香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

COMP 4901B

Large Language Models

Data Synthesis

Junxian He

Nov 5, 2025

HW3 self-train ST

no training

Recap: Reasoning Model

Can you explain RLVR for deep reasoning models

3:07 PM

DeepSeek-R1

Thinking...

> We are going to explain RLVR (Reinforced Language Models with Vision for Reward) for deep reasoning models.
> However, note that RLVR is a specific method introduced in a paper. Let me recall the details.
>
> The paper "Reinforced Language Models with Vision for Reward" (RLVR) was introduced by researchers from Google DeepMind and UC Berkeley.
> The main idea is to use visual information to provide rewards for training language models to perform complex, multi-step reasoning tasks.
>
> Why is this needed?
>
> - Traditional language models are trained on large text corpora but may struggle with tasks that require grounded reasoning (like solving puzzles that involve visual or spatial understanding).
> - RLVR proposes to use reinforcement learning (RL) where the reward signal comes from a vision model that checks the correctness of the reasoning steps by interpreting images.
>
> Applications:
>
> - The paper demonstrates applications in tasks such as solving Raven's Progressive Matrices, geometric reasoning, and other visual puzzles.
>
> Challenges:
>
> - Training such a model is complex because it requires generating images and text together.
> - The vision model must be robust to provide meaningful rewards.
>
> This is a high-level explanation. For more details, one should refer to the original paper:
> "Reinforced Language Models with Vision for Reward" by Zheng et al. (2023)
>
> Let me know if you have any specific questions.

RLVR (Reinforced Language Models with Vision for Reward) is a framework designed to enhance **deep reasoning models** by integrating **reinforcement learning (RL)** with **visual supervision**. It addresses challenges in tasks requiring multi-step reasoning, particularly those involving visual inputs (e.g., puzzles, diagrams, or multimodal queries). Below is a breakdown:

Core Idea

Traditional language models (LMs) often struggle with **complex, multi-step reasoning** due to:

- Lack of grounded feedback during training.
- Difficulty verifying intermediate reasoning steps.
- Sparse rewards (only final outcomes are evaluated).

RLVR tackles this by:

1. **Generating step-by-step reasoning traces** (e.g., chain-of-thought).
2. **Using vision models** to provide **dense rewards** for each step based on visual alignment.
3. **Fine-tuning the LM** using reinforcement learning to maximize cumulative rewards.

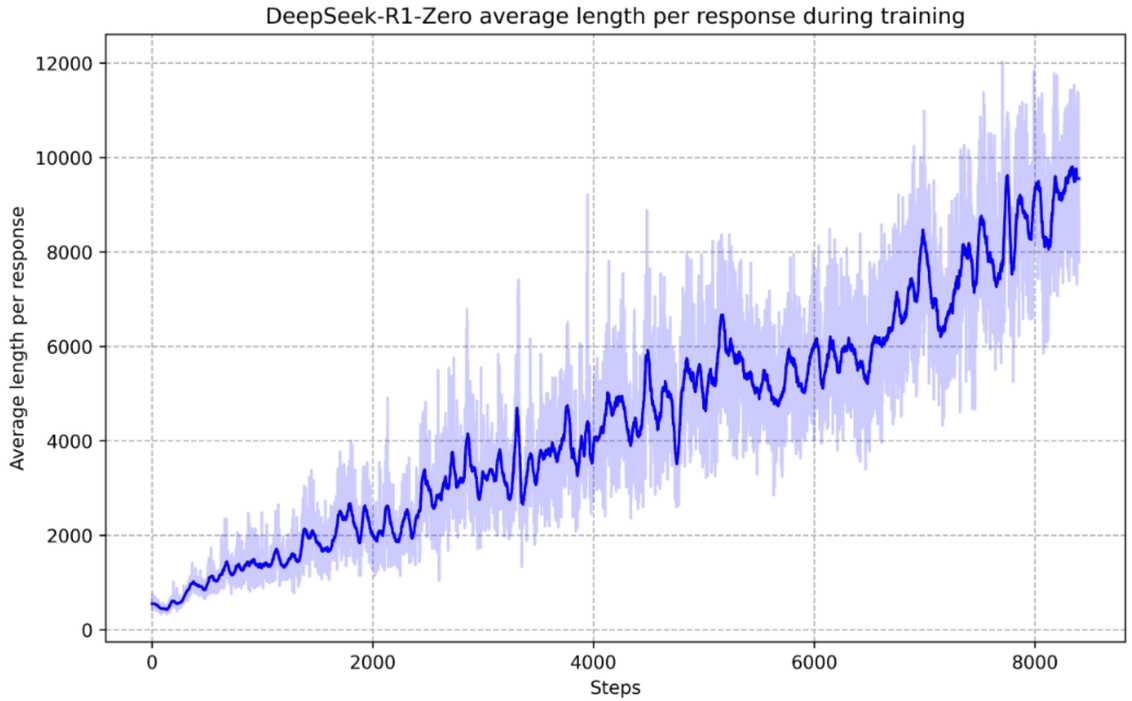
Recap: Long CoT Patterns Emerging in RLVR Training

DeepSeek-R1 and Kimi-k1.5 choose the extremely simple recipe — Reinforcement Learning

Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

Recap: Long CoT Patterns Emerging in RLVR Training

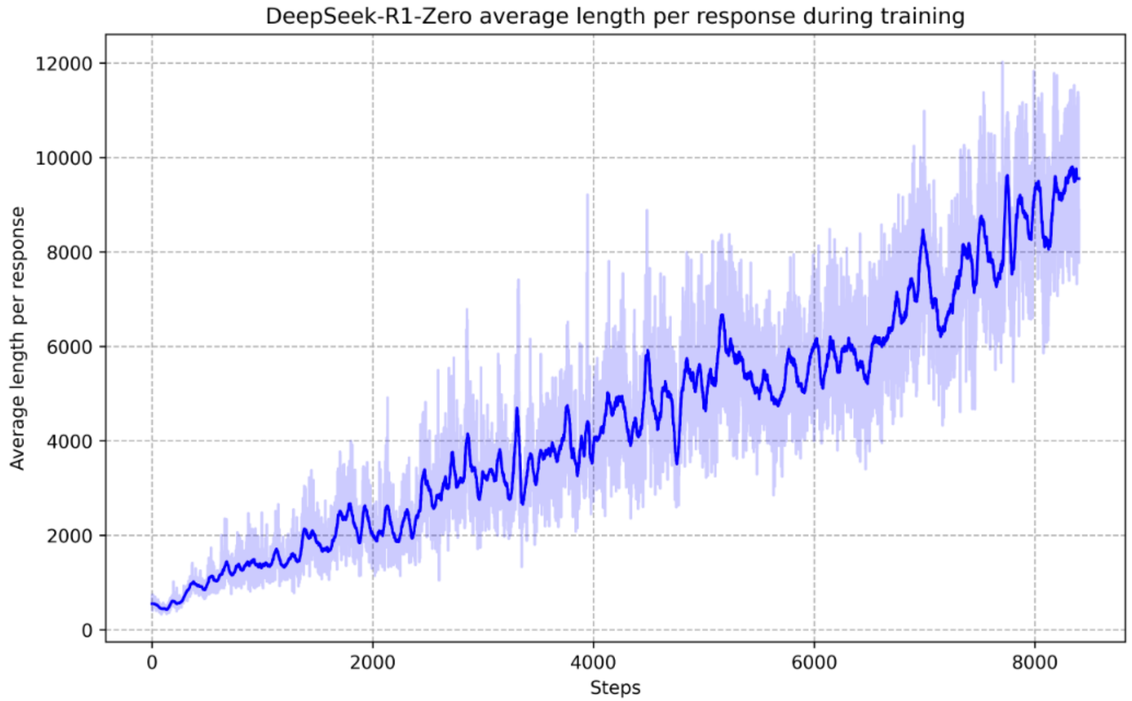
DeepSeek-R1 and Kimi-k1.5 choose the extremely simple recipe — Reinforcement Learning



Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

Recap: Long CoT Patterns Emerging in RLVR Training

DeepSeek-R1 and Kimi-k1.5 choose the extremely simple recipe — Reinforcement Learning



Model naturally develops long CoT during RL

Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

Recap: Self-Reflection Emerging in RL Training

Example:

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$

Recap: Self-Reflection Emerging in RL Training

Example:

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

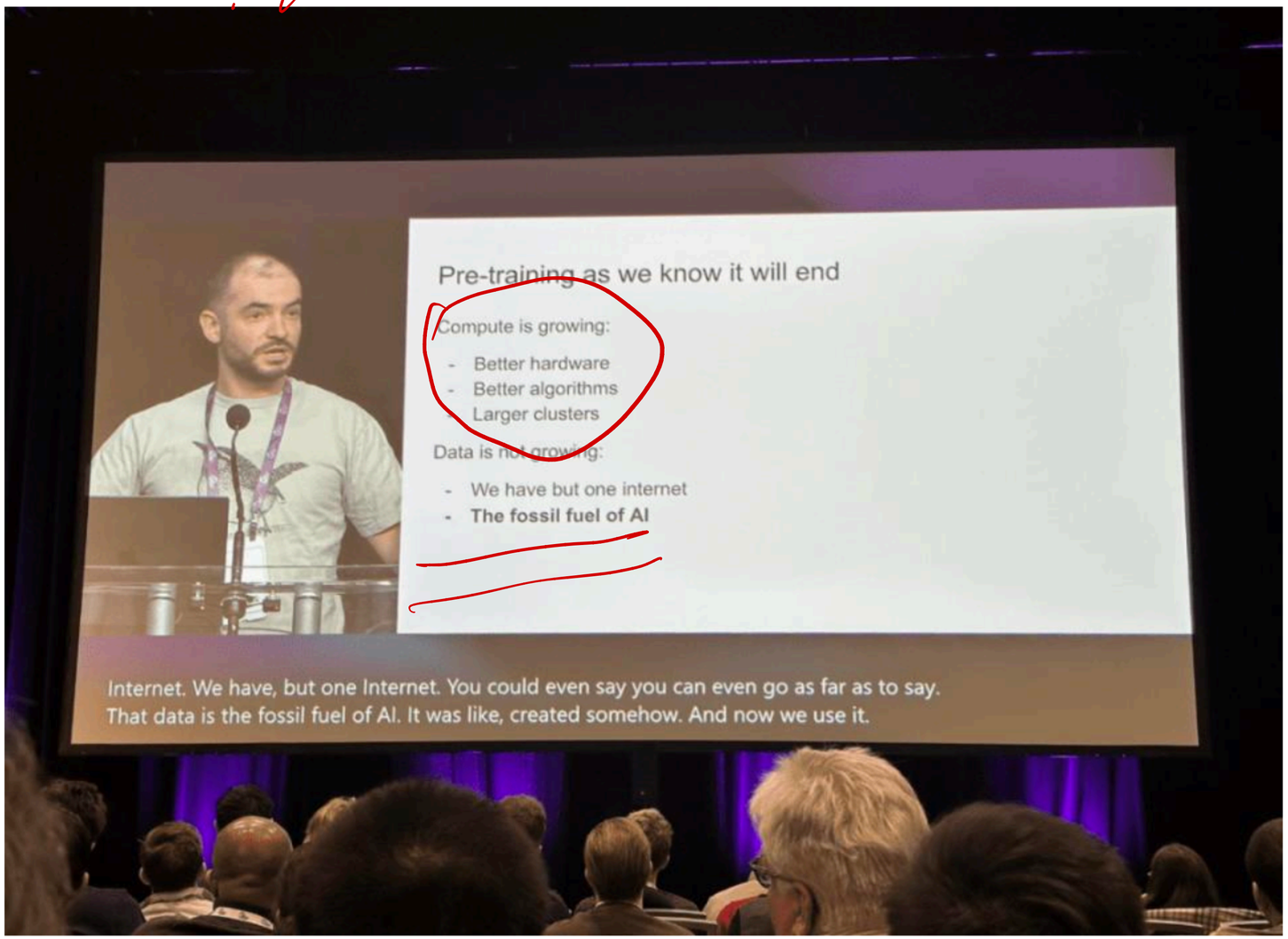
$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$

Without reward model, long CoT and self-reflection just naturally emerge.

Why Synthetic Data



Data is not growing and we are hitting the data wall

Synthetic Data in Pretraining

Prompts

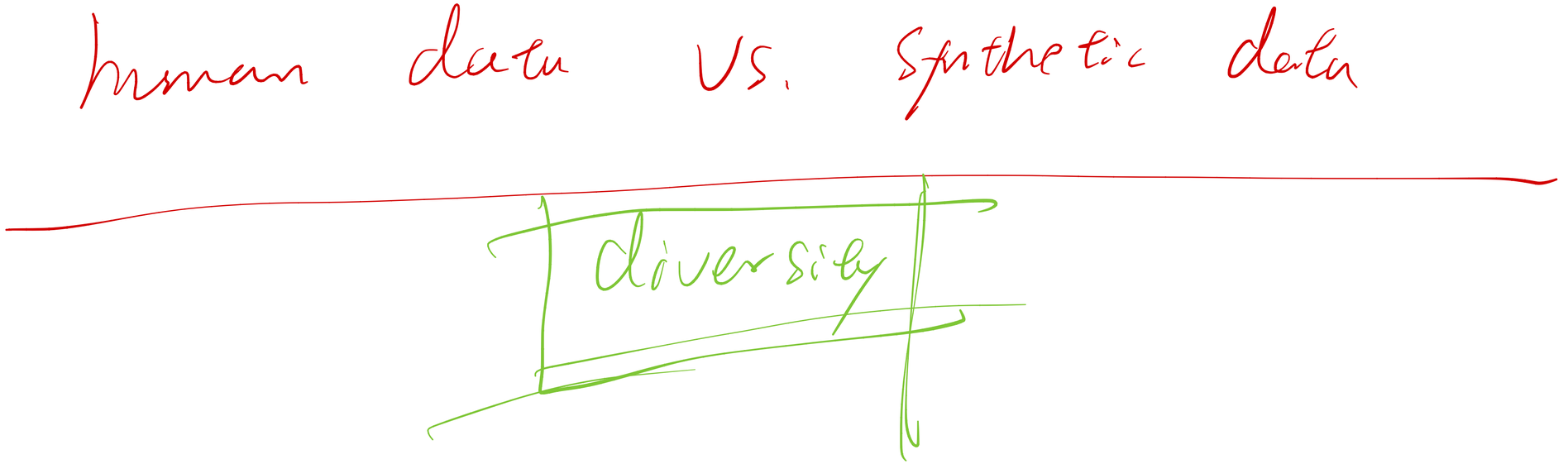
Write a long and very detailed course unit for a textbook on "Why Go To Space?" intended for **young children.** We are currently writing the first chapter: "1. Introduction". You will be writing the first sub-unit for this chapter. Write the new sub-unit titled "1.1. Why do we spend billions of dollars exploring space?" while trying to be: ...(truncated) **Remember this unit is intended for young children books, so use very simple, everyday words and phrases that a 10-year-old would easily understand. Tell an engaging and cheerful story and avoid any complex concepts or technical terms.**	Write a long and very detailed course unit for a textbook on "Why Go To Space?" intended for **professionals and researchers in the field.** We are currently writing the first chapter: "1. Introduction". You will be writing the first sub-unit for this chapter. Write the new sub-unit titled "1.1. Why do we spend billions of dollars exploring space?" while trying to be: ...(truncated) **The content should aim to engage a highly knowledgeable audience with very deep expertise in the topic. Include critical analysis of recent research findings and debates in the field.**	Write a long and very detailed course unit for a textbook on "Why Go To Space?" intended for **high school students.** We are currently writing the first chapter: "1. Introduction". You will be writing the first sub-unit for this chapter. Write the new sub-unit titled "1.1. Why do we spend billions of dollars exploring space?" while trying to be: ...(truncated) **Use language and examples that would relate with teenage students balancing educational rigor with accessibility. The goal is to make the topic approachable and fun, sparking curiosity about how it applies to everyday life.**

Huggingface. Cosmopedia: how to create large-scale synthetic data for pre-training. 2024

How to Obtain Prompts?

Diversity is the biggest issue for large scale data synthesis

Human data is naturally diverse

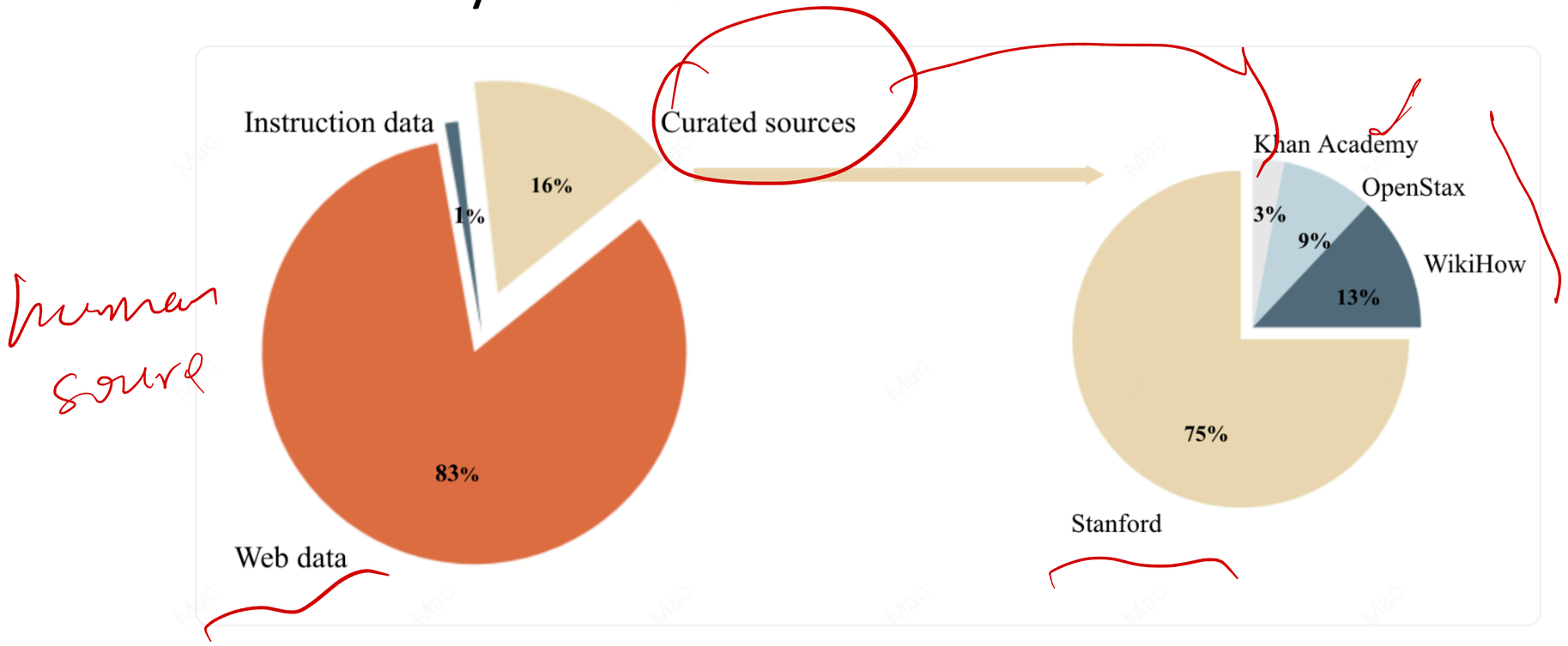


Huggingface. Cosmopedia: how to create large-scale synthetic data for pre-training. 2024

How to Obtain Prompts?

Diversity is the biggest issue for large scale data synthesis

Human data is naturally diverse



Huggingface. Cosmopedia: how to create large-scale synthetic data for pre-training. 2024

Synthetic Data in Pretraining

Web extract/seed sample:
The Cardiovascular BioImaging Core offers the latest echo technology, including real-time three-dimensional (3D) and 4D, and speckle tracking imaging. Real-time 3D and speckle tracking echocardiography are new technologies that give accurate measures of regional and global cardiac function. These technologies rival information obtained by more expensive modalities (like cardiac MRI) and have… (truncated)

Topic:
Medicine

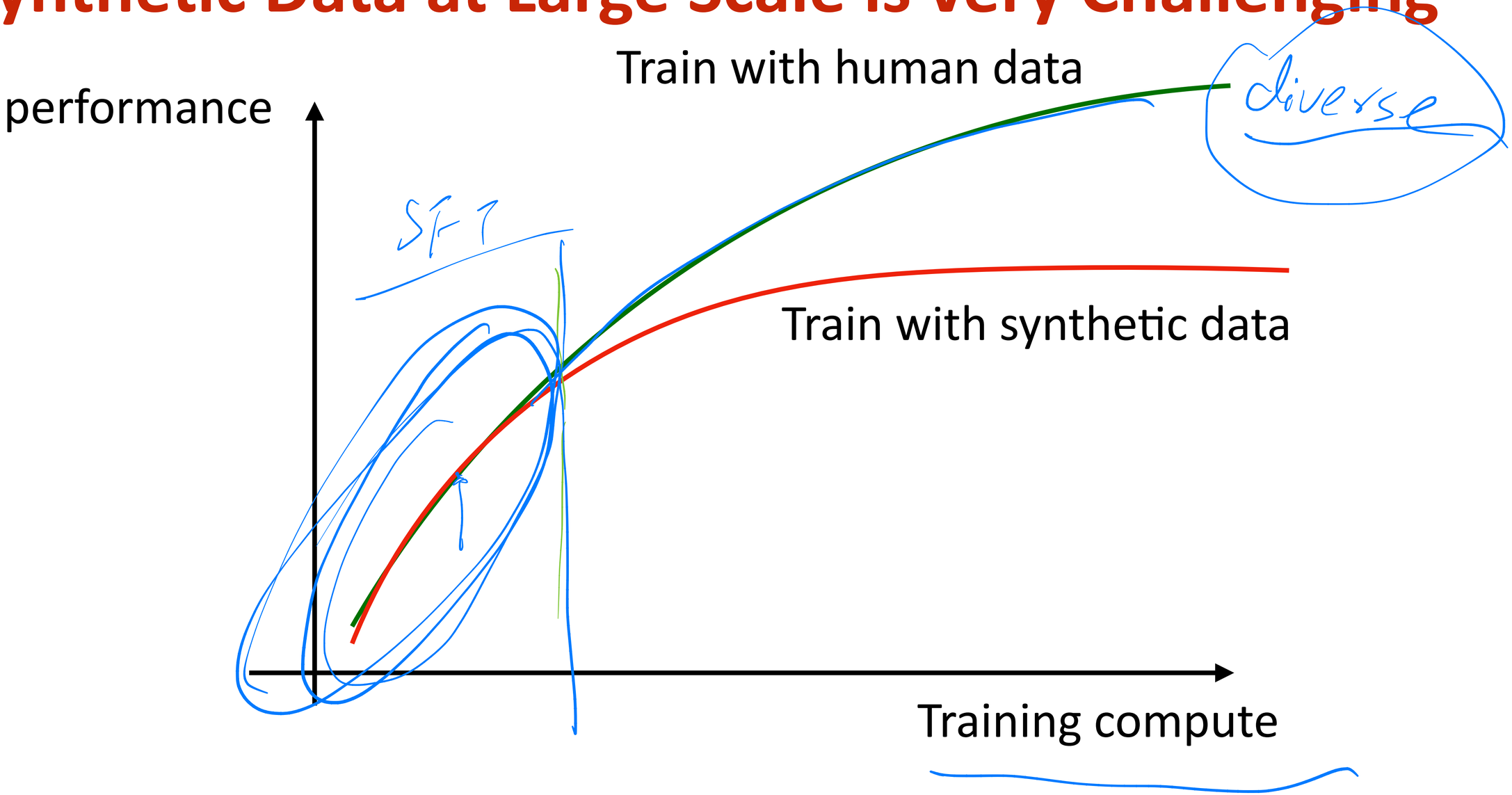
Prompt:

Here is an extract from a webpage: **"The Cardiovascular BioImaging Core offers the latest echo technology, including real-time three-dimensional (3D) and 4D, and speckle tracking imaging. Real-time 3D and speckle tracking echocardiography are new technologies that give accurate measures of regional and global cardiac function. These technologies rival information obtained by more expensive modalities (like cardiac MRI) and have..(truncated)."**.
Write an informative and insightful blog post that expands upon the extract above, within the context of **"Medicine"**.

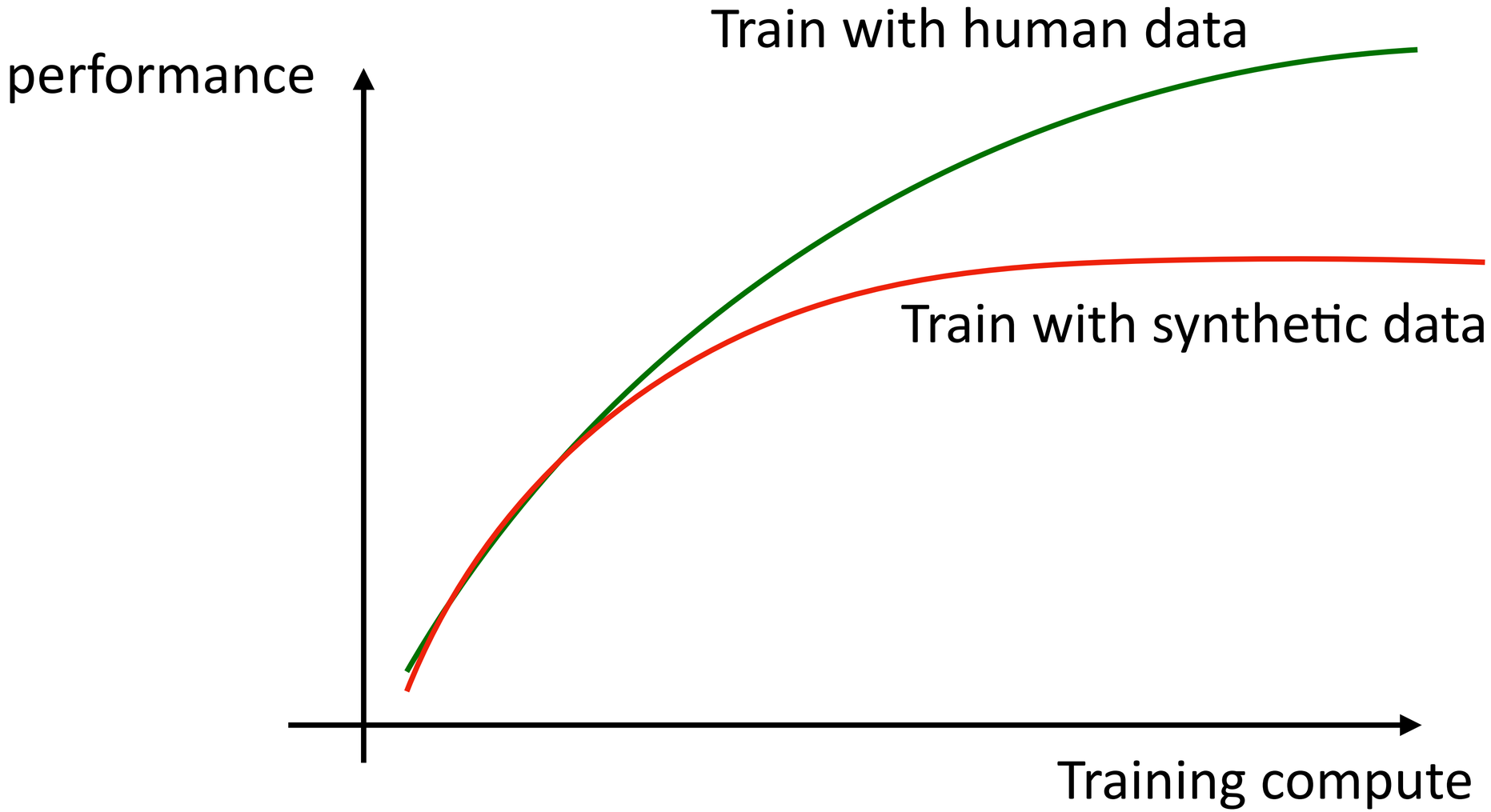
Your post should delve into the nuances of the topic, offering fresh perspectives and deeper analysis.
Aim to:
- Inform: Provide valuable, well-researched information that educates the reader.
- Engage: Write in a conversational tone that connects with the audience, making complex ideas accessible.
- Illustrate: Use examples, anecdotes, or personal experiences to bring the topic to life. Do not give a title and do not start with sentences like "Have you ever..." or "Hello dear readers..", simply write the content without these introductory phrases.

Huggingface. Cosmopedia: how to create large-scale synthetic data for pre-training. 2024

Synthetic Data at Large Scale is very Challenging

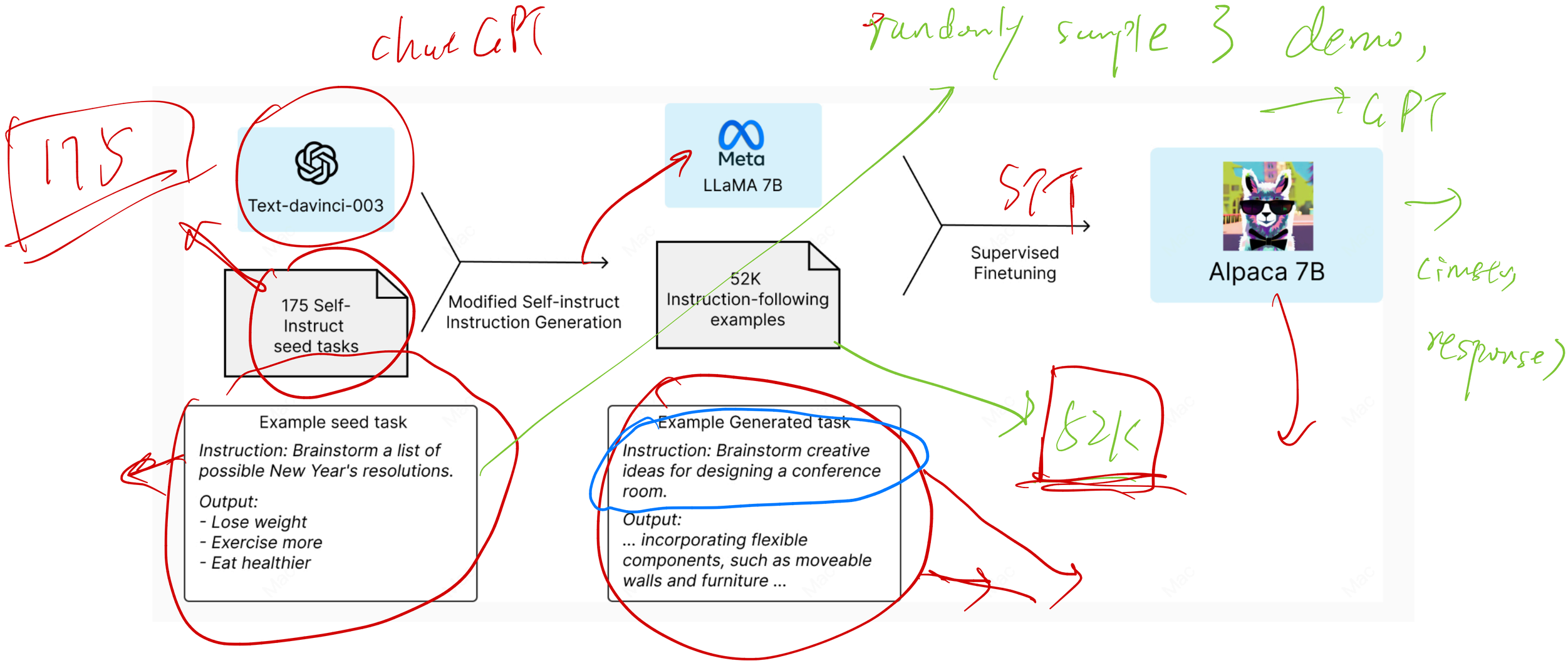


Synthetic Data at Large Scale is very Challenging



Diversity is the biggest issue here, no diversity, no scaling

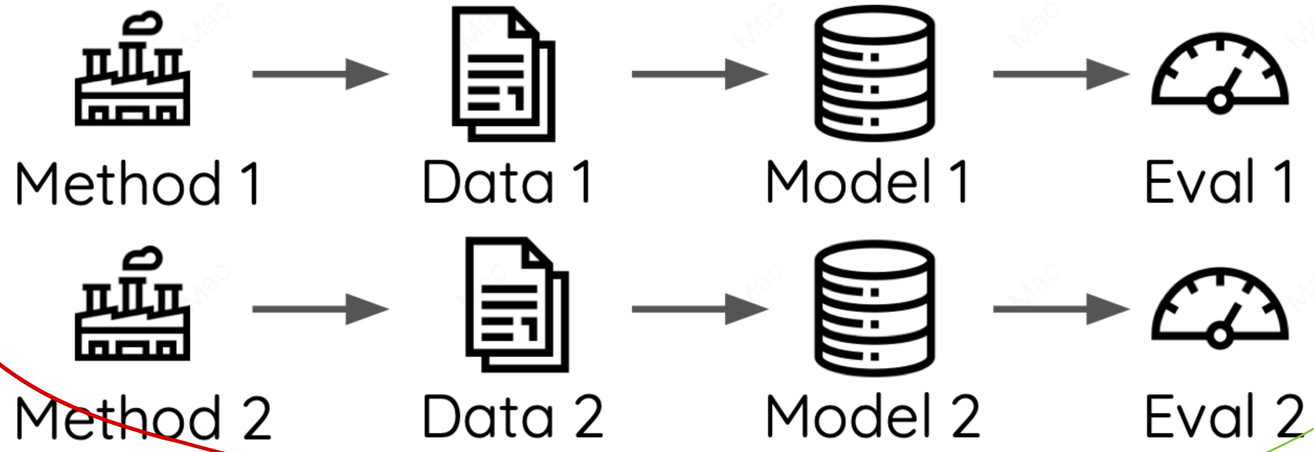
Synthetic Data at SFT is Very Successful



Taori et al. Alpaca: A Strong, Replicable Instruction-Following Model. 2023

How to Evaluate Synthetic Data

- **Extrinsic:** Does it help in a downstream task?



- **Intrinsic:** What are the characteristics of the data or generation process?

Intrinsic Evaluation of Synthetic Data

- Questions regarding whether the data is correct, judged by manual or automatic methods
- E.g. Self-Instruct manually annotates:

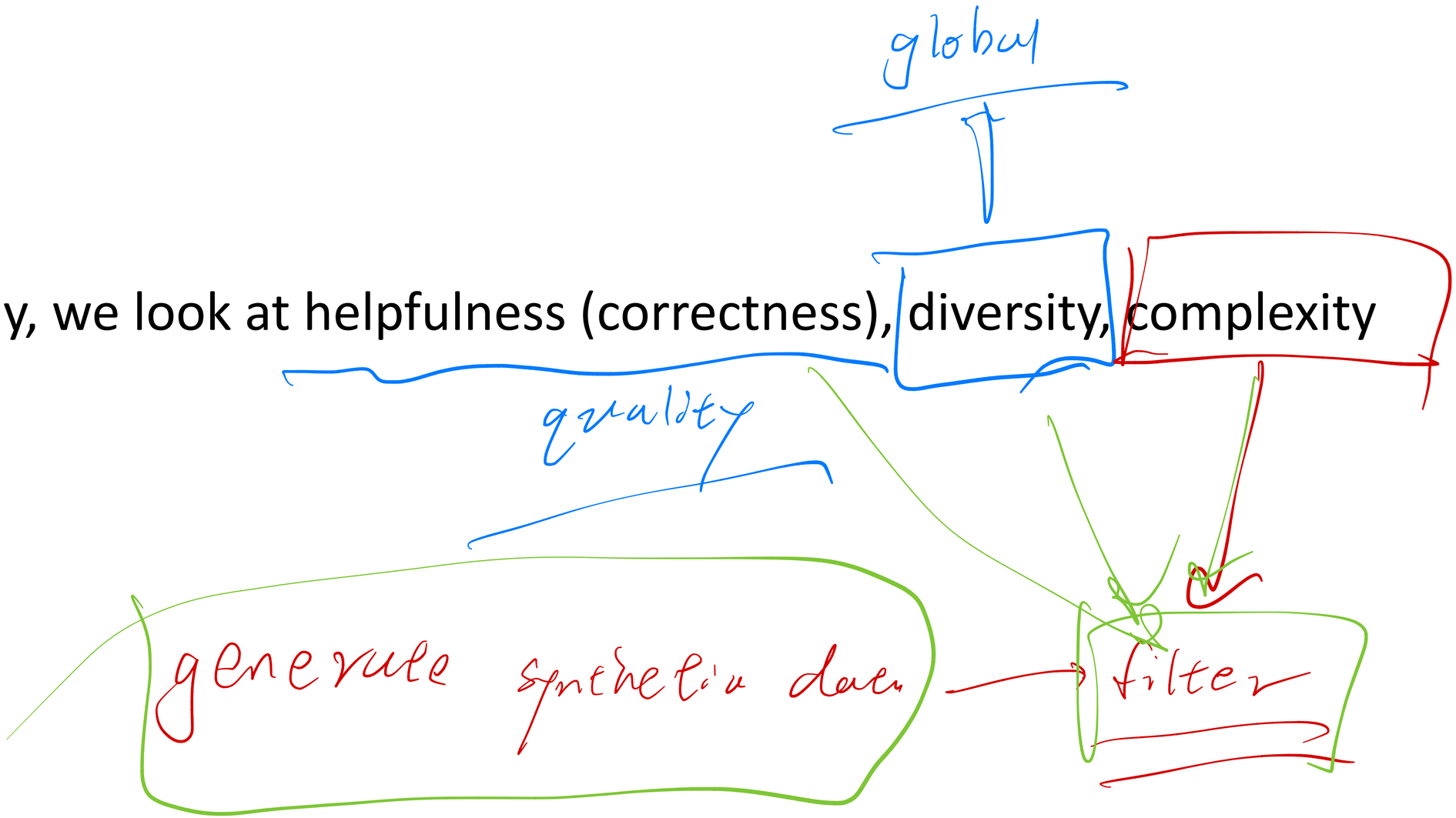
Quality Review Question	Yes %
Does the instruction describe a valid task?	92%
Is the input appropriate for the instruction?	79%
Is the output a correct and acceptable response to the instruction and input?	58%
All fields are valid	54%

Summarize

Wang et al. Self-Instruct: Aligning Language Models with Self-Generated Instructions. 2022

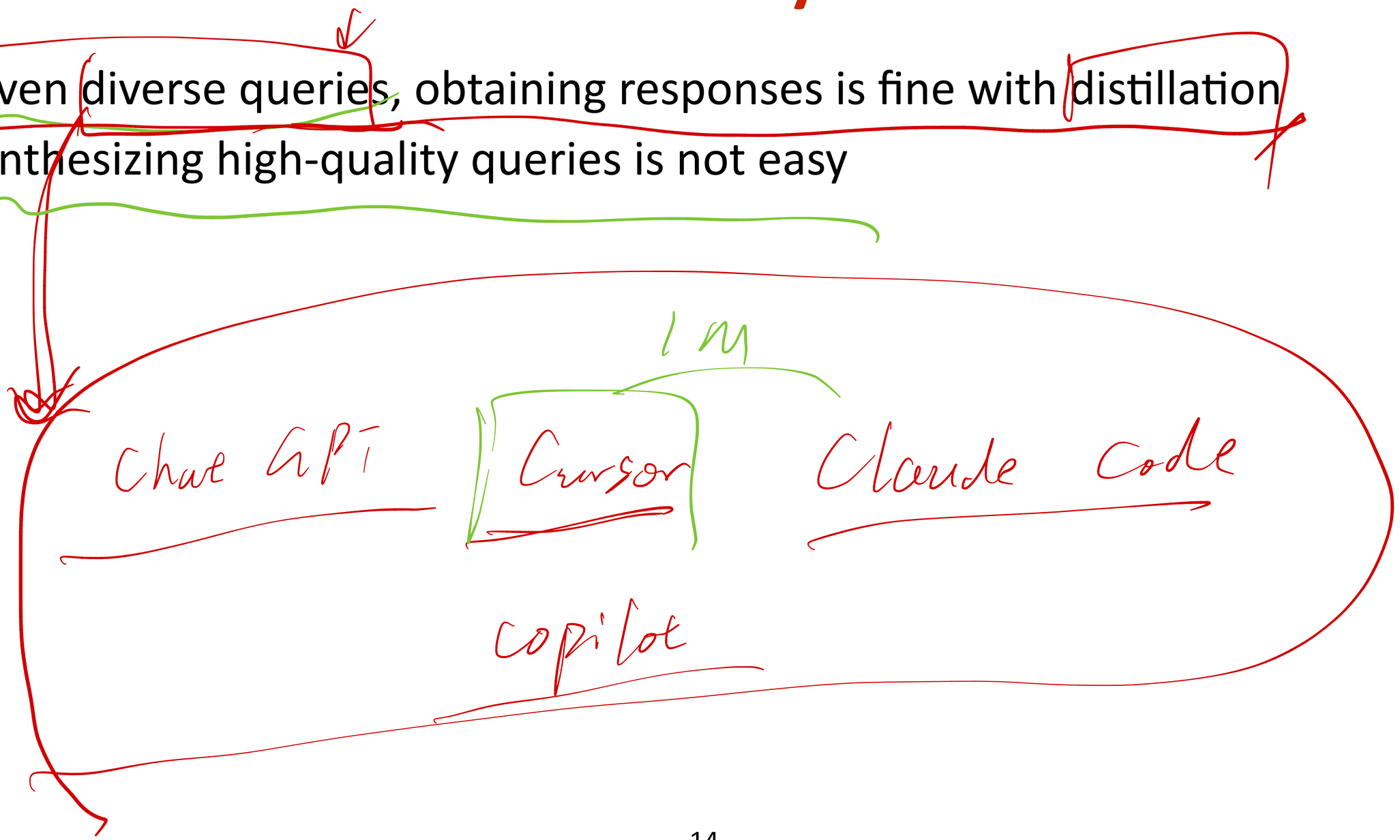
Intrinsic Evaluation of Synthetic Data

Generally, we look at helpfulness (correctness), diversity, complexity

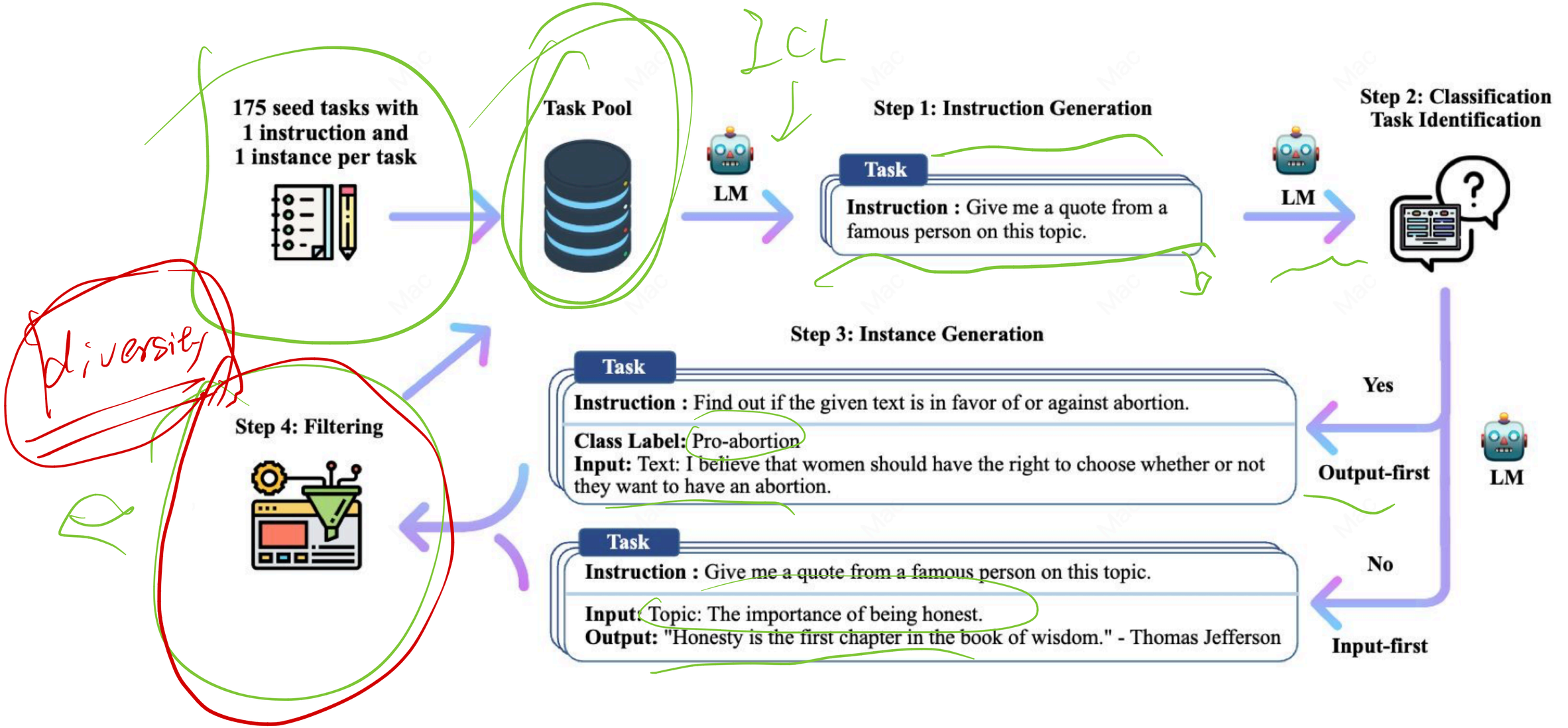


How to Generate Synthetic Data

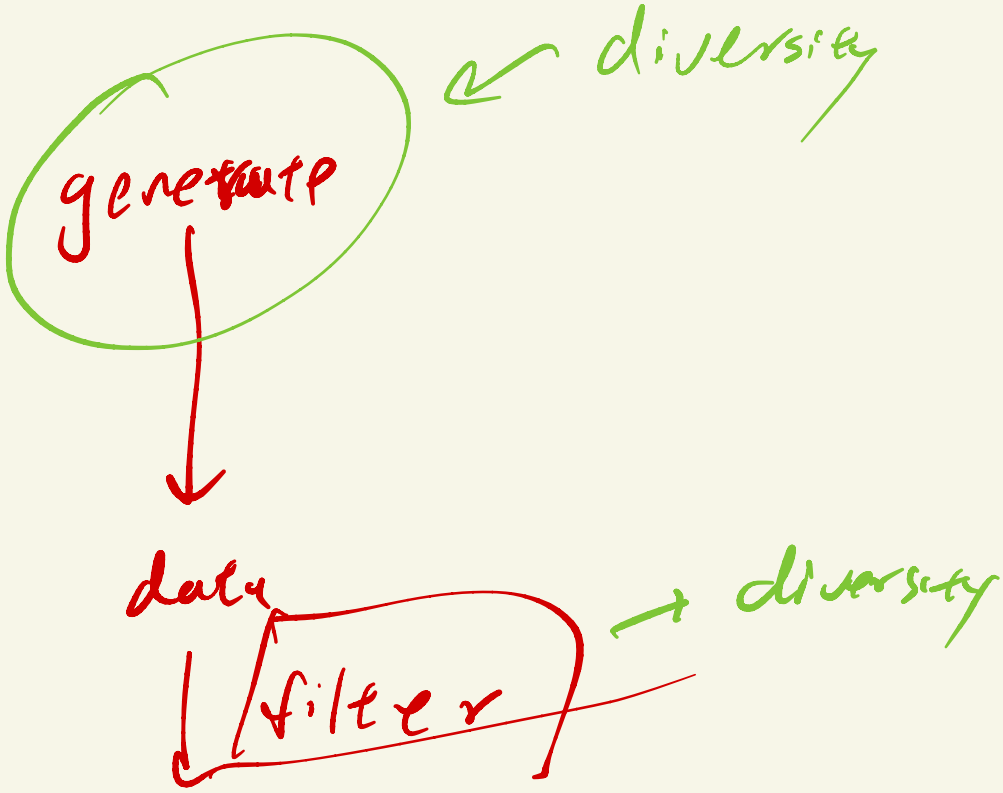
- Given diverse queries, obtaining responses is fine with distillation
- Synthesizing high-quality queries is not easy



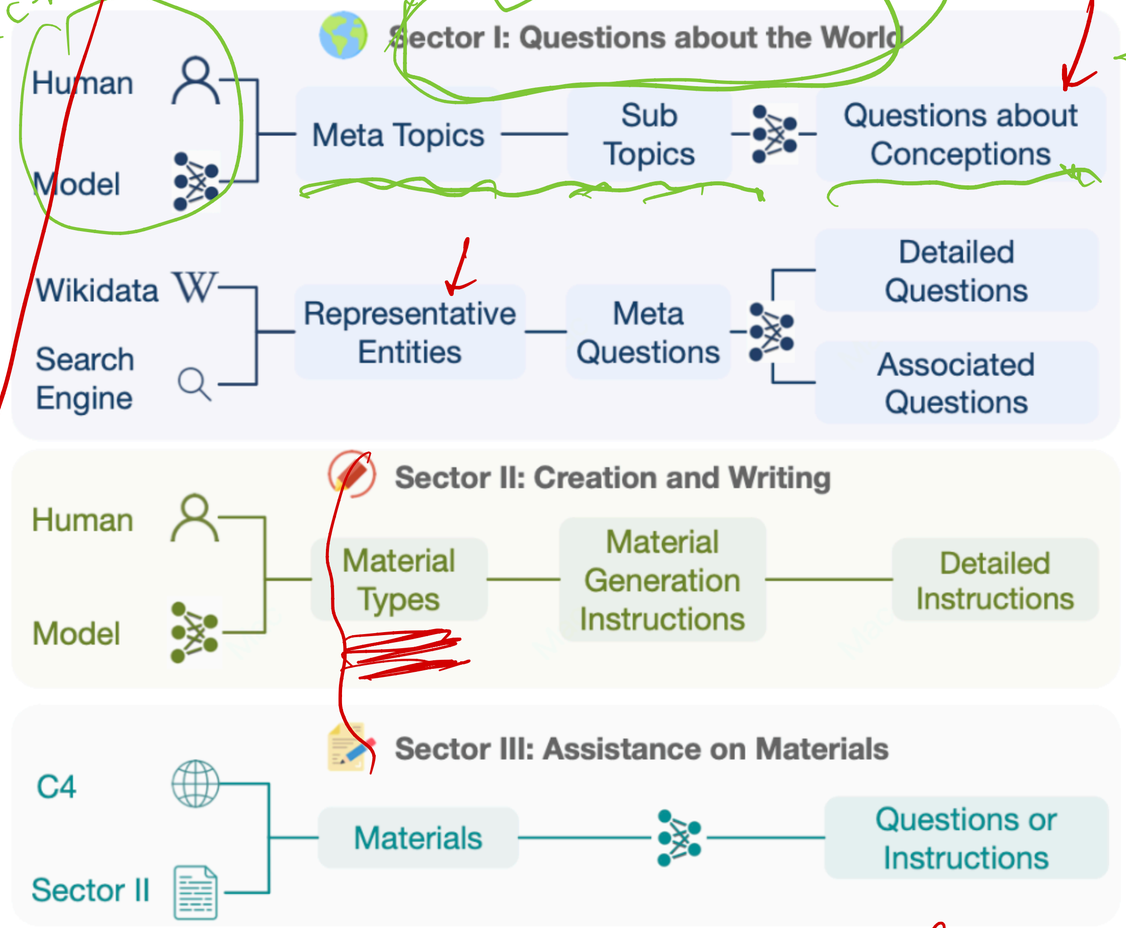
How to Create Synthetic Data

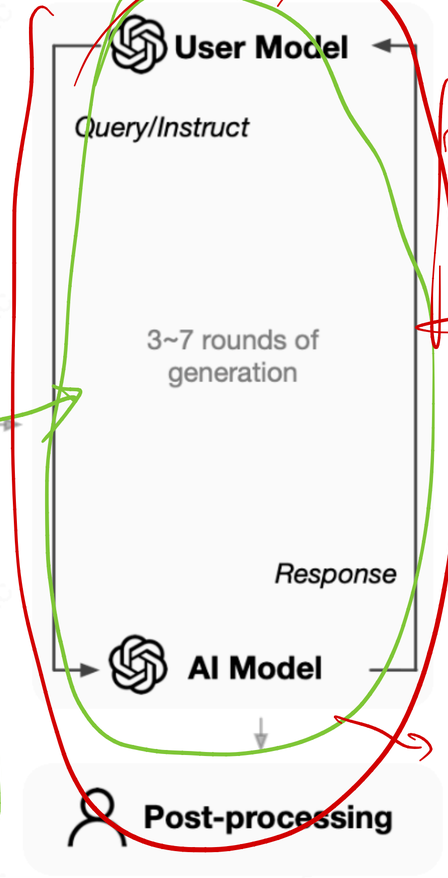


Wang et al. Self-Instruct: Aligning Language Models with Self-Generated Instructions. 2022


diversity
generate
data
filter
diversity

How to Create Synthetic Data





Ding et al. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. 2023

How to Create Synthetic Data

Technology	Health and wellness	Travel and adventure
Food and drink	Art and culture	Science and innovation
Fashion and style	Relationships and dating	Sports and fitness
Nature and the environment	Music and entertainment	Politics and current events
Education and learning	Money and finance	Work and career
Philosophy and ethics	History and nostalgia	Social media and communication
Creativity and inspiration	Personal growth and development	Spirituality and faith
Pop culture and trends	Beauty and self-care	Family and parenting
Entrepreneurship and business	Literature and writing	Gaming and technology
Mindfulness and meditation	Diversity and inclusion	Travel and culture exchange

Table 2: 30 meta-concepts used to generate the first sector of UltraChat data.

Articles and Blog Posts	Job Application Material	Stories
Legal Documents and Contracts	Poems	Educational Content
Screenplays	Scripts for Language Learning	Technical Documents and Reports
Marketing Materials	Social Media Posts	Personal Essays
Emails	Scientific Papers and Summaries	Speeches and Presentations
Recipes and Cooking Instructions	News Articles	Song Lyrics
Product Descriptions and Reviews	Programs and Code	

Table 3: 20 types of text materials used for sector 2 and 3 UltraChat generation.

explicit

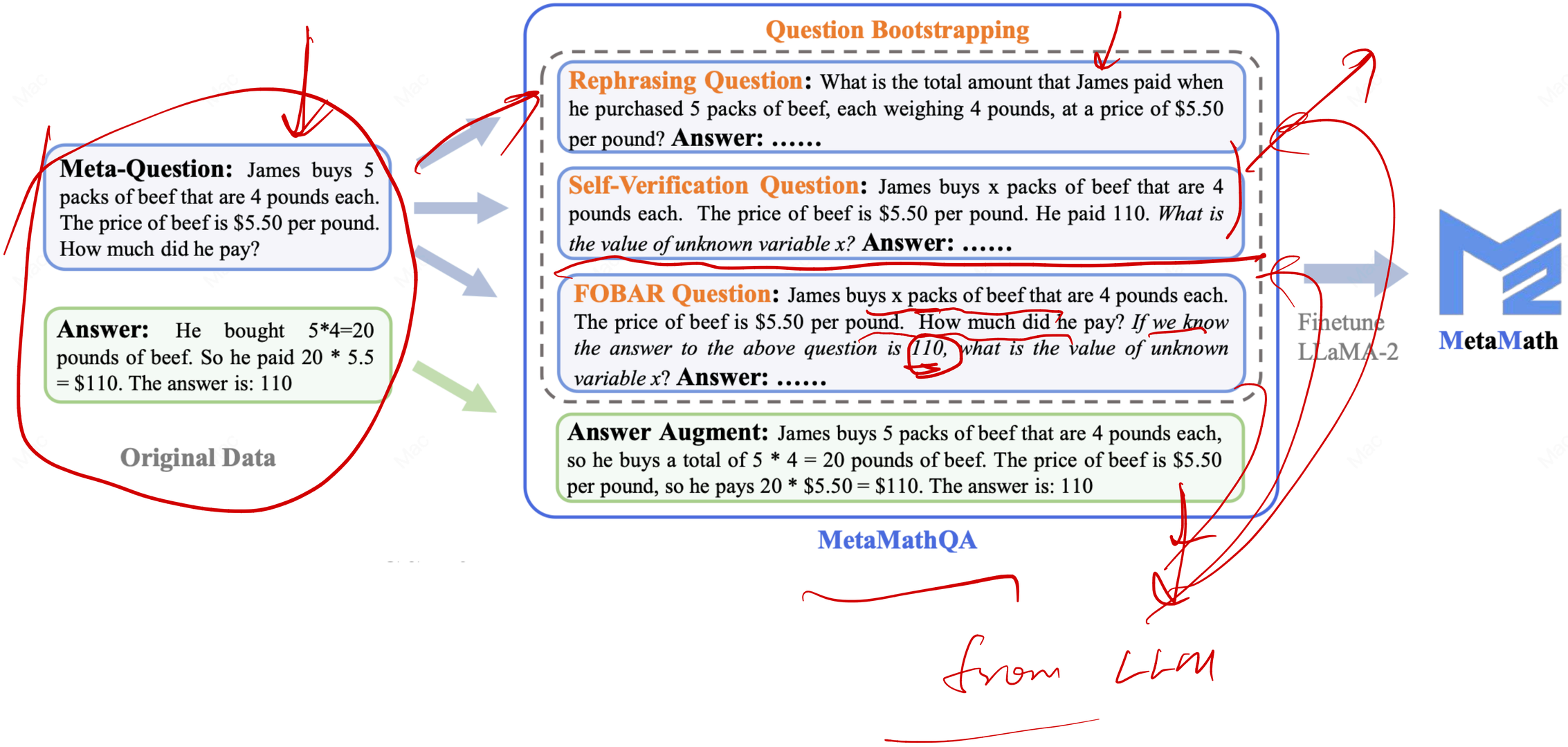
Ding et al. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. 2023

How to Create Synthetic Data

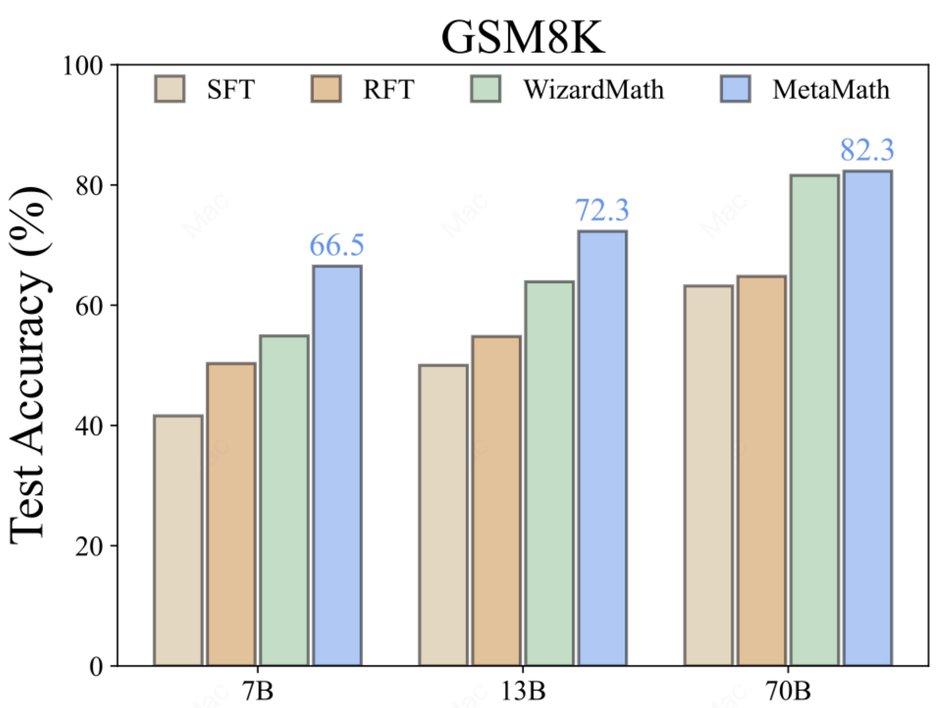
Model	Vicuna Set	Commonsense		World Knowledge		Professional Knowledge		Ability		Writing	Overall
		Easy	Moderate	Easy	Difficult	Physics	Biology	Math	Reasoning		
Dolly-12B	6.61	7.77	7.90	8.53	8.50	8.57	8.53	6.43	5.13	6.36	7.15
MPT-7B	8.38	8.17	**9.07**	8.30	8.87	8.57	8.87	8.80	7.53	7.76	8.32
OpenAssistant-12B	8.40	8.97	8.70	9.57	8.23	8.67	8.80	8.80	8.17	7.81	8.47
Baize-13B	8.36	9.03	8.87	9.37	8.97	8.83	8.93	8.50	8.57	7.90	8.57
Alpaca-7B	8.05	9.50	8.83	9.67	9.17	8.60	8.80	9.10	7.80	8.16	8.60
Koala-13B	8.60	9.53	8.93	9.77	9.23	9.10	**9.33**	8.90	8.70	8.34	8.88
Vicuna-13B	8.63	9.53	9.03	9.63	9.27	9.00	9.27	9.10	9.10	8.51	8.96
ChatGPT	**8.79**	**9.77**	**9.07**	9.77	9.30	9.07	9.27	**9.37**	**9.63**	**8.63**	**9.12**
UltraLLaMa-13B	8.70	9.70	9.03	**9.90**	**9.33**	**9.17**	9.27	9.27	8.87	8.51	9.02

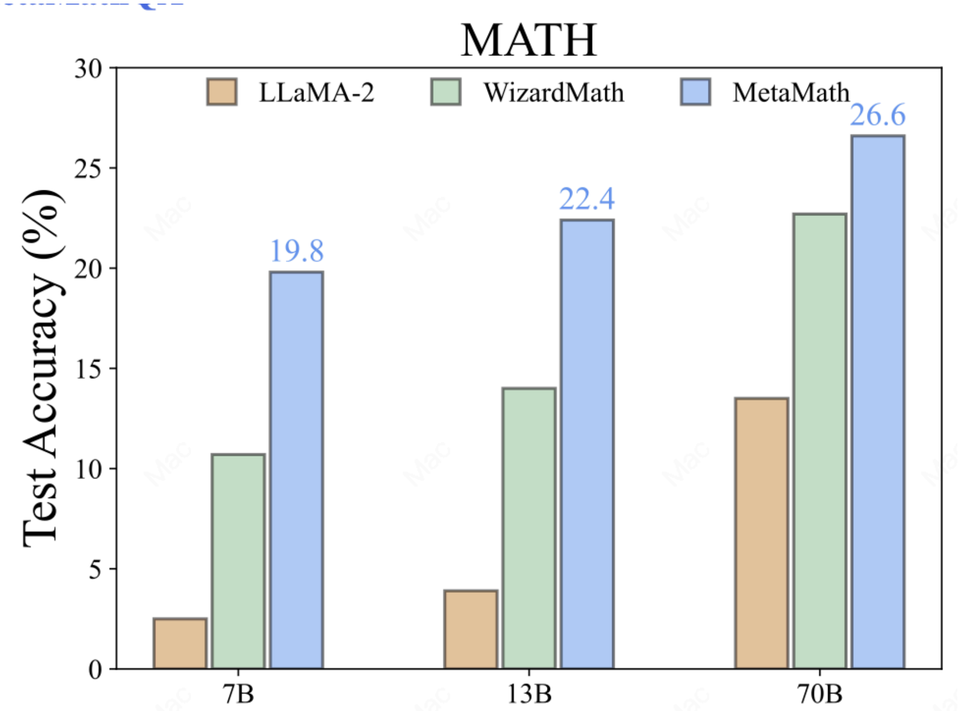
Ding et al. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. 2023

How to Create Synthetic Data

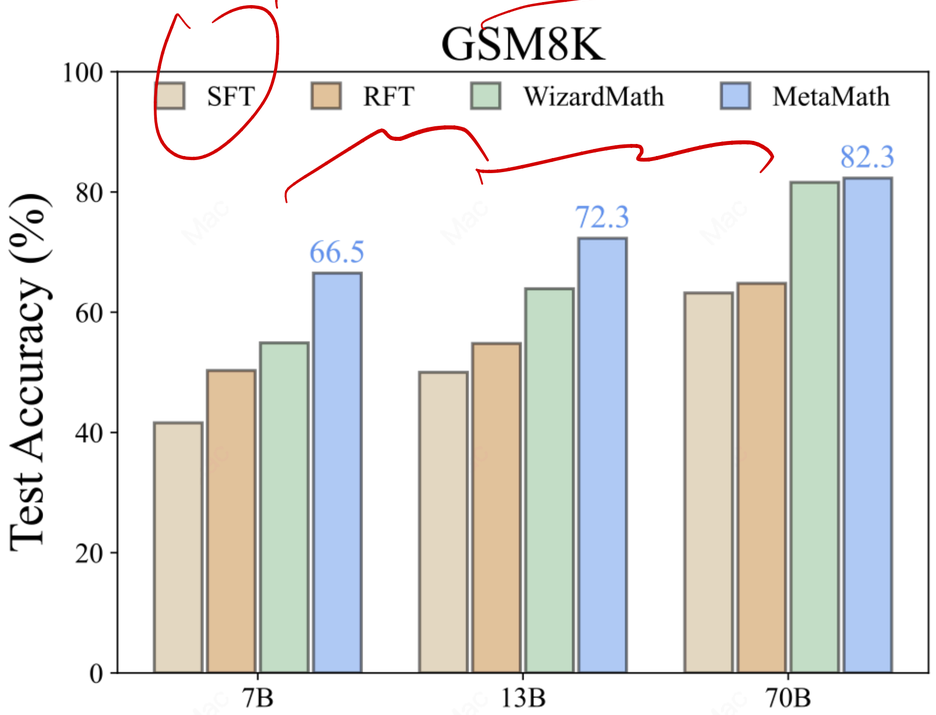


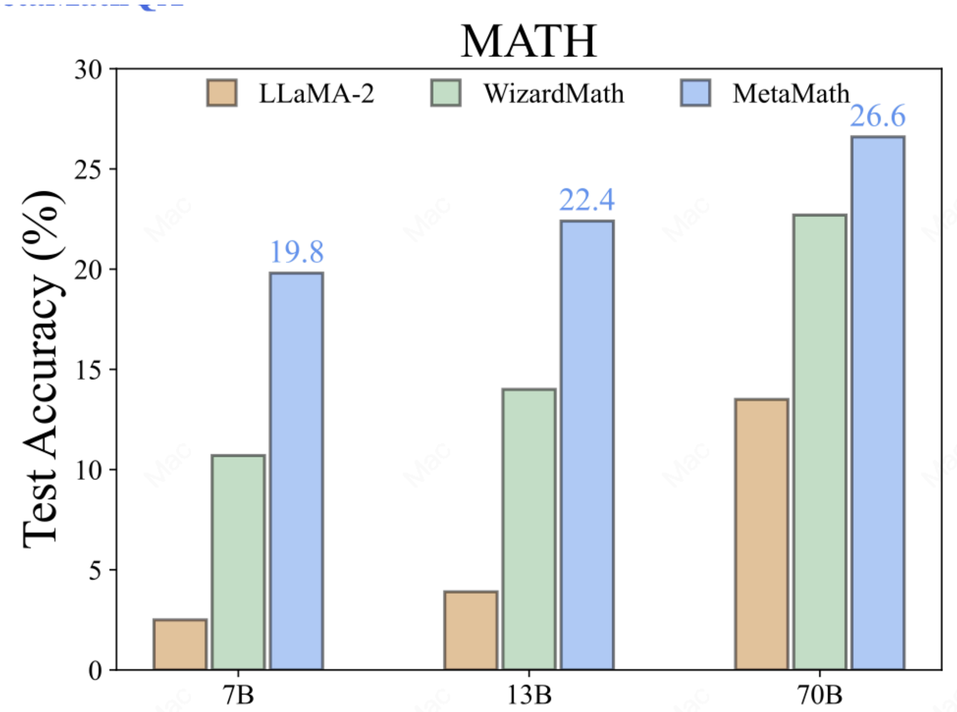
How to Create Synthetic Data





How to Create Synthetic Data

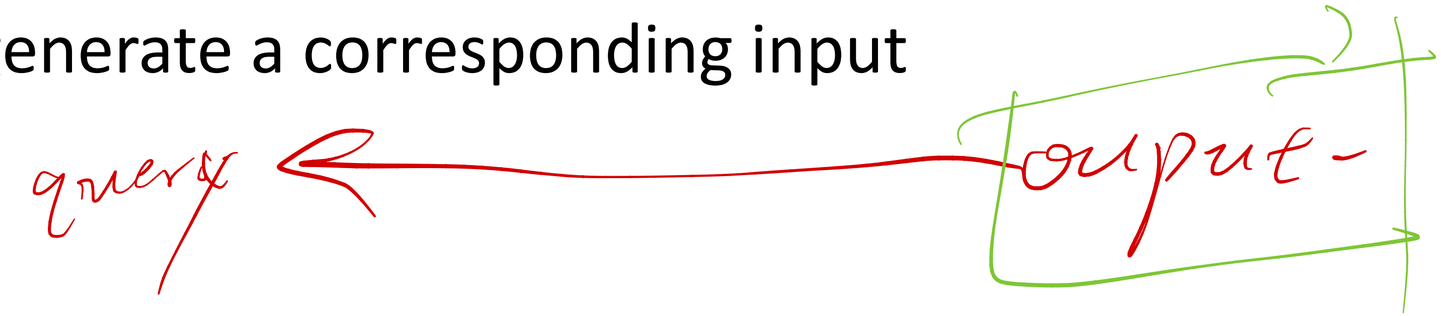




Better than Training on Human Data

Back-translation for Synthetic Data Generation

Given an output, generate a corresponding input

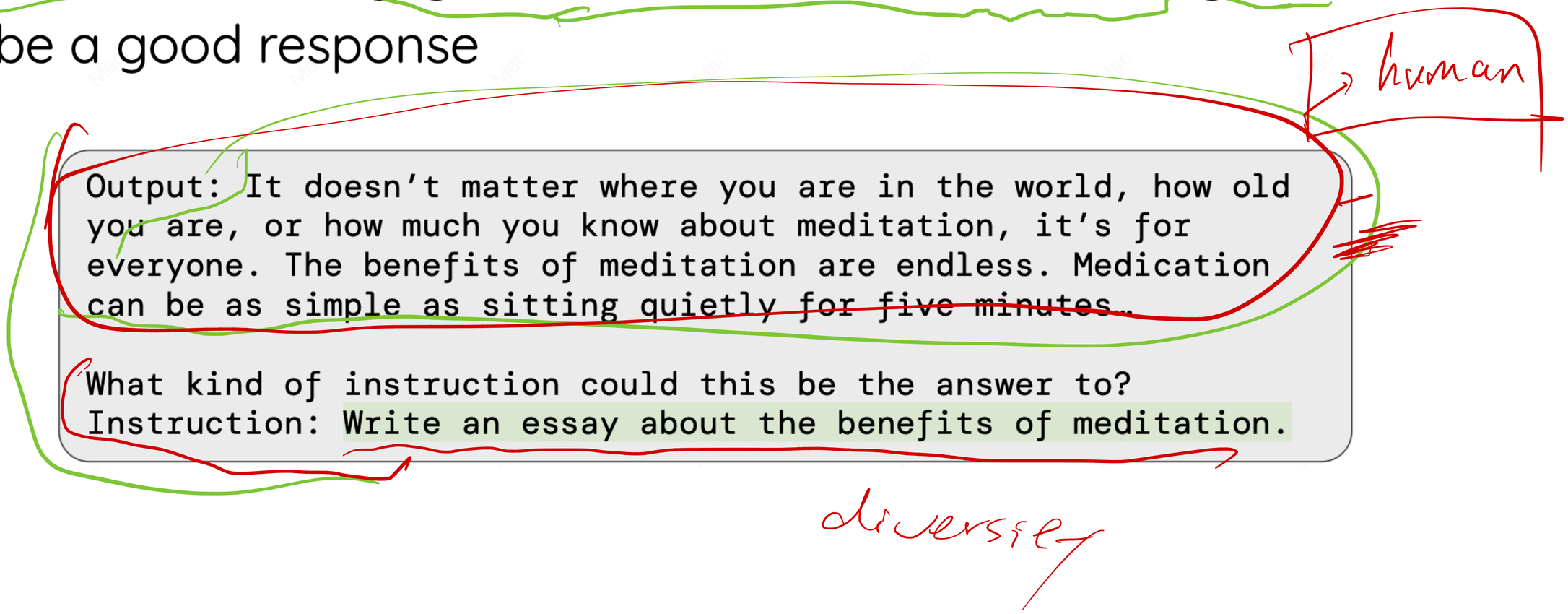


Li et al. Self-Alignment with Instruction Backtranslation. 2023

Back-translation for Synthetic Data Generation

Given an output, generate a corresponding input

Given web text y, generate instruction x for which y would be a good response



Li et al. Self-Alignment with Instruction Backtranslation. 2023

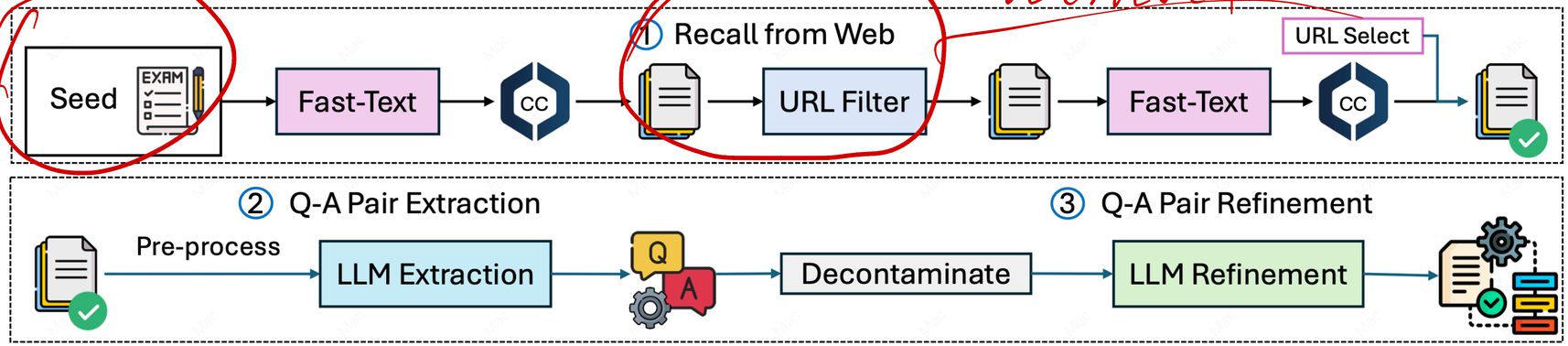
Transformation of Existing Data

Use or retrieve existing data, then transform it into an example under the desired task

Transformation of Existing Data

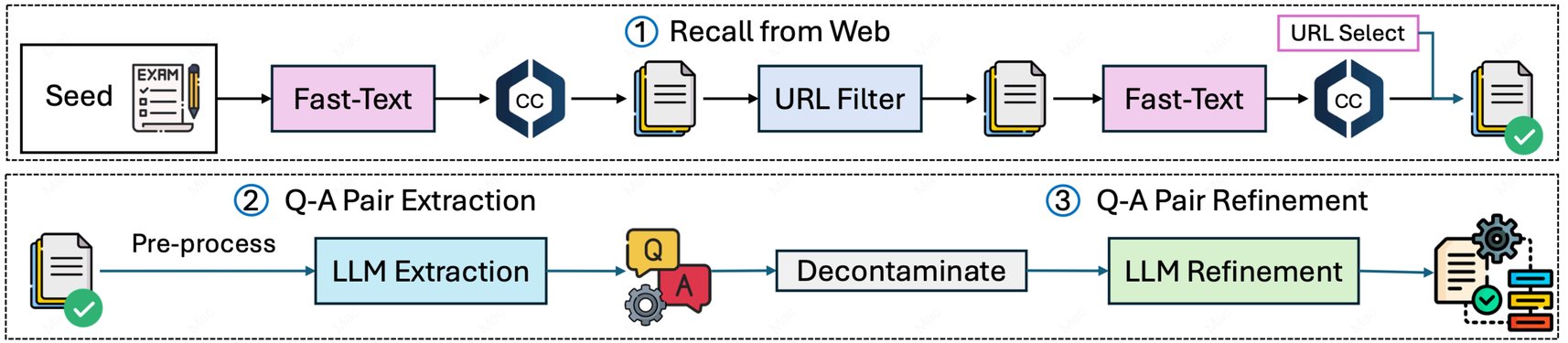
Yue et al. MAmmoTH2: Scaling Instructions from the Web. 2024

Transformation of Existing Data



Yue et al. MAmmoTH2: Scaling Instructions from the Web. 2024

Transformation of Existing Data



Raw Docs *Unformatted Text, Site Information, Ads*

Topics Science\nAnatomy&Physiology\nAstronomy\nAstrophysics \nBiology\nChemistry \n...Socratic Meta...Featured Answers
How do you simplify #((u^4v^3)/(u^2v^-1)^4)^0# and write it using only positive exponents?
Answer by NickTheTurtle (Apr 1, 2017)
Explanation:\nAnything raised to the \(0^{th}\) power is simply 1. \n\nRelated Questions\n What is the quotient of powers property? \n\n\n\n How do you simplify expressions using the quotient rule?...\nImpact of this question\n1274 views around the world #Apps\niOS\nAndroid\nLinks\n[Privacy](#)\n[Terms](#)\n[Help](#)

Extracted QA *Formatted QA but lacking detailed solutions*

Question: How do you simplify $(u^4\,v^3/(u^2\,v^{-1}\,)^4)^0$ and write it using only positive exponents?
Answer: Explanation: Anything to the $0th$ power is just simply 1.

Rewritten QA *Formatted QA augmented with detailed solutions*

Question:
How do you simplify $(u^4\,v^3/(u^2\,v^{-1}\,)^4)^0$ and write it using only positive exponents?

Answer:
To simplify the expression $(u^4\,v^3/(u^2\,v^{-1}\,)^4)^0$ and rewrite it using only positive exponents, we start by evaluating the expression from the innermost operation outward.

First, consider the exponent of zero on the entire fraction:
Any expression raised to the power of zero is equal to 1 (provided the base is not zero). *Therefore*, $(u^4\,v^3/(u^2\,v^{-1}\,)^4)^0 = 1$

This simplification makes the other calculations unnecessary because raising any non-zero expression to the power of zero will always result in 1.

So, the simplified expression is simply: 1

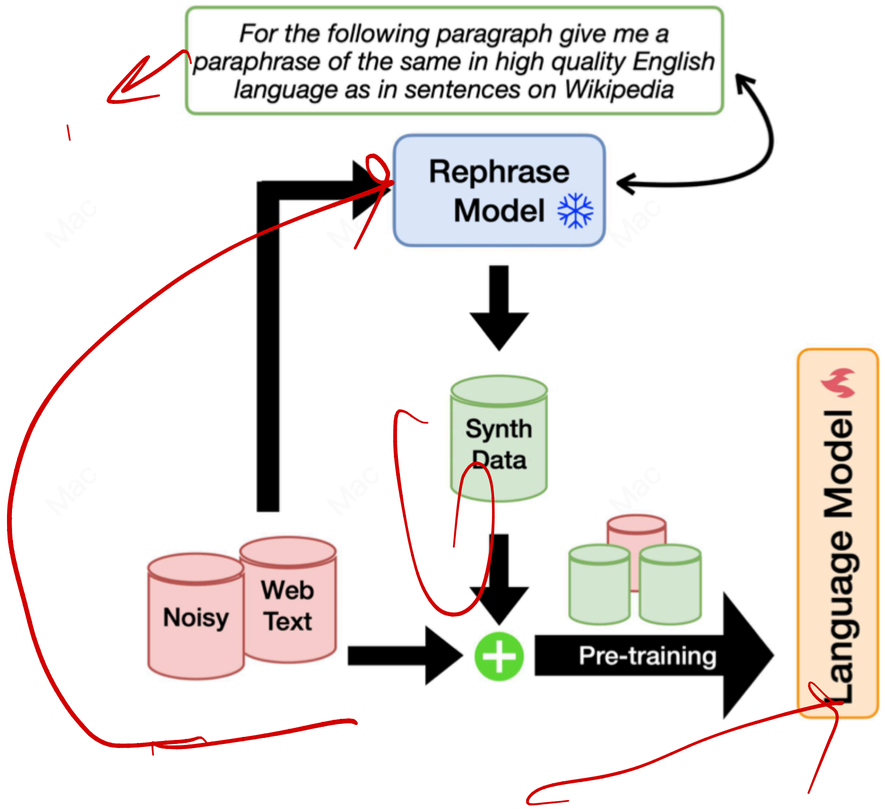
diversity

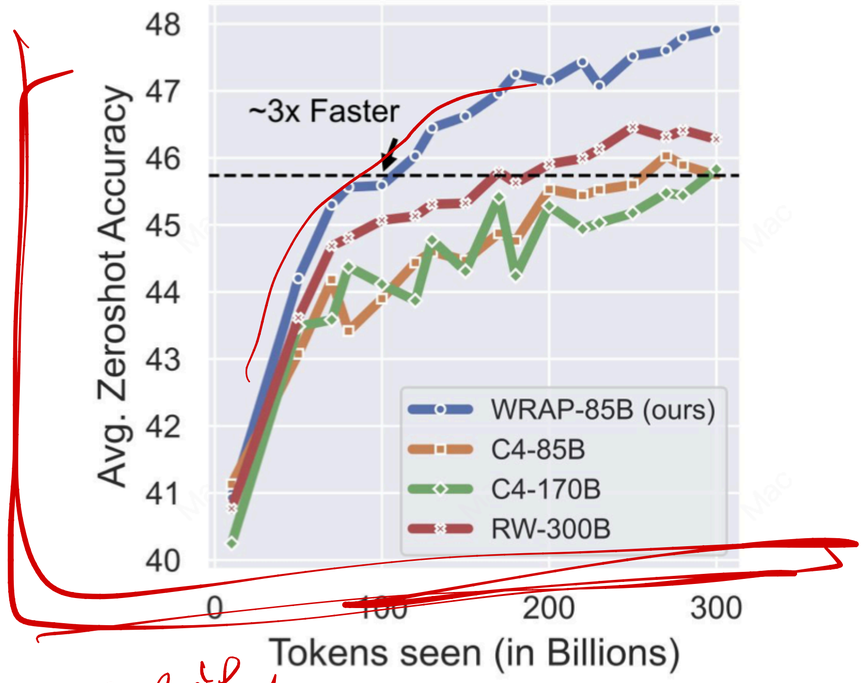
quality

Yue et al. MAmmoTH2: Scaling Instructions from the Web. 2024

Rephrasing documents for pretraining

Use LMs to paraphrase noisy web text to create new data!





Maini et al. Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling. 2024

How to Filter Synthetic Data

- Correctness Filtering
- Quality Filtering
- Diversity Filtering

Quality Filtering

Can simply asking LLM to judge the quality

You are evaluating a response that has been submitted for a particular task, using a specific set of standards. Below is the data:

[BEGIN DATA]
***
[Task]: {task}
***
[Submission]: {submission}
***
[Criterion]: helpfulness:
"1": "Not helpful - The generated text is completely irrelevant, unclear, or incomplete. It does not provide any useful information to the user."
"2": "Somewhat helpful - The generated text has some relevance to the user's question, but it may be unclear or incomplete. It provides only partial information, or the information provided may not be useful for the user's needs."
"3": "Moderately helpful - The generated text is relevant to the user's question, and it provides a clear and complete answer. However, it may lack detail or explanation that would be helpful for the user."
"4": "Helpful - The generated text is quite relevant to the user's question, and it provides a clear, complete, and detailed answer. It offers additional information or explanations that are useful for the user. However, some of the points of the response are somewhat repetitive or could be combined for greater clarity and concision"
"5": "Very helpful - The generated text is highly relevant to the user's question, and it provides a clear, complete, and detailed answer. It offers additional information, explanations, or analogies that are not only useful but also insightful and valuable to the user. However, the structured of the response is not well-organized and there is no clear progression or logical sequence of different points in the response."
"6": "Highly helpful - The generated text provides a clear, complete, and detailed answer. It offers additional information or explanations that are not only useful but also insightful and valuable to the user. The response is also in a logical and easy-to-follow manner by explicitly using headings, bullet points, or numbered lists to break up the information and make it easier to read."
***
[END DATA]

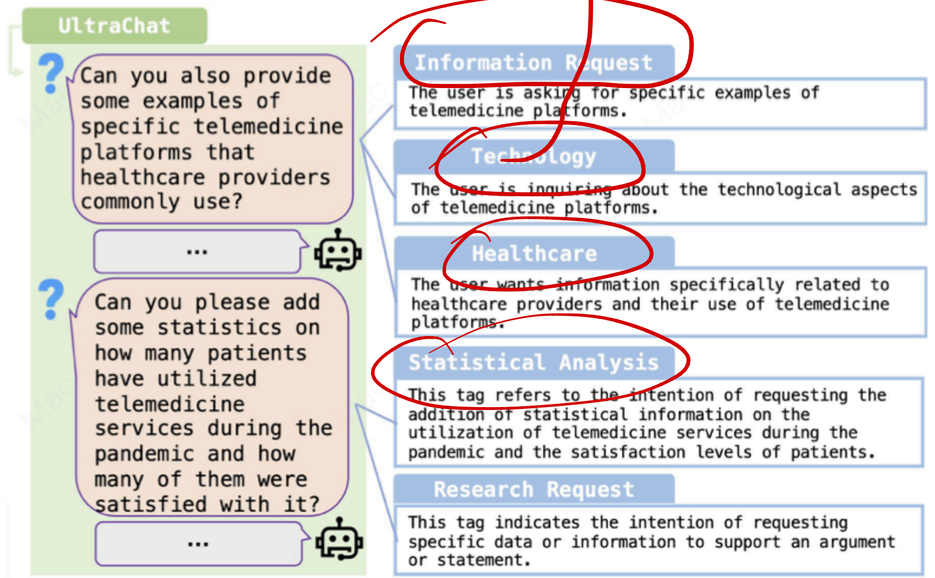
Does the submission meet the criterion? First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print the choice only from "1, 2, 3, 4, 5, 6" (without quotes or punctuation) on its own line corresponding to the correct answer. At the end, repeat just the selected choice again by itself on a new line.

Zhou et al. Less is More for Alignment. 2023.

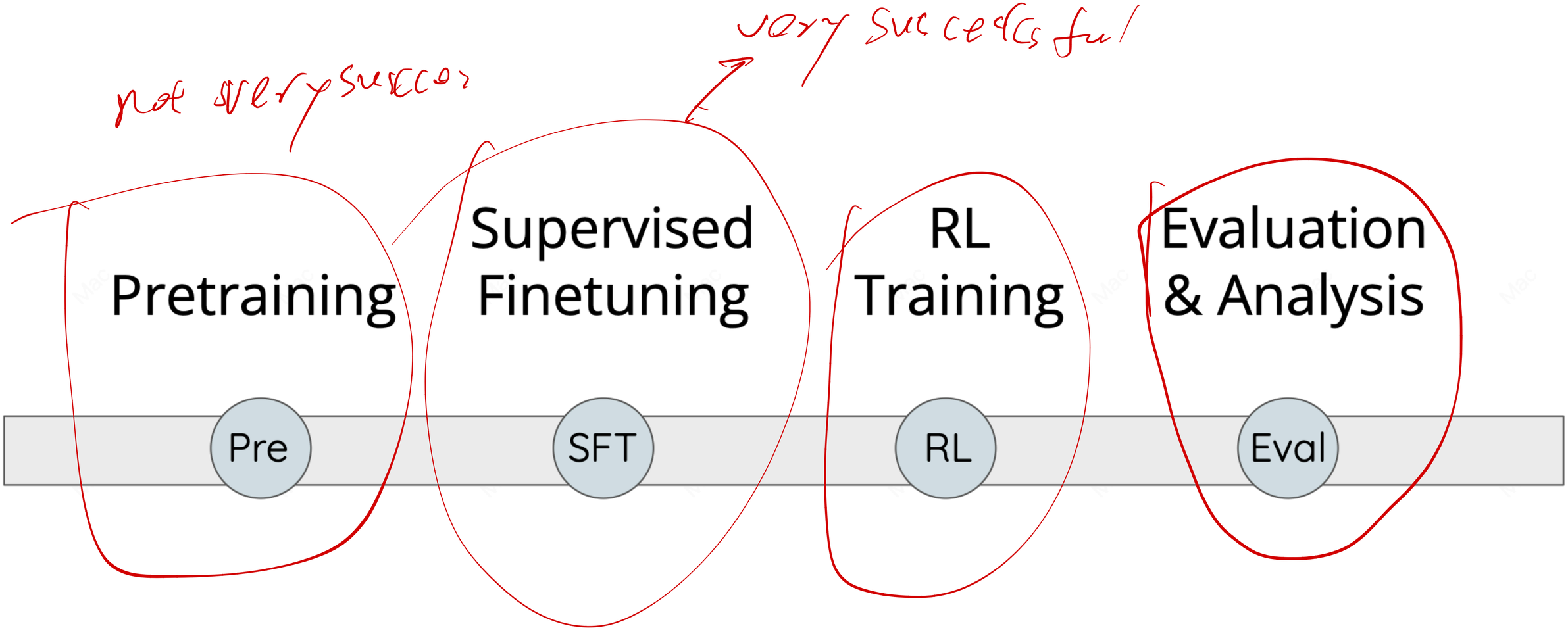
Diversity Filtering

Filter similar examples as defined by

- Rouge-L (Self-Instruct; Impossible Distillation)
- Embedding similarity (QDIT, DiverseEvol, DEITA)
- Semantic tags (#InsTag)



How Do We Use Synthetic Data



Synthetic Data is Naturally Rooted in RL

Synthetic data in SFT are typically from other models, while synthetic data in RL are from the model to be optimized

policy model

Chu et al. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. 2025

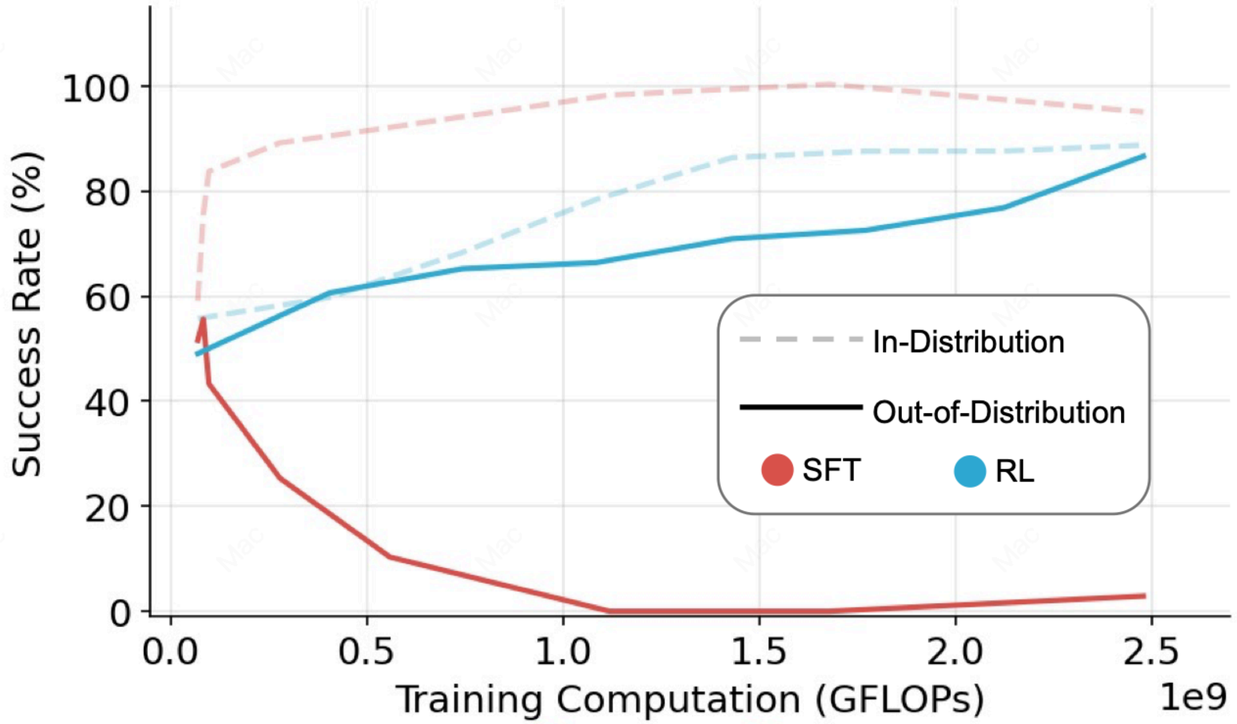
Synthetic Data is Naturally Rooted in RL

Synthetic data in SFT are typically from other models, while synthetic data in RL are from the model to be optimized What is the difference?

Chu et al. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. 2025

Synthetic Data is Naturally Rooted in RL

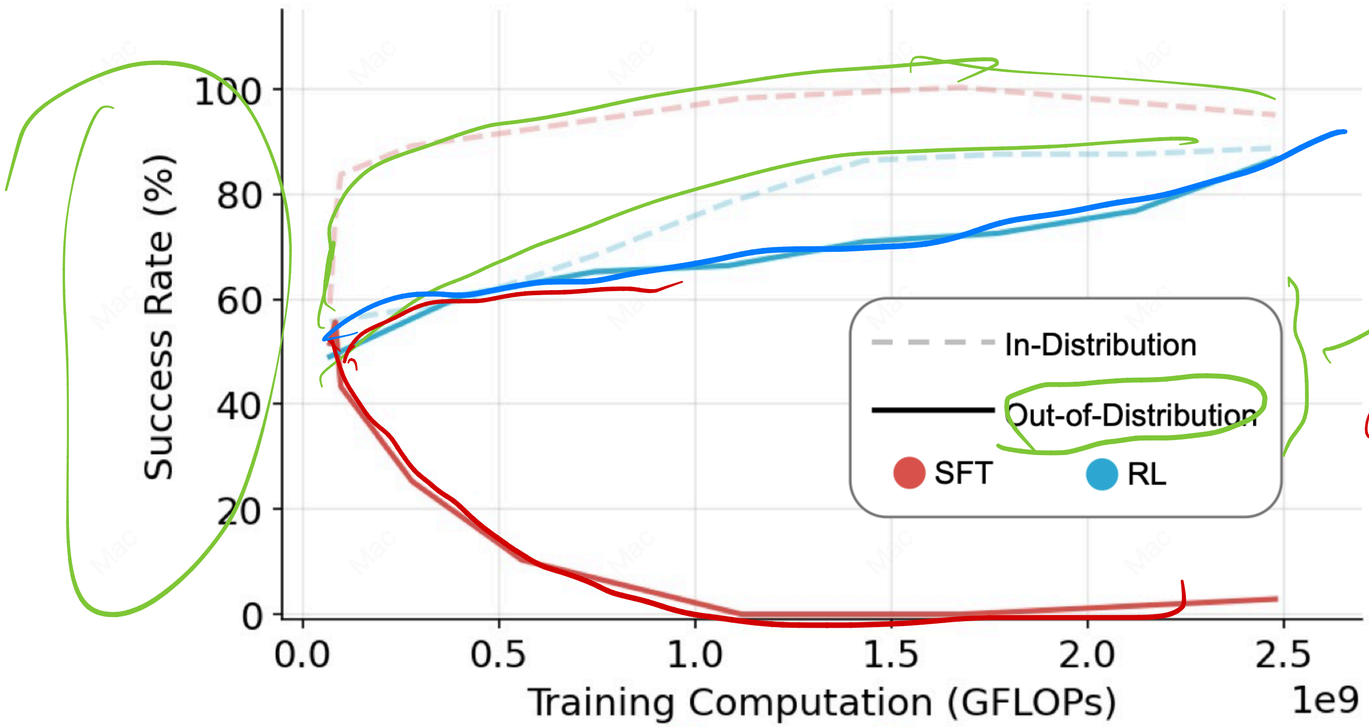
Synthetic data in SFT are typically from other models, while synthetic data in RL are from the model to be optimized What is the difference?



Chu et al. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. 2025

Synthetic Data is Naturally Rooted in RL

Synthetic data in SFT are typically from other models, while synthetic data in RL are from the model to be optimized What is the difference?



train: math

test:

in-distribu: math

ood-of distribution:

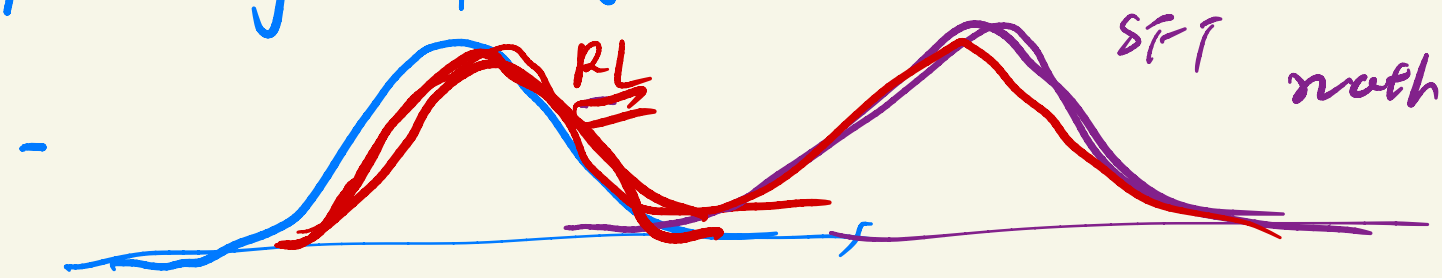
IFEval

SFT Memorizes, RL Generalizes:

Chu et al. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. 2025

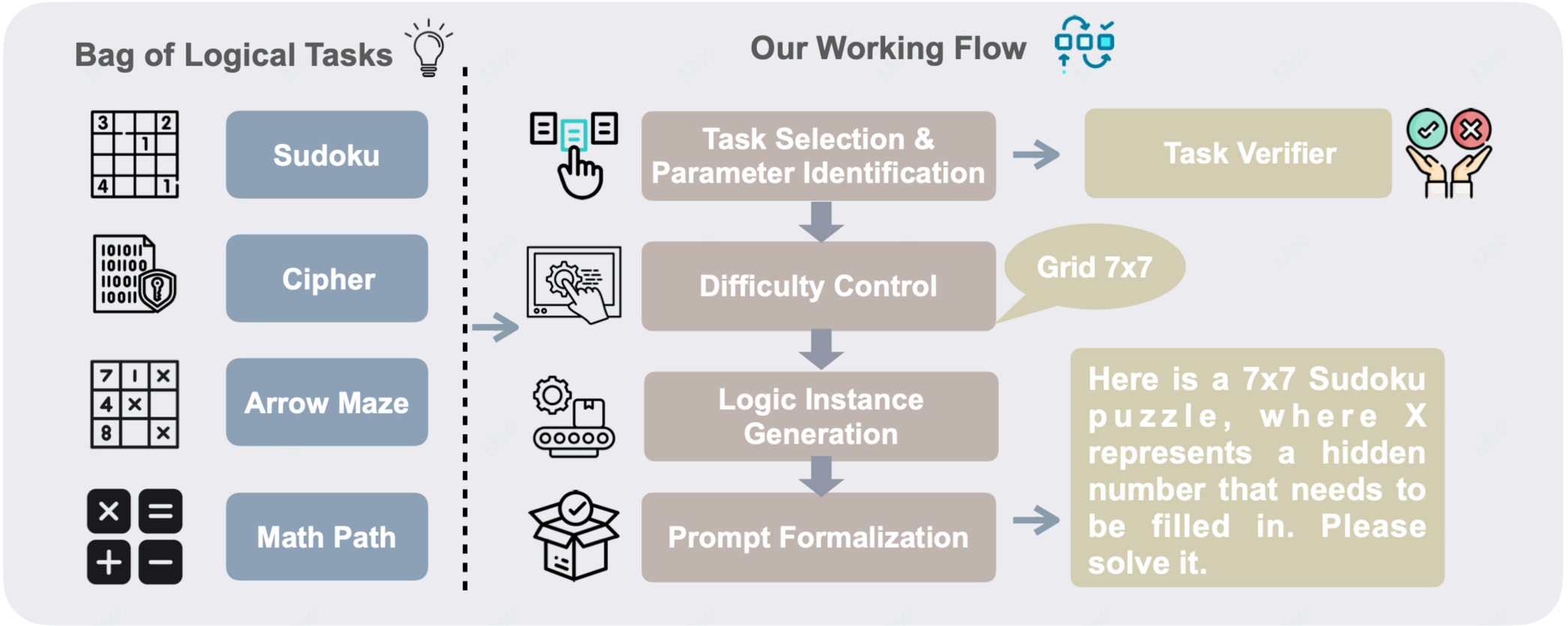
Why SFT memorize. RL generalizes

after pretraining:



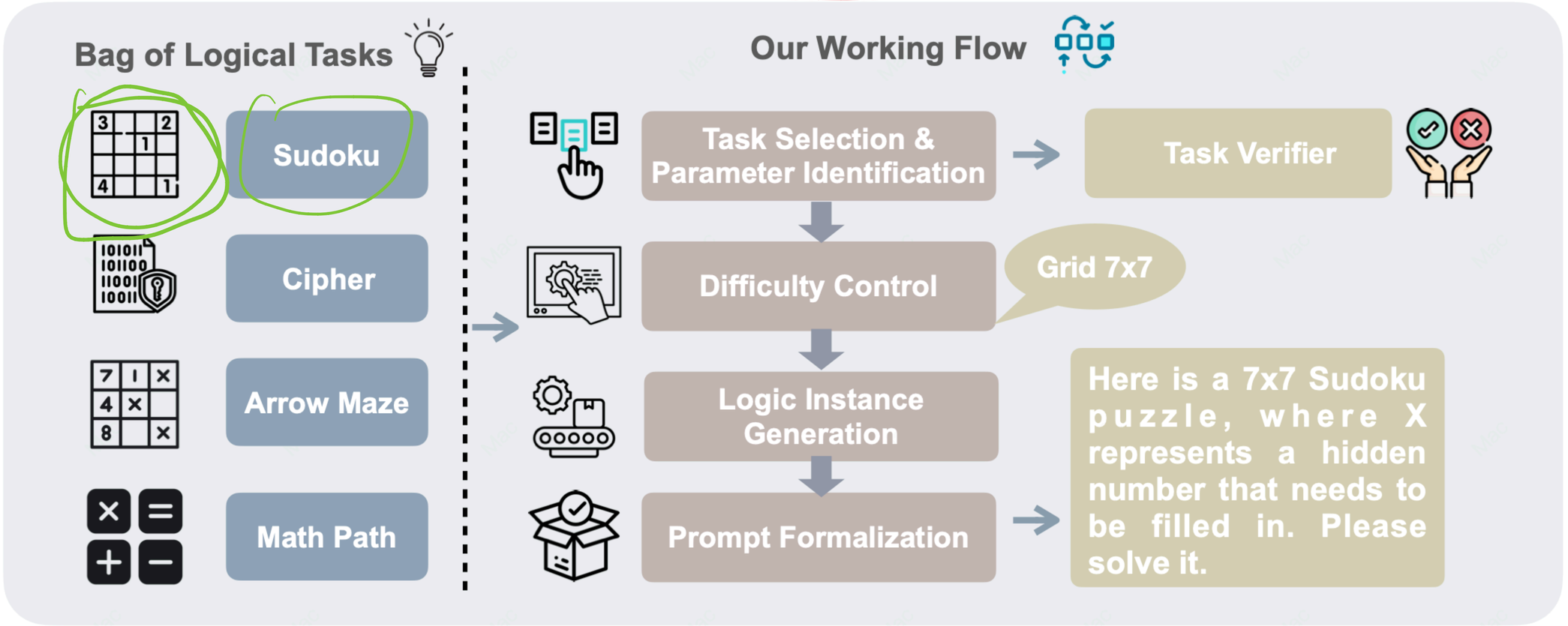
SFT: forget

Synthetic (query, answer) Pairs in RL



Liu et al. SynLogic: Synthesizing Verifiable Reasoning Data at Scale for Learning Logical Reasoning and Beyond. 2025

Synthetic (query, answer) Pairs in RL



RL only needs a query and verification script

Liu et al. SynLogic: Synthesizing Verifiable Reasoning Data at Scale for Learning Logical Reasoning and Beyond. 2025

Thank You!