.9	14.0	28.9
Mistral-7B-MMIQC	2.3M	37.4	75.4	28.5	38.0	9.4	16.2	34.2
Mistral-7B-MathScale	2.0M	35.2	74.8	21.8	–	–	–	–
Mistral-7B-KPMath-Plus	1.6M	**46.8**	82.1	–	–	–	–	–
Mistral-7B-VRT	0.59M	38.7	82.3	24.2	35.6	8.7	16.2	34.3
DART-Math-Mistral-7B (Uniform)	0.59M	43.5 ↑4.8	**82.6** ↑0.3	26.9 ↑2.7	42.0 ↑6.4	13.2 ↑4.5	16.4 ↑0.2	37.4 ↑3.1
DART-Math-Mistral-7B (Prop2Diff)	0.59M	45.5 ↑6.8	81.1 ↓1.2	**29.4** ↑5.2	**45.1** ↑9.5	**14.7** ↑6.0	**17.0** ↑0.8	**38.8** ↑4.5
Llama3-8B-ICL	–	21.2	51.0	19.9	27.4	4.2	19.8	23.9
Llama3-8B-MetaMath	0.40M	32.5	77.3	20.6	35.0	5.5	13.8	30.8
Llama3-8B-MMIQC	2.3M	39.5	77.6	**29.5**	41.0	9.6	16.2	35.6
Llama3-8B-VRT	0.59M	39.7	81.7	23.9	41.7	9.3	14.9	35.2
DART-Math-Llama3-8B (Uniform)	0.59M	45.3 ↑5.6	**82.5** ↑0.8	27.1 ↑3.2	**48.2** ↑6.5	13.6 ↑4.3	15.4 ↑0.5	38.7 ↑3.5
DART-Math-Llama3-8B (Prop2Diff)	0.59M	**46.6** ↑6.9	81.1 ↓0.6	28.8 ↑4.9	48.0 ↑6.3	**14.5** ↑5.2	**19.4** ↑4.5	**39.7** ↑4.5

Table 2: Main results on mathematical benchmarks. College, DM, Olympiad, Theorem denote the CollegeMath, DeepMind-Mathematics, OlympiadBench-Math, TheoremQA benchmarks respectively. We annotate the absolute accuracy change compared to the VRT baseline within the same base model. Bold means the best score within the respective base model. ICL, MetaMath, MMIQC, and VRT baselines are from our own runs, while other numbers are copied from the respective papers or reports. For WizardMath and Xwin-Math, we take the public model checkpoints and evaluate ourselves using their official CoT prompt. †: For Xwin-Math, we take the best public models that are based on Llama2 (Touvron et al., 2023), which is not a very fair comparison with others.

All Main Results

2 in-domain benchmarks

+

4 challenging

out-of-domain benchmarks

+

Model	# Samples	In-Domain MATH	In-Domain GSM8K	Out-of-Domain College	Out-of-Domain DM	Out-of-Domain Olympiad	Out-of-Domain Theorem	Out-of-Domain AVG
GPT-4-Turbo (24-04-09)	–	73.4	94.5	–	–	–	48.4	–
GPT-4 (0314)	–	52.6	94.7	24.4	–	–	–	–
Claude-3-Opus	–	60.1	95.0	–	–	–	–	–
Gemini 1.5 Pro	–	67.7	–	–	–	–	–	–
70B General Base Model								
Llama2-70B-Xwin-Math-V1.1†	1.4M	52.5	90.2	33.1	58.0	16.3	14.9	44.2
Llama3-70B-ICL	–	44.0	80.1	33.5	51.7	10.8	27.0	41.2
Llama3-70B-MetaMath	0.40M	44.9	88.0	31.9	53.2	11.6	21.9	41.9
Llama3-70B-MMIQC	2.3M	49.4	89.3	37.6	60.4	15.3	23.5	45.9
Llama3-70B-VRT	0.59M	53.1	90.3	36.8	62.8	19.3	**28.6**	48.5
DART-Math-Llama3-70B (Uniform)	0.59M	54.9 ↑1.8	**90.4** ↑0.1	**38.5** ↑1.7	**64.1** ↑1.3	19.1 ↓0.2	27.4 ↓1.2	49.1 ↑0.6
DART-Math-Llama3-70B (Prop2Diff)	0.59M	**56.1** ↑3.0	89.6 ↓0.7	37.9 ↑1.1	**64.1** ↑1.3	**20.0** ↑0.7	28.2 ↓0.4	**49.3** ↑0.8
7B Math-Specialized Base Model								
DeepSeekMath-7B-ICL	–	35.5	64.2	34.7	45.2	9.3	23.5	35.4
DeepSeekMath-7B-Instruct	0.78M	46.9	82.7	37.1	52.2	14.2	28.1	43.5
DeepSeekMath-7B-MMIQC	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DeepSeekMath-7B-KPMath-Plus	1.6M	48.8	83.9	–	–	–	–	–
DeepSeekMath-7B-VRT	0.59M	53.0	**88.2**	**41.9**	60.2	19.1	27.2	48.3
DART-Math-DSMath-7B (Uniform)	0.59M	52.9 ↓0.1	**88.2**	40.1 ↓1.8	60.2	21.3 ↑2.2	**32.5** ↑5.3	49.2 ↑0.9
DART-Math-DSMath-7B (Prop2Diff)	0.59M	**53.6** ↑0.6	86.8 ↓1.4	40.7 ↓1.2	**61.6** ↑1.4	**21.7** ↑2.6	32.2 ↑5.0	**49.4** ↑1.1
7-8B General Base Model								
Llama2-7B-Xwin-Math-V1.1†	1.4M	45.5	84.9	27.6	43.0	10.5	15.0	37.8
Mistral-7B-ICL	–	16.5	45.9	17.9	23.5	3.7	14.2	20.3
Mistral-7B-WizardMath-V1.1 (RL)	–	32.3	80.4	23.1	38.4	7.7	16.6	33.1
Mistral-7B-MetaMath	0.40M	29.8	76.5	19.3	28.0	5.9	14.0	28.9
Mistral-7B-MMIQC	2.3M	37.4	75.4	28.5	38.0	9.4	16.2	34.2
Mistral-7B-MathScale	2.0M	35.2	74.8	21.8	–	–	–	–
Mistral-7B-KPMath-Plus	1.6M	**46.8**	82.1	–	–	–	–	–
Mistral-7B-VRT	0.59M	38.7	82.3	24.2	35.6	8.7	16.2	34.3
DART-Math-Mistral-7B (Uniform)	0.59M	43.5 ↑4.8	**82.6** ↑0.3	26.9 ↑2.7	42.0 ↑6.4	13.2 ↑4.5	16.4 ↑0.2	37.4 ↑3.1
DART-Math-Mistral-7B (Prop2Diff)	0.59M	45.5 ↑6.8	81.1 ↓1.2	**29.4** ↑5.2	**45.1** ↑9.5	**14.7** ↑6.0	**17.0** ↑0.8	**38.8** ↑4.5
Llama3-8B-ICL	–	21.2	51.0	19.9	27.4	4.2	19.8	23.9
Llama3-8B-MetaMath	0.40M	32.5	77.3	20.6	35.0	5.5	13.8	30.8
Llama3-8B-MMIQC	2.3M	39.5	77.6	**29.5**	41.0	9.6	16.2	35.6
Llama3-8B-VRT	0.59M	39.7	81.7	23.9	41.7	9.3	14.9	35.2
DART-Math-Llama3-8B (Uniform)	0.59M	45.3 ↑5.6	**82.5** ↑0.8	27.1 ↑3.2	**48.2** ↑6.5	13.6 ↑4.3	15.4 ↑0.5	38.7 ↑3.5
DART-Math-Llama3-8B (Prop2Diff)	0.59M	**46.6** ↑6.9	81.1 ↓0.6	28.8 ↑4.9	48.0 ↑6.3	**14.5** ↑5.2	**19.4** ↑4.5	**39.7** ↑4.5

Table 2: Main results on mathematical benchmarks. College, DM, Olympiad, Theorem denote the CollegeMath, DeepMind-Mathematics, OlympiadBench-Math, TheoremQA benchmarks respectively. We annotate the absolute accuracy change compared to the VRT baseline within the same base model. Bold means the best score within the respective base model. ICL, MetaMath, MMIQC, and VRT baselines are from our own runs, while other numbers are copied from the respective papers or reports. For WizardMath and Xwin-Math, we take the public model checkpoints and evaluate ourselves using their official CoT prompt. †: For Xwin-Math, we take the best public models that are based on Llama2 (Touvron et al., 2023), which is not a very fair comparison with others.

All Main Results

2 in-domain benchmarks

+

4 challenging

out-of-domain benchmarks

+

4 base models

Model	# Samples	In-Domain MATH	In-Domain GSM8K	Out-of-Domain College	Out-of-Domain DM	Out-of-Domain Olympiad	Out-of-Domain Theorem	Out-of-Domain AVG
GPT-4-Turbo (24-04-09)	–	73.4	94.5	–	–	–	48.4	–
GPT-4 (0314)	–	52.6	94.7	24.4	–	–	–	–
Claude-3-Opus	–	60.1	95.0	–	–	–	–	–
Gemini 1.5 Pro	–	67.7	–	–	–	–	–	–
70B General Base Model								
Llama2-70B-Xwin-Math-V1.1†	1.4M	52.5	90.2	33.1	58.0	16.3	14.9	44.2
Llama3-70B-ICL	–	44.0	80.1	33.5	51.7	10.8	27.0	41.2
Llama3-70B-MetaMath	0.40M	44.9	88.0	31.9	53.2	11.6	21.9	41.9
Llama3-70B-MMIQC	2.3M	49.4	89.3	37.6	60.4	15.3	23.5	45.9
Llama3-70B-VRT	0.59M	53.1	90.3	36.8	62.8	19.3	**28.6**	48.5
DART-Math-Llama3-70B (Uniform)	0.59M	54.9 ↑1.8	**90.4** ↑0.1	**38.5** ↑1.7	**64.1** ↑1.3	19.1 ↓0.2	27.4 ↓1.2	49.1 ↑0.6
DART-Math-Llama3-70B (Prop2Diff)	0.59M	**56.1** ↑3.0	89.6 ↓0.7	37.9 ↑1.1	**64.1** ↑1.3	**20.0** ↑0.7	28.2 ↓0.4	**49.3** ↑0.8
7B Math-Specialized Base Model								
DeepSeekMath-7B-ICL	–	35.5	64.2	34.7	45.2	9.3	23.5	35.4
DeepSeekMath-7B-Instruct	0.78M	46.9	82.7	37.1	52.2	14.2	28.1	43.5
DeepSeekMath-7B-MMIQC	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DeepSeekMath-7B-KPMath-Plus	1.6M	48.8	83.9	–	–	–	–	–
DeepSeekMath-7B-VRT	0.59M	53.0	**88.2**	**41.9**	60.2	19.1	27.2	48.3
DART-Math-DSMath-7B (Uniform)	0.59M	52.9 ↓0.1	**88.2**	40.1 ↓1.8	60.2	21.3 ↑2.2	**32.5** ↑5.3	49.2 ↑0.9
DART-Math-DSMath-7B (Prop2Diff)	0.59M	**53.6** ↑0.6	86.8 ↓1.4	40.7 ↓1.2	**61.6** ↑1.4	**21.7** ↑2.6	32.2 ↑5.0	**49.4** ↑1.1
7-8B General Base Model								
Llama2-7B-Xwin-Math-V1.1†	1.4M	45.5	84.9	27.6	43.0	10.5	15.0	37.8
Mistral-7B-ICL	–	16.5	45.9	17.9	23.5	3.7	14.2	20.3
Mistral-7B-WizardMath-V1.1 (RL)	–	32.3	80.4	23.1	38.4	7.7	16.6	33.1
Mistral-7B-MetaMath	0.40M	29.8	76.5	19.3	28.0	5.9	14.0	28.9
Mistral-7B-MMIQC	2.3M	37.4	75.4	28.5	38.0	9.4	16.2	34.2
Mistral-7B-MathScale	2.0M	35.2	74.8	21.8	–	–	–	–
Mistral-7B-KPMath-Plus	1.6M	**46.8**	82.1	–	–	–	–	–
Mistral-7B-VRT	0.59M	38.7	82.3	24.2	35.6	8.7	16.2	34.3
DART-Math-Mistral-7B (Uniform)	0.59M	43.5 ↑4.8	**82.6** ↑0.3	26.9 ↑2.7	42.0 ↑6.4	13.2 ↑4.5	16.4 ↑0.2	37.4 ↑3.1
DART-Math-Mistral-7B (Prop2Diff)	0.59M	45.5 ↑6.8	81.1 ↓1.2	**29.4** ↑5.2	**45.1** ↑9.5	**14.7** ↑6.0	**17.0** ↑0.8	**38.8** ↑4.5
Llama3-8B-ICL	–	21.2	51.0	19.9	27.4	4.2	19.8	23.9
Llama3-8B-MetaMath	0.40M	32.5	77.3	20.6	35.0	5.5	13.8	30.8
Llama3-8B-MMIQC	2.3M	39.5	77.6	**29.5**	41.0	9.6	16.2	35.6
Llama3-8B-VRT	0.59M	39.7	81.7	23.9	41.7	9.3	14.9	35.2
DART-Math-Llama3-8B (Uniform)	0.59M	45.3 ↑5.6	**82.5** ↑0.8	27.1 ↑3.2	**48.2** ↑6.5	13.6 ↑4.3	15.4 ↑0.5	38.7 ↑3.5
DART-Math-Llama3-8B (Prop2Diff)	0.59M	**46.6** ↑6.9	81.1 ↓0.6	28.8 ↑4.9	48.0 ↑6.3	**14.5** ↑5.2	**19.4** ↑4.5	**39.7** ↑4.5

Table 2: Main results on mathematical benchmarks. College, DM, Olympiad, Theorem denote the CollegeMath, DeepMind-Mathematics, OlympiadBench-Math, TheoremQA benchmarks respectively. We annotate the absolute accuracy change compared to the VRT baseline within the same base model. Bold means the best score within the respective base model. ICL, MetaMath, MMIQC, and VRT baselines are from our own runs, while other numbers are copied from the respective papers or reports. For WizardMath and Xwin-Math, we take the public model checkpoints and evaluate ourselves using their official CoT prompt. †: For Xwin-Math, we take the best public models that are based on Llama2 (Touvron et al., 2023), which is not a very fair comparison with others.

All Main Results

2 in-domain benchmarks

+

4 challenging
out-of-domain benchmarks

+

4 base models
of different kinds

Model	# Samples	In-Domain MATH	In-Domain GSM8K	Out-of-Domain College	Out-of-Domain DM	Out-of-Domain Olympiad	Out-of-Domain Theorem	Out-of-Domain AVG
GPT-4-Turbo (24-04-09)	–	73.4	94.5	–	–	–	48.4	–
GPT-4 (0314)	–	52.6	94.7	24.4	–	–	–	–
Claude-3-Opus	–	60.1	95.0	–	–	–	–	–
Gemini 1.5 Pro	–	67.7	–	–	–	–	–	–
70B General Base Model								
Llama2-70B-Xwin-Math-V1.1†	1.4M	52.5	90.2	33.1	58.0	16.3	14.9	44.2
Llama3-70B-ICL	–	44.0	80.1	33.5	51.7	10.8	27.0	41.2
Llama3-70B-MetaMath	0.40M	44.9	88.0	31.9	53.2	11.6	21.9	41.9
Llama3-70B-MMIQC	2.3M	49.4	89.3	37.6	60.4	15.3	23.5	45.9
Llama3-70B-VRT	0.59M	53.1	90.3	36.8	62.8	19.3	**28.6**	48.5
`DART-Math-Llama3-70B` (Uniform)	0.59M	54.9 ↑1.8	**90.4** ↑0.1	**38.5** ↑1.7	**64.1** ↑1.3	19.1 ↓0.2	27.4 ↓1.2	49.1 ↑0.6
`DART-Math-Llama3-70B` (Prop2Diff)	0.59M	**56.1** ↑3.0	89.6 ↓0.7	37.9 ↑1.1	**64.1** ↑1.3	**20.0** ↑0.7	28.2 ↓0.4	**49.3** ↑0.8
7B Math-Specialized Base Model								
DeepSeekMath-7B-ICL	–	35.5	64.2	34.7	45.2	9.3	23.5	35.4
DeepSeekMath-7B-Instruct	0.78M	46.9	82.7	37.1	52.2	14.2	28.1	43.5
DeepSeekMath-7B-MMIQC	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DeepSeekMath-7B-KPMath-Plus	1.6M	48.8	83.9	–	–	–	–	–
DeepSeekMath-7B-VRT	0.59M	53.0	**88.2**	**41.9**	60.2	19.1	27.2	48.3
`DART-Math-DSMath-7B` (Uniform)	0.59M	52.9 ↓0.1	**88.2**	40.1 ↓1.8	60.2	21.3 ↑2.2	**32.5** ↑5.3	49.2 ↑0.9
`DART-Math-DSMath-7B` (Prop2Diff)	0.59M	**53.6** ↑0.6	86.8 ↓1.4	40.7 ↓1.2	**61.6** ↑1.4	**21.7** ↑2.6	32.2 ↑5.0	**49.4** ↑1.1
7-8B General Base Model								
Llama2-7B-Xwin-Math-V1.1†	1.4M	45.5	84.9	27.6	43.0	10.5	15.0	37.8
Mistral-7B-ICL	–	16.5	45.9	17.9	23.5	3.7	14.2	20.3
Mistral-7B-WizardMath-V1.1 (RL)	–	32.3	80.4	23.1	38.4	7.7	16.6	33.1
Mistral-7B-MetaMath	0.40M	29.8	76.5	19.3	28.0	5.9	14.0	28.9
Mistral-7B-MMIQC	2.3M	37.4	75.4	28.5	38.0	9.4	16.2	34.2
Mistral-7B-MathScale	2.0M	35.2	74.8	21.8	–	–	–	–
Mistral-7B-KPMath-Plus	1.6M	**46.8**	82.1	–	–	–	–	–
Mistral-7B-VRT	0.59M	38.7	82.3	24.2	35.6	8.7	16.2	34.3
`DART-Math-Mistral-7B` (Uniform)	0.59M	43.5 ↑4.8	**82.6** ↑0.3	26.9 ↑2.7	42.0 ↑6.4	13.2 ↑4.5	16.4 ↑0.2	37.4 ↑3.1
`DART-Math-Mistral-7B` (Prop2Diff)	0.59M	45.5 ↑6.8	81.1 ↓1.2	**29.4** ↑5.2	**45.1** ↑9.5	**14.7** ↑6.0	**17.0** ↑0.8	**38.8** ↑4.5
Llama3-8B-ICL	–	21.2	51.0	19.9	27.4	4.2	19.8	23.9
Llama3-8B-MetaMath	0.40M	32.5	77.3	20.6	35.0	5.5	13.8	30.8
Llama3-8B-MMIQC	2.3M	39.5	77.6	**29.5**	41.0	9.6	16.2	35.6
Llama3-8B-VRT	0.59M	39.7	81.7	23.9	41.7	9.3	14.9	35.2
`DART-Math-Llama3-8B` (Uniform)	0.59M	45.3 ↑5.6	**82.5** ↑0.8	27.1 ↑3.2	**48.2** ↑6.5	13.6 ↑4.3	15.4 ↑0.5	38.7 ↑3.5
`DART-Math-Llama3-8B` (Prop2Diff)	0.59M	**46.6** ↑6.9	81.1 ↓0.6	28.8 ↑4.9	48.0 ↑6.3	**14.5** ↑5.2	**19.4** ↑4.5	**39.7** ↑4.5

Table 2: Main results on mathematical benchmarks. College, DM, Olympiad, Theorem denote the CollegeMath, DeepMind-Mathematics, OlympiadBench-Math, TheoremQA benchmarks respectively. We annotate the absolute accuracy change compared to the VRT baseline within the same base model. Bold means the best score within the respective base model. ICL, MetaMath, MMIQC, and VRT baselines are from our own runs, while other numbers are copied from the respective papers or reports. For WizardMath and Xwin-Math, we take the public model checkpoints and evaluate ourselves using their official CoT prompt. †: For Xwin-Math, we take the best public models that are based on Llama2 (Touvron et al., 2023), which is not a very fair comparison with others.

All Main Results

2 in-domain benchmarks

+

4 challenging
out-of-domain benchmarks

+

4 base models
of different kinds

↓

Model	# Samples	In-Domain MATH	In-Domain GSM8K	Out-of-Domain College	Out-of-Domain DM	Out-of-Domain Olympiad	Out-of-Domain Theorem	Out-of-Domain AVG
GPT-4-Turbo (24-04-09)	–	73.4	94.5	–	–	–	48.4	–
GPT-4 (0314)	–	52.6	94.7	24.4	–	–	–	–
Claude-3-Opus	–	60.1	95.0	–	–	–	–	–
Gemini 1.5 Pro	–	67.7	–	–	–	–	–	–
70B General Base Model								
Llama2-70B-Xwin-Math-V1.1†	1.4M	52.5	90.2	33.1	58.0	16.3	14.9	44.2
Llama3-70B-ICL	–	44.0	80.1	33.5	51.7	10.8	27.0	41.2
Llama3-70B-MetaMath	0.40M	44.9	88.0	31.9	53.2	11.6	21.9	41.9
Llama3-70B-MMIQC	2.3M	49.4	89.3	37.6	60.4	15.3	23.5	45.9
Llama3-70B-VRT	0.59M	53.1	90.3	36.8	62.8	19.3	**28.6**	48.5
DART-Math-Llama3-70B (Uniform)	0.59M	54.9 ↑1.8	**90.4** ↑0.1	**38.5** ↑1.7	**64.1** ↑1.3	19.1 ↓0.2	27.4 ↓1.2	49.1 ↑0.6
DART-Math-Llama3-70B (Prop2Diff)	0.59M	**56.1** ↑3.0	89.6 ↓0.7	37.9 ↑1.1	**64.1** ↑1.3	**20.0** ↑0.7	28.2 ↓0.4	**49.3** ↑0.8
7B Math-Specialized Base Model								
DeepSeekMath-7B-ICL	–	35.5	64.2	34.7	45.2	9.3	23.5	35.4
DeepSeekMath-7B-Instruct	0.78M	46.9	82.7	37.1	52.2	14.2	28.1	43.5
DeepSeekMath-7B-MMIQC	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DeepSeekMath-7B-KPMath-Plus	1.6M	48.8	83.9	–	–	–	–	–
DeepSeekMath-7B-VRT	0.59M	53.0	**88.2**	**41.9**	60.2	19.1	27.2	48.3
DART-Math-DSMath-7B (Uniform)	0.59M	52.9 ↓0.1	**88.2**	40.1 ↓1.8	60.2	21.3 ↑2.2	**32.5** ↑5.3	49.2 ↑0.9
DART-Math-DSMath-7B (Prop2Diff)	0.59M	**53.6** ↑0.6	86.8 ↓1.4	40.7 ↓1.2	**61.6** ↑1.4	**21.7** ↑2.6	32.2 ↑5.0	**49.4** ↑1.1
7-8B General Base Model								
Llama2-7B-Xwin-Math-V1.1†	1.4M	45.5	84.9	27.6	43.0	10.5	15.0	37.8
Mistral-7B-ICL	–	16.5	45.9	17.9	23.5	3.7	14.2	20.3
Mistral-7B-WizardMath-V1.1 (RL)	–	32.3	80.4	23.1	38.4	7.7	16.6	33.1
Mistral-7B-MetaMath	0.40M	29.8	76.5	19.3	28.0	5.9	14.0	28.9
Mistral-7B-MMIQC	2.3M	37.4	75.4	28.5	38.0	9.4	16.2	34.2
Mistral-7B-MathScale	2.0M	35.2	74.8	21.8	–	–	–	–
Mistral-7B-KPMath-Plus	1.6M	**46.8**	82.1	–	–	–	–	–
Mistral-7B-VRT	0.59M	38.7	82.3	24.2	35.6	8.7	16.2	34.3
DART-Math-Mistral-7B (Uniform)	0.59M	43.5 ↑4.8	**82.6** ↑0.3	26.9 ↑2.7	42.0 ↑6.4	13.2 ↑4.5	16.4 ↑0.2	37.4 ↑3.1
DART-Math-Mistral-7B (Prop2Diff)	0.59M	45.5 ↑6.8	81.1 ↓1.2	**29.4** ↑5.2	**45.1** ↑9.5	**14.7** ↑6.0	**17.0** ↑0.8	**38.8** ↑4.5
Llama3-8B-ICL	–	21.2	51.0	19.9	27.4	4.2	19.8	23.9
Llama3-8B-MetaMath	0.40M	32.5	77.3	20.6	35.0	5.5	13.8	30.8
Llama3-8B-MMIQC	2.3M	39.5	77.6	**29.5**	41.0	9.6	16.2	35.6
Llama3-8B-VRT	0.59M	39.7	81.7	23.9	41.7	9.3	14.9	35.2
DART-Math-Llama3-8B (Uniform)	0.59M	45.3 ↑5.6	**82.5** ↑0.8	27.1 ↑3.2	**48.2** ↑6.5	13.6 ↑4.3	15.4 ↑0.5	38.7 ↑3.5
DART-Math-Llama3-8B (Prop2Diff)	0.59M	**46.6** ↑6.9	81.1 ↓0.6	28.8 ↑4.9	48.0 ↑6.3	**14.5** ↑5.2	**19.4** ↑4.5	**39.7** ↑4.5

Table 2: Main results on mathematical benchmarks. College, DM, Olympiad, Theorem denote the CollegeMath, DeepMind-Mathematics, OlympiadBench-Math, TheoremQA benchmarks respectively. We annotate the absolute accuracy change compared to the VRT baseline within the same base model. Bold means the best score within the respective base model. ICL, MetaMath, MMIQC, and VRT baselines are from our own runs, while other numbers are copied from the respective papers or reports. For WizardMath and Xwin-Math, we take the public model checkpoints and evaluate ourselves using their official CoT prompt. †: For Xwin-Math, we take the best public models that are based on Llama2 (Touvron et al., 2023), which is not a very fair comparison with others.

All Main Results

2 in-domain benchmarks

+

4 challenging
out-of-domain benchmarks

+

4 base models
of different kinds

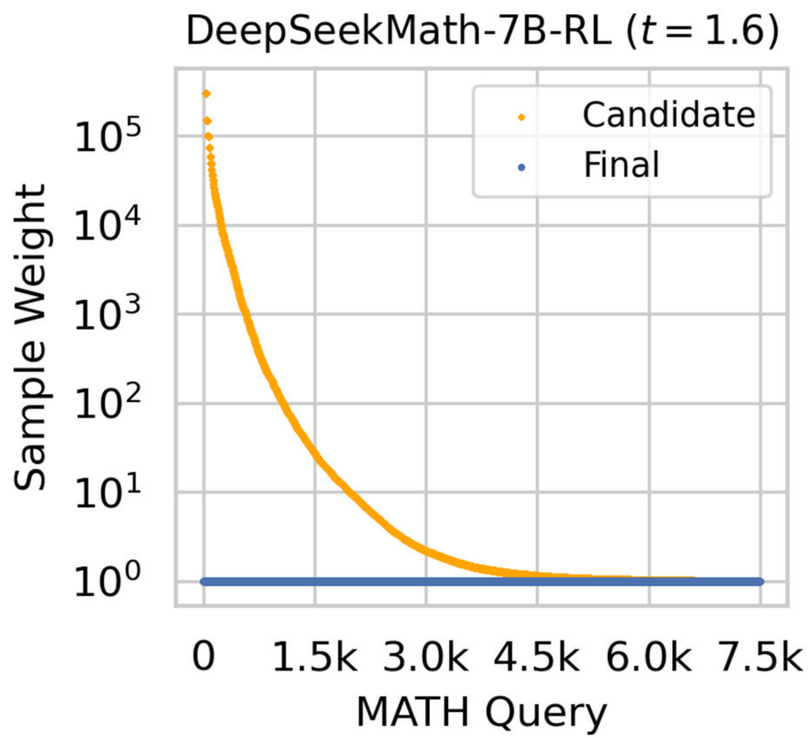
↓

DART is effective!

Model	# Samples	In-Domain MATH	GSM8K	Out-of-Domain College	DM	Olympiad	Theorem	AVG
GPT-4-Turbo (24-04-09)	–	73.4	94.5	–	–	–	48.4	–
GPT-4 (0314)	–	52.6	94.7	24.4	–	–	–	–
Claude-3-Opus	–	60.1	95.0	–	–	–	–	–
Gemini 1.5 Pro	–	67.7	–	–	–	–	–	–
70B General Base Model								
Llama2-70B-Xwin-Math-V1.1†	1.4M	52.5	90.2	33.1	58.0	16.3	14.9	44.2
Llama3-70B-ICL	–	44.0	80.1	33.5	51.7	10.8	27.0	41.2
Llama3-70B-MetaMath	0.40M	44.9	88.0	31.9	53.2	11.6	21.9	41.9
Llama3-70B-MMIQC	2.3M	49.4	89.3	37.6	60.4	15.3	23.5	45.9
Llama3-70B-VRT	0.59M	53.1	90.3	36.8	62.8	19.3	**28.6**	48.5
DART-Math-Llama3-70B (Uniform)	0.59M	54.9 ↑1.8	**90.4** ↑0.1	**38.5** ↑1.7	**64.1** ↑1.3	19.1 ↓0.2	27.4 ↓1.2	49.1 ↑0.6
DART-Math-Llama3-70B (Prop2Diff)	0.59M	**56.1** ↑3.0	89.6 ↓0.7	37.9 ↑1.1	**64.1** ↑1.3	**20.0** ↑0.7	28.2 ↓0.4	**49.3** ↑0.8
7B Math-Specialized Base Model								
DeepSeekMath-7B-ICL	–	35.5	64.2	34.7	45.2	9.3	23.5	35.4
DeepSeekMath-7B-Instruct	0.78M	46.9	82.7	37.1	52.2	14.2	28.1	43.5
DeepSeekMath-7B-MMIQC	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DeepSeekMath-7B-KPMath-Plus	1.6M	48.8	83.9	–	–	–	–	–
DeepSeekMath-7B-VRT	0.59M	53.0	**88.2**	**41.9**	60.2	19.1	27.2	48.3
DART-Math-DSMath-7B (Uniform)	0.59M	52.9 ↓0.1	**88.2**	40.1 ↓1.8	60.2	21.3 ↑2.2	**32.5** ↑5.3	49.2 ↑0.9
DART-Math-DSMath-7B (Prop2Diff)	0.59M	**53.6** ↑0.6	86.8 ↓1.4	40.7 ↓1.2	**61.6** ↑1.4	**21.7** ↑2.6	32.2 ↑5.0	**49.4** ↑1.1
7-8B General Base Model								
Llama2-7B-Xwin-Math-V1.1†	1.4M	45.5	84.9	27.6	43.0	10.5	15.0	37.8
Mistral-7B-ICL	–	16.5	45.9	17.9	23.5	3.7	14.2	20.3
Mistral-7B-WizardMath-V1.1 (RL)	–	32.3	80.4	23.1	38.4	7.7	16.6	33.1
Mistral-7B-MetaMath	0.40M	29.8	76.5	19.3	28.0	5.9	14.0	28.9
Mistral-7B-MMIQC	2.3M	37.4	75.4	28.5	38.0	9.4	16.2	34.2
Mistral-7B-MathScale	2.0M	35.2	74.8	21.8	–	–	–	–
Mistral-7B-KPMath-Plus	1.6M	**46.8**	82.1	–	–	–	–	–
Mistral-7B-VRT	0.59M	38.7	82.3	24.2	35.6	8.7	16.2	34.3
DART-Math-Mistral-7B (Uniform)	0.59M	43.5 ↑4.8	**82.6** ↑0.3	26.9 ↑2.7	42.0 ↑6.4	13.2 ↑4.5	16.4 ↑0.2	37.4 ↑3.1
DART-Math-Mistral-7B (Prop2Diff)	0.59M	45.5 ↑6.8	81.1 ↓1.2	**29.4** ↑5.2	**45.1** ↑9.5	**14.7** ↑6.0	**17.0** ↑0.8	**38.8** ↑4.5
Llama3-8B-ICL	–	21.2	51.0	19.9	27.4	4.2	19.8	23.9
Llama3-8B-MetaMath	0.40M	32.5	77.3	20.6	35.0	5.5	13.8	30.8
Llama3-8B-MMIQC	2.3M	39.5	77.6	**29.5**	41.0	9.6	16.2	35.6
Llama3-8B-VRT	0.59M	39.7	81.7	23.9	41.7	9.3	14.9	35.2
DART-Math-Llama3-8B (Uniform)	0.59M	45.3 ↑5.6	**82.5** ↑0.8	27.1 ↑3.2	**48.2** ↑6.5	13.6 ↑4.3	15.4 ↑0.5	38.7 ↑3.5
DART-Math-Llama3-8B (Prop2Diff)	0.59M	**46.6** ↑6.9	81.1 ↓0.6	28.8 ↑4.9	48.0 ↑6.3	**14.5** ↑5.2	**19.4** ↑4.5	**39.7** ↑4.5

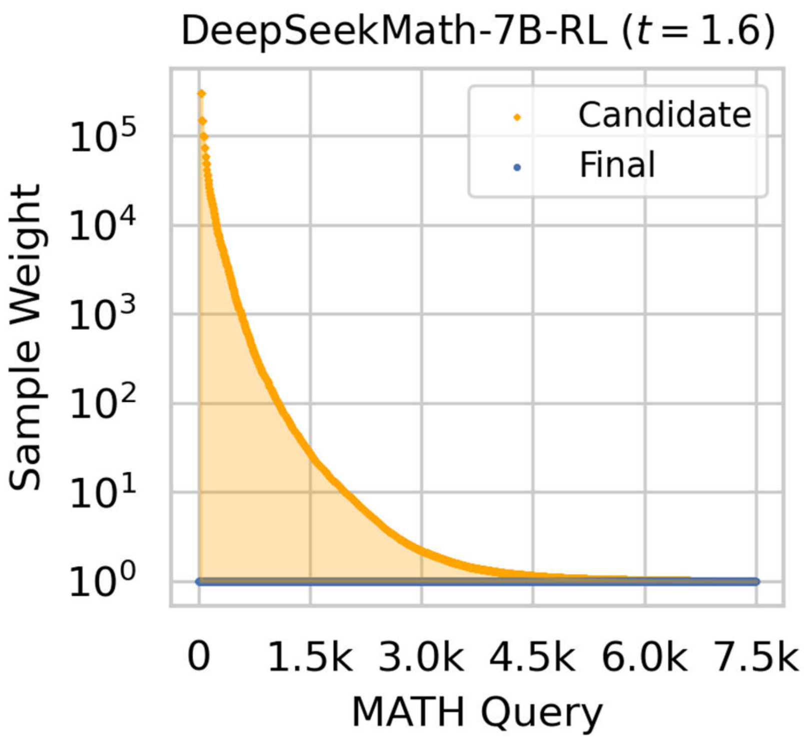
Table 2: Main results on mathematical benchmarks. College, DM, Olympiad, Theorem denote the CollegeMath, DeepMind-Mathematics, OlympiadBench-Math, TheoremQA benchmarks respectively. We annotate the absolute accuracy change compared to the VRT baseline within the same base model. Bold means the best score within the respective base model. ICL, MetaMath, MMIQC, and VRT baselines are from our own runs, while other numbers are copied from the respective papers or reports. For WizardMath and Xwin-Math, we take the public model checkpoints and evaluate ourselves using their official CoT prompt. †: For Xwin-Math, we take the best public models that are based on Llama2 (Touvron et al., 2023), which is not a very fair comparison with others.

Wait! What about the Synthesis Cost?



Wait! What about the Synthesis Cost?

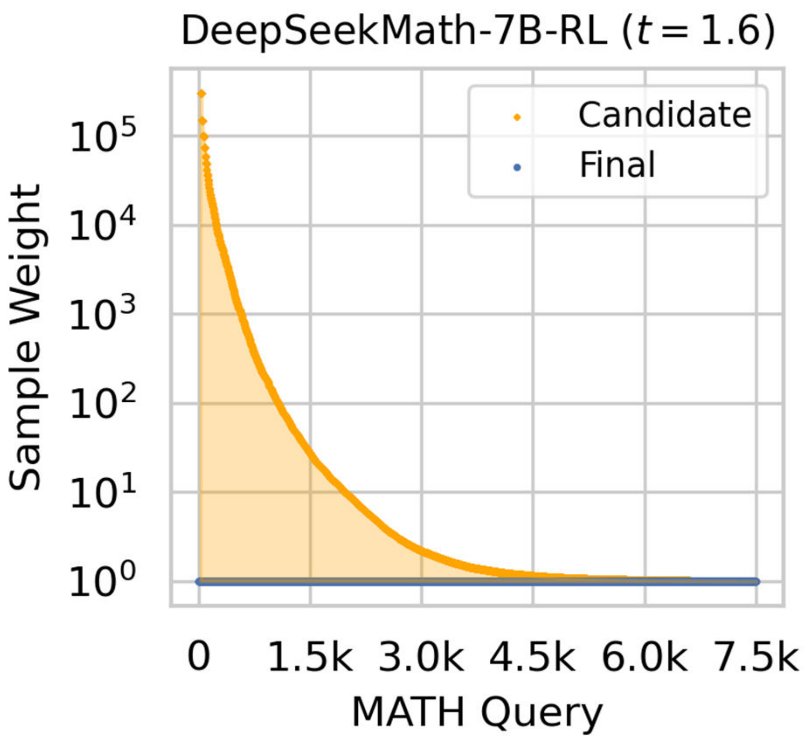
The synthesis cost can be huge?



Wait! What about the Synthesis Cost?

The synthesis cost can be huge?

But the final cost should be acceptable!

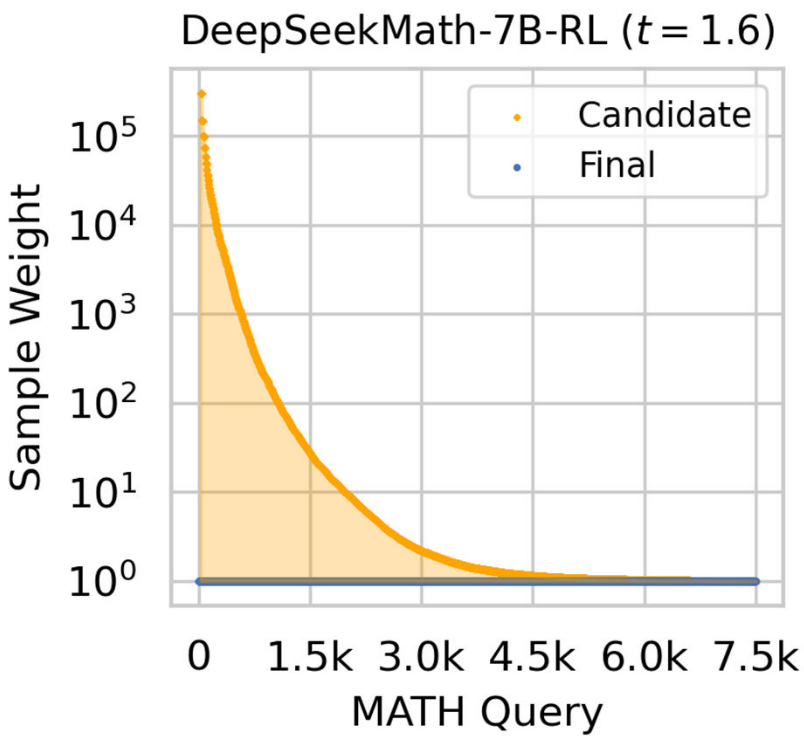


Wait! What about the Synthesis Cost?

The synthesis cost can be huge?

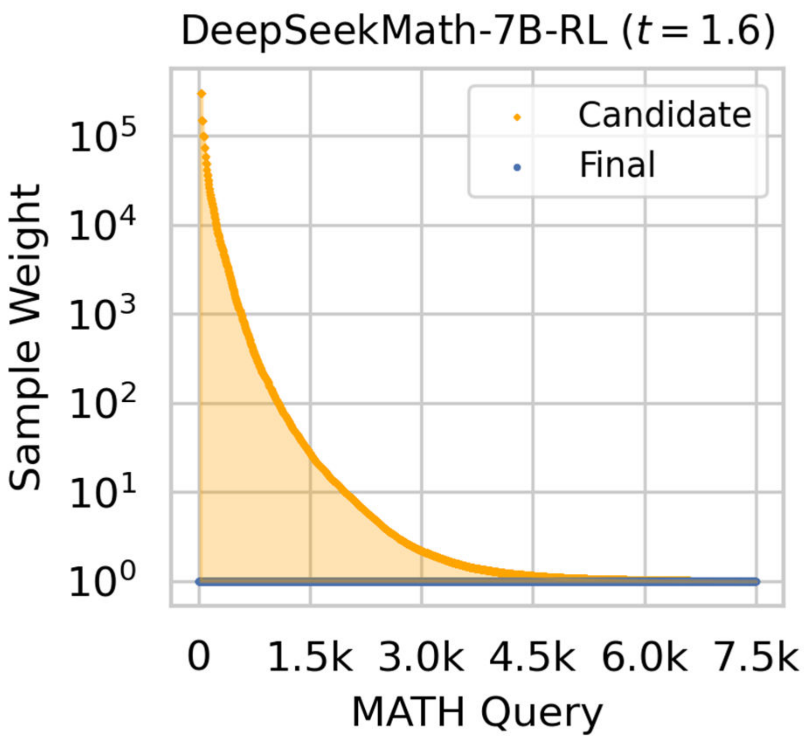
But the final cost should be acceptable!

1. The synthesis cost is **one-time** and **amortizable** by numerous training runs, e.g.,



Wait! What about the Synthesis Cost?

The synthesis cost can be huge?



But the final cost should be acceptable!

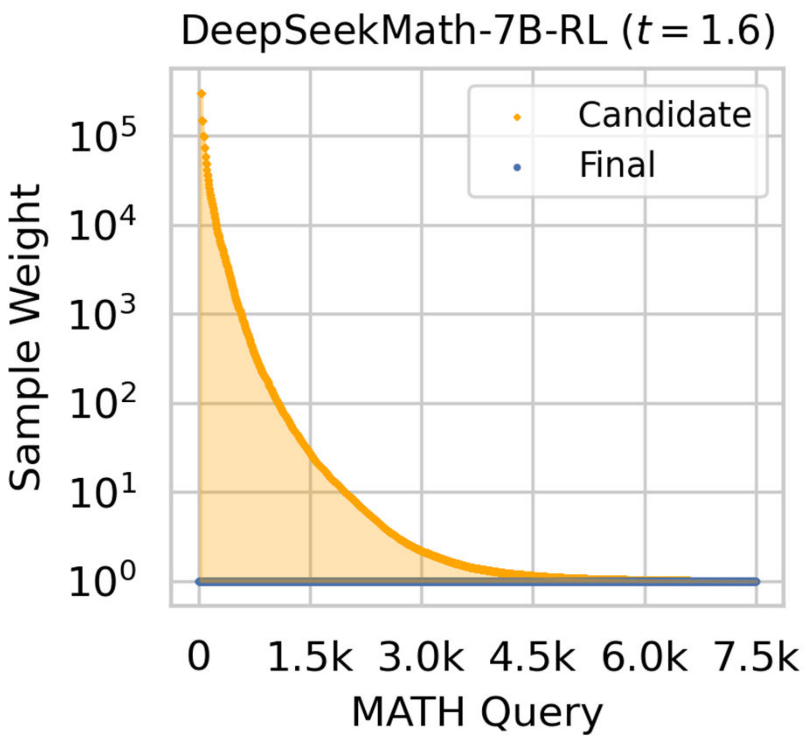
1. The synthesis cost is **one-time** and **amortizable** by numerous training runs, e.g.,
   a. Data mixture into new dataset

Wait! What about the Synthesis Cost?

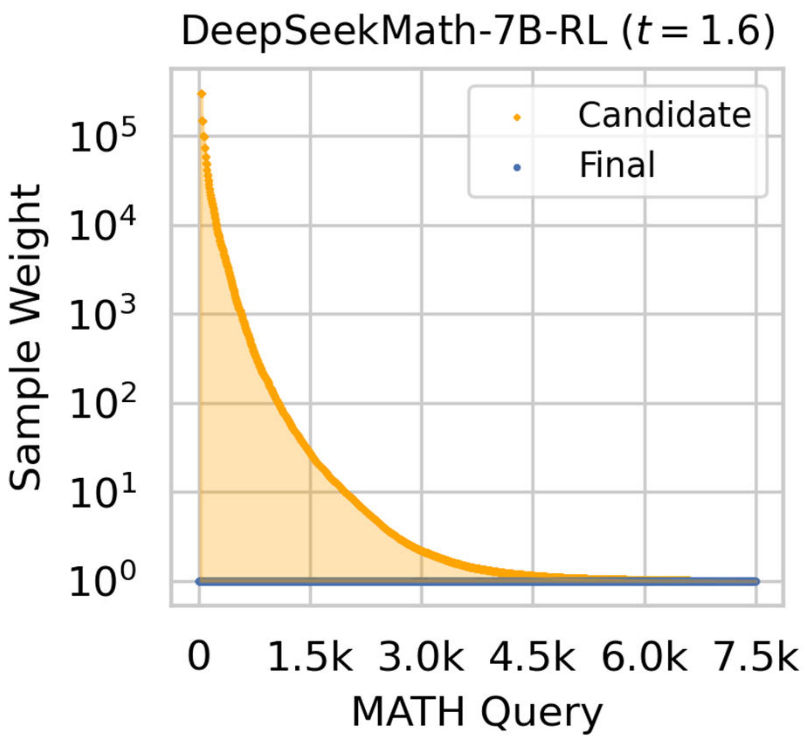
The synthesis cost can be huge?

But the final cost should be acceptable!

1. The synthesis cost is **one-time** and **amortizable** by numerous training runs, e.g.,
   a. Data mixture into new dataset
   b. Hyperparameter search

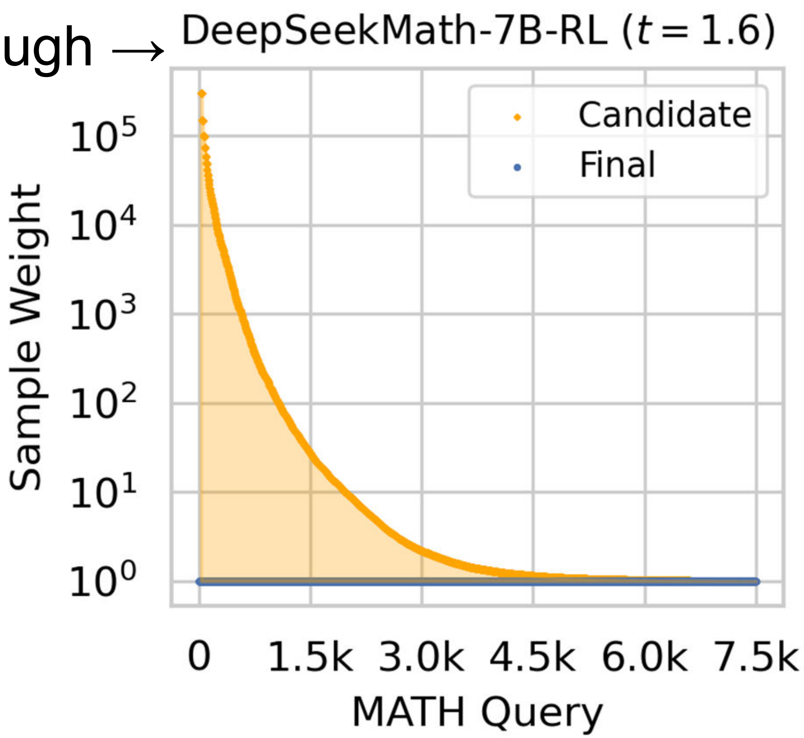


The final cost should be acceptable!



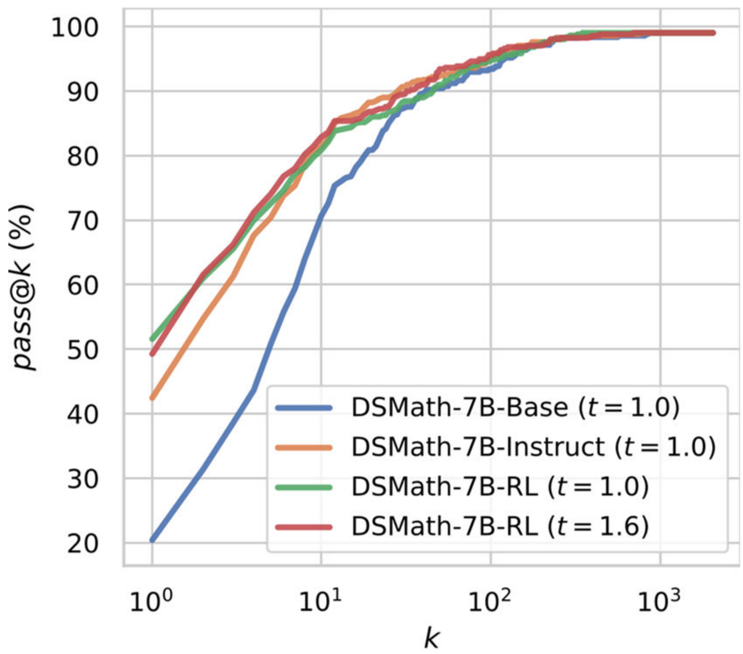
The final cost should be acceptable!

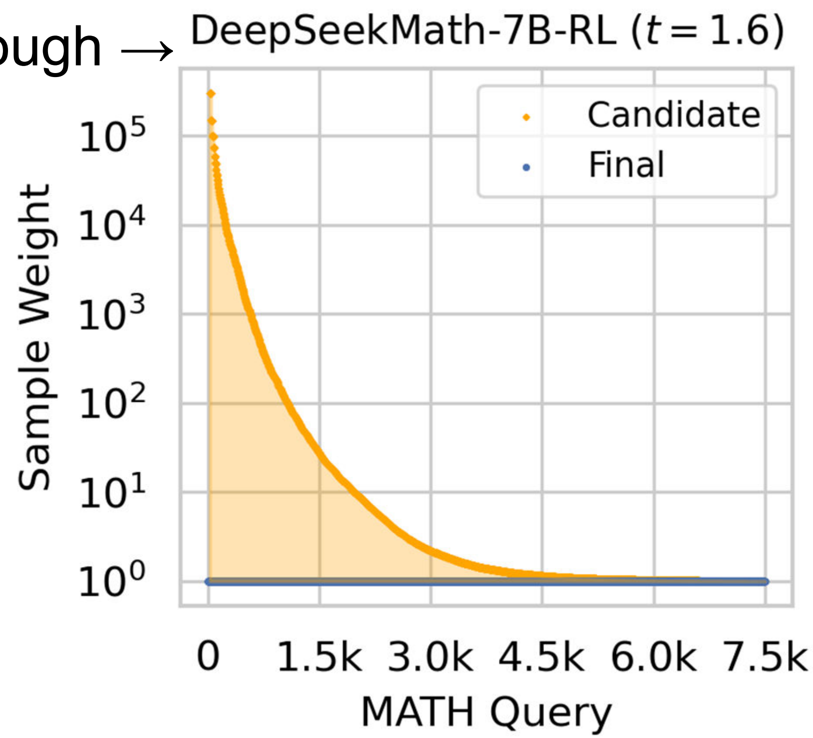
2. **Small** open-weight model is enough →



The final cost should be acceptable!

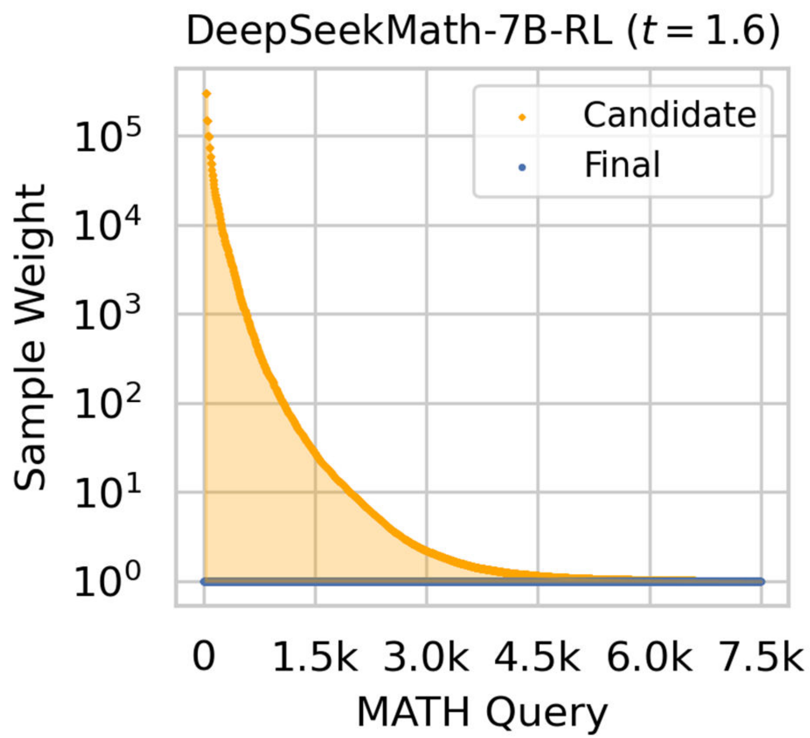
2. **Small** open-weight model is enough →





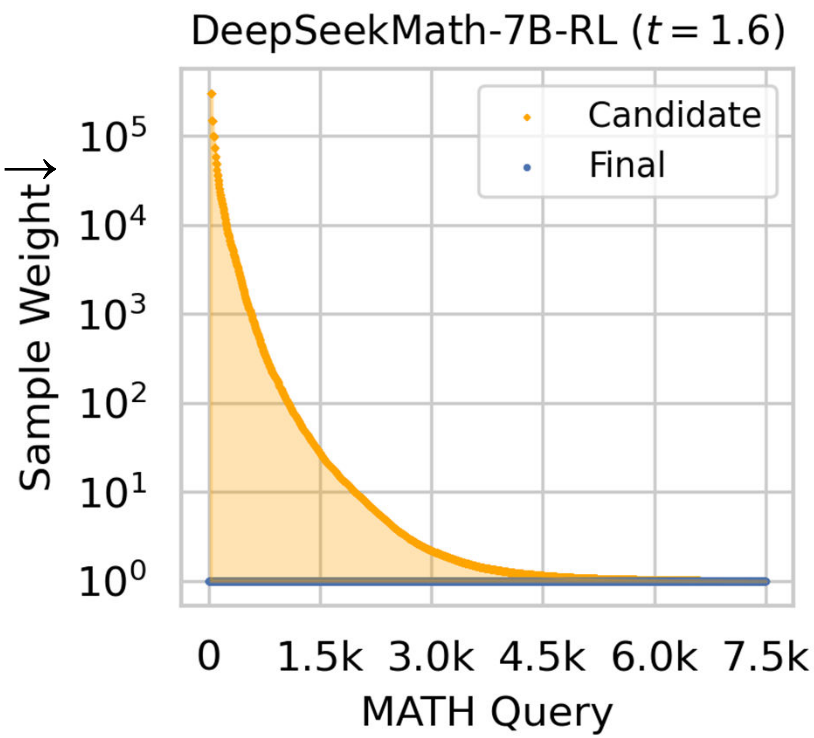
E.g., *DeepSeekMath-7B* models can achieve **almost perfect coverage** on the *MATH500* test set.
→ Expensive proprietary models like *ChatGPT* widely used before are not necessary.

The final cost should be acceptable!



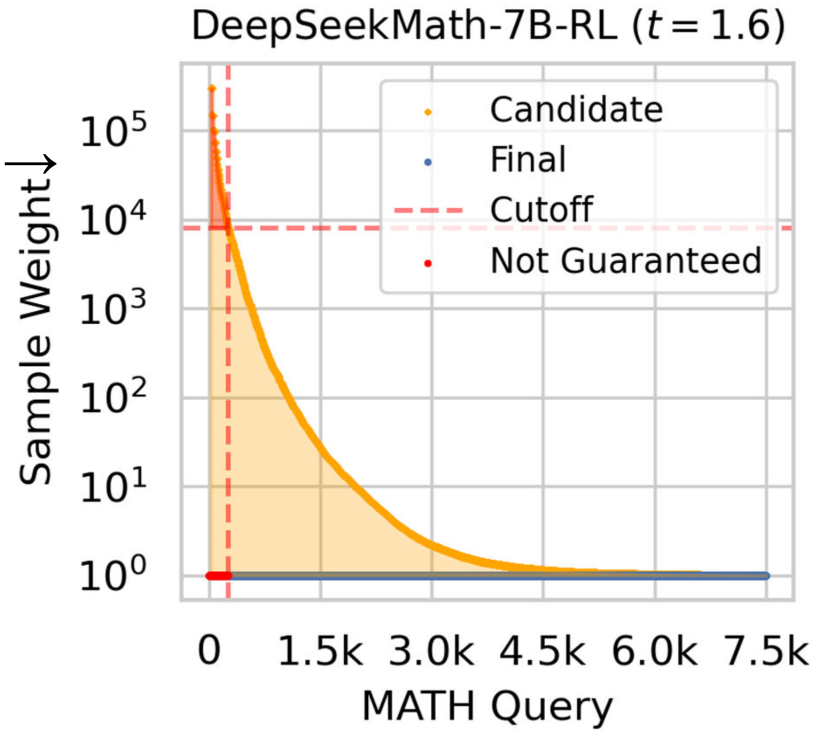
The final cost should be acceptable!

3. We can **cut off extremely-hard** part →



The final cost should be acceptable!

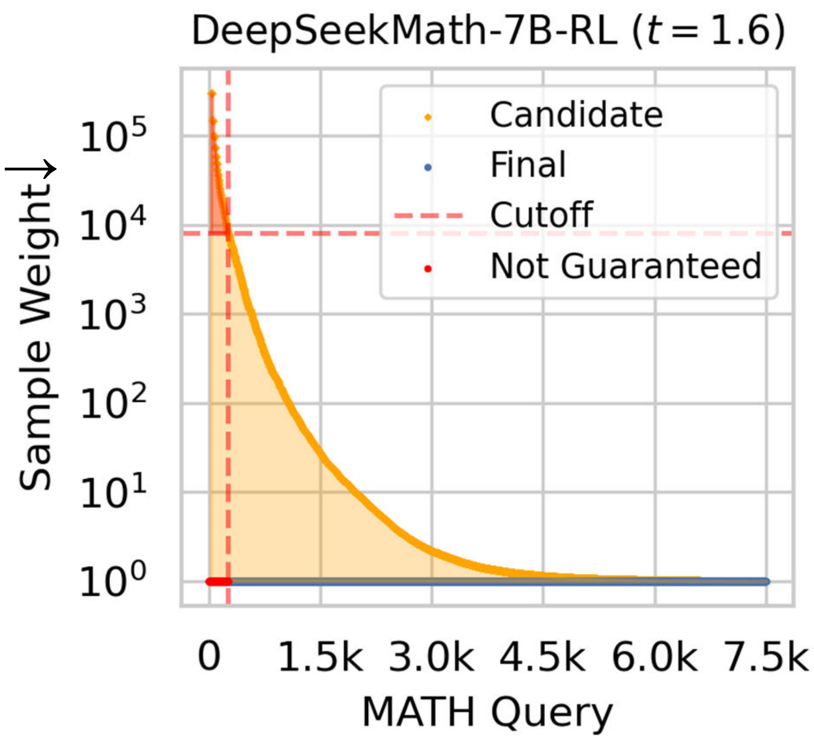
3. We can **cut off extremely-hard** part →



The final cost should be acceptable!

3. We can **cut off extremely-hard** part →

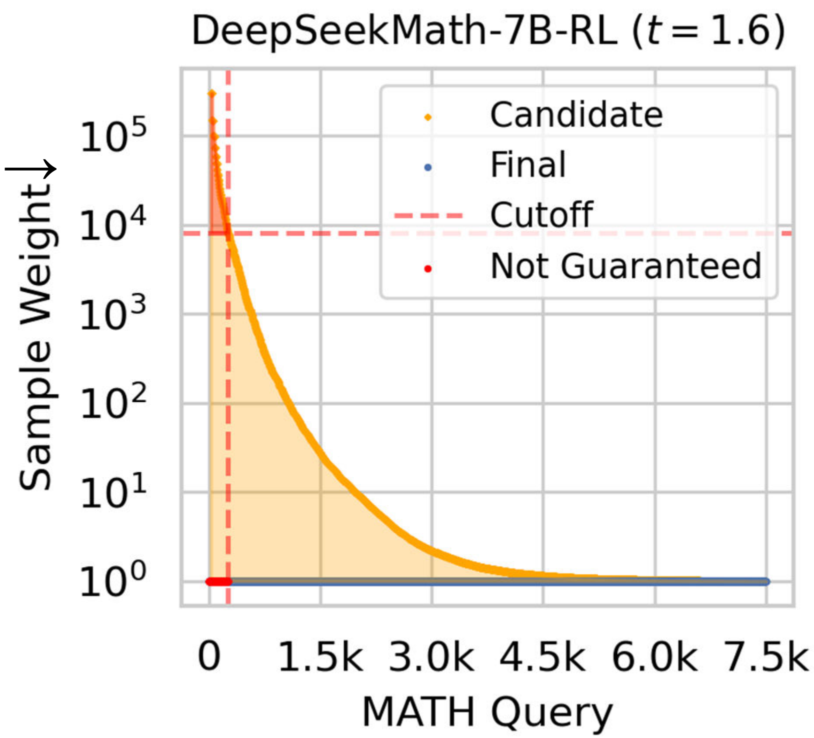
The save is significant (note the **log** scale),



The final cost should be acceptable!

3. We can **cut off extremely-hard** part →

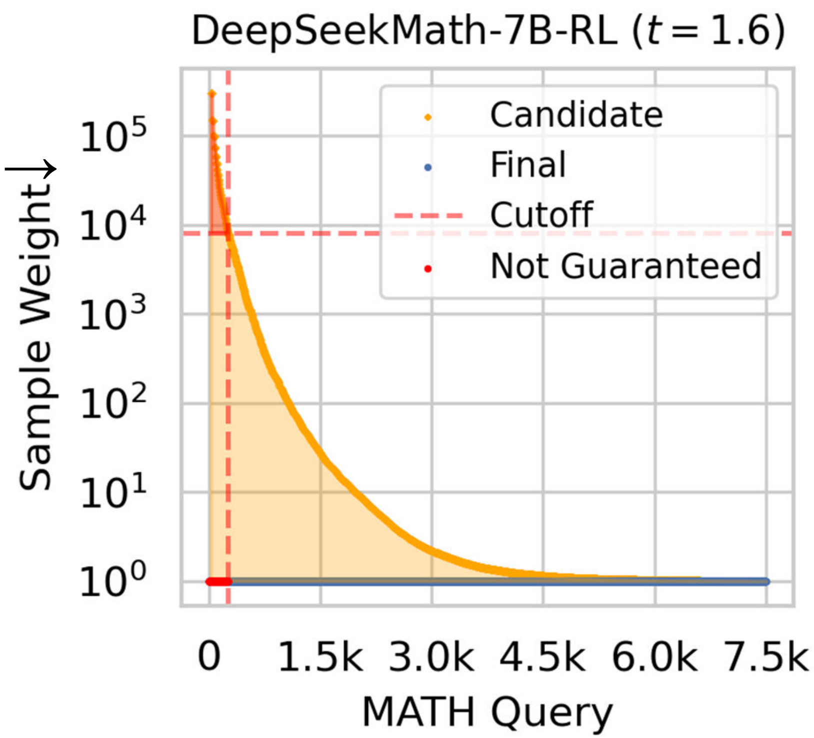
The save is significant (note the **log** scale), while **the loss should be minimal**:



The final cost should be acceptable!

3. We can **cut off extremely-hard** part →

The save is significant (note the **log** scale), while **the loss should be minimal**:

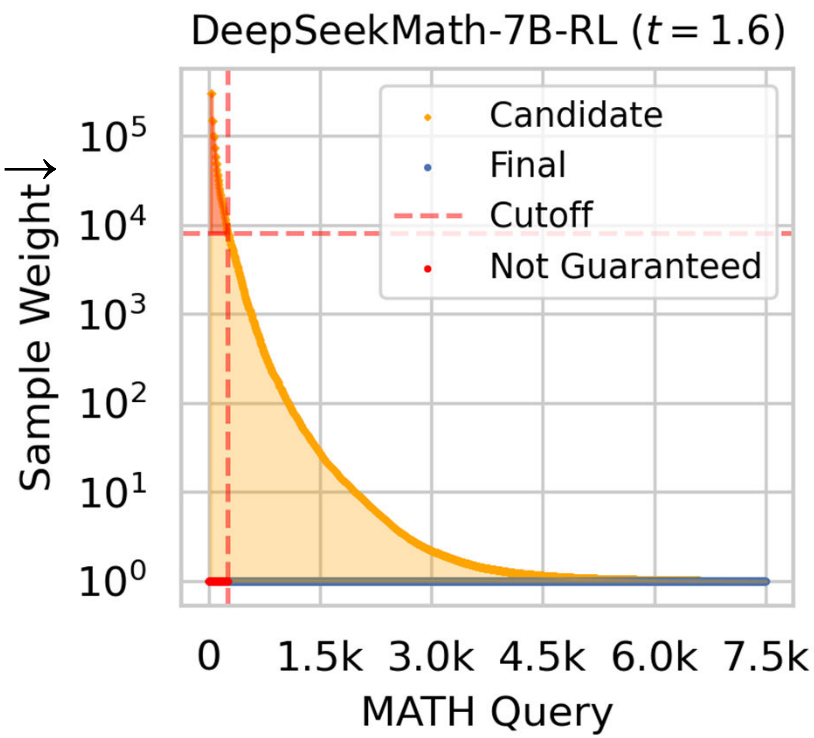


The final cost should be acceptable!

3. We can **cut off extremely-hard** part →

The save is significant (note the **log** scale), while **the loss should be minimal**:

1. The dropped portion is small

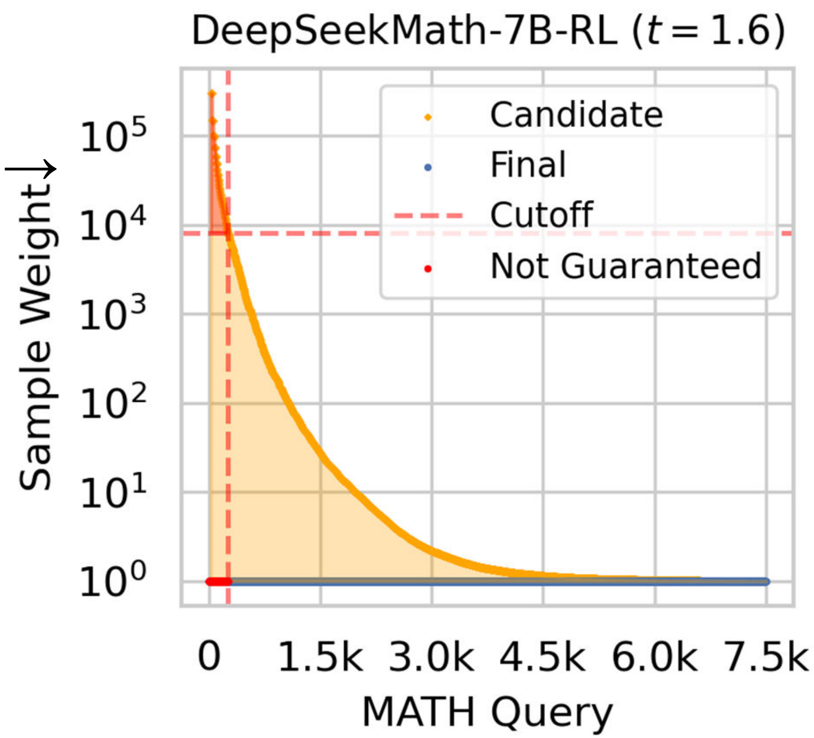


The final cost should be acceptable!

3. We can **cut off extremely-hard** part →

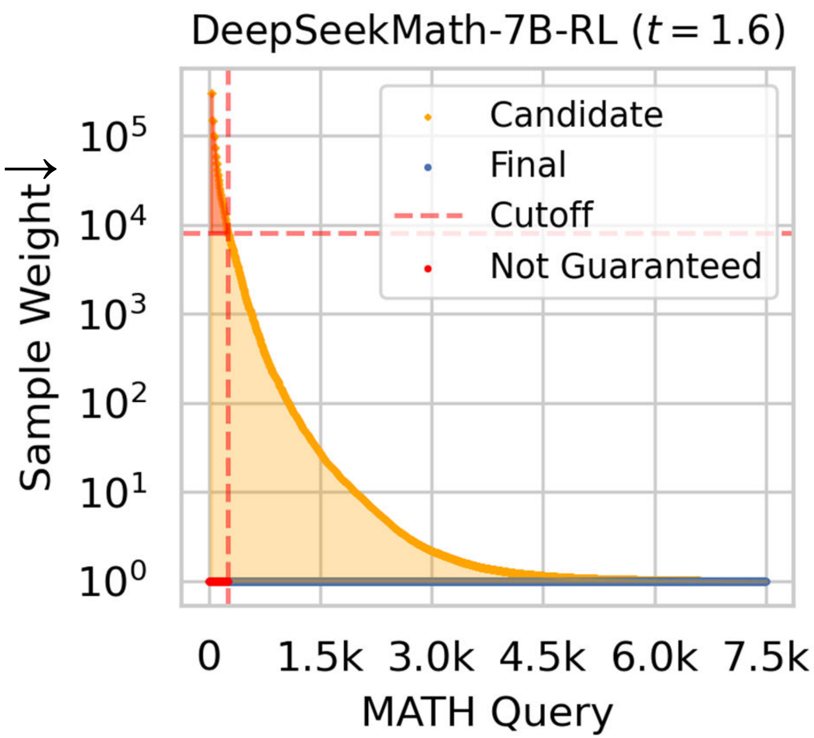
The save is significant (note the **log** scale), while **the loss should be minimal**:

1. The dropped portion is small



The final cost should be acceptable!

3. We can **cut off extremely-hard** part →



The save is significant (note the **log** scale), while **the loss should be minimal**:

1. The dropped portion is small

1. The hardest part is usually noisy

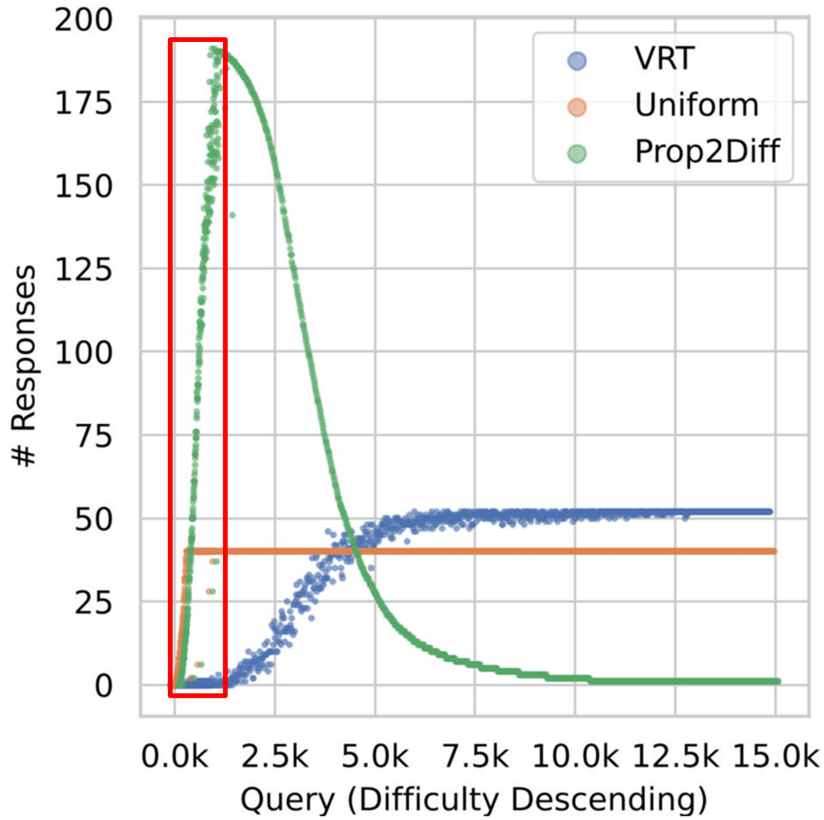
The final cost should be acceptable!

3. We can **cut off extremely-hard** part

The save is significant (note the **log** scale), while **the loss should be minimal**:

1. The dropped portion is small

1. The hardest part is usually noisy



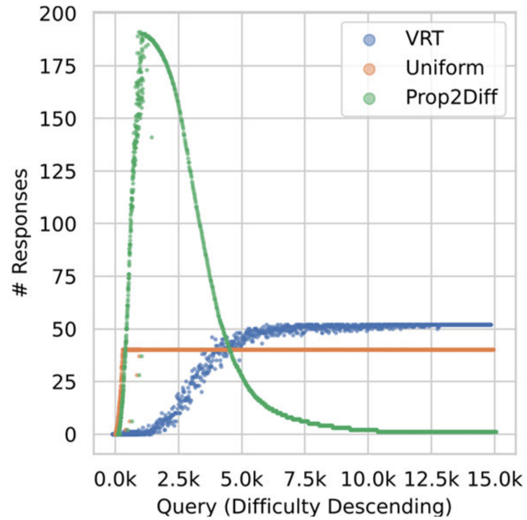
Summary

Summary

1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.

Summary

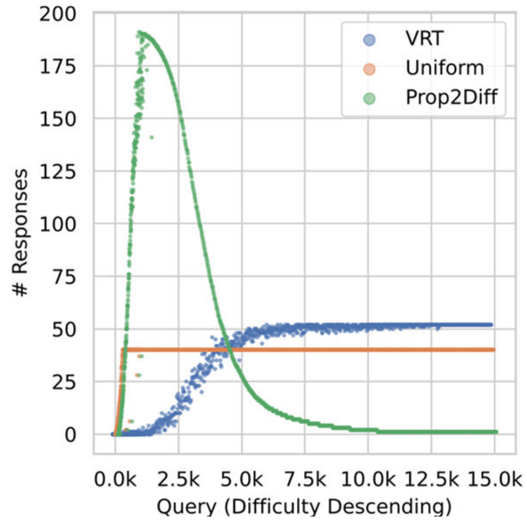
1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.



Summary

1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.

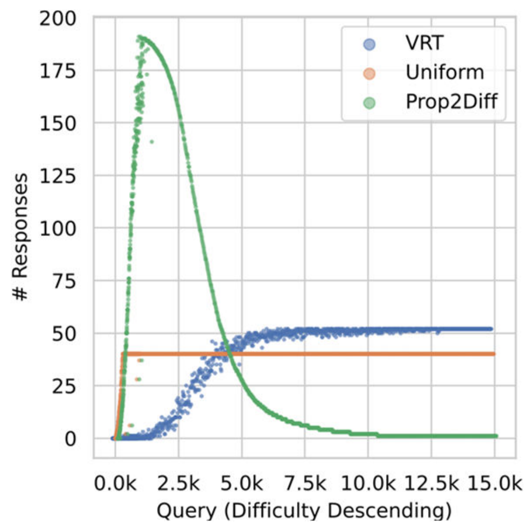
2. Observation: Severe **biases towards easier queries** exist in these synthetic datasets.



Summary

1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.

2. Observation: Severe **biases towards easier queries** exist in these synthetic datasets.

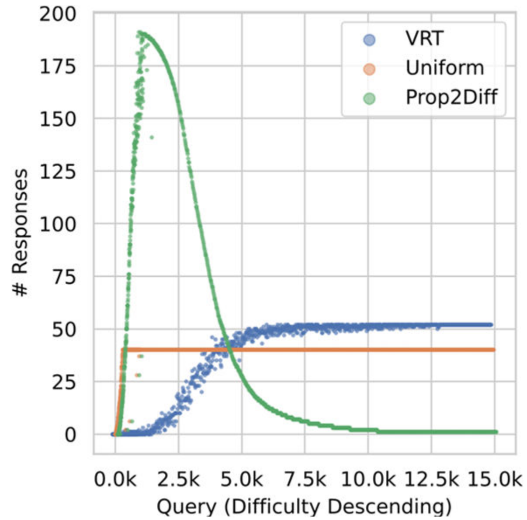


Summary

1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.

2. Observation: Severe **biases towards easier queries** exist in these synthetic datasets.

3. Attribution: Such biases arise from **Vanilla Rejection Tuning**,

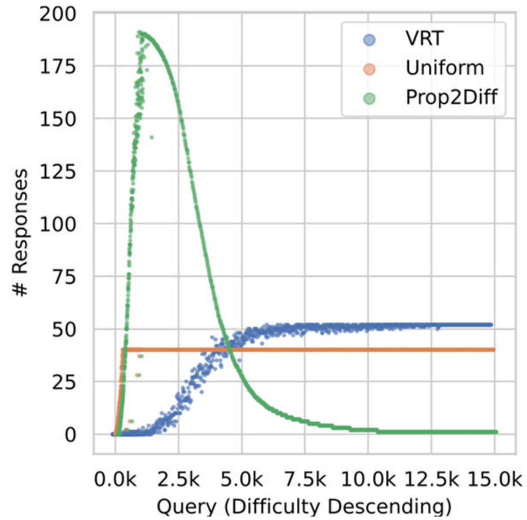


Summary

1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.

2. Observation: Severe **biases towards easier queries** exist in these synthetic datasets.

3. Attribution: Such biases arise from **Vanilla Rejection Tuning**, which only controls the **candidate** distribution in rejection sampling.

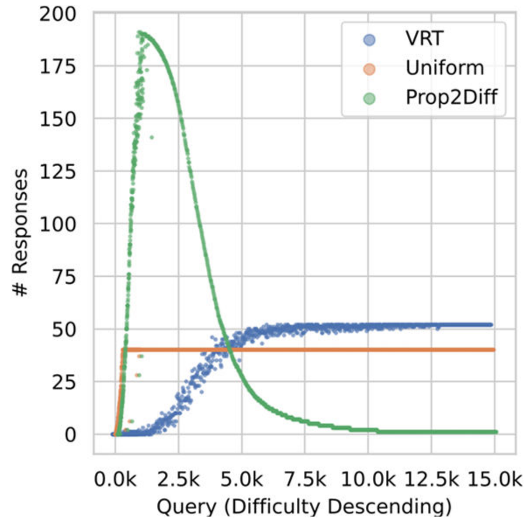


Summary

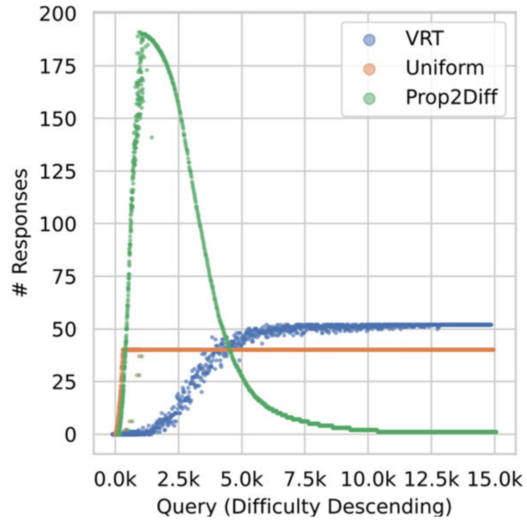
1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.

2. Observation: Severe **biases towards easier queries** exist in these synthetic datasets.

3. Attribution: Such biases arise from **Vanilla Rejection Tuning**, which only controls the **candidate** distribution in rejection sampling.



Summary



1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.

2. Observation: Severe **biases towards easier queries** exist in these synthetic datasets.

3. Attribution: Such biases arise from **Vanilla Rejection Tuning**, which only controls the **candidate** distribution in rejection sampling.

4. Method: **Difficulty-Aware Rejection Tuning (*DART*)** eliminates such biases by explicitly controlling the **final** distribution for training.

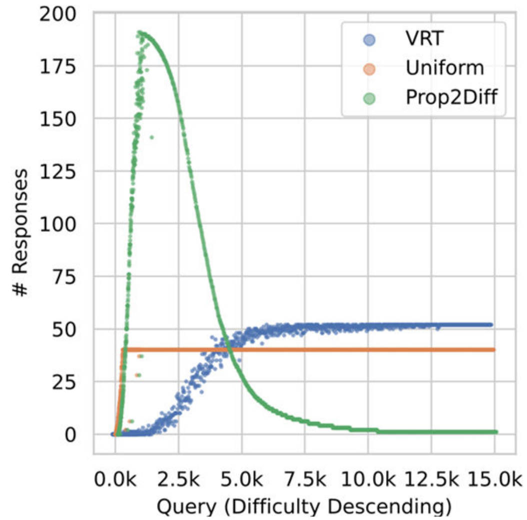
Summary

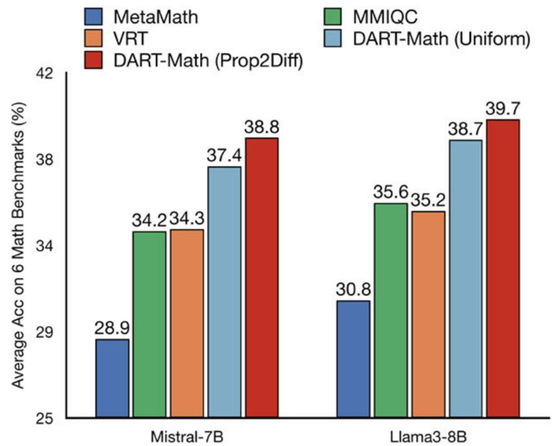
1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.

2. Observation: Severe **biases towards easier queries** exist in these synthetic datasets.

3. Attribution: Such biases arise from **Vanilla Rejection Tuning**, which only controls the **candidate** distribution in rejection sampling.

4. Method: **Difficulty-Aware Rejection Tuning (*DART*)** eliminates such biases by explicitly controlling the **final** distribution for training.





Summary

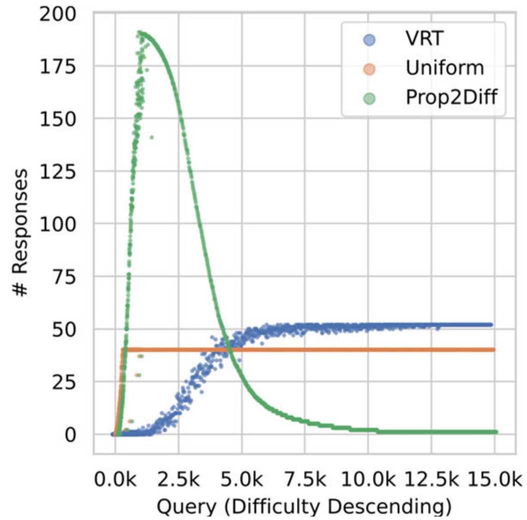
1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.

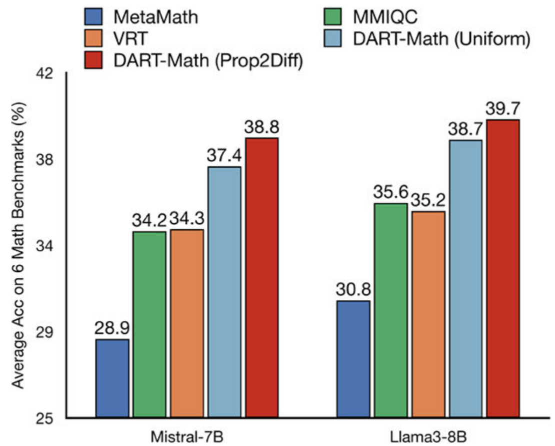
2. Observation: Severe **biases towards easier queries** exist in these synthetic datasets.

3. Attribution: Such biases arise from **Vanilla Rejection Tuning**, which only controls the **candidate** distribution in rejection sampling.

4. Method: **Difficulty-Aware Rejection Tuning (*DART*)** eliminates such biases by explicitly controlling the **final** distribution for training.

5. Results: *DART-Math* produce new **SotA and data-efficient** public instruction tuning datasets for mathematical problem-solving,





Summary

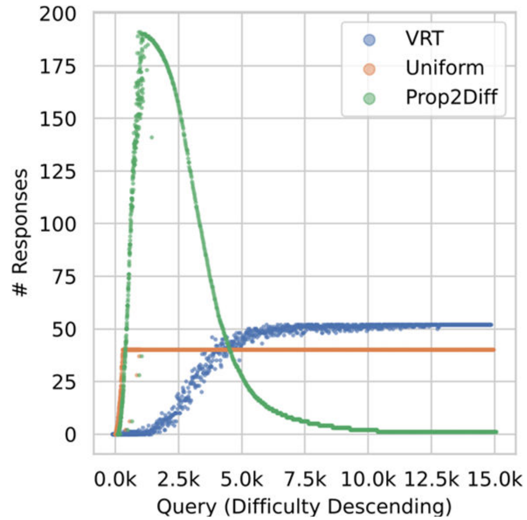
1. Background: **Synthetic instruction tuning datasets** achieve SotA for mathematical problem-solving.

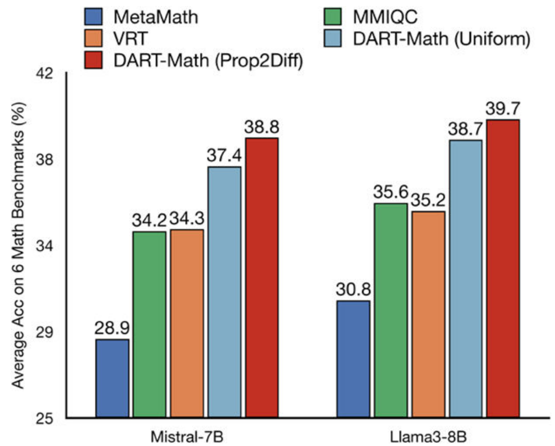
2. Observation: Severe **biases towards easier queries** exist in these synthetic datasets.

3. Attribution: Such biases arise from **Vanilla Rejection Tuning**, which only controls the **candidate** distribution in rejection sampling.

4. Method: **Difficulty-Aware Rejection Tuning (*DART*)** eliminates such biases by explicitly controlling the **final** distribution for training.

5. Results: *DART-Math* produce new **SotA and data-efficient** public instruction tuning datasets for mathematical problem-solving, without reliance on expensive proprietary models like *ChatGPT*.





Insights

Insights

1. Interpretability: **Enough difficult data are critical** for complex reasoning tasks like mathematical problem-solving, which needs **explicit control** to ensure the balance.

Insights

1. Interpretability: **Enough difficult data are critical** for complex reasoning tasks like mathematical problem-solving, which needs **explicit control** to ensure the balance.

Insights

1. Interpretability: **Enough difficult data are critical** for complex reasoning tasks like mathematical problem-solving, which needs **explicit control** to ensure the balance.

1. Scalability: **Scaling up training compute further and more cost-efficiently** in data synthesis by allocating budget **adaptive to data difficulty**.

Thanks!

📝 Paper@arXiv | 🤗 Datasets&Models@HF | 🐱 Code@GitHub | 🏆 Published@NeurIPS 2024

🐦 Thread@X(Twitter) | 🐶 中文博客@知乎 | 📊 Leaderboard@PapersWithCode | 📑 BibTeX