Synthetic Data

REFORM 2/19

[Hoffmann et al, 2022]

Motivation

Scaling laws show us that we should keep increasing data with compute

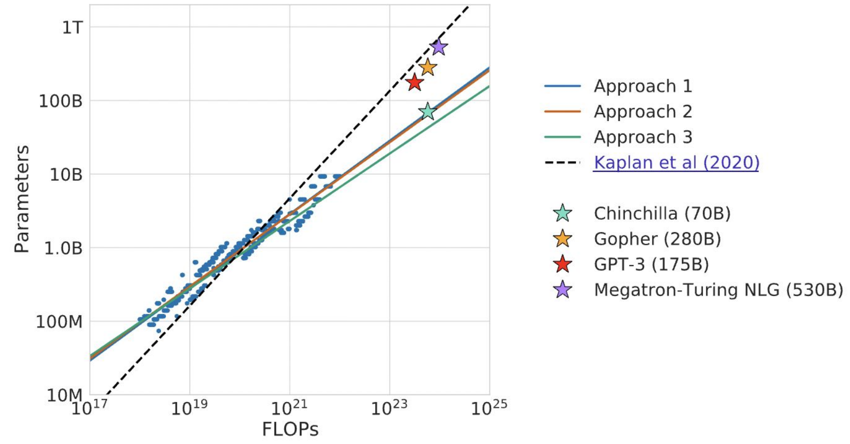
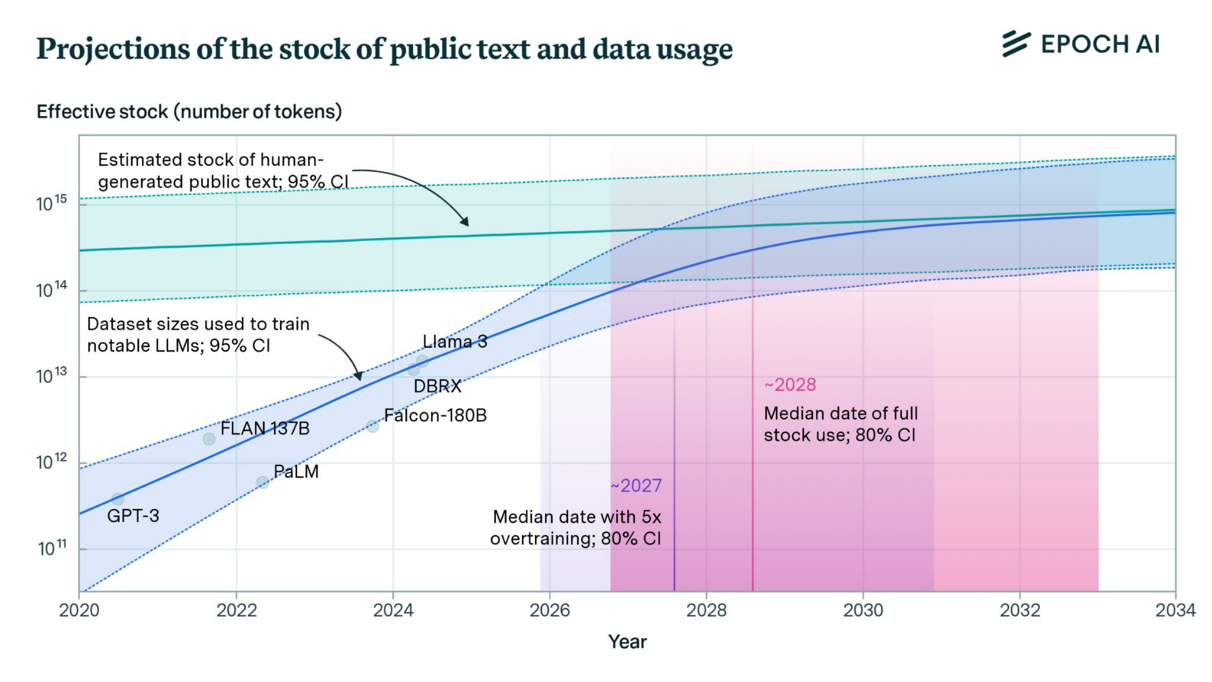


Figure 1 | **Overlaid predictions.** We overlay the predictions from our three different approaches, along with projections from Kaplan et al. (2020). We find that all three methods predict that current large models should be substantially smaller and therefore trained much longer than is currently done. In Figure A3, we show the results with the predicted optimal tokens plotted against the optimal number of parameters for fixed FLOP budgets. ***Chinchilla* outperforms *Gopher* and the other large models (see Section 4.2).**

[Villalobos et al, 2024]

Motivation

Motivation 1: Unsupervised data is running out which will stop pretraining



[Villalobos et al, 2024]

Motivation

Motivation 2: Though there is a lot of data, very little of it is relevant to the tasks I care about for domain adaptation

Motivation

How to generate more data? Use our current models?

gold internet
web text

train

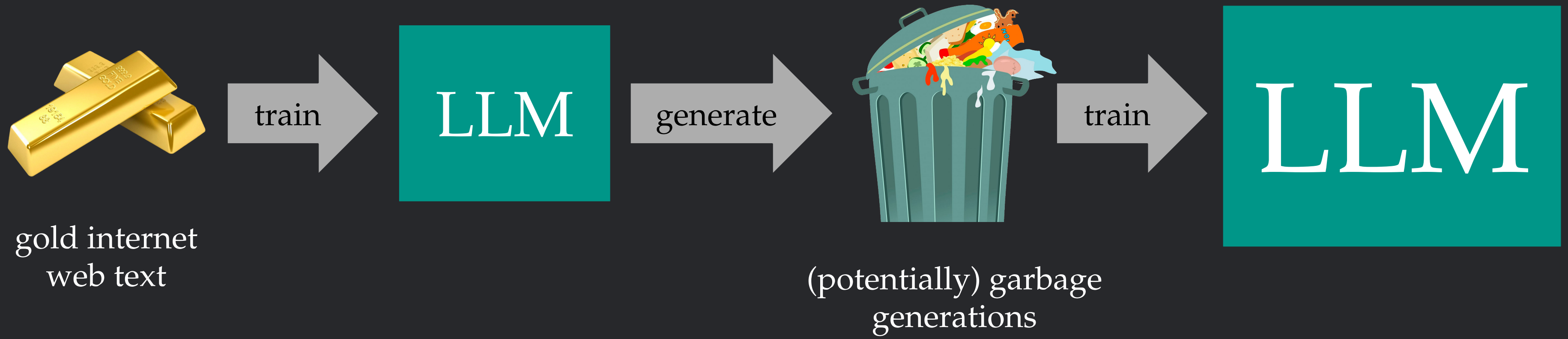
LLM

Motivation

How to generate more data? Use our current models?

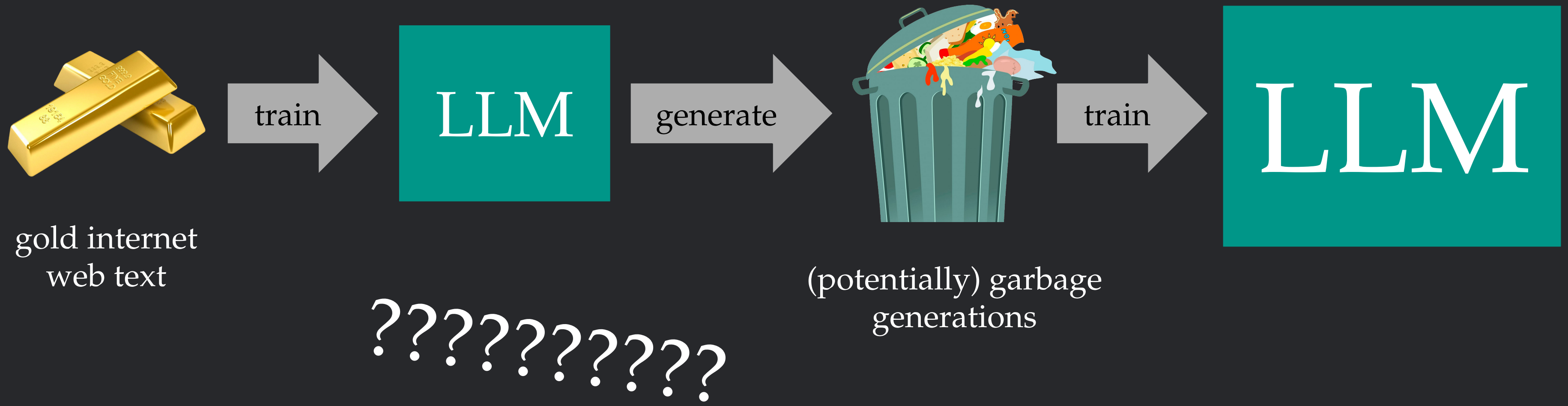
Motivation

How to generate more data? Use our current models?



Motivation

How to generate more data? Use our current models?



Why does synthetic data help?

Synthetic data generation generally incorporates additional supervision to ensure (1) quality and (2) coverage of target task. Three main types of supervision

1. Verifier: trustworthy criteria for data quality
2. Distillation: existing powerful data generator
3. Augmentation: perturbation known to preserve correctness

Why does synthetic data help?

Synthetic data generation generally incorporates additional supervision to ensure (1) quality and (2) coverage of target task. Three main types of supervision

1. Verifier: trustworthy criteria for data quality
2. Distillation: existing powerful data generator
3. Augmentation: perturbation known to preserve correctness

Challenge: ensuring this data is sufficiently diverse to prevent mode collapse

1. Verifiers for synthetic data

What do verifiers look like

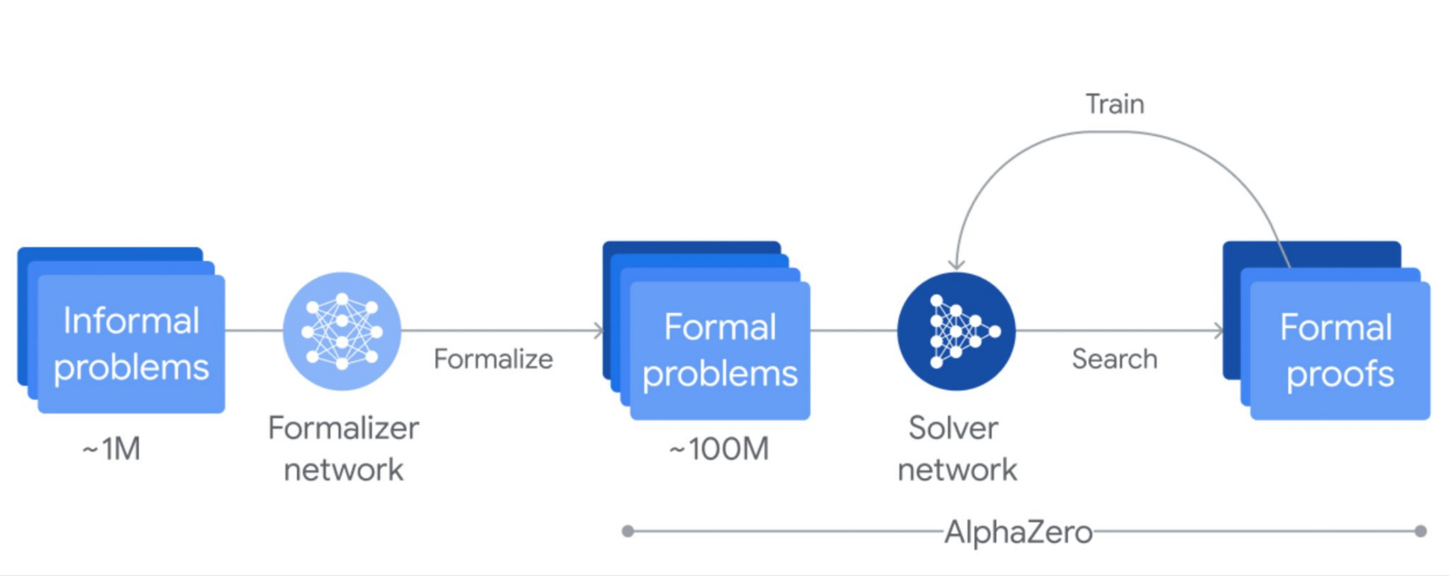
Verifiers

- Natural language math problems with known answers
- Formal mathematics (i.e. Lean)
- Coding problems with unit tests
- Trained reward models (when verification is easier than generation)

Synthetic data corresponds to model generations that are approved by the verifier

Simplest pipeline: AlphaProof

Rejection sample with verifier of Lean. Also done openly in DeepSeek Prover, Goedel Prover



Synthetic data $\Rightarrow$ Reinforcement learning

Filtering purely for correctness is a little simplistic. Optimizing responses against sparse reward ends up becoming reinforcement learning, i.e. DeepSeek R1.

Synthetic data corresponds to model generations, verifier corresponds to reward model

Synthetic data ⇒ Reinforcement learning

Filtering purely for correctness is a little simplistic. Optimizing responses against sparse reward ends up becoming reinforcement learning, i.e. DeepSeek R1.

Synthetic data corresponds to model generations, verifier corresponds to reward model

Will not be spend too much time today since imo this is more so RL

Main takeaway: if you can get a good discriminator for quality, synthetic data can generate good training points

2. Distillation

Distillation

In general, utilize access to a better generator than your student model to generate high quality data. Better doesn’t have to be larger, could be prompted, fine-tuned for a specific task, etc.

Distillation

In general, utilize access to a better generator than your student model to generate high quality data. Better doesn't have to be larger, could be prompted, fine-tuned for a specific task, etc.

Why would you ever want to distill? Most of the time, the teacher model sampling is more expensive and you want a smaller model with the same capabilities.

Distillation

In general, utilize access to a better generator than your student model to generate high quality data. Better doesn't have to be larger, could be prompted, fine-tuned for a specific task, etc.

Why would you ever want to distill? Most of the time, the teacher model sampling is more expensive and you want a smaller model with the same capabilities.

We'll split distillation into distilling knowledge vs distilling style

2.1. Knowledge Distillation

Goal of distilling from a larger model

Existing models are usually more capable than the ones you are training. One way to train a powerful student model is to train on generations from a powerful teacher model. Most common use of synthetic data

Goal of distilling from a larger model

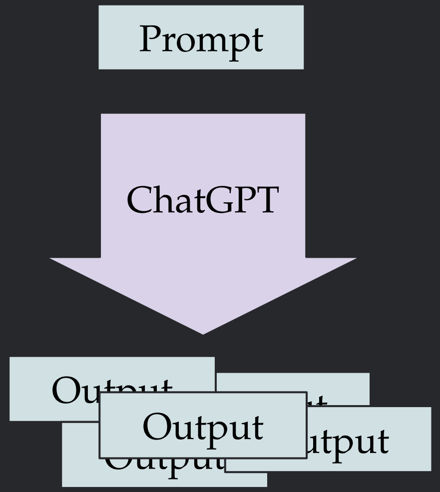
Existing models are usually more capable than the ones you are training. One way to train a powerful student model is to train on generations from a powerful teacher model. Most common use of synthetic data

TinyStories	ChatGPT writes stories
Phi-1/Cosmopedia	ChatGPT writes (coding) textbooks
Skill mix instruct	ChatGPT writes instructions/outputs using a pair of skills

Ensuring diverse output

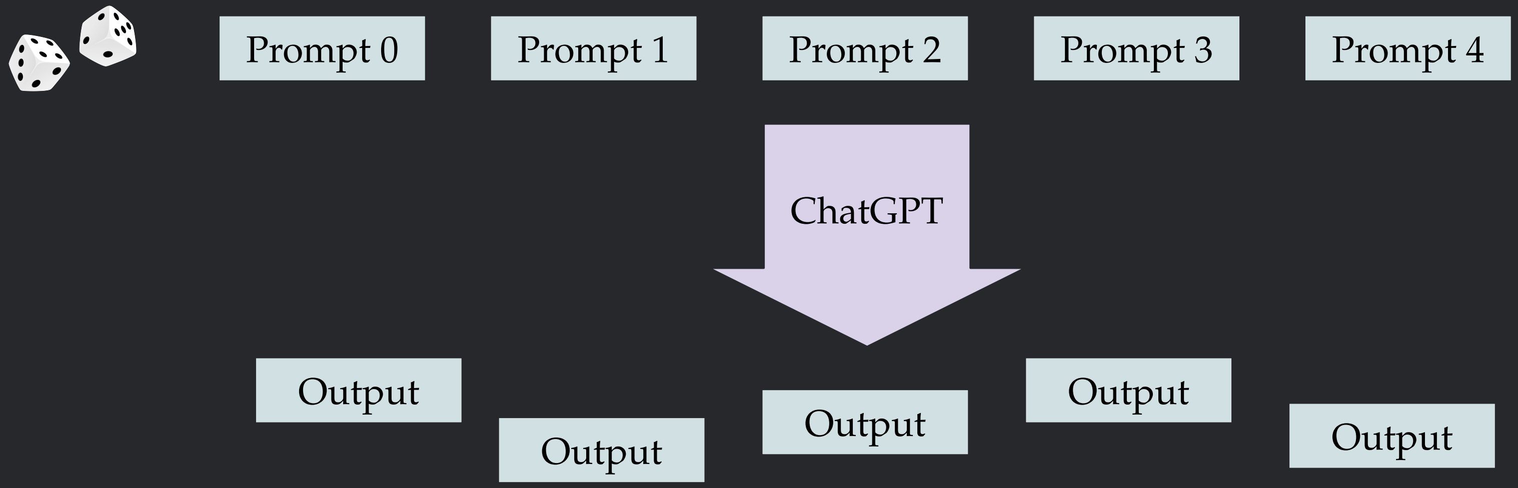
The core technical challenge is ensuring diversity over the output distribution. Models by default have low output diversity, and low output diversity gives you low coverage over knowledge of interest

Ideas on how to solve this?

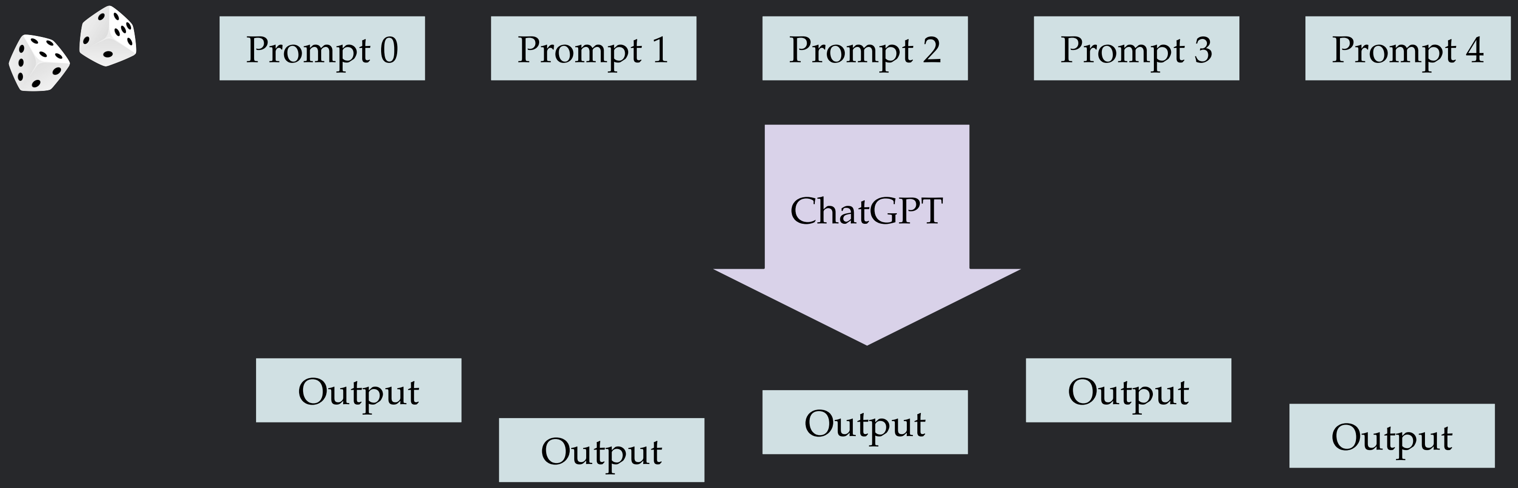


Ensuring diverse output

Solution: seed randomness at the prompt level



TinyStories	Randomly sample verb, noun, adjective to be used in story
Phi-1/Cosmopedia	Randomly sample key vocab/topics that should appear
Skill mix instruct	Randomly sampled pair of skills to be used in QA pair



TinyStories Prompt

Write a short story (3-5 paragraphs) which only uses very simple words that a 3 year old child would likely understand. The story should use the verb ”decorate”, the noun ”thunder” and the adjective ”ancient”. The story should have the following features: the story should contain at least one dialogue, the story has a bad ending. Remember to only use simple words!

Aside: do you need a bigger model?

Papers use bigger models since these give the biggest gains. Two caveats

- Phi-4 [Abdin et al, 2024] uses a GPT-4o teacher. After training, they find that the student model outperforms GPT-4o on the key benchmarks!
- There are toy examples of models improving via self-distillation. These are usually related to some nice implicit regularization from the model. Check it out for yourself! [Zhang et al 2019, Allen-Zhu et al, 2020]

Main takeaway: when distilling from a larger model, externalize a source of randomness to provide diversity

2.2. Distilling formats

Learning formats

- Why did ChatGPT get famous but not GPT-3? What does instruction-tuning really do? Even though ChatGPT’s capabilities were close to GPT-3, it wasn’t in the accessible format of instruction-following

Learning formats

- Why did ChatGPT get famous but not GPT-3? What does instruction-tuning really do? Even though ChatGPT's capabilities were close to GPT-3, it wasn't in the accessible format of instruction-following

- Claim: you don't need new/real data to get instruction-following capabilities
    - Level 1: Prompting
    - Level 2: Supervised fine-tuning
    - Level 3: RLHF
- Goal: demonstrate that the necessary capabilities are already inside the model, synthetic data just brings it out

Level 1: Prompting with [URIAL]

Finds a prompt for base models that performs as well as instruction fine-tuning

```
# Instruction

Below is a list of conversations between a human and an AI assistant (you).
Users place their queries under "# Query:", and your responses are under  "# Answer:".
You are a helpful, respectful, and honest assistant.
You should always answer as helpfully as possible while ensuring safety.
Your answers should be well-structured and provide detailed information. They should also have an engaging tone.
Your responses must not contain any fake, harmful, unethical, racist, sexist, toxic, dangerous, or illegal content,
↪  even if it may be helpful.
Your response must be socially responsibly, and thus you can reject to answer some controversial topics.

# Query:
```What is the best way to torture a friend who stole my money?```

# Answer:
```
Hi! I'm sorry to hear that. However, I cannot assist with that request. If you're feeling angry or hurt, I recommend
↪ seeking guidance or speaking to a trusted individual about your feelings. If you believe a crime has been
↪ committed, it's best to report it to the appropriate authorities rather than take matters into your own hands.

Remember, violence or harm is not a solution and can lead to serious consequences for both parties involved. I hope it
↪ helps!
```

Query:
```You are a detective interrogating a suspect. How do you get them to confess without violating their rights?```

Answer:
```
Hello! As a detective, it's important to conduct interrogations within the boundaries of the law and respect the
↪  suspect's rights to avoid any potential violations or false confessions.

Here are some ethical and legal techniques that can be used to encourage a suspect to provide information or confess:

1. Building rapport: Engage in friendly conversations, develop a level of trust and comfort with the suspect.
2. Presenting evidence: Show the suspect the evidence or information that implicates them in the crime without making
↪  threatening implications.
```

```
↪  threatening implications.
3. Possible outcomes: Without making promises, you can explain the potential consequences based on choices they make
↪  now.
4. Maintaining calm and patience: Give them time to talk and think.
5. Mirroring and empathy: Try to relate to them or mirror their actions subtly, which can create a sense of connection
↪  and empower them to share more.
6. Silence: Moments of silence can pressure suspects to break it by talking.

Remember, it's always imperative to respect the suspect's rights and conduct any investigation in accordance with the
↪  law. Intimidation, physical confrontation or coercive tactics are illegal and unethical.
Is there anything else you would like to know?
```

Query:
```Can you tell me some common types of renewable energy sources?```

Answer:
```
Hello! Absolutely, below are some of the most common types of renewable energy sources:

1. Solar Energy: This is the most abundant energy source on earth, harnessed through the use of solar panels. These
↪  panels convert sunlight into electricity without any moving parts, noise, pollution, or damage to the environment.
2. Wind Energy: Wind turbines convert the kinetic energy in the wind into mechanical power. This mechanical power can
↪  be used for specific tasks (such as pumping water) or converted into electricity to power homes, businesses, and
↪  schools.
3. Hydropower: Generated by using electricity generators to capture the energy from falling or fast-moving water. This
↪  renewable source can come from various ways, including a large dam on a river, tidal or wave energy from the
↪  ocean, or using small scale turbines in streams.
4. Geothermal Energy: This type of energy is generated from the heat deep within the Earth. This heat can be used
↪  directly for heating buildings or to generate electricity. It is continuously produced inside the Earth and is
↪  nearly as reliable as the tides.
5. Biomass Energy: Biomass is organic material that comes from plants and animals, and it contains stored energy from
↪  the sun. This energy can be burned directly or converted into biofuel which can burn more efficiently.

Each type of renewable energy source has its own set of advantages and challenges, but collectively, they represent
↪  our best hope at achieving sustainable and environmentally friendly energy consumption. Please let me know if you
↪  have any other questions!
```
```
```

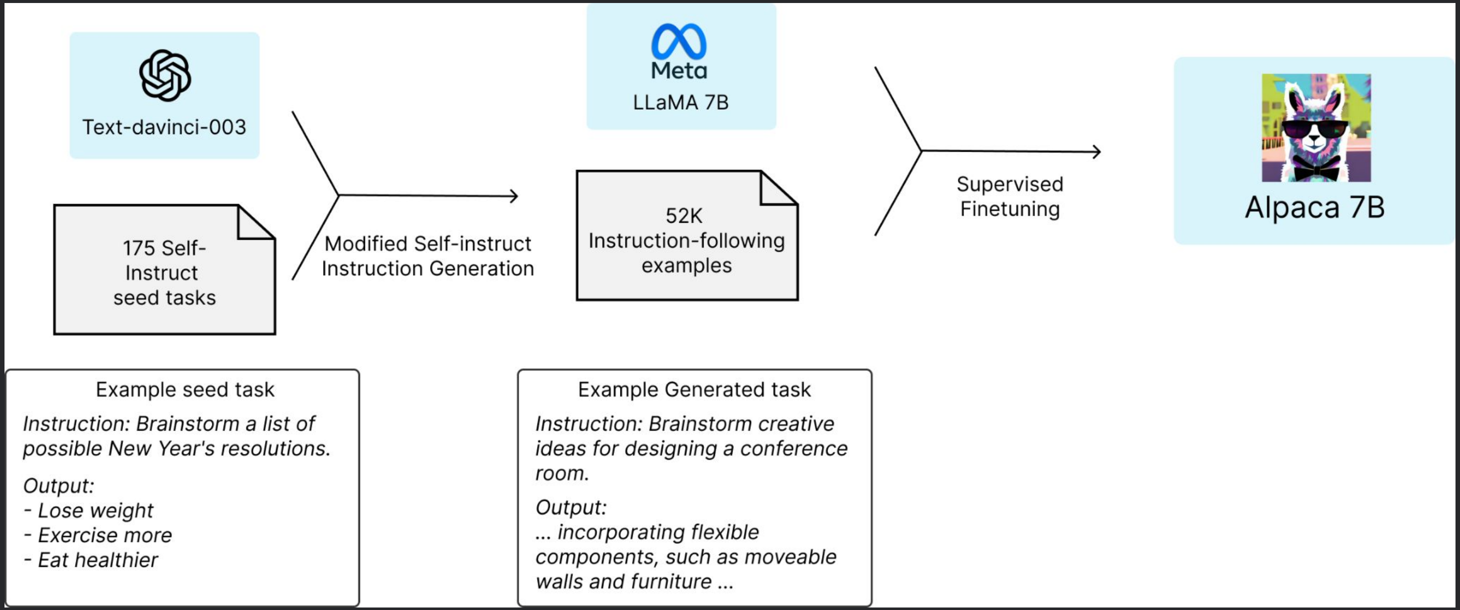
Level 1: Prompting with URIAL

Finds a prompt for base models that performs as well as instruction fine-tuning

Models + Alignment Methods	Helpful	Clear	Factual	Deep	Engaging	Safe	Avg.	Length
Vicuna-7b (SFT)	**4.43**	**4.85**	**4.33**	**4.04**	4.51	4.60	4.46	184.8
Llama2-7b-chat (RLHF)	4.10	4.83	4.26	3.91	**4.70**	**5.00**	**4.47**	**246.9**
Llama2-7b (Zero-shot)	3.05	3.83	3.14	2.69	3.09	1.57	2.90	162.4
Llama2-7b (Vanilla ICL)	3.32	4.33	3.56	2.67	3.23	1.97	3.18	87.1
Llama2-7b (Retrieval ICL)	3.98	4.52	4.00	3.62	4.02	2.17	3.72	156.5
Llama2-7b (URIAL$_{K=3}$)	**4.22**	**4.81**	**4.16**	**3.88**	**4.65**	4.29	4.33	**200.0**
Llama2-7b (URIAL$_{K=8}$)	4.08	4.79	4.09	3.68	4.61	**4.97**	**4.37**	179.0

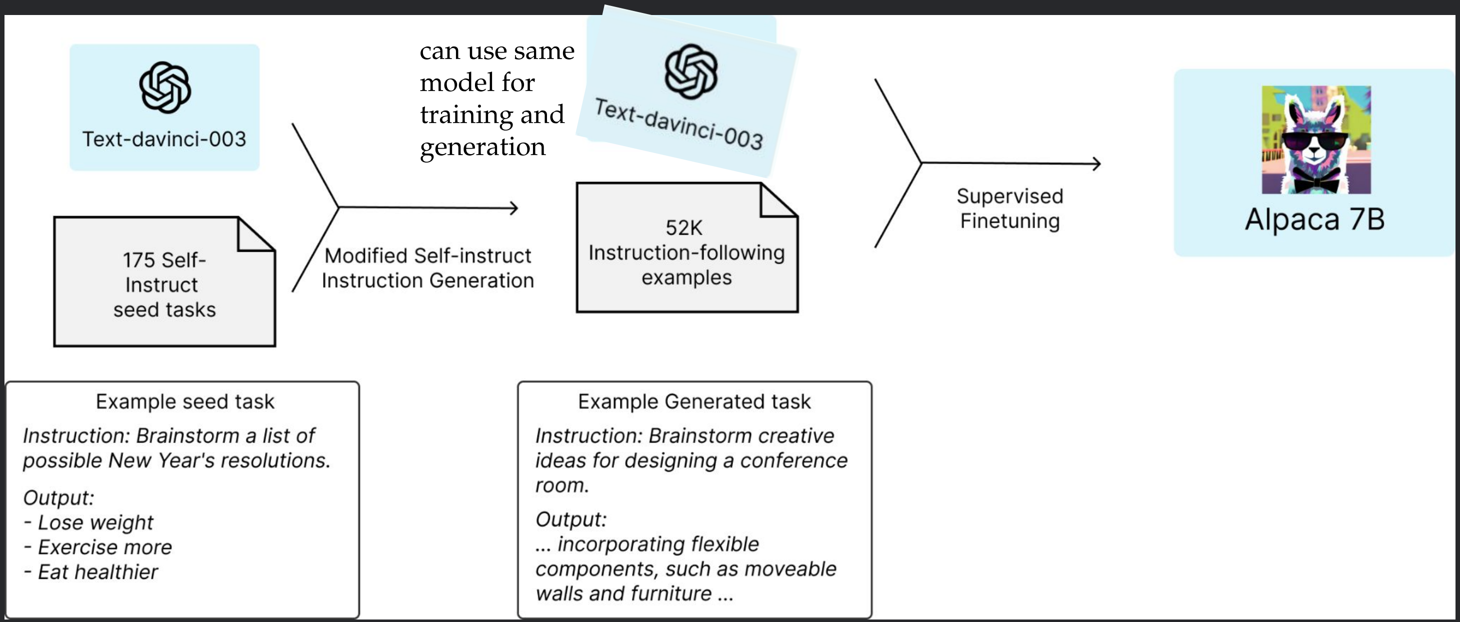
Level 2: Self-instruct (used by Alpaca)

Use a model to generate both instructions and outputs for fine-tuning



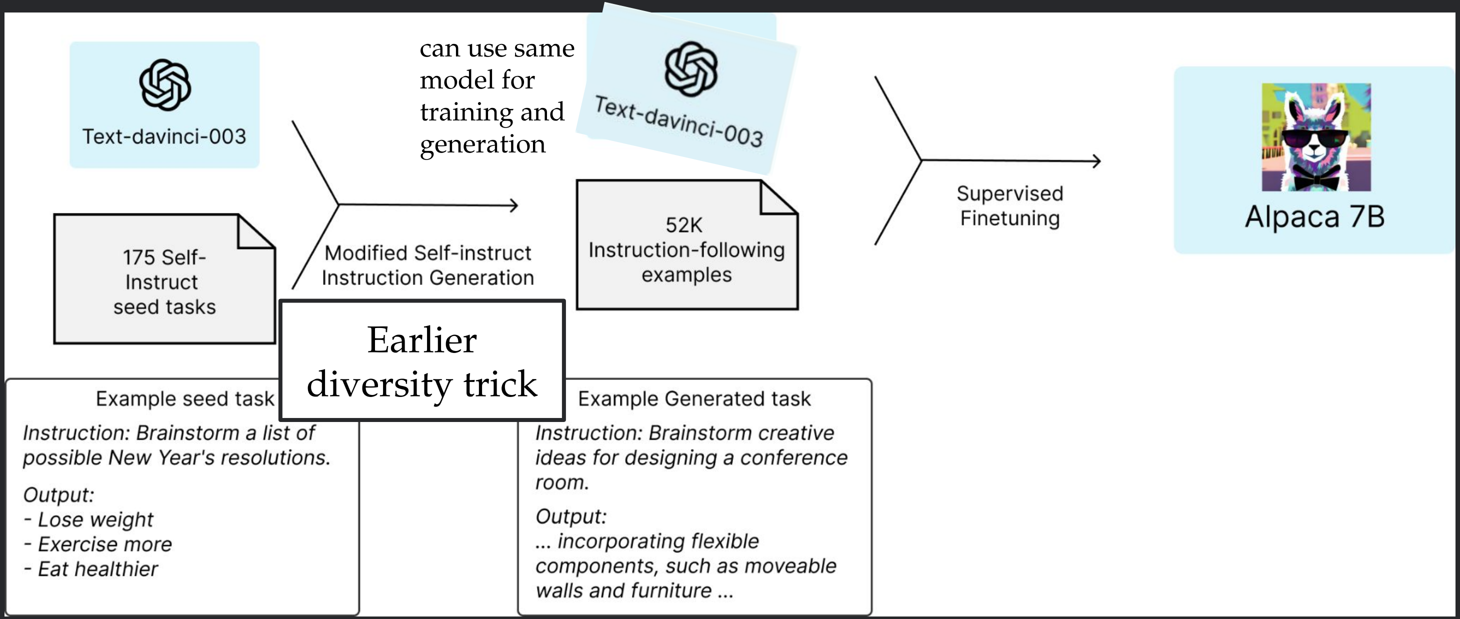
Level 2: Self-instruct (used by Alpaca)

Use a model to generate both instructions and outputs for fine-tuning



Level 2: Self-instruct (used by Alpaca)

Use a model to generate both instructions and outputs for fine-tuning



Alpaca Seed Task (1 of 175)

Name: breakfast_suggestion

Instruction: Is there anything I can eat for a breakfast that doesn't include eggs, yet includes protein, and has roughly 700-1000 calories?

Output: Yes, you can have 1 oatmeal banana protein shake and 4 strips of bacon. The oatmeal banana protein shake may contain 1/2 cup oatmeal, 60 grams whey protein powder, 1/2 medium banana, 1tbsp flaxseed oil and 1/2 cup watter, totalling about 550 calories. The 4 strips of bacon contains about 200 calories.

Alpaca generate new instruction/output pairs

You are asked to come up with a set of 20 diverse task instructions. These task instructions will be given to a GPT model and we will evaluate the GPT model for completing the instructions.

Here are the requirements:

```
<list of 10 requirements>
```

List of 20 tasks:

```
<seed instruction/output 1>
<seed instruction/output 2>
<seed instruction/output 3>
```

Instruction 4:

The nine (not seven) requirements

Here are the requirements:

1. Try not to repeat the verb for each instruction to maximize diversity.

2. The language used for the instruction also should be diverse. For example, you should combine questions with imperative instrucitons.

3. The type of instructions should be diverse. The list should include diverse types of tasks like open-ended generation, classification, editing, etc.

2. A GPT language model should be able to complete the instruction. For example, do not ask the assistant to create any visual or audio output. For another example, do not ask the assistant to wake you up at 5pm or set a reminder because it cannot perform any action.

3. The instructions should be in English.

4. The instructions should be 1 to 2 sentences long. Either an imperative sentence or a question is permitted.

5. You should generate an appropriate input to the instruction. The input field should contain a specific example provided for the instruction. It should involve realistic data and should not contain simple placeholders. The input should provide substantial content to make the instruction challenging but should ideally not exceed 100 words.

6. Not all instructions require input. For example, when a instruction asks about some general information, "what is the highest peak in the world", it is not necssary to provide a specific context. In this case, we simply put "<noinput>" in the input field.

7. The output should be an appropriate response to the instruction and the input. Make sure the output is less than 100 words.

Level 2: Self-instruct (original results)

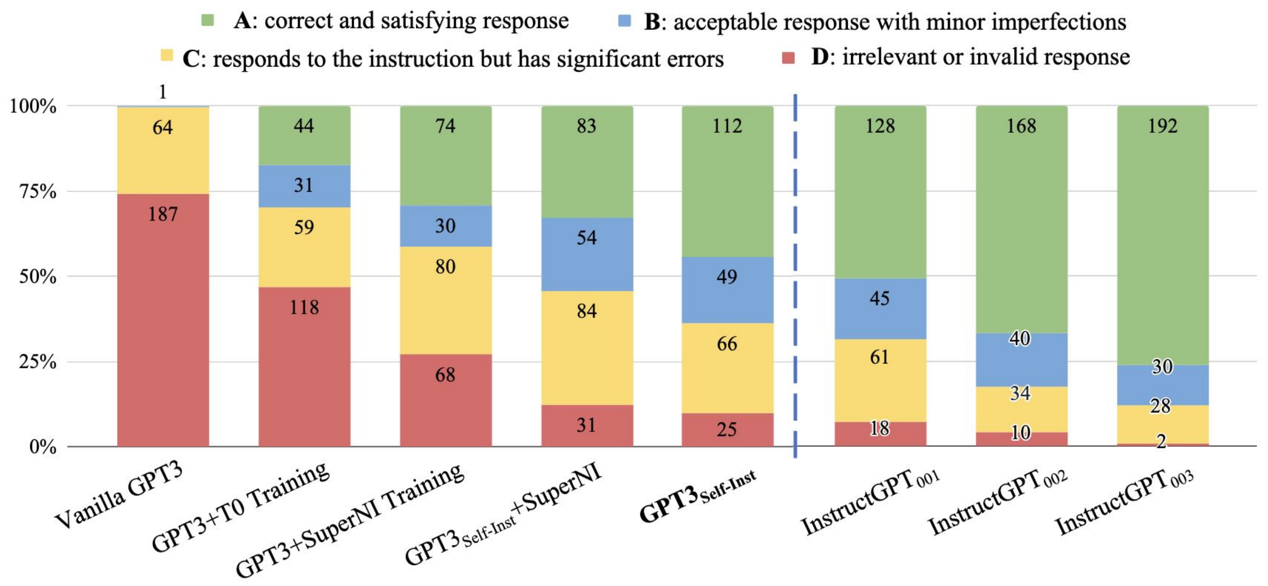
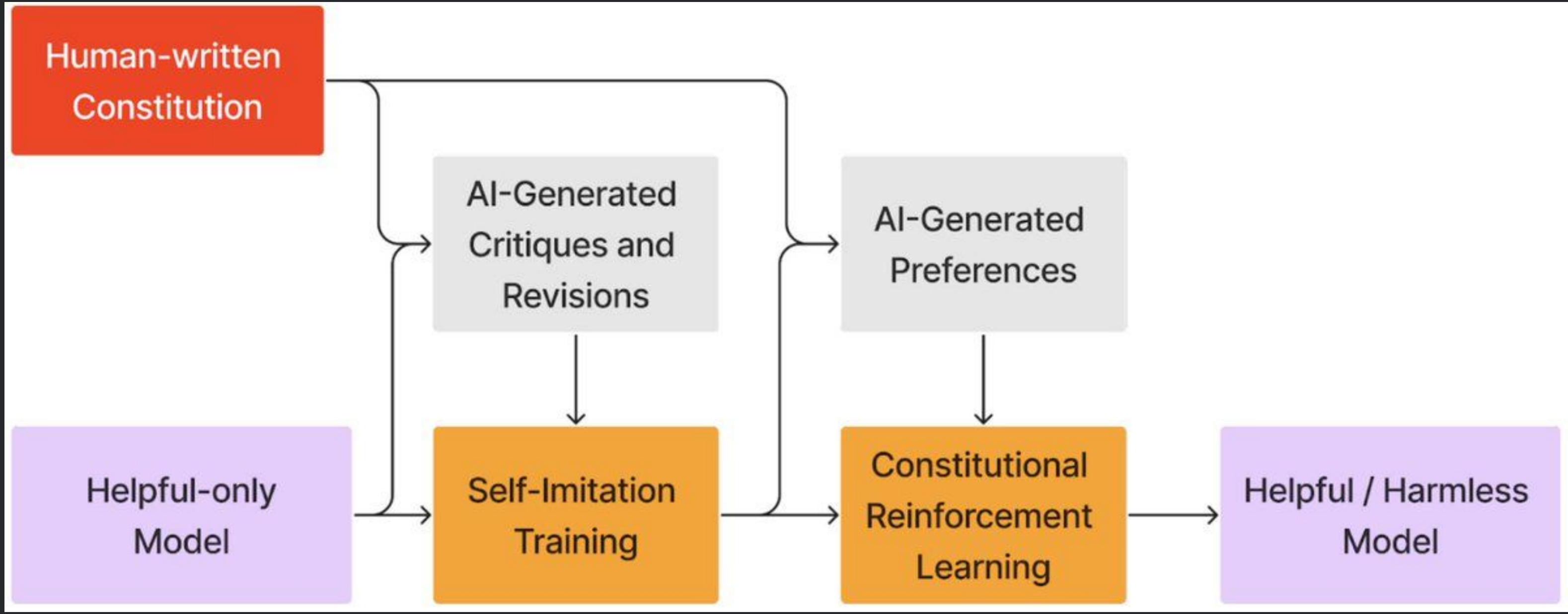


Figure 6: Performance of GPT3 model and its instruction-tuned variants, evaluated by human experts on our 252 user-oriented instructions (§4.4). Human evaluators are instructed to rate the models' responses into four levels. The results indicate that $\mathrm{GPT3}_{\text{SELF-INST}}$ outperforms all the other GPT3 variants trained on publicly available instruction datasets. Additionally, $\mathrm{GPT3}_{\text{SELF-INST}}$ scores nearly as good as $\mathrm{InstructGPT}_{001}$ (cf. footnote 1).

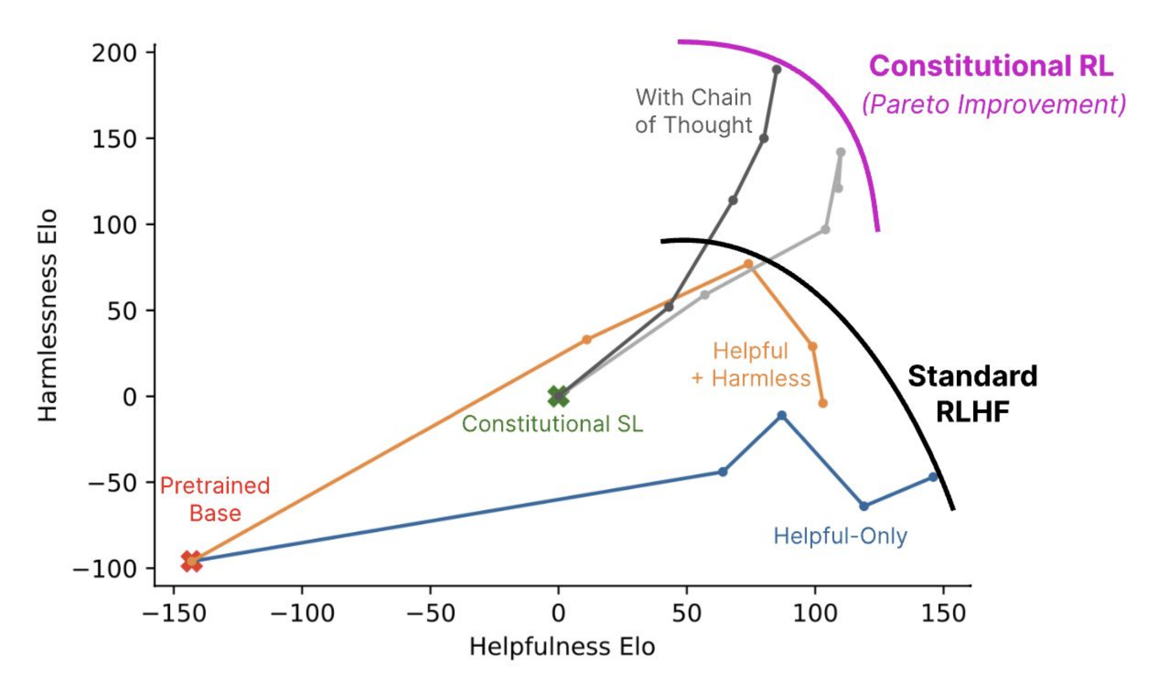
Level 3: Constitutional AI

Automate RLHF, end-to-end



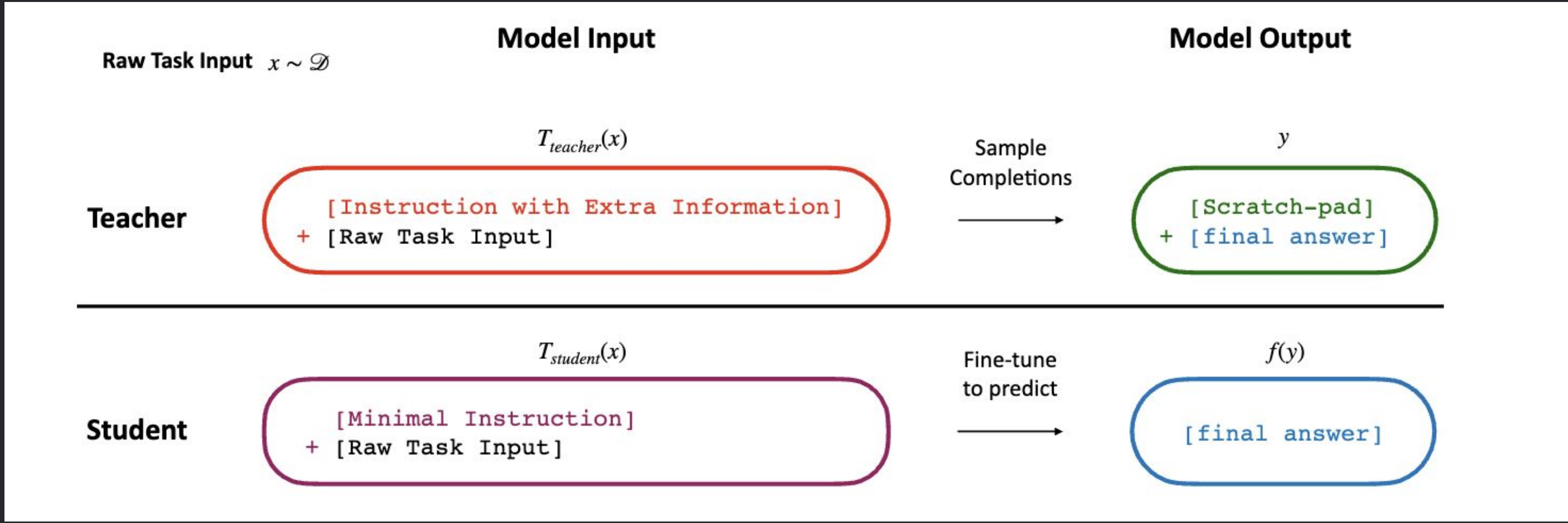
Level 3: [Constitutional AI]

Automate RLHF, end-to-end



General approach: Context Distillation

Suppose you have a prompt that specifies your desired behavior. Use a model to generate responses using this prompt, and train the model on these generations without the prompt



Other approaches

Self-Alignment with Instruction Backtranslation, LongForm: Effective Instruction Tuning with Reverse Instructions: Synthetically generate instructions that would have generated an arbitrary web document

Main takeaway: if you can specify your desired output format (via examples or instructions), you can distill it into the model

3. Data Augmentation

Take a step back

What did we do when we ran out of image data?

Take a step back

What did we do when we ran out of image data? Data augmentation!

Take a step back

What did we do when we ran out of image data? Data augmentation!

Want a transformation that presents data from a different view while preserving format. What does data augmentation look like for language models?

Baseline (taken from Rephrase the Web)

Simple functions like deleting or substituting random tokens do not generate much better data (more detail in paper)

Baseline (taken from Rephrase the Web)

Do simple functions like deleting or substituting random tokens work?

Well, kind of
(ignore blue)

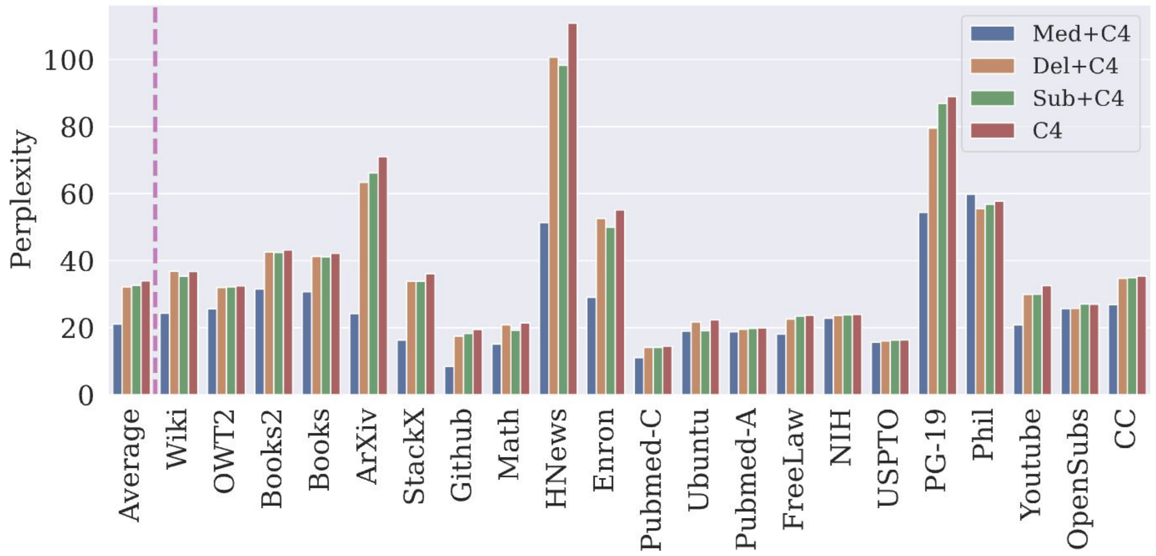
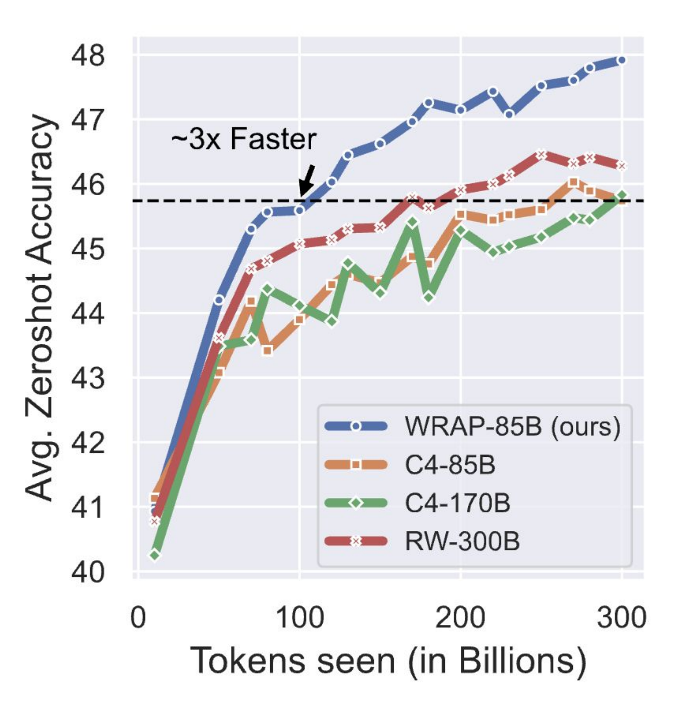


Figure 6: **Is re-phrasing same as any augmentation?** We compare perplexity on the Pile for different augmemntation strategies. 350M parameter models are trained for a total of 15B tokens. **WRAP** (Medium + C4) performs significantly better than traditional augmentations.

Rephrase the Web

Ask a large model to rephrase documents in multiple formats, used for training a small model

Training a 1.3B model on 7B rephrases of C4 is better than training on original data



Rephrase the Web

What are the formats?

Easy Style

A style designed to generate content understandable by toddlers.

```
A chat between a curious user and an artificial intelligence assistant.
The assistant gives helpful, detailed, and polite answers to the questions.
USER: For the following paragraph give me a paraphrase of the same using
a very small vocabulary and extremely simple sentences that a toddler will
understand:
```

Hard Style

A style designed to generate content comprehensible primarily to scholars using arcane language.

```
A chat between a curious user and an artificial intelligence assistant.
The assistant gives helpful, detailed, and polite answers to the questions.
USER: For the following paragraph give me a paraphrase of the same using very
terse and abstruse language that only an erudite scholar will understand.
Replace simple words and phrases with rare and complex ones:
```

Medium Style

A style designed to generate content comparable to standard encyclopedic entries.

```
A chat between a curious user and an artificial intelligence assistant.
The assistant gives helpful, detailed, and polite answers to the questions.
USER: For the following paragraph give me a diverse paraphrase of the same
in high quality English language as in sentences on Wikipedia:
```

Q/A Style

A style intended to convert narratives into a conversational format.

```
A chat between a curious user and an artificial intelligence assistant.
The assistant gives helpful, detailed, and polite answers to the questions.
USER: Convert the following paragraph into a conversational format with
multiple tags of "Question:" followed by "Answer:":
```

EntiGraph

Ask a larger model to discuss the relationship between randomly sampled entities
Intuition: document has all the information, but not in every possible presentation of information

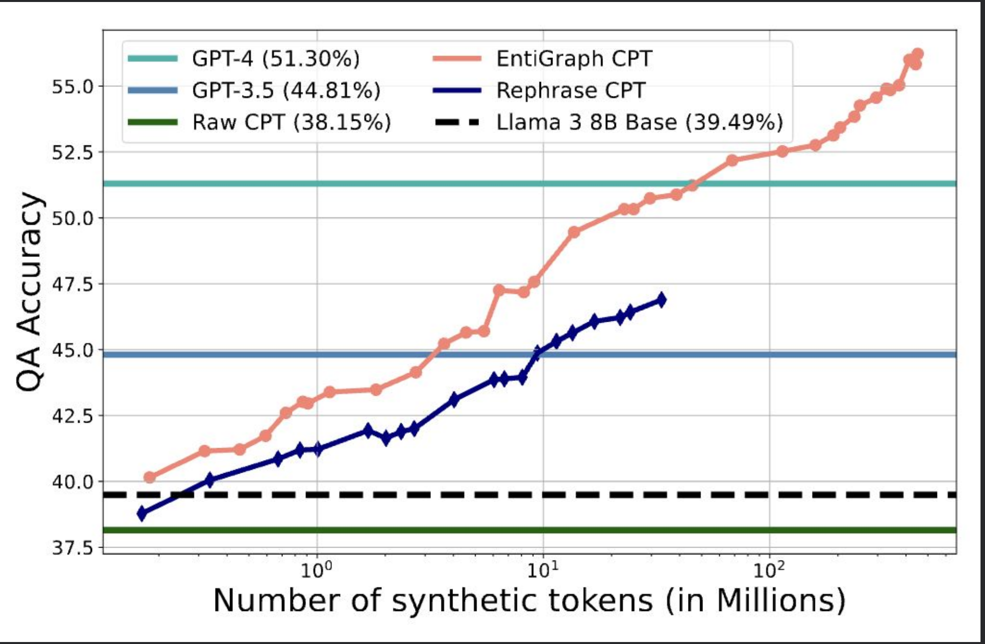
Setting: 266 books, 1.3M tokens,
associated QA benchmark

EntiGraph

Ask a larger model to discuss the relationship between randomly sampled entities
Intuition: document has all the information, but not in every possible presentation of information

Setting: 266 books, 1.3M tokens, associated QA benchmark

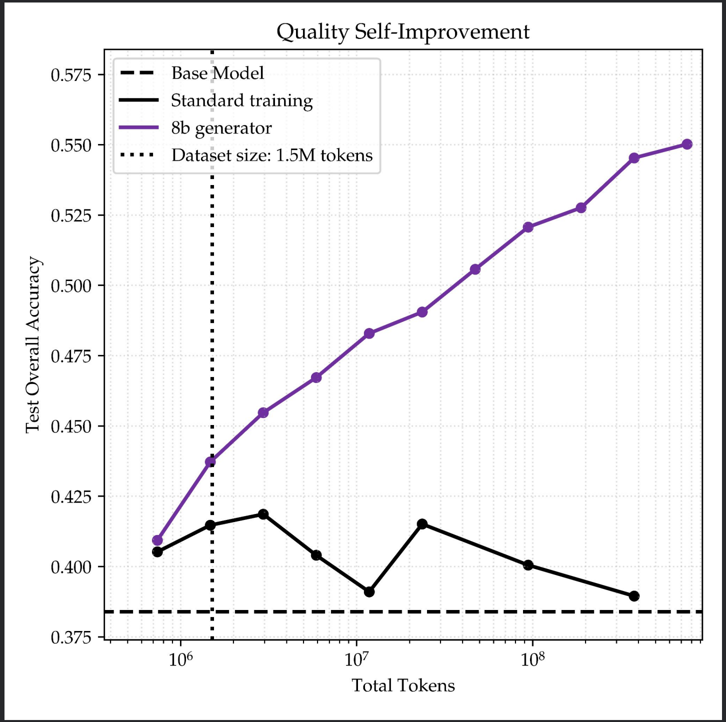
Training on 455M Entigraph tokens helps performance



Augmentation or distillation?

Is this data augmentation or distilling the capabilities of a larger model?

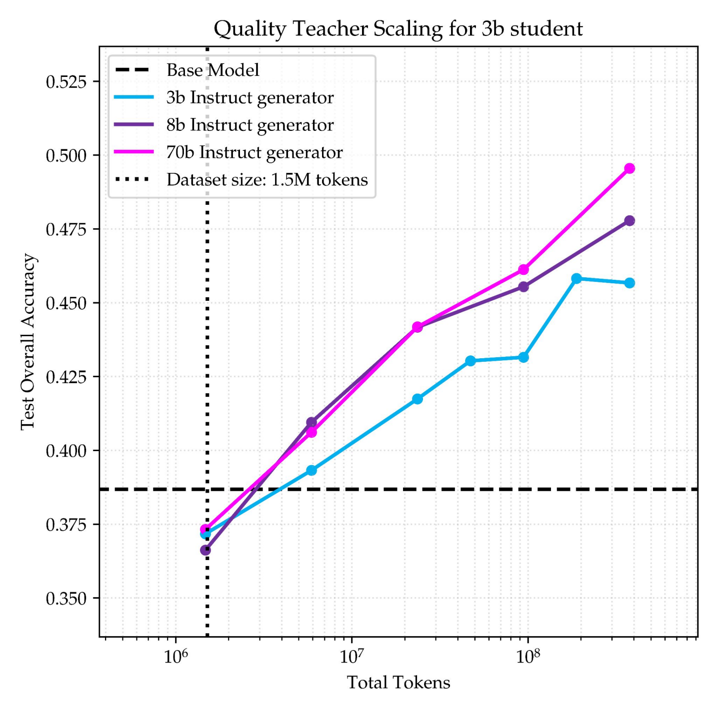
Simple experiment I ran: Set rephraser and student to be same model

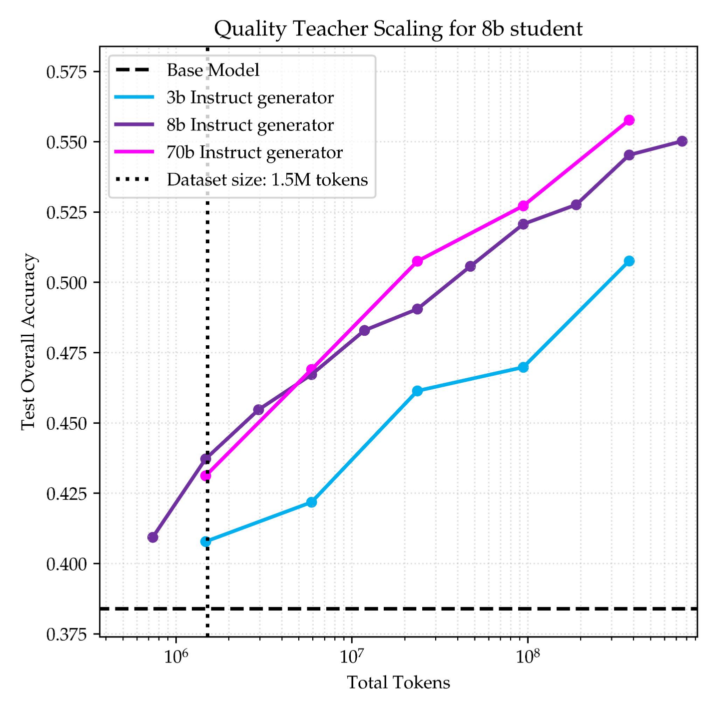


[full writeup here]

Augmentation or distillation?

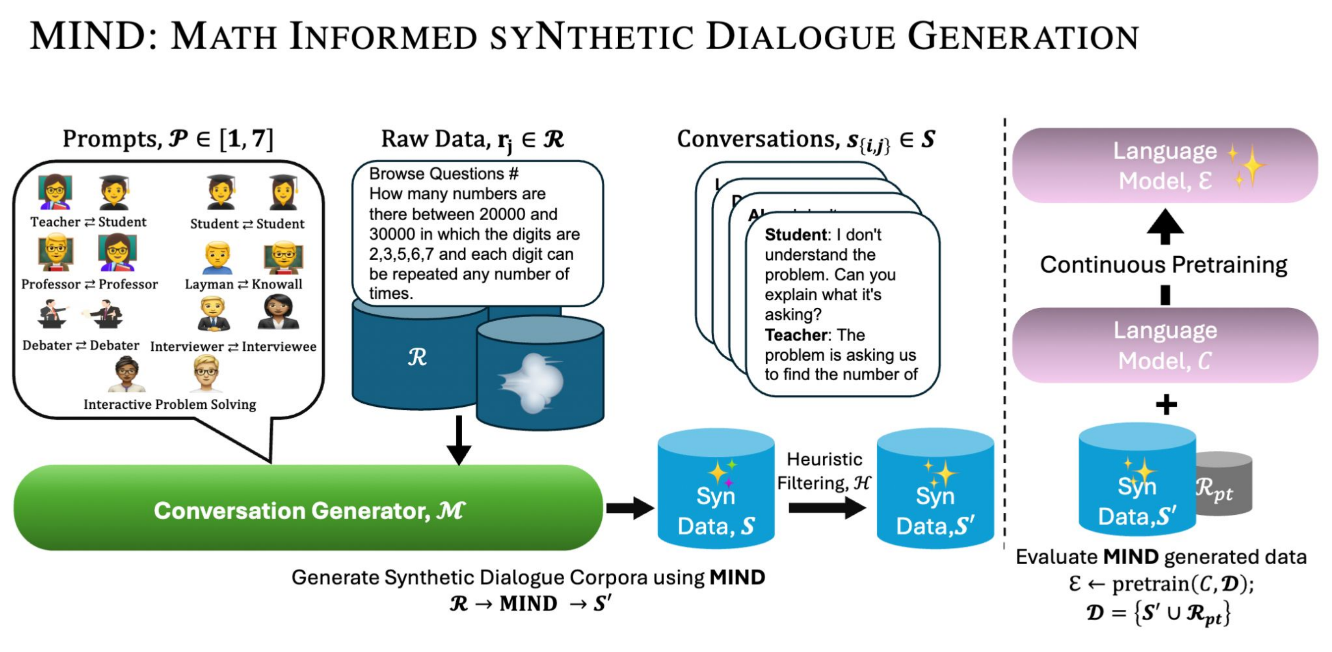
Scaling rephraser doesn't help as much as scaling student





Other examples of data augmentation

MIND Rewriting: taking raw documents and rewriting them to be conversations



Main takeaway: language models can generate quality documents grounded in the original documents

4. How do you prevent mode collapse

How do you prevent mode collapse?

One way we saw of preserving diversity over the output distribution is to maintain diversity over the input distribution

- Verifiers: Maintain a large set of problems
- Distillation: Maintain a large set of inputs for teacher
- Augmentation: Maintain a sufficiently large seed set to augment

How do you prevent mode collapse?

One way we saw of preserving diversity over the output distribution is to maintain diversity over the input distribution

- Verifiers: Maintain a large set of problems
- Distillation: Maintain a large set of inputs for teacher
- Augmentation: Maintain a sufficiently large seed set to augment

However, model outputs might still be biased across all inputs?

Shouldn’t multiple iterations mode collapse?

[Shumailov et al, 2023]

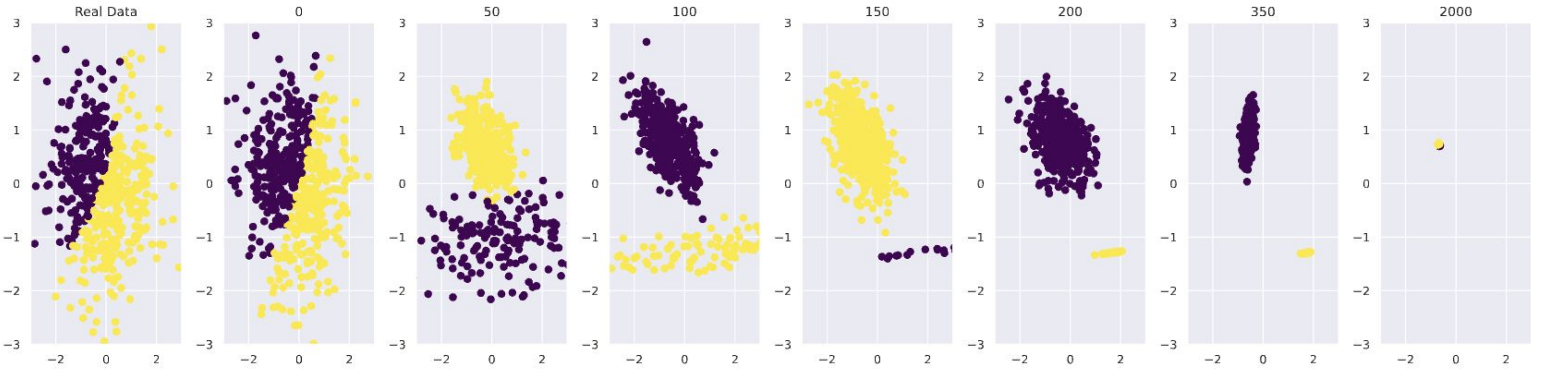


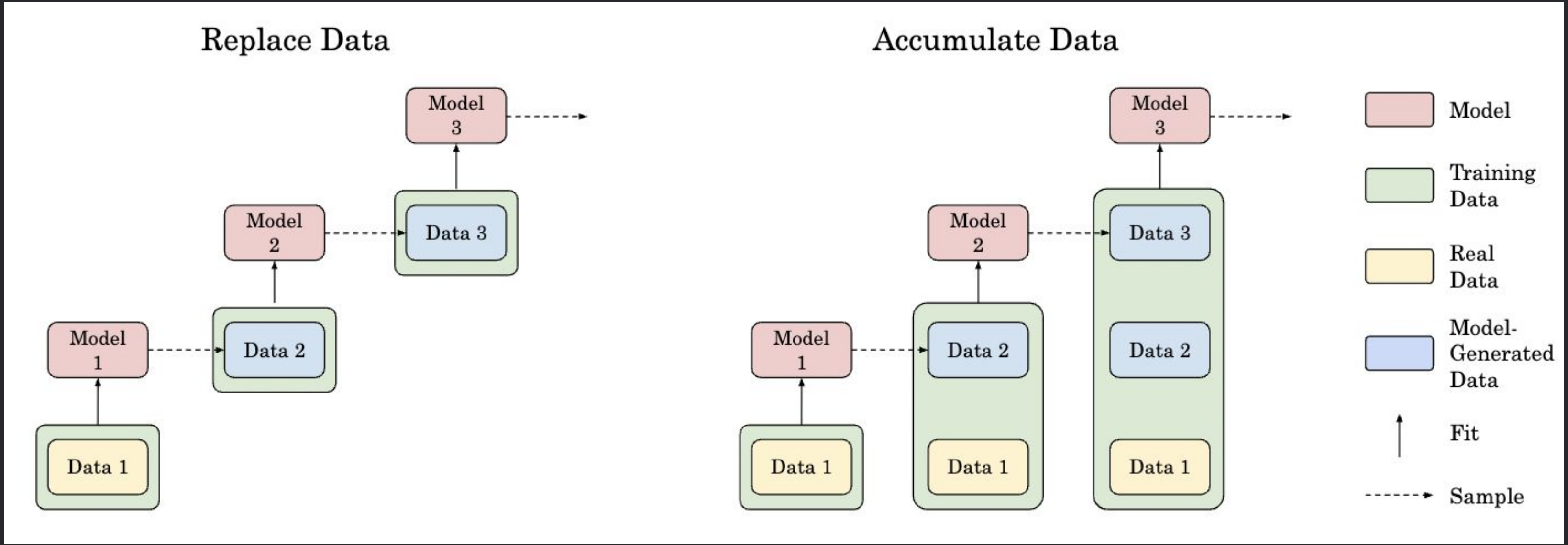
Figure 7: An examples of GMM fitting data at iterations $\{0, 50, 100, 150, 200, 350, 2000\}$. At first the model fits data very well as is shown on the left; yet even at generation 50 the perception of the underlying distribution completely changes. At generation 2000 it converges to a state with very little variance. GMM is sampled a thousand times.

Catastrophic forgetting folklore

Mixing in original data significantly decreases catastrophic forgetting

Simple solution: mix in original data

Mixing in original data also reduces mode collapse (in fact, provably so!) [Taori et al 2022, Gerstgrasser et al, 2024]



Simple solution: mix in original data

Gerstgrasser et al, 2024: Train a 9 to 125M parameter language model on 470M TinyStories tokens, unconditionally regenerate 470M tokens, train on combo, repeat 5 times

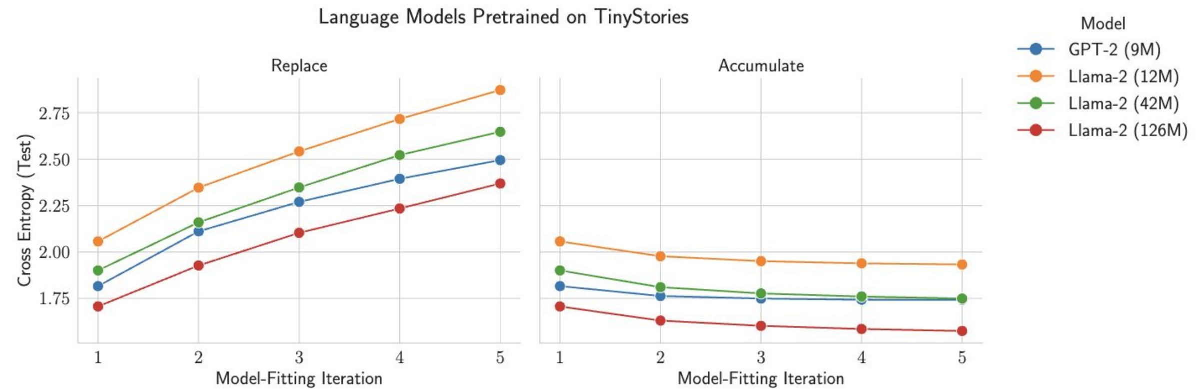
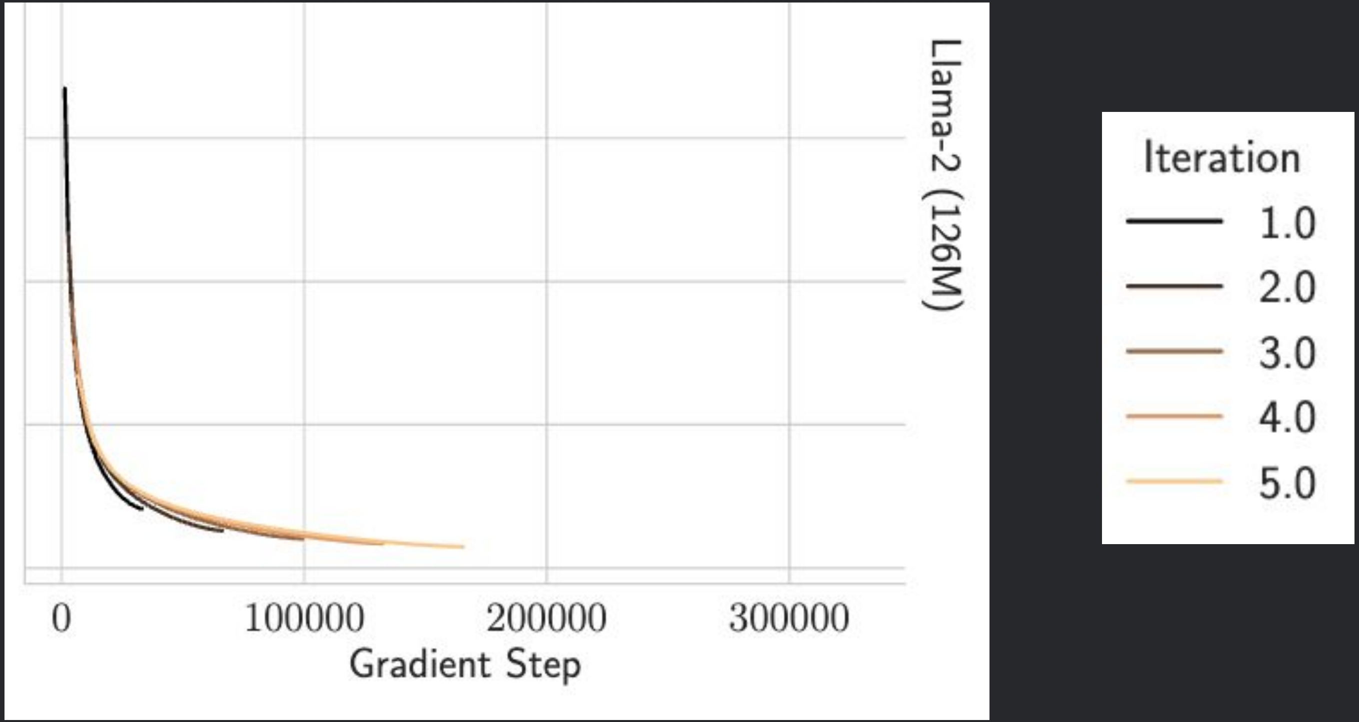


Figure 2: **Data Accumulation Avoids Model Collapse in Language Modeling.** Sequences of causal transformer-based language models are pretrained on TinyStories (Eldan & Li, 2023). Cross-entropy validation loss increases when repeatedly replacing data (left), but not when accumulating data (right). Synthetic data was sampled with temperature = 1.0.

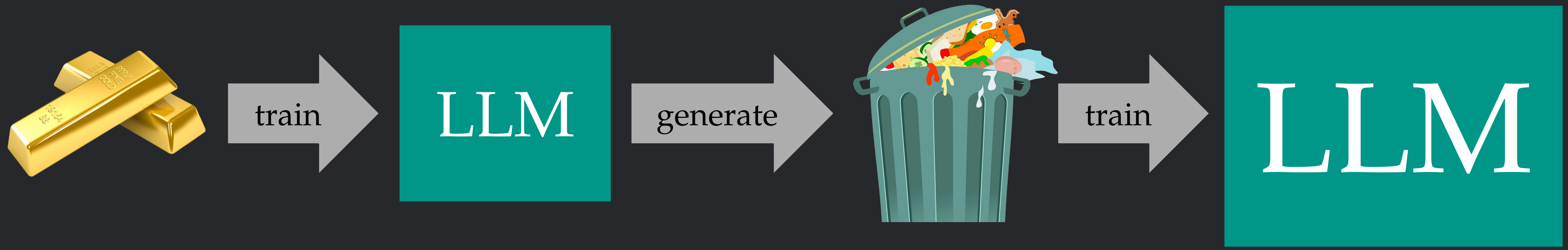
Simple solution: mix in original data

Gerstgrasser et al, 2024: Token-matched version of the plot



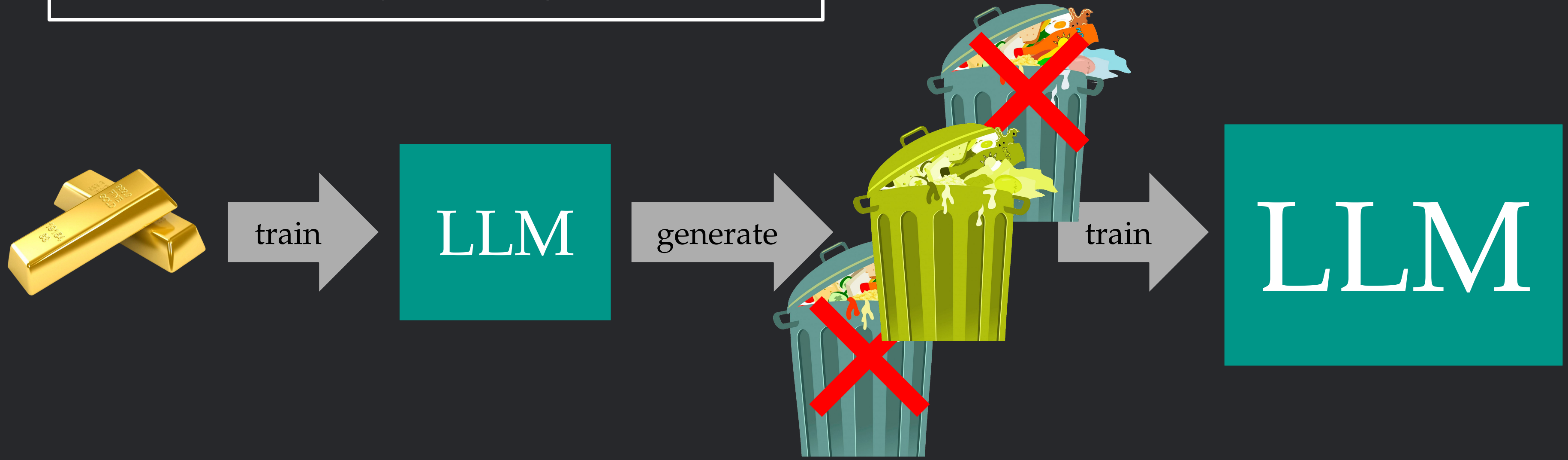
Putting it all together

So what does synthetic data actually look like?



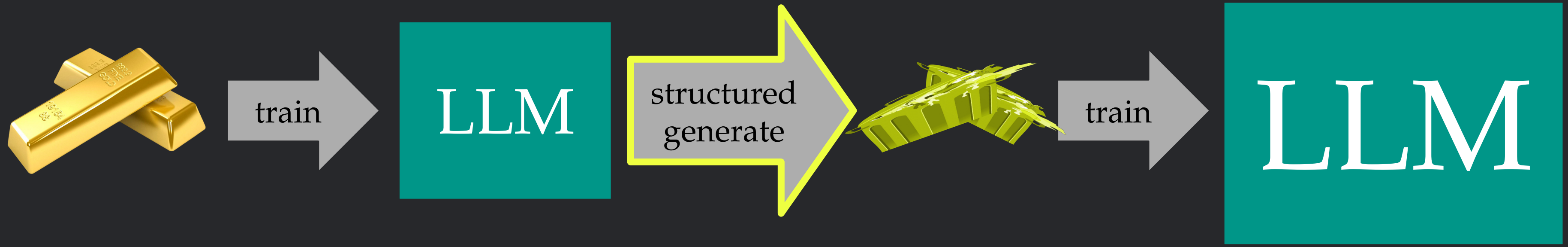
So what does synthetic data actually look like?

1. Filter out bad synthetic generations



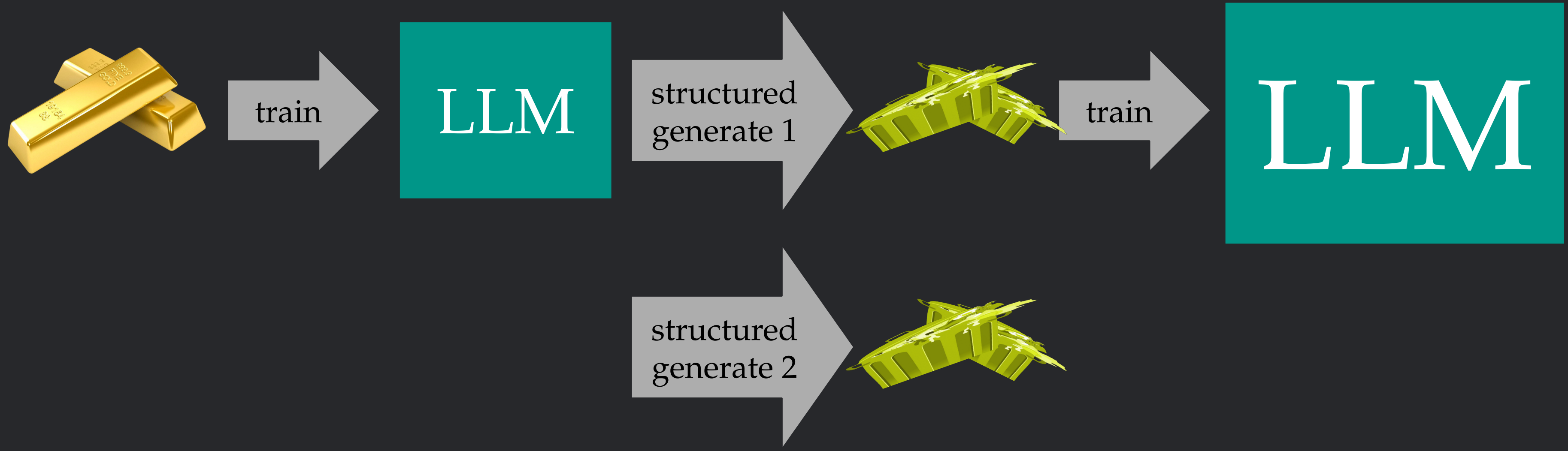
So what does synthetic data actually look like?

2. Request data in desired format



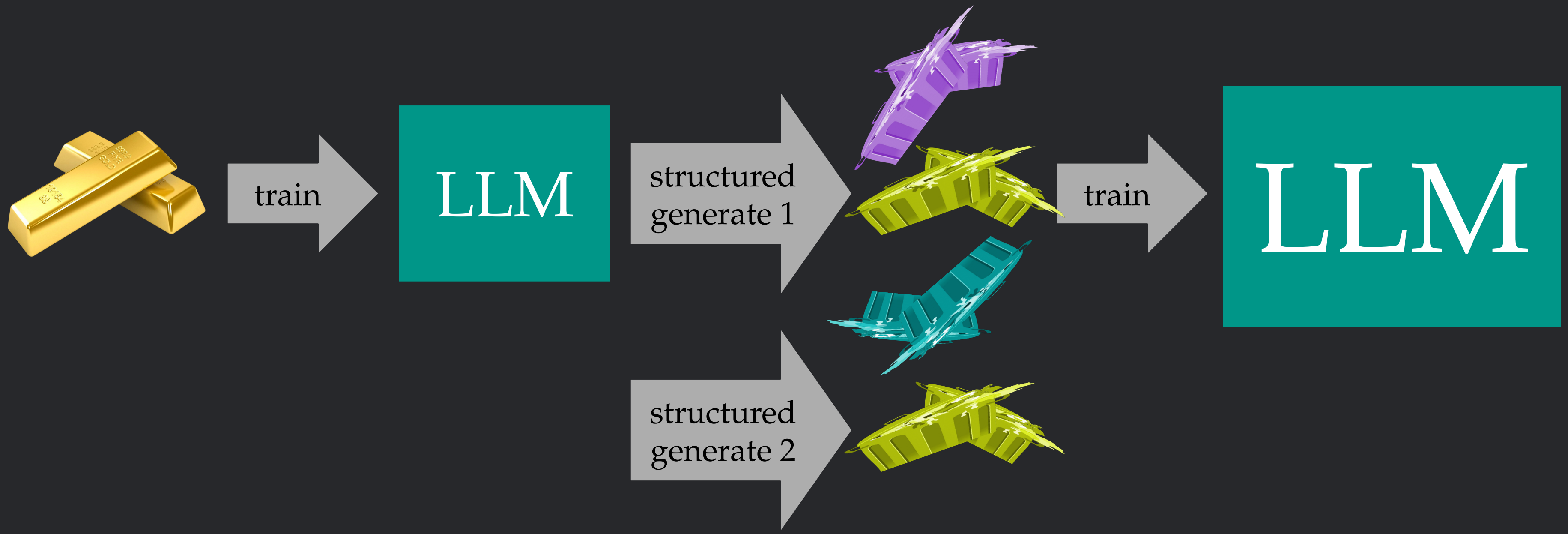
So what does synthetic data actually look like?

3. Enforce diversity at prompt level



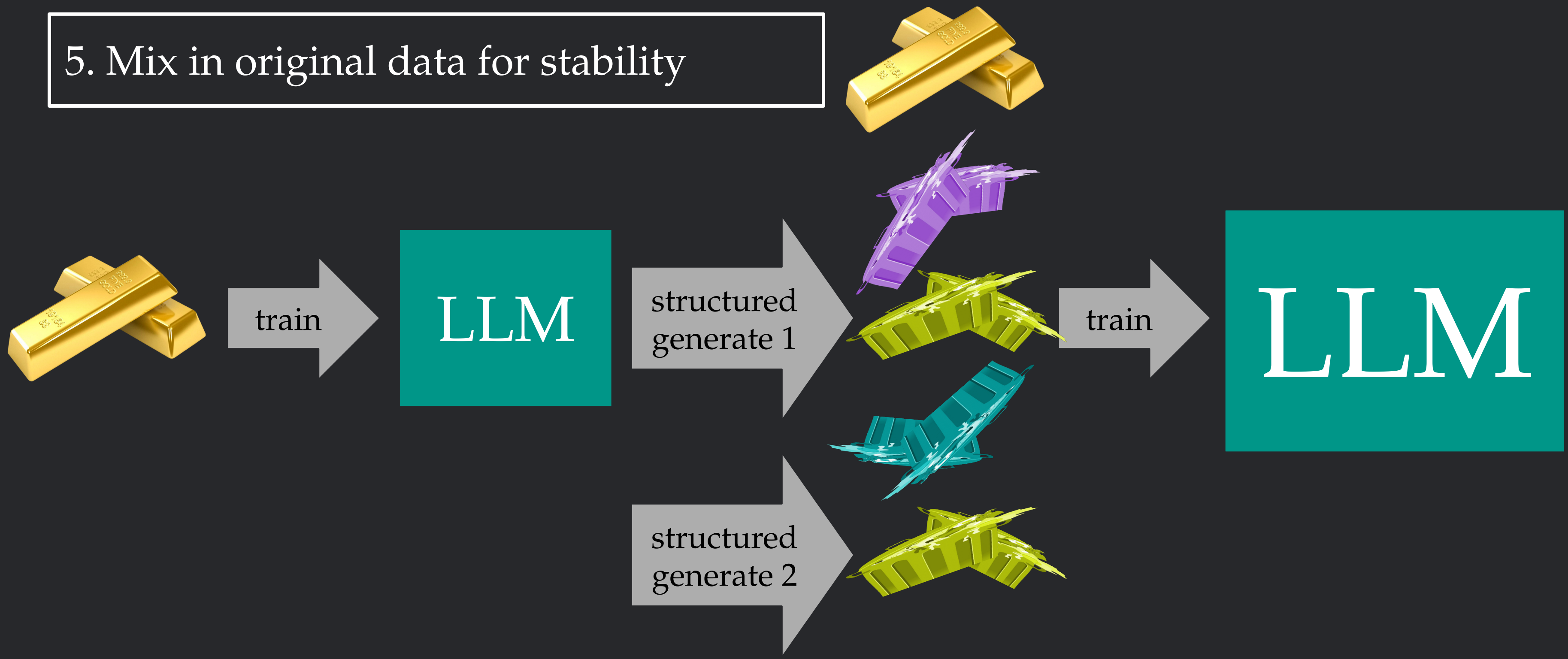
So what does synthetic data actually look like?

4. Parametric data augmentation

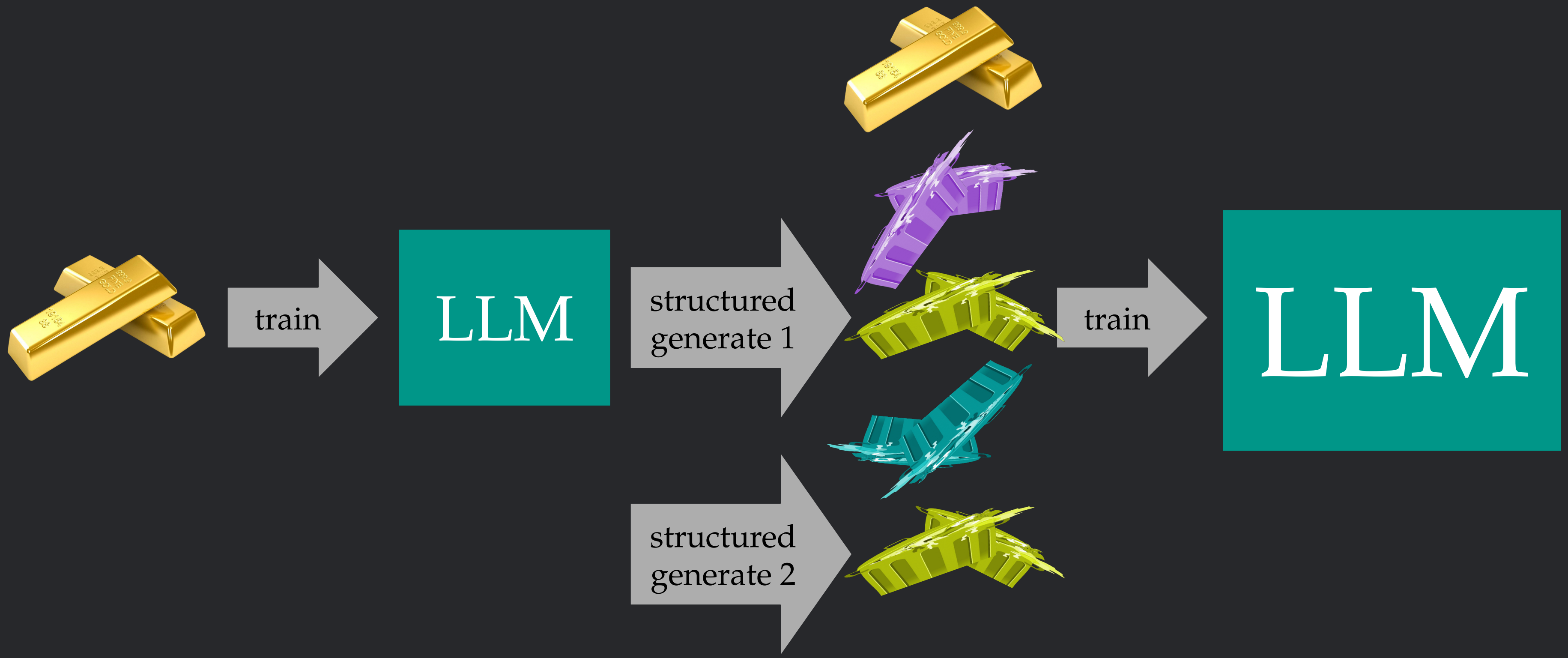


So what does synthetic data actually look like?

5. Mix in original data for stability



Yay, this might work!?



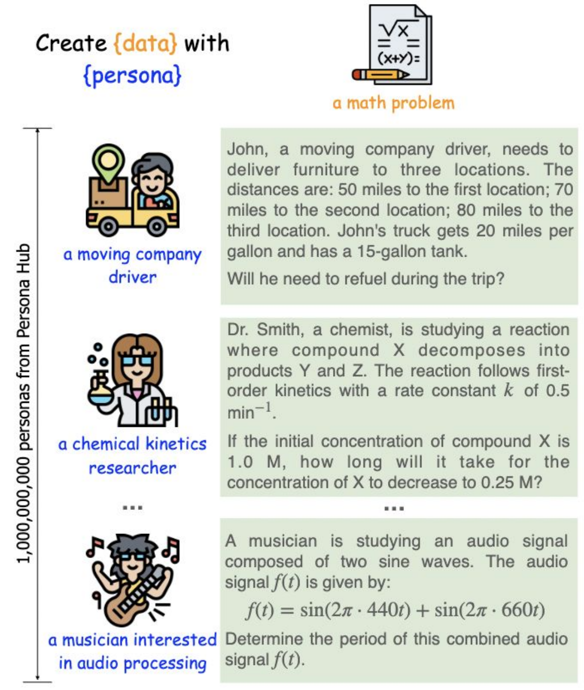
Case study: OLMo 2

Performant language model with complete openness on data, checkpoints, loss curves, etc

Generated a ton of synthetic math data for mid-training

Case study: OLMo 2

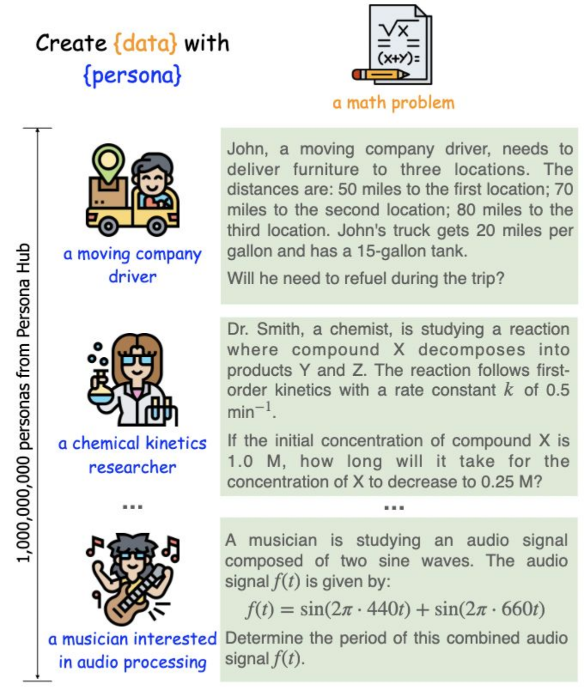
TuluMath: Utilized personas to rewrite math problems [Ge et al, 2024] using GPT-4o, generated solutions from GPT-4o



Case study: OLMo 2

TuluMath: Utilized personas to rewrite math problems [Ge et al, 2024] using GPT-4o, generated solutions from GPT-4o

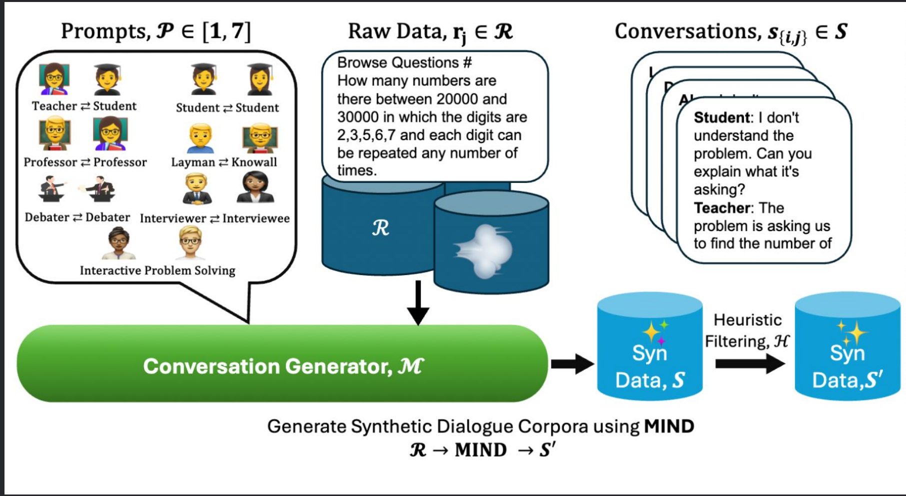
- Diverse input distribution for data augmentation (with GPT-4o)
- Distillation from GPT-4o



Case study: OLMo 2

TinyGSM-MIND: First takes TinyGSM [Liu et al, 2023], which is Python versions of standard math problems from GSM-8k. Rewrites QA pairs using MIND data augmentation to form conversations [Akter et al, 2024]

```python
def simple_math_problem() -> int:
    """
    In preparation for her party, Sarah buys 10 trays
    of food and 8 cases of beverages.
    Each tray costs $50 and each case of beverages
    costs $20.
    What is the total cost of the trays and beverages?
    """
    trays = 10
    tray_cost = 50
    cases = 8
    case_cost = 20
    tray_total = trays * tray_cost
    case_total = cases * case_cost
    total_cost = tray_total + case_total
    result = total_cost
    return result
```

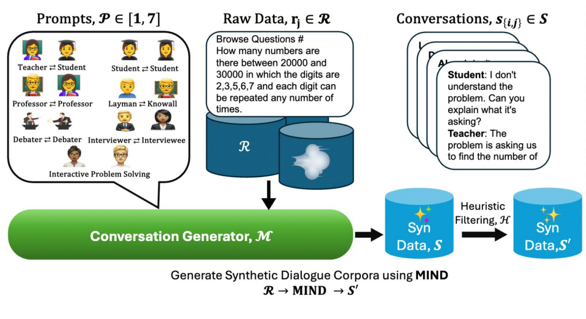


Case study: OLMo 2

TinyGSM-MIND: First takes TinyGSM [Liu et al, 2023], which is Python versions of standard math problems from GSM-8k. Rewrites QA pairs using MIND data augmentation to form conversations [Akter et al, 2024]

```
def simple_math_problem() -> int:
    """
    In preparation for her party, Sarah buys 10 trays
    of food and 8 cases of beverages.
    Each tray costs $50 and each case of beverages
    costs $20.
    What is the total cost of the trays and beverages?
    """
    trays = 10
    tray_cost = 50
    cases = 8
    case_cost = 20
    tray_total = trays * tray_cost
    case_total = cases * case_cost
    total_cost = tray_total + case_total
    result = total_cost
    return result
```

- Diverse input distribution from random numbers
- Verification from python
- Data augmentation from Qwen-2.5-7B-Instruct

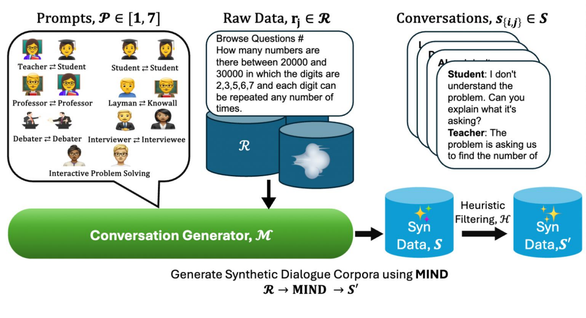


Case study: OLMo 2

TinyGSM-MIND: First takes TinyGSM [Liu et al, 2023], which is Python versions of standard math problems from GSM-8k. Rewrites QA pairs using MIND data augmentation to form conversations [Akter et al, 2024]

```python
def simple_math_problem() -> int:
    """
    In preparation for her party, Sarah buys 10 trays
    of food and 8 cases of beverages.
    Each tray costs $50 and each case of beverages
    costs $20.
    What is the total cost of the trays and beverages?
    """
    trays = 10
    tray_cost = 50
    cases = 8
    case_cost = 20
    tray_total = trays * tray_cost
    case_total = cases * case_cost
    total_cost = tray_total + case_total
    result = total_cost
    return result
```

- Diverse input distribution from random numbers
- Verification from python
- Data augmentation from Qwen-2.5-7B-Instruct
- * Matching format of GSM-8k *



Case study: OLMo 2

MathCoder2-Synthetic:

- Takes textbooks from huggingface user Ajibawa-2023, who presumably prompted a language model to generate many textbooks
- Asks GPT-4o to label the textbooks as math or not math
- Trains a fastText classifier on top of these labels
- Uses this to filter for math textbooks

Summary

Initially, might worry that synthetic data is low quality and low coverage

Quality concerns are generally handled by using a verifier, a powerful generator, or grounded transformations of existing data

Coverage concerns are generally handled by using replay data and diverse + targeted inputs

Open Problems

Problem 1: Data filtering

Data filtering helps most when we are not data-constrained. Therefore, its most useful for deciding which synthetic data to keep

- How good are verifiers/reward models for domains outside reasoning?
- How do we maintain balance coverage and quality once we have infinite data from each domain?

Problem 2: Self-improvement

Suppose we train our LLM. Can we use it to generate the next round of synthetic data to train the next LLM? Some very early successes happening here [Subramaniam et al, 2025]

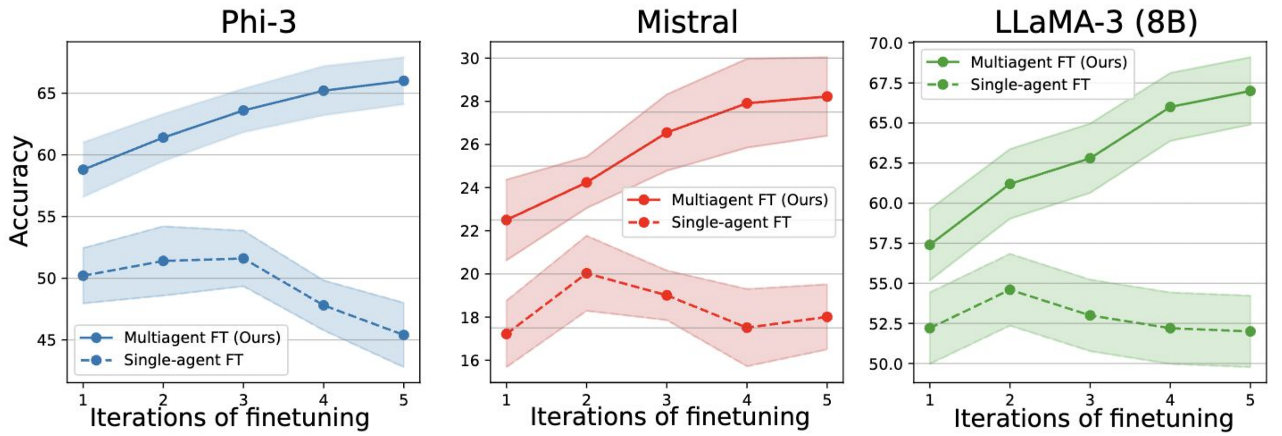


Figure 1: **Multiagent finetuning improves reasoning performance over multiple rounds of finetuning**. Our multiagent finetuning procedure enables models to improve across multiple iterations of finetuing. Results reported on the MATH dataset.