Building LLM Reasoners

Lecture 6: SFT, RLHF

Greg Durrett

Administrative details and recap

Supervised Fine-Tuning

SFT History

Understanding SFT Data

RLHF

DPO

Putting it together: Olmo3, OpenThoughts

Administrative details and recap

Supervised Fine-Tuning

SFT History

Understanding SFT Data

RLHF

DPO

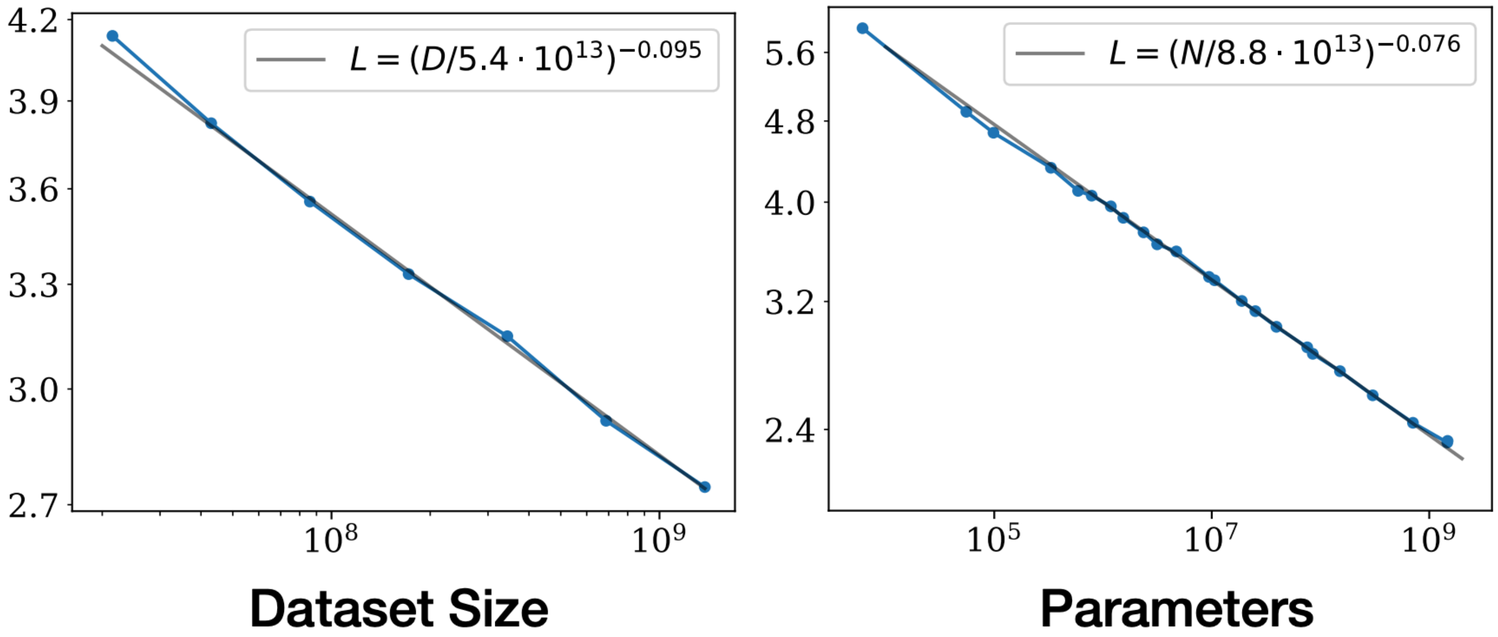
Putting it together: Olmo3, OpenThoughts

Administrative details and recap

Administrivia

- Assignment 2 due next week (+ quiz)
- Assignment 3 released next week
- Midterm in two weeks

Scaling Laws

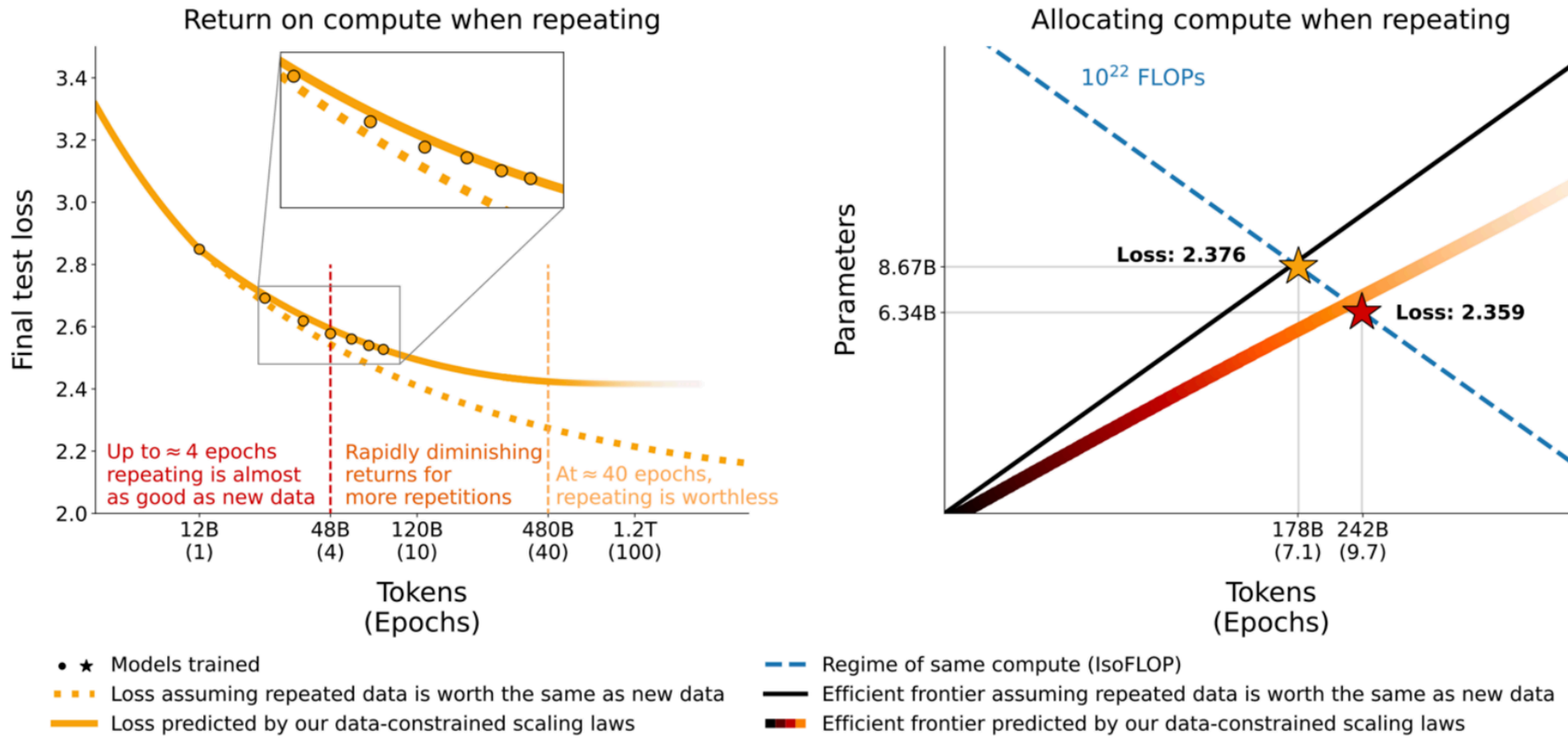


- What could scaling laws do for us?

Kaplan et al. (2020)

Q1: When to train on repeated data?

In practice, we have finite data – how does repeating examples affect scaling?



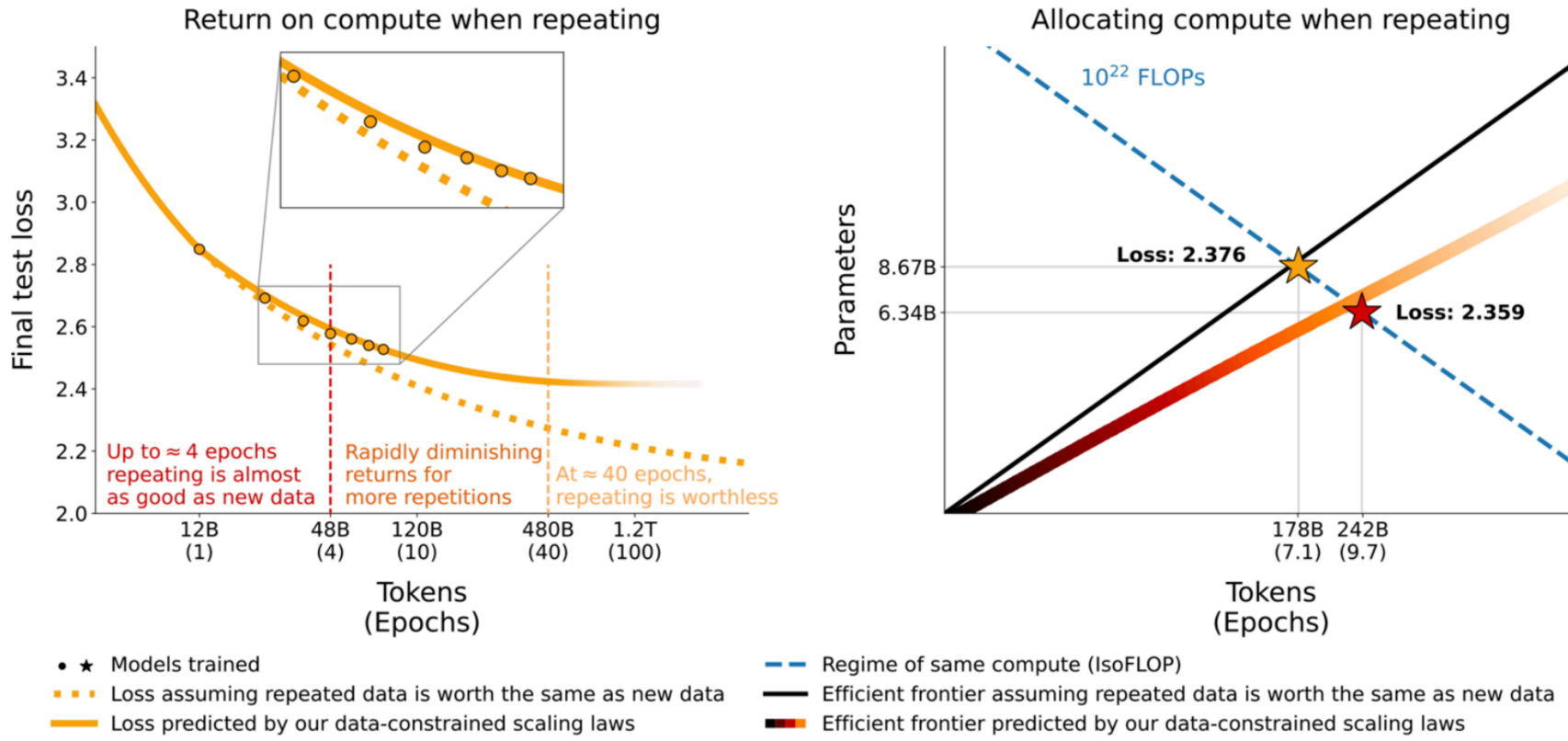
Muennighoff et al. (2023)

$$D' = U_D + U_D R_D^* (1 - e^{\frac{-R_D}{R_D^*}}) \ .$$

D' = Effective data
Ud = Unique tokens
Rd* = Constant
Rd = Repetition

Q1: When to train on repeated data?

In practice, we have finite data – how does repeating examples affect scaling?



Muennighoff et al. (2023)

$$D' = U_D + U_D R_D^* (1 - e^{\frac{-R_D}{R_D^*}}) \ .$$

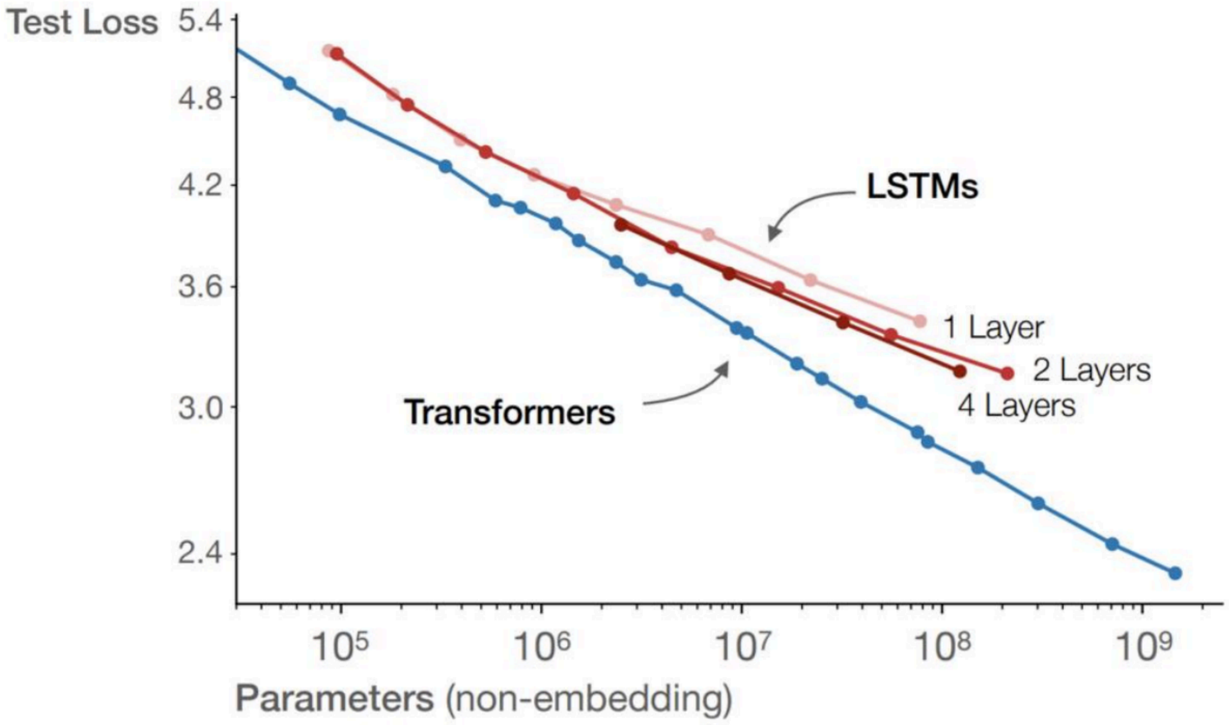
D' = Effective data
Ud = Unique tokens
Rd* = Constant
Rd = Repetition

Q2: What architecture to use?

Q: Are transformers better than LSTMs?

Brute force way: spend tens of millions to train a LSTM GPT-3

Scaling law way:



[Kaplan+ 2021]

Q3: How big a model to train?

Pick a range of FLOP budgets, vary the total parameter count, take the min over these convex shapes. The minima form a power law.

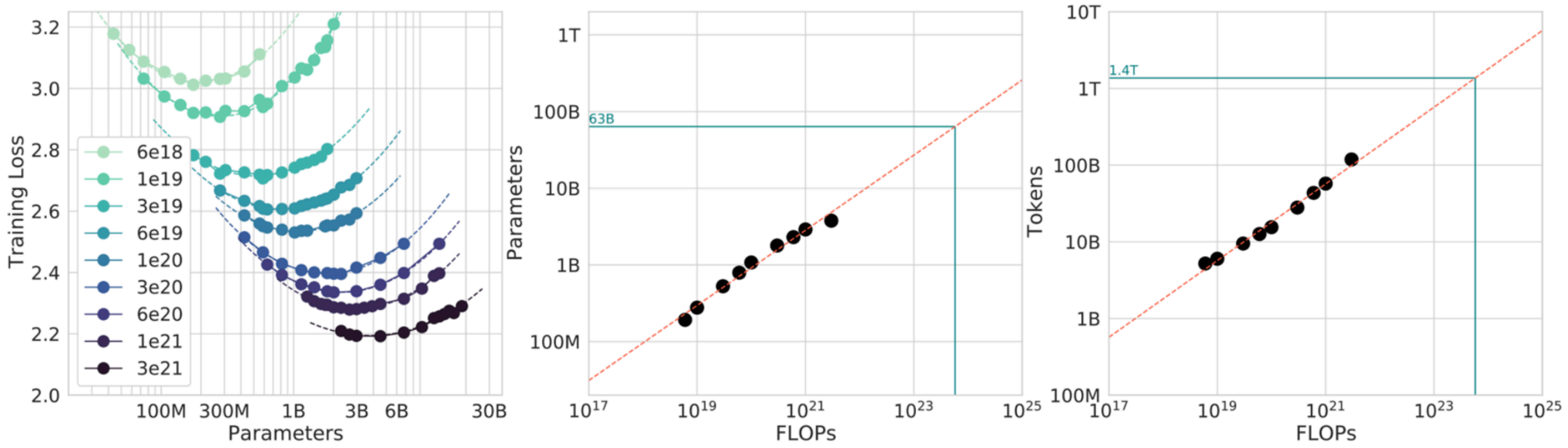
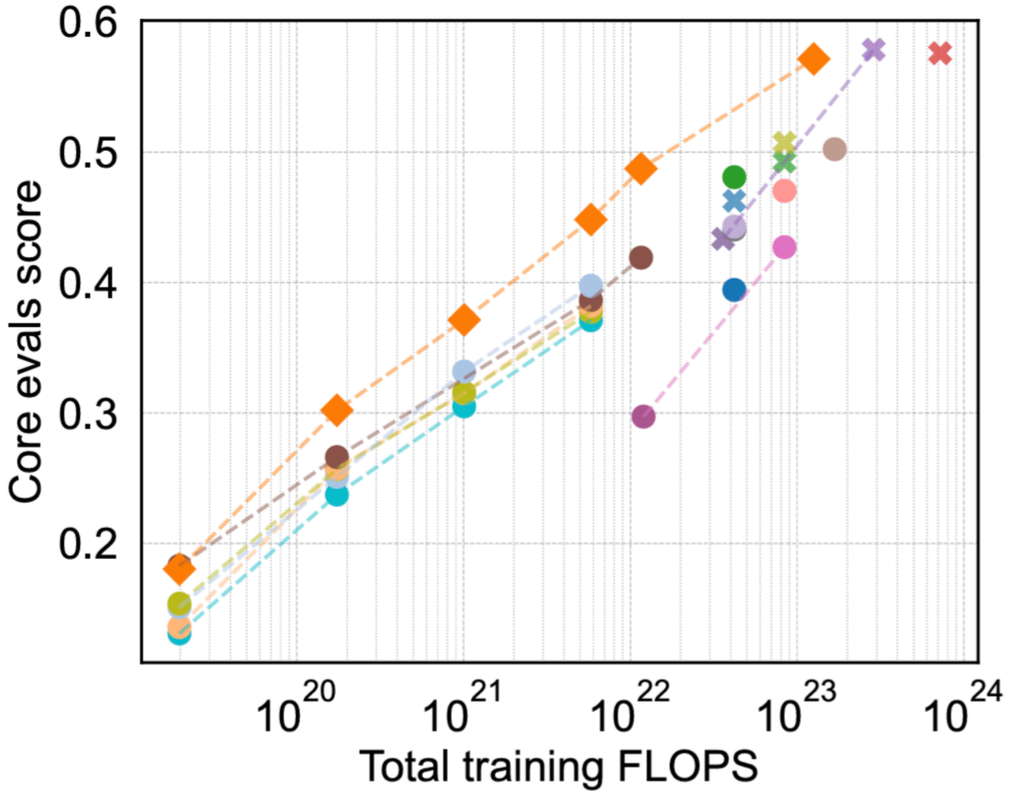
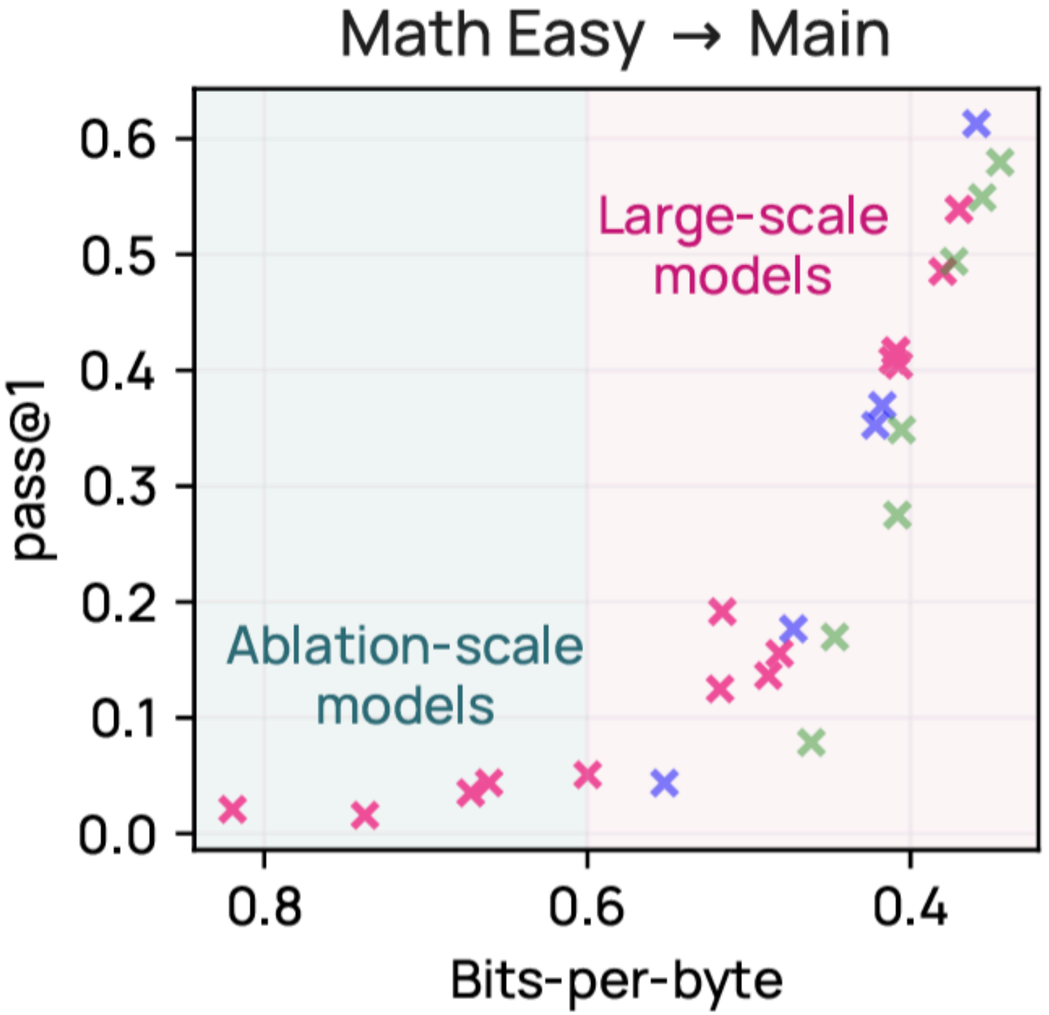


Figure 3 | **IsoFLOP curves.** For various model sizes, we choose the number of training tokens such that the final FLOPs is a constant. The cosine cycle length is set to match the target FLOP count. We find a clear valley in loss, meaning that for a given FLOP budget there is an optimal model to train (**left**). Using the location of these valleys, we project optimal model size and number of tokens for larger models (**center** and **right**). In green, we show the estimated number of parameters and tokens for an *optimal* model trained with the compute budget of *Gopher*.

In practice



DCLM



Olmo 3

Administrative details and recap

Supervised Fine-Tuning

SFT History

Understanding SFT Data

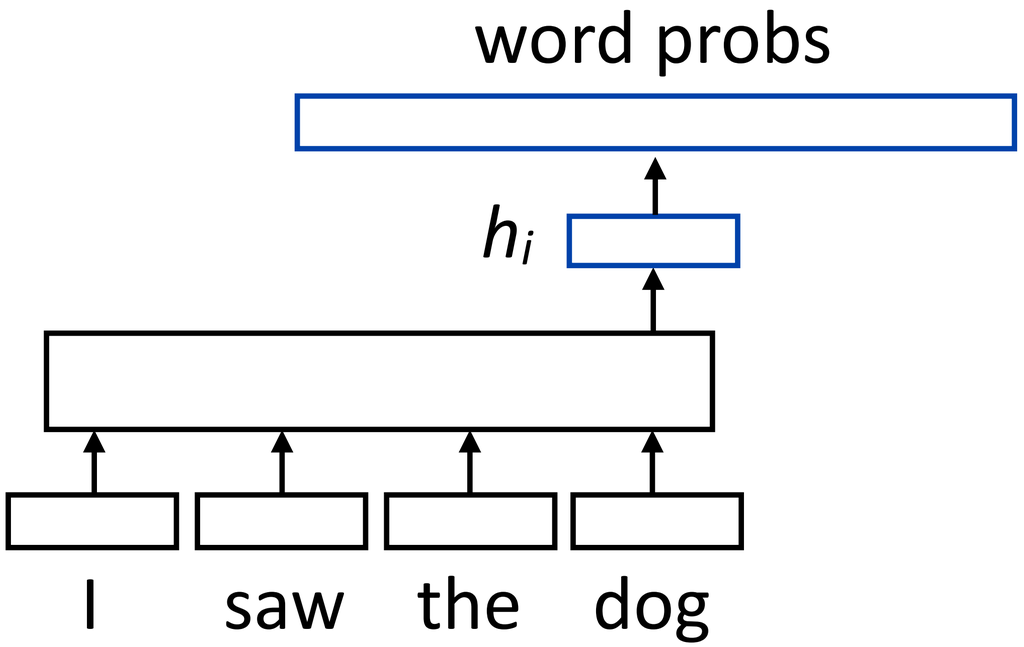
RLHF

DPO

Putting it together: Olmo3, OpenThoughts

Supervised Fine-Tuning

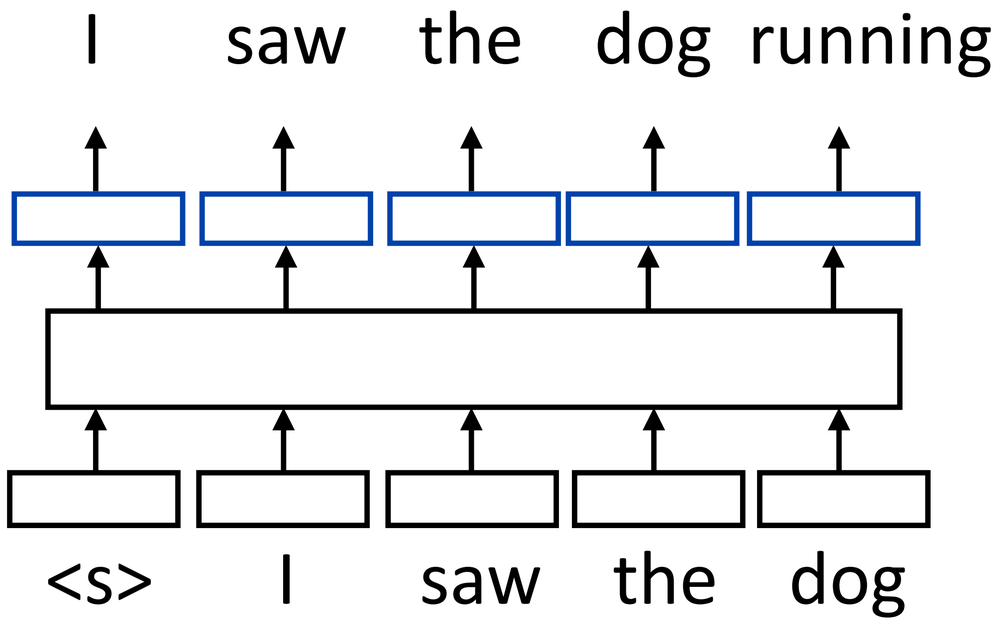
Transformer Language Modeling



$$P(w|\text{context}) = \text{softmax}(W\mathbf{h}_i)$$

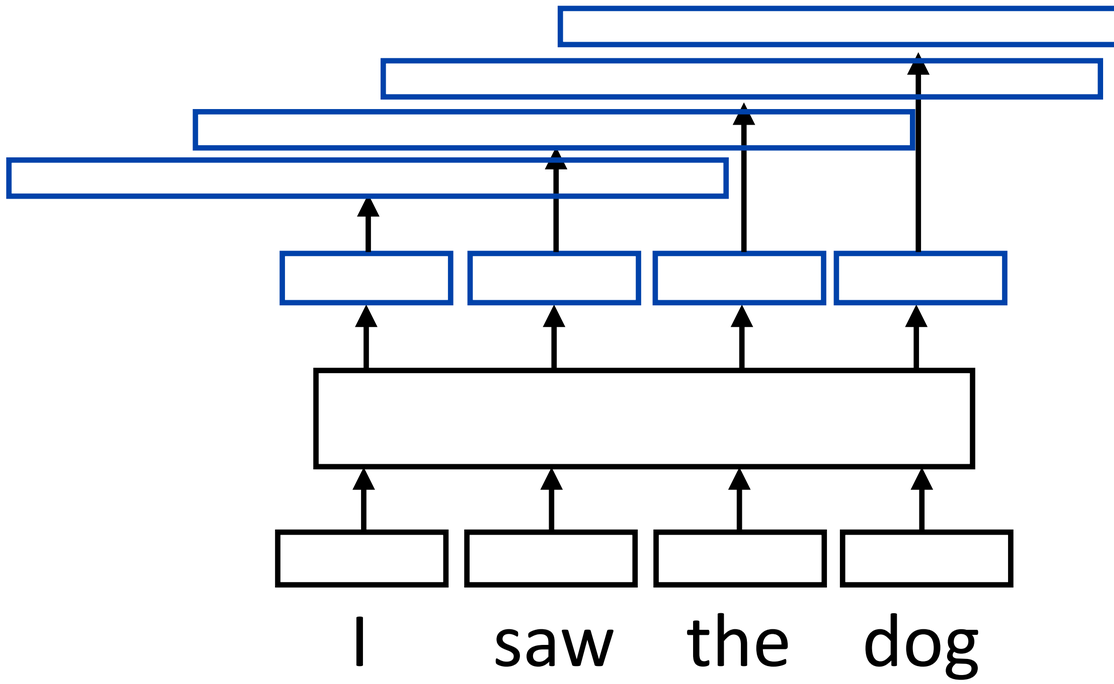
W is a (vocab size) x (hidden size) matrix

Training Transformer LMs



- Input is a sequence of words, output is those words shifted by one
- Allows us to train on predictions across several timesteps simultaneously

Training Transformer LMs



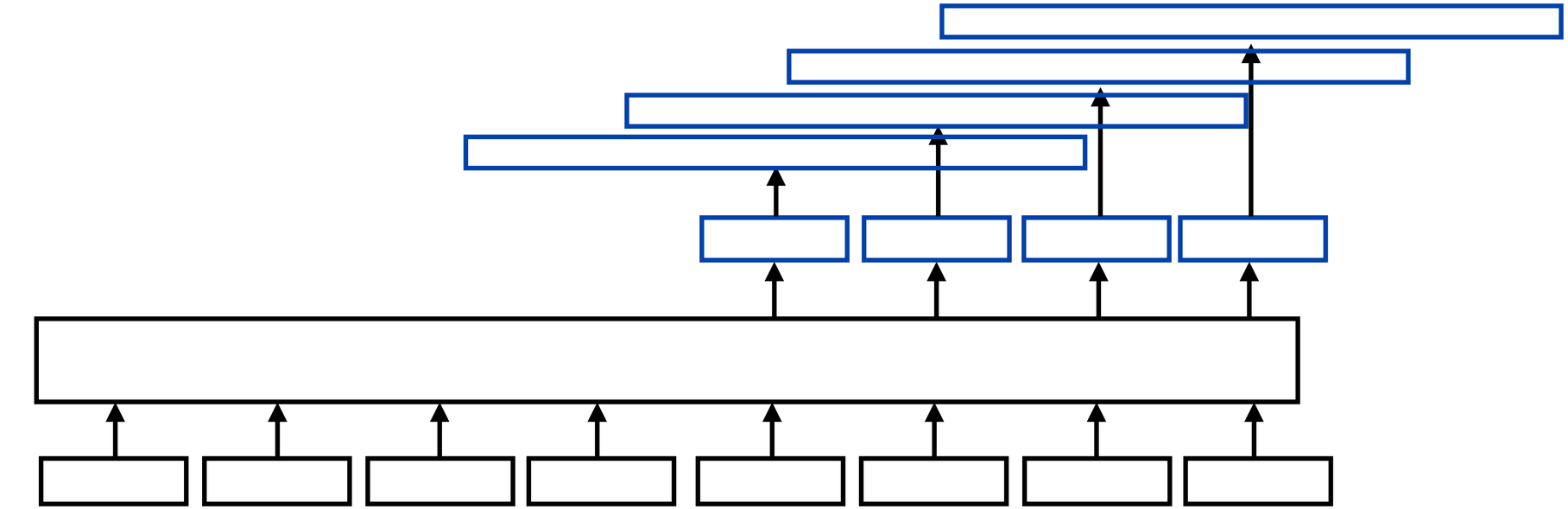
P(w|context)

loss = — log P(w*|context)

Total loss = sum of negative log likelihoods at each position

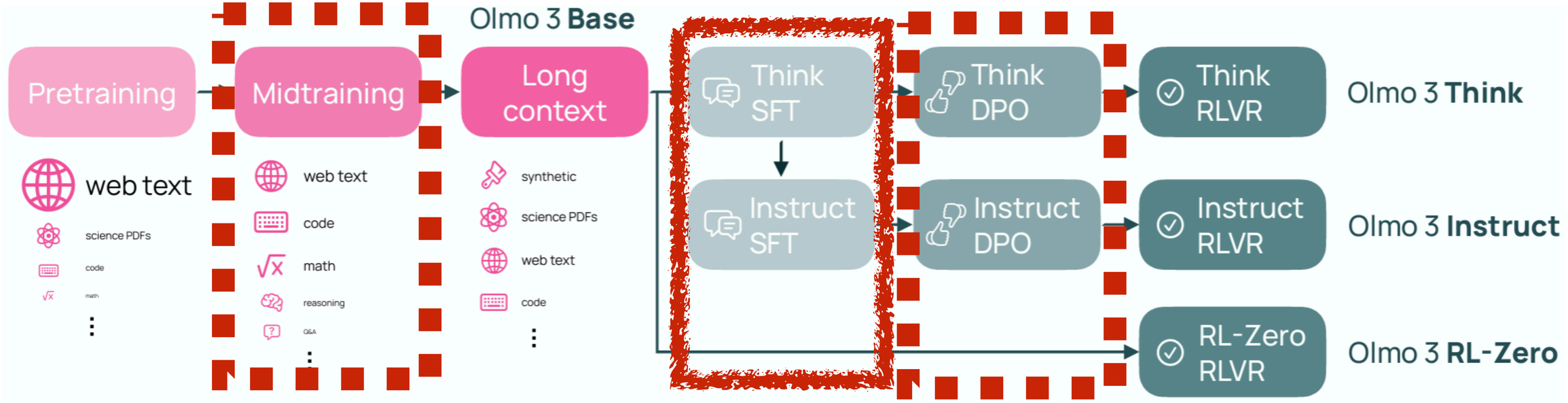
- Parallel inference across several tokens at training time, but at decoding time, tokens are generated one at a time

What is SFT?



- Same as language model training, but only compute loss on the response. Learn to respond to questions given prompts.

When is SFT?



Kind of SFT!
(similar data but train over all tokens, not just responses)

SFT!

Kind of SFT!
(requires labeled data but different objective)

Why is SFT?

- LLM pre-training can be considered either unsupervised or self-supervised
  - Unsupervised: we're just learning a function p(x) representing our data (the web)
  - Self-supervised: it's supervised prediction (predict the next word), but the labels are "free"

- This isn't the same as supervised learning, where we tell a model exactly what outputs we want it to return. **SFT lets us do this.**

- Where is pre-training on the web going to differ from what we want?

What was the "shot heard round the world"?

The shot that was heard round the world was fired in Lexington, MA. It was fired because, at the time, America was still under British rule. The British soldiers had been taking firewood by throwing it down from the top of their fort at Lexington. Someone who was not a soldier, by the name of Joseph Warren (who had ties to the Sons of Liberty or the Sons of Freedom) shouted to the British to stop doing it. The British soldiers ignored him, not

What do you think of this completion?

What was the "shot heard round the world"?

"What to the slave is the 4th of July?"

What was the sideline occupation of Thomas Jefferson's wife, Martha, and their enslaved girl, Sally Hemmings?

What was the name of the first African American regiment from Vermont?

What do you think of this one?

Samples from GPT-3

Why SFT? Safety

Truthful

Unbiased

Non-toxic

Harmless (e.g., don’t help users build a bomb)

More later in the semester!

Roadmap

- SFT is useful for telling a model to respond in a particular way (answer this question in this style)

- Let’s understand how SFT was used historically, then look at some data

Administrative details and recap

Supervised Fine-Tuning

SFT History

Understanding SFT Data

RLHF

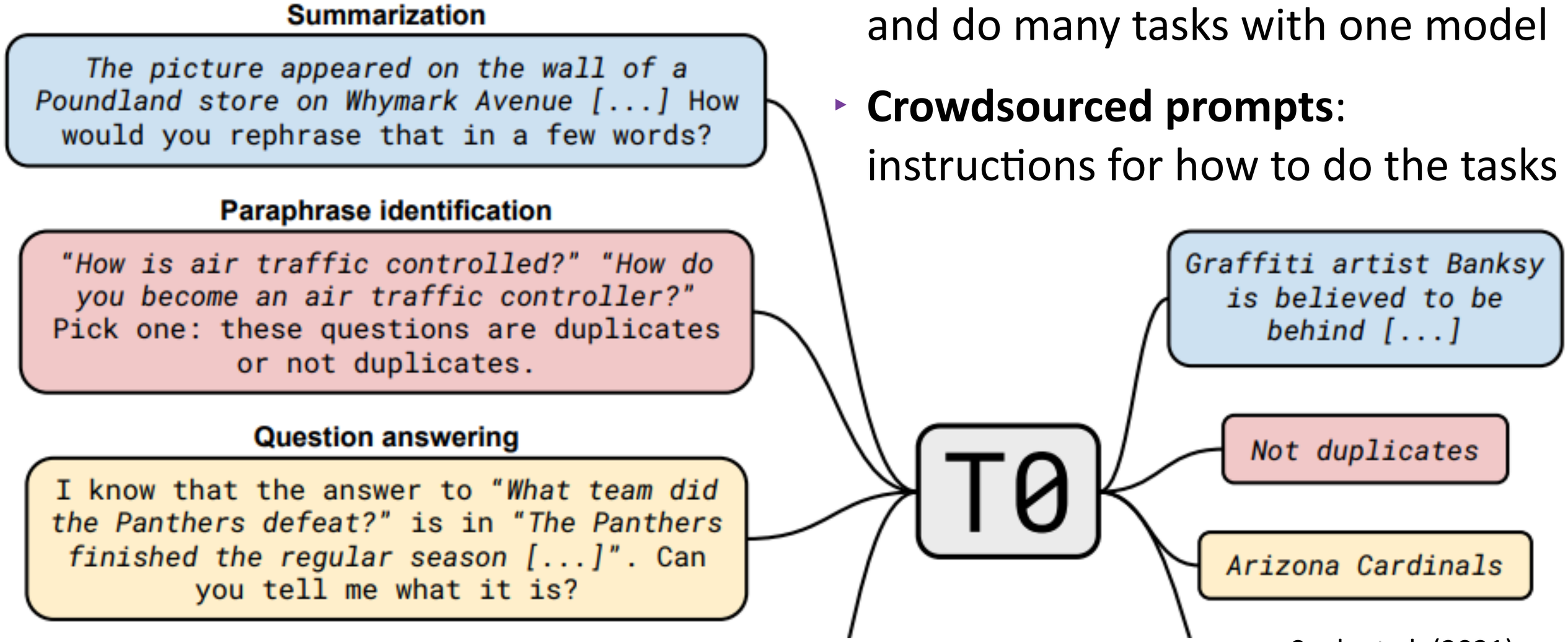
DPO

Putting it together: Olmo3, OpenThoughts

SFT History

Task Generalization: T0

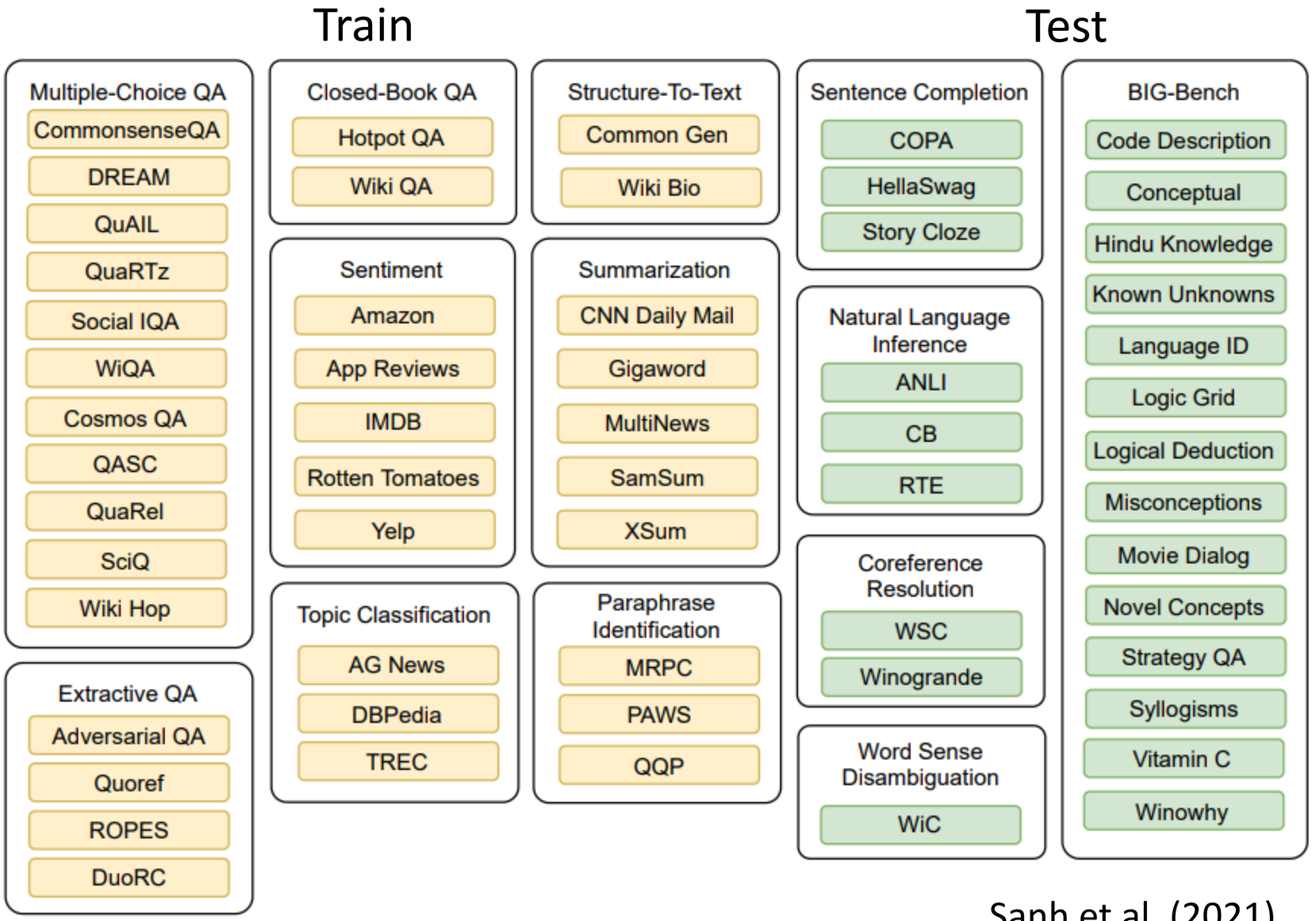
- T0: tries to deliver on the goal of T5 and do many tasks with one model
- **Crowdsourced prompts**: instructions for how to do the tasks



Sanh et al. (2021)

Task Generalization

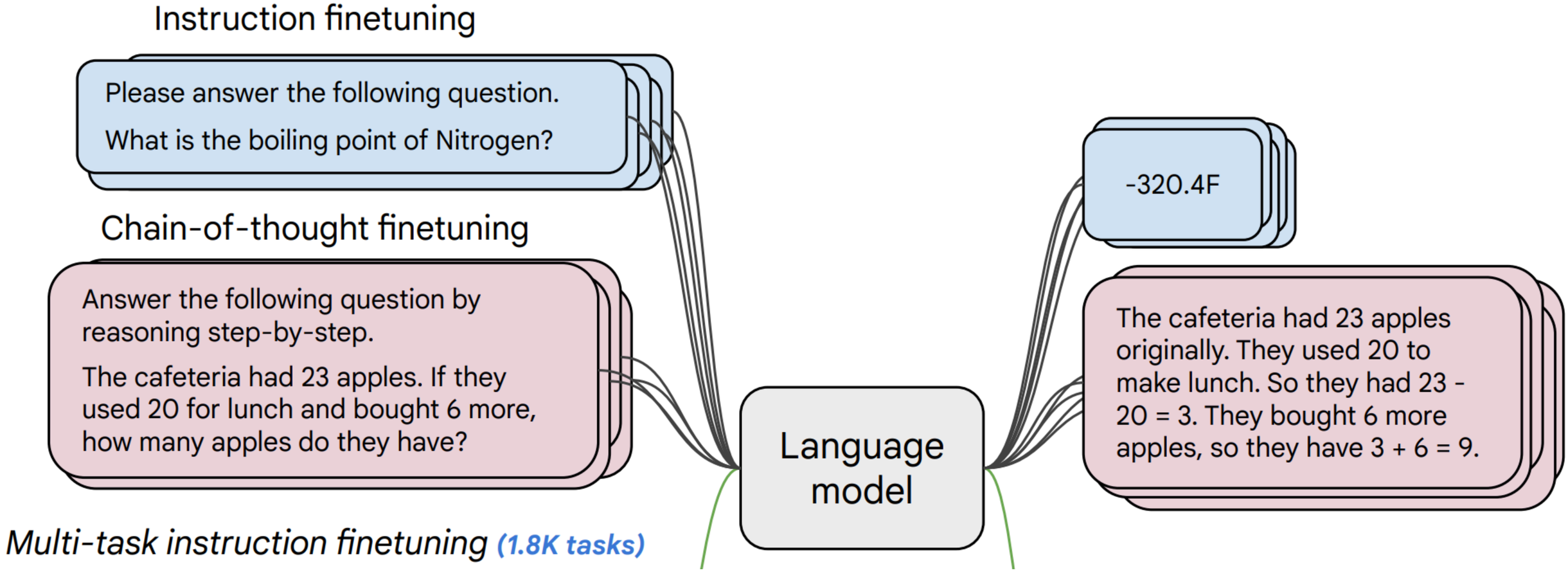
- Pre-train: T5 task
- Train: a collection of tasks with prompts. **This uses existing labeled training data**
- Test: a new task specified only by a new prompt. **No training data in this task**



Sanh et al. (2021)

Flan-PaLM

- Flan-PaLM (October 20, 2022): 1800 tasks, 540B parameter model fine-tuned on many tasks after pre-training



Chung et al. (2022)

Flan-PaLM

- Flan-PaLM (October 20, 2022): 1800 tasks, 540B parameter model
- MMLU task (Hendrycks et al., 2020): 57 high school/college/professional exams:

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is

(A) 9.8 m/s² ✔
(B) more than 9.8 m/s² ✘
(C) less than 9.8 m/s² ✘
(D) Cannot say unless the speed of throw is given. ✘

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a

(A) pair of points ✘
(B) circle ✘
(C) half-line ✘
(D) line ✔

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

Chung et al. (2022)

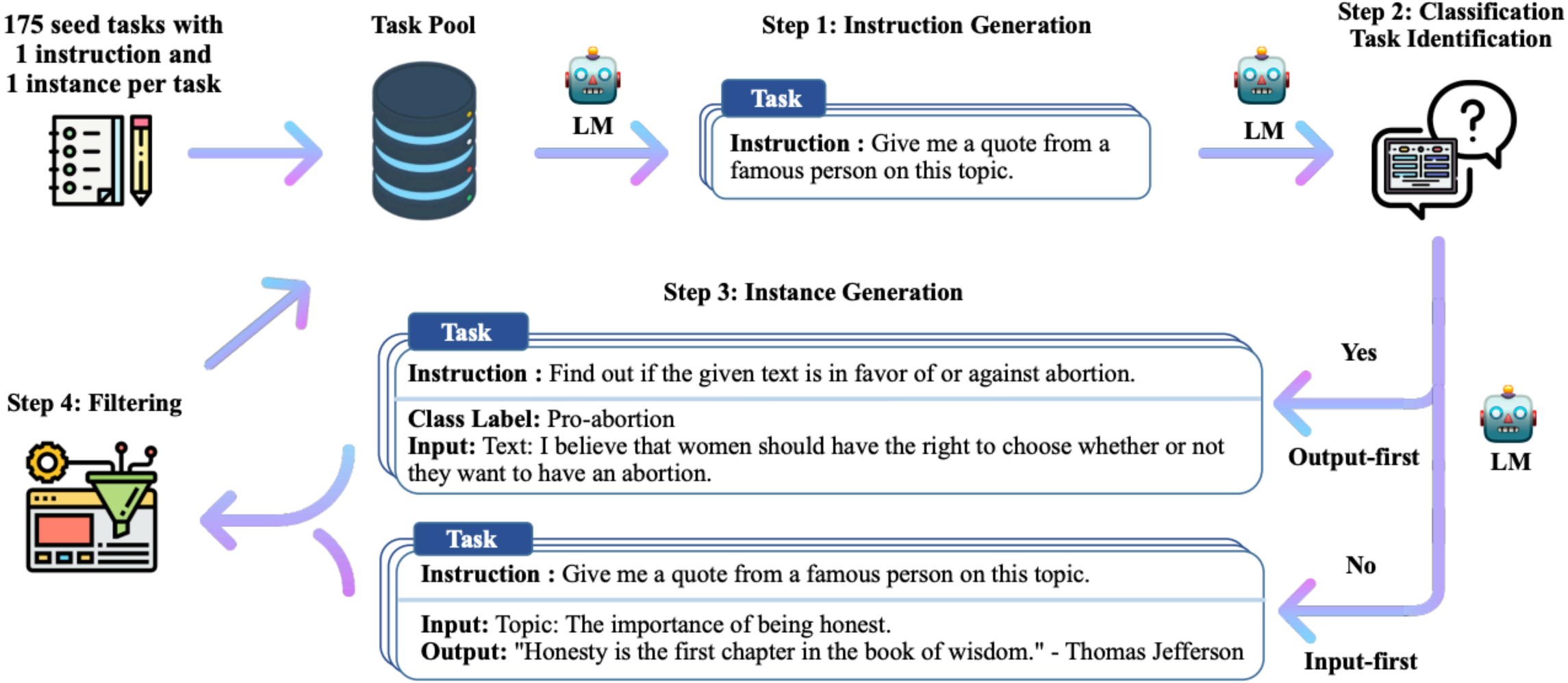
Flan-PaLM

- Flan-PaLM (October 20, 2022): 1800 tasks, 540B parameter model
- MMLU task (Hendrycks et al., 2020): 57 high school/college/professional exams:

-	Random	25.0
-	Average human rater	34.5
May 2020	GPT-3 5-shot	43.9
Mar. 2022	Chinchilla 5-shot	67.6
Apr. 2022	PaLM 5-shot	69.3
Oct. 2022	**Flan-PaLM 5-shot**	**72.2**
	Flan-PaLM 5-shot: CoT + SC	**75.2**
-	Average human expert	89.8

Chung et al. (2022)

Self-Instruct/Alpaca



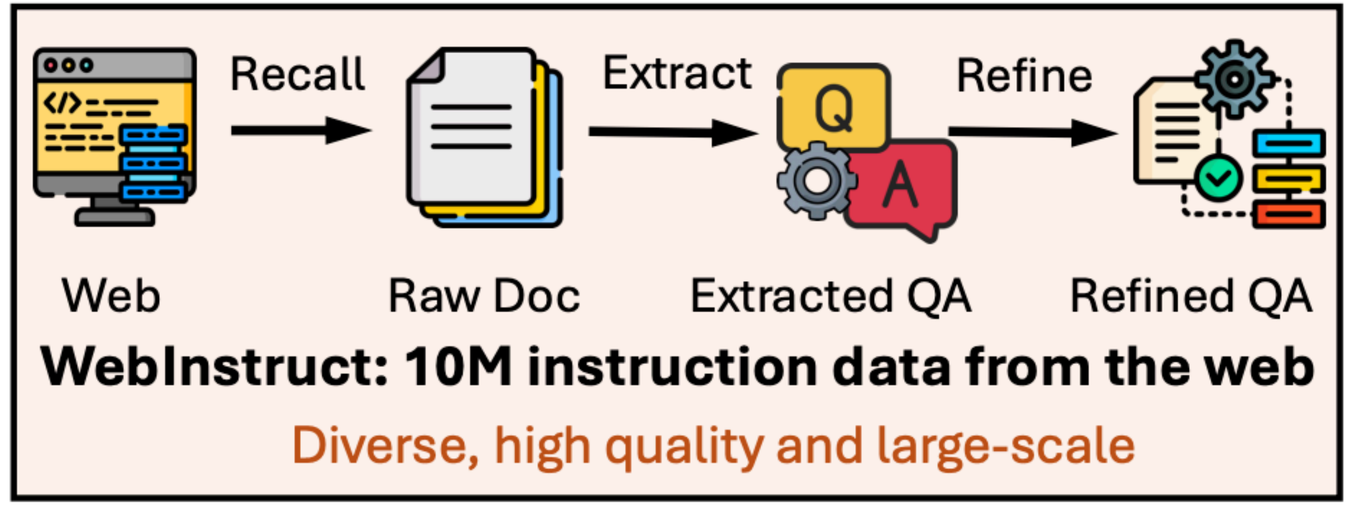
- Fine-tune Llama on 52k outputs with answers generated by text-davinci-003

Yizhong Wang et al. (2023) Self-Instruct

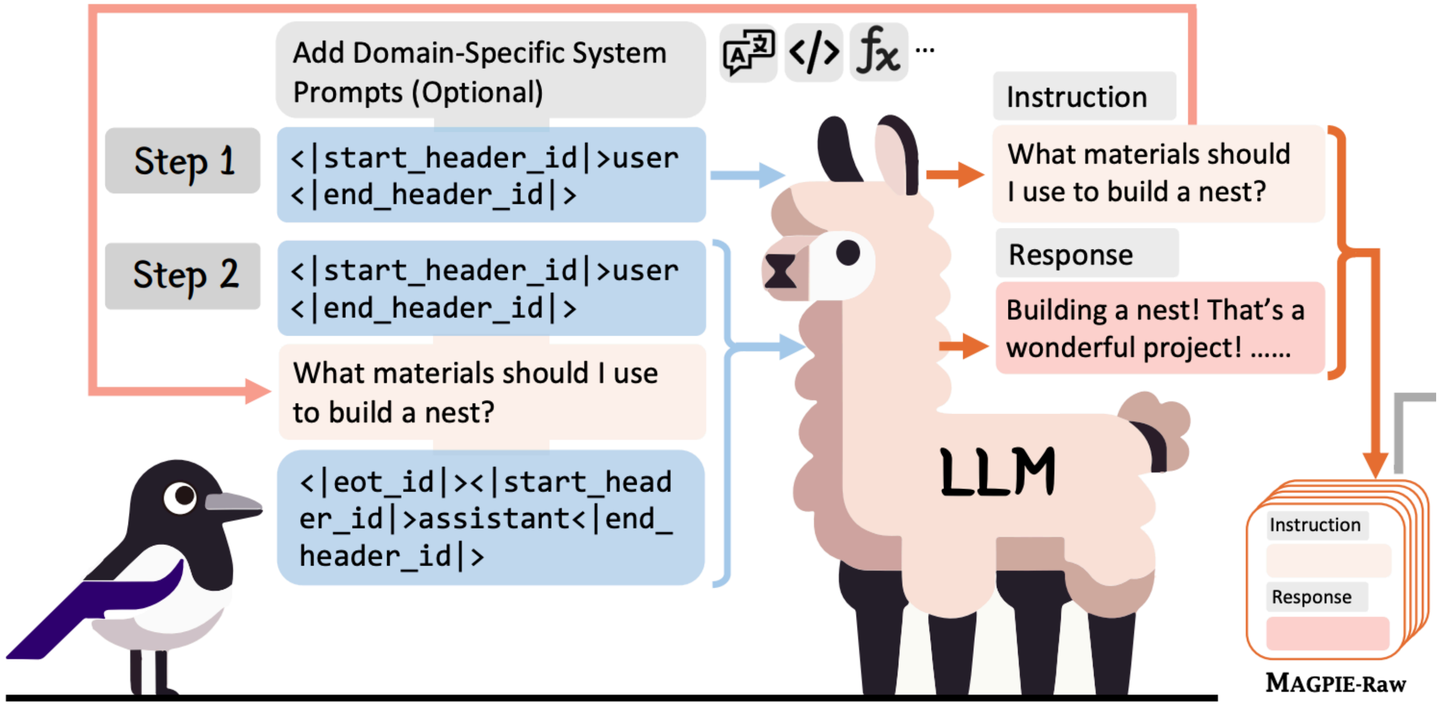
Ronen Taori et al. (2023) Alpaca

“Found” Instruction Data

- MAmmoTH2: extract instruction data from the web (using LLMs to reformulate it)



- MAGPIE: generate user prompts and then the responses from scratch using an LLM, then filter them and train on that data



Where are we now?

- Many sources of data:
  - Existing labeled dataset (T0, Flan-PaLM)
  - Outputs of larger models (Alpaca)
  - Rearranged data from the web (MAmmoTH2)
- How do these differ?

Administrative details and recap

Supervised Fine-Tuning

SFT History

Understanding SFT Data

RLHF

DPO

Putting it together: Olmo3, OpenThoughts

Understanding SFT Data

Slide credit: Tatsu Hashimoto

FLAN

Stephanie - Can you finalize the attached and have it signed. I need to initial it, but it needs to be signed by Brad Richter. Thanks. Write a subject line for this email.	Ronald Chisholm LOI
Ahold to Sell Spain Operations to Permira (AP) AP - The Dutch supermarket retailer Ahold, seeking to streamline global operations and reduce debt, said Sunday it will sell its holdings in Spain to Permira Funds for about #36;849 million. What is this text about? OPTIONS: - World - Sports - Business - Science/Tech	Business
Write highlights for this article: Sauntering down leafy avenues past typical Dutch step-gabled buildings, carpets of crocuses, the occasional cyclist whooshing quietly by and the sun bouncing off the canal, I can think of few more pleasant places to spend a spring weekend than The Hague. […]The most prized Dutch colony was apparently run from The Hague (aka 'the Widow of Indonesia'). For summer: the excellent Grapes & Olives and the Café de Bieb on Veenkade will have boats on a newly renovated bit of canal, with music and boat trips planned, including a tour at the Kompaan brewery (I can vouch for the product). My art gorging finishes at the Panorama, an extraordinary, 14 ft-high circular panorama of The Hague's seaside resort, Scheveningen, painted by Hendrik Mesdag. I take a 15-minute tram to the place itself and gorge on scallops and sole at Catch by Simonis on the harbour front. The Hague has netted a new fan. British Airways (0844 493 0787, www.ba.com) flies to Rotterdam Den Haag (half-hour taxi) from £100 return. For tours with Remco Dörr, email remco.dorr@hotmail.com. Double rooms at the Carlton Ambassador (00 31 70 363 0363, www.carlton.nl/ambassador) cost from £84, room only. More information on travel in The Netherlands at www.holland.com.	The least known of the Dutch cities, The Hague was a village until 1806. It owes its growth to Louis Bonaparte, Napoleon's brother, who ruled here. The city has a wealth of art, including Vermeer's 'Girl With A Pearl Earring'
Here is some data about a restaurant: name = Aromi, eatType = coffee shop, food = English, customer rating = 5 out of 5, area = city centre. Write a sentence that includes the following data about a restaurant	In the city centre there is a coffee shop with a customer rating of 5 out of 5 called Aromi which serves English food.

Slide credit: Tatsu Hashimoto

Alpaca

Give three tips for staying healthy.	1.Eat a balanced diet and make sure to include plenty of fruits and vegetables. 2. Exercise regularly to keep your body active and strong. 3. Get enough sleep and maintain a consistent sleep schedule.
What does “algorithm” mean?	An algorithm is a step-by-step set of instructions used to solve a problem or achieve a goal. It is a process of breaking down complex tasks into smaller, more manageable steps. Algorithms can be used to find the shortest path from one place to another, sort data, or render data from an immense database in a few seconds.
Find the average number in a list.	def avg_list(nums): total = 0 for n in nums: total += n return total / len(nums) print(avg_list([4, 3, 6, 10, 8])) # Output 6.4

Slide credit: Tatsu Hashimoto

OpenAssistant

Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.	"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. […]. Overall, the concept of monopsony is essential to understanding the dynamics of labor markets and the impact of market power on workers. Further research is needed to understand the extent and impact of monopsonies on the economy and to develop policies to address this issue. References: Bivens, J., & Mishel, L. (2013). The Pay of Corporate Executives and Financial Professionals as Evidence of Rents in Top 1 Percent Incomes. Journal of Economic Perspectives, 27(3), 57-78.
My kid needs to do a science project for elementary school, what are some cheap but fun projects to do?	There are many cheap and fun science projects that elementary school kids can do! Here are a few ideas: 1. Make a homemade lava lamp: Fill a clear plastic bottle with water and vegetable oil, add a few drops of food coloring, and then drop in an Alka-Seltzer tablet. The reaction will cause bubbles to rise and fall like a lava lamp. […]7. Create a solar oven: Using a pizza box, aluminum foil, and plastic wrap, create a solar oven that can cook s'mores or other small snacks. 8. Explore magnets: Using magnets and a variety of objects, have your child experiment with magnetism to see what objects are attracted to magnets and which ones are not. These are just a few ideas to get you started! There are many more cheap and fun science projects that your child can do with household items and a little creativity.

Slide credit: Tatsu Hashimoto

Response Lengths Vary

Table 1: Instruction datasets investigated in this work. CoT and FLAN V2 are sampled to 100K to match the sizes of other datasets. We report the average number of conservation turns ($\bar{N}_{rounds}$), average length of prompts ($\bar{L}_{prompt}$), average length of completion ($\bar{L}_{completion}$).

Datasets	**Sourced from**	**# Instances**	$\bar{N}_{rounds}$	$\bar{L}_{prompt}$	$\bar{L}_{completion}$
SuperNI [48]	NLP datasets + Human-written Instructions	96,913	1.0	291.1	38.7
CoT [50]	NLP datasets + Human-written CoTs	100,000	1.0	266.0	53.2
Flan V2 [31]	NLP datasets + Human-written Instructions	100,000	1.0	355.7	31.2
Dolly [12]	Human-written from scratch	15,011	1.0	118.1	91.3
Open Assistant 1 [26]	Human-written from scratch	34,795	1.6	34.8	212.5
Self-instruct [47]	Generated w/ vanilla GPT3 LM	82,439	1.0	41.5	29.3
Unnatural Instructions [23]	Generated w/ Davinci-002	68,478	1.0	107.8	23.6
Alpaca [43]	Generated w/ Davinci-003	52,002	1.0	27.8	64.6
Code-Alpaca [6]	Generated w/ Davinci-003	20,022	1.0	35.6	67.8
GPT4-Alpaca [36]	Generated w/ Davinci-003 + GPT4	52,002	1.0	28.0	161.8
Baize [52]	Generated w/ ChatGPT	210,311	3.1	17.6	52.8
ShareGPT[3]	User prompts + outputs from various models	168,864	3.2	71.0	357.8

Yizhong Wang et al. (2023) TULU

Impact of these Datasets

Table 3: Comparison of different instruction tuning datasets, showing that different instruction-tuning datasets can excel in different aspects, and mixtures perform best on average. Cells are blue if the finetuning boosts the vanilla LLAMA performance, and orange if the finetuning hurts the performance.

	MMLU (factuality)	GSM (reasoning)	BBH (reasoning)	TydiQA (multilinguality)	Codex-Eval (coding)	AlpacaEval (open-ended)	Average
	EM (0-shot)	EM (8-shot, CoT)	EM (3-shot, CoT)	F1 (1-shot, GP)	P@10 (0-shot)	Win % vs Davinci-003	
Vanilla LLaMa 13B	42.3	14.5	39.3	43.2	28.6	-	-
+SuperNI	49.7	4.0	4.5	**50.2**	12.9	4.2	20.9
+CoT	44.2	40.0	41.9	47.8	23.7	6.0	33.9
H +Flan V2	**50.6**	20.0	40.8	47.2	16.8	3.2	29.8
H +Dolly	45.6	18.0	28.4	46.5	31.0	13.7	30.5
H +Open Assistant 1	43.3	15.0	39.6	33.4	31.9	58.1	36.9
+Self-instruct	30.4	11.0	30.7	41.3	12.5	5.0	21.8
+Unnatural Instructions	46.4	8.0	33.7	40.9	23.9	8.4	26.9
+Alpaca	45.0	9.5	36.6	31.1	29.9	21.9	29.0
G +Code-Alpaca	42.5	13.5	35.6	38.9	34.2	15.8	30.1
G +GPT4-Alpaca	46.9	16.5	38.8	23.5	**36.6**	63.1	37.6
+Baize	43.7	10.0	38.7	33.6	28.7	21.9	29.4
G +ShareGPT	49.3	27.0	40.4	30.5	34.1	**70.5**	42.0
H +Human data mix.	50.2	38.5	39.6	47.0	25.0	35.0	39.2
H+G +Human+GPT data mix.	49.3	**40.5**	**43.3**	45.6	35.9	56.5	**45.2**

Yizhong Wang et al. (2023) TULU

Administrative details and recap

Supervised Fine-Tuning

SFT History

Understanding SFT Data

RLHF

DPO

Putting it together: Olmo3, OpenThoughts

RLHF

RLHF: Introduction

- “Reinforcement learning from human feedback”
- In 2022, fine-tuning on labeled datasets was commonplace (T5/T0, Flan, other task-specific models from 2019-2022)
- These models could do many tasks, but didn’t feel broadly useful, and they weren’t “chatbots” as we think of them now
- Big shift with ChatGPT in November 2022: RLHF-tuned models

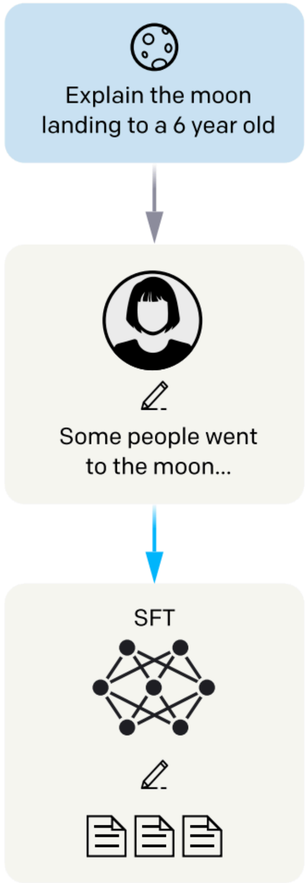
RLHF

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

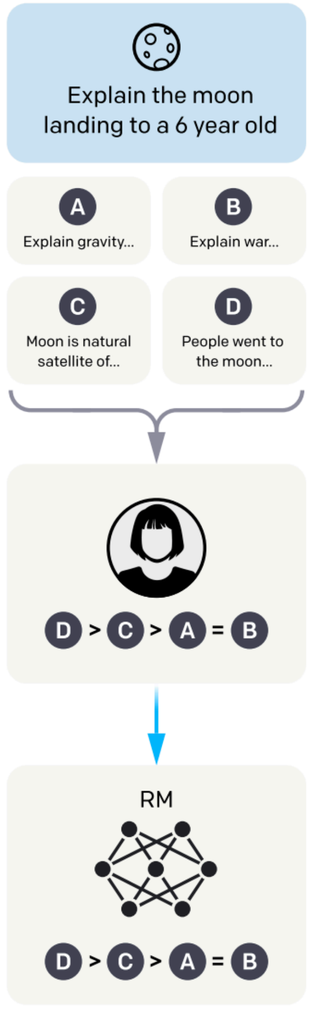


Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



- Apply this approach to optimizing outputs from large language models
- Step 3 (not shown): do RL with this policy

Ouyang et al. (2022)

Learning Reward Models

- Input **x**: *who was the US president during World War II?*
- Outputs $\mathbf{y}^+$: *Franklin D. Roosevelt, Harry Truman*

- Classical RL: assign some value +3 to this output

- Should we just get humans to label rewards? What scale do we use? What score should this get?

Ouyang et al. (2022)

Learning Reward Models

- Input **x**: *who was the US president during World War II?*
- Outputs $\mathbf{y}^+$: *Franklin D. Roosevelt, Harry Truman*

  $\mathbf{y}^-$: *Herbert Hoover, Franklin D. Roosevelt, Harry Truman*

$$P(y^+ \succ y^- \mid \mathbf{x}) = \frac{\exp(r(y^+, \mathbf{x}))}{\exp(r(y^+, \mathbf{x})) + \exp(r(y^-, \mathbf{x}))}$$

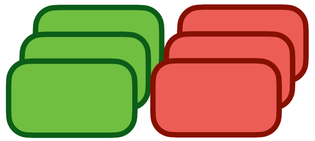
- Bradley-Terry model: turns scores into log probabilities of 1 being preferred to 2. Same as logistic regression where we classify pairs as 1 > 2 or 2 < 1, but we learn a continuous scoring function

Ouyang et al. (2022)

Learning Reward Models

- Input **x**: *who was the US president during World War II?*
- Outputs **y**$^{+}$: *Franklin D. Roosevelt, Harry Truman*

  y^{-}: *Herbert Hoover, Franklin D. Roosevelt, Harry Truman*



Lots of (**y**$^{+}$,**y**$^{-}$) pairs

$$\rightarrow \; P(y^+ \succ y^- \mid \mathbf{x}) = \frac{\exp(r(y^+, \mathbf{x}))}{\exp(r(y^+, \mathbf{x})) + \exp(r(y^-, \mathbf{x}))}$$

- Outcome: reward model *r*(y, **x**) returning real-valued scores

Ouyang et al. (2022)

RLHF

- Goal: find a policy π_θ (LM parameters) that optimizes the following:

$$R(\mathbf{x}, y) = r(\mathbf{x}, y) - \lambda D_{\mathrm{KL}}(\pi_\theta(y \mid \mathbf{x}) \| \pi_\theta^{\mathrm{SFT}}(y \mid \mathbf{x}))$$

get high reward

stay close to an initial SFT policy

- This is called *proximal policy optimization* (PPO)
- Important to regularize towards the SFT policy! Reward models are not stable enough to make things work
- PPO has some details in its implementation: it's an *advantage actor-critic* model, so there's a separate value function that gets learned

Christiano et al. (2017)

RLHF

Table 1: Distribution of use case categories from our API prompt dataset.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: """ {summary} """ This is the outline of the commercial for that play: """

- For OpenAI, RLHF data is collected from their API. **Very different from instruct-tuning datasets**

Ouyang et al. (2022)

Administrative details and recap

Supervised Fine-Tuning

SFT History

Understanding SFT Data

RLHF

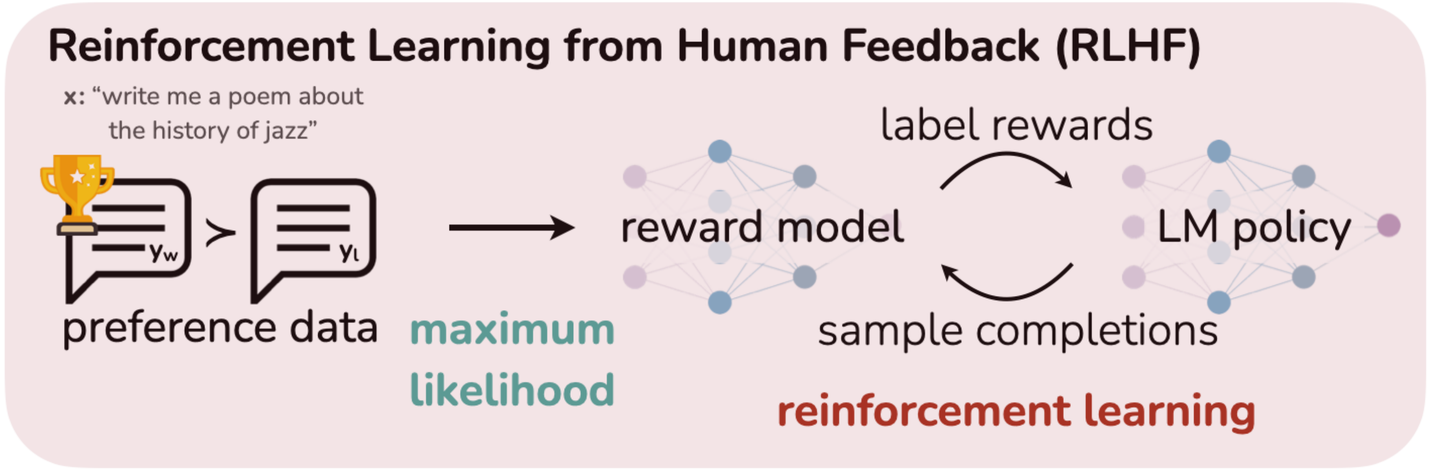
DPO

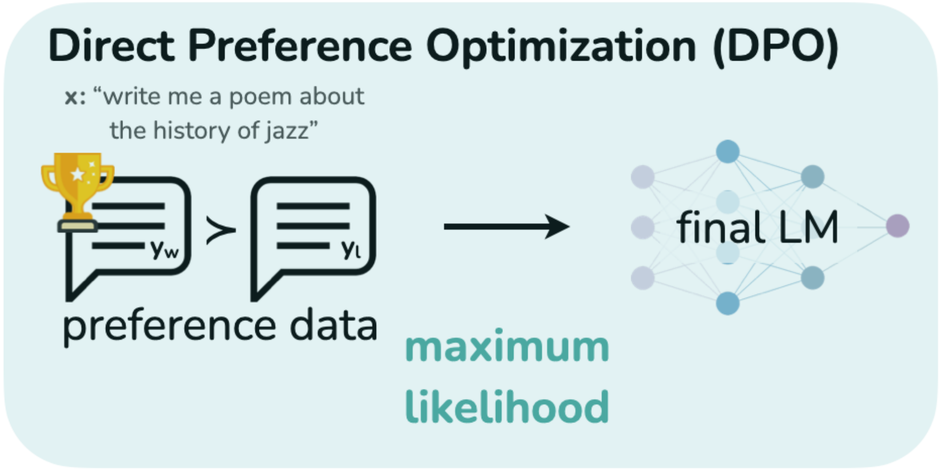
Putting it together: Olmo3, OpenThoughts

DPO

Direct Preference Optimization (DPO)

- Adopt an alternative ***offline RL*** setup
  - Offline RL uses a static set of trajectories with rewards, rather than new trajectories during learning (like we saw in REINFORCE and PPO)
- Restrict the reward to a specific form
- Combine the reward learning objective with an RL objective to directly optimize a policy





Slide credit: Yoav Artzi

Direct Preference Optimization (DPO)

- DPO starts with a very similar RL objective to PPO

$$\arg\max_\theta E_{\bar{x}\sim\mathcal{D},\bar{y}\sim\pi_\theta(\bar{y}|\bar{x})}\left[r(\bar{x},\bar{y}) - \beta\mathrm{KL}[\pi_\theta(\bar{y}\,|\,\bar{x}),\pi_{\text{ref}}(\bar{y}\,|\,\bar{x})]\right]$$

  - Where π_{ref} is the SFT policy before we fine-tune it with preference data

Maximize the expected reward according to our prompt data and policy

Penalize for the distribution getting further from the pre-RL distribution

Slide credit: Yoav Artzi

Direct Preference Optimization (DPO)

- DPO starts with a very similar RL objective to PPO

$$\arg\max_\theta E_{\bar{x}\sim\mathcal{D},\bar{y}\sim\pi_\theta(\bar{y}|\bar{x})}\left[r(\bar{x},\bar{y}) - \beta\mathrm{KL}[\pi_\theta(\bar{y}\,|\,\bar{x}),\pi_{\text{ref}}(\bar{y}\,|\,\bar{x})]\right]$$

  - Where π_{ref} is the SFT policy before we fine-tune it with preference data

- The optimal policy takes this form (according to theoretical results from RL)

$$\pi^*(\bar{y}\,|\,\bar{x}) = \frac{1}{Z(\bar{x})}\pi_{\text{ref}}(\bar{y}\,|\,\bar{x})\exp\left(\frac{1}{\beta}r(\bar{x},\bar{y})\right)$$

- We can rearrange that to give:

$$r(\bar{x},\bar{y}) = \beta\log\frac{\pi^*(\bar{y}\,|\,\bar{x})}{\pi_{\text{ref}}(\bar{y}\,|\,\bar{x})} + \beta\log Z(\bar{x})$$

- Combine this with Bradley-Terry and…

Slide credit: Yoav Artzi

Direct Preference Optimization (DPO)

- Through some manipulation, it can be shown that the optimal policy π^* for RLHF satisfies the preference model

$$p^*(y_1 \succ y_2 \mid x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

ref = SFT policy. preferred output should be more likely under our learned policy than under reference, dispreferred output should be less likely

- We can now learn the policy directly to optimize the log likelihood of the preference data in a fashion that looks like supervised learning:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

Rafailov et al. (2023)

Direct Preference Optimization (DPO)

- The DPO gradient is:

$$\nabla \mathcal{L}_{\text{DPO}}(\theta) =$$

$$-\beta E_{(\bar{x},\bar{y}_w,\bar{y}_l)\sim\mathcal{D}}\left[\sigma\left(\hat{r}_\theta(\bar{x},\bar{y}_l) - \hat{r}_\theta(\bar{x},\bar{y}_w)\right)\left[\nabla \log \pi_\theta(\bar{y}_w \mid \bar{x}) - \nabla \log \pi_\theta(\bar{y}_l \mid \bar{x})\right]\right]$$

β functions like a "learning rate" following the strength of the KL constraint

Per-example weight: higher weight when the reward model is wrong

Increase likelihood of preferred example

Decrease likelihood of dispreferred example

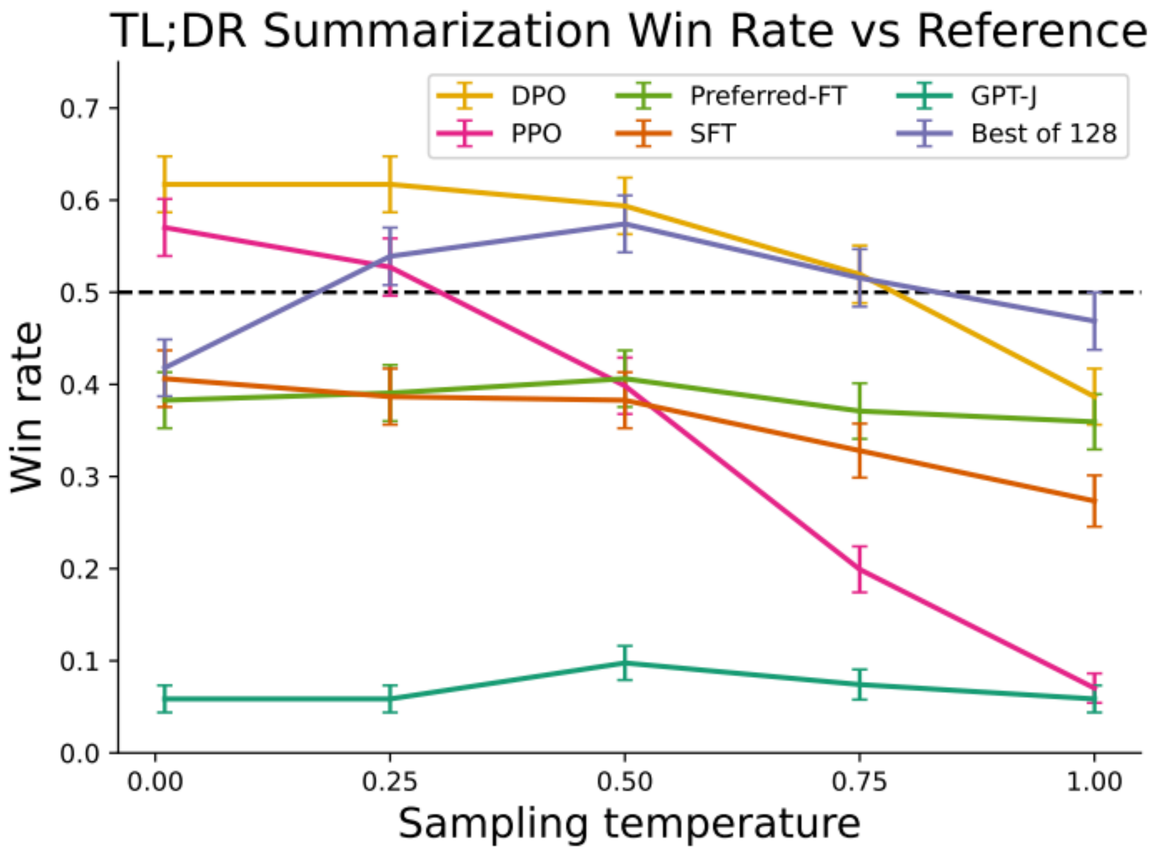
where $\hat{r}(\bar{x},\bar{y}) = \beta \log \frac{\pi_\theta(\bar{y} \mid \bar{x})}{\pi_{\text{ref}}(\bar{y} \mid \bar{x})}$

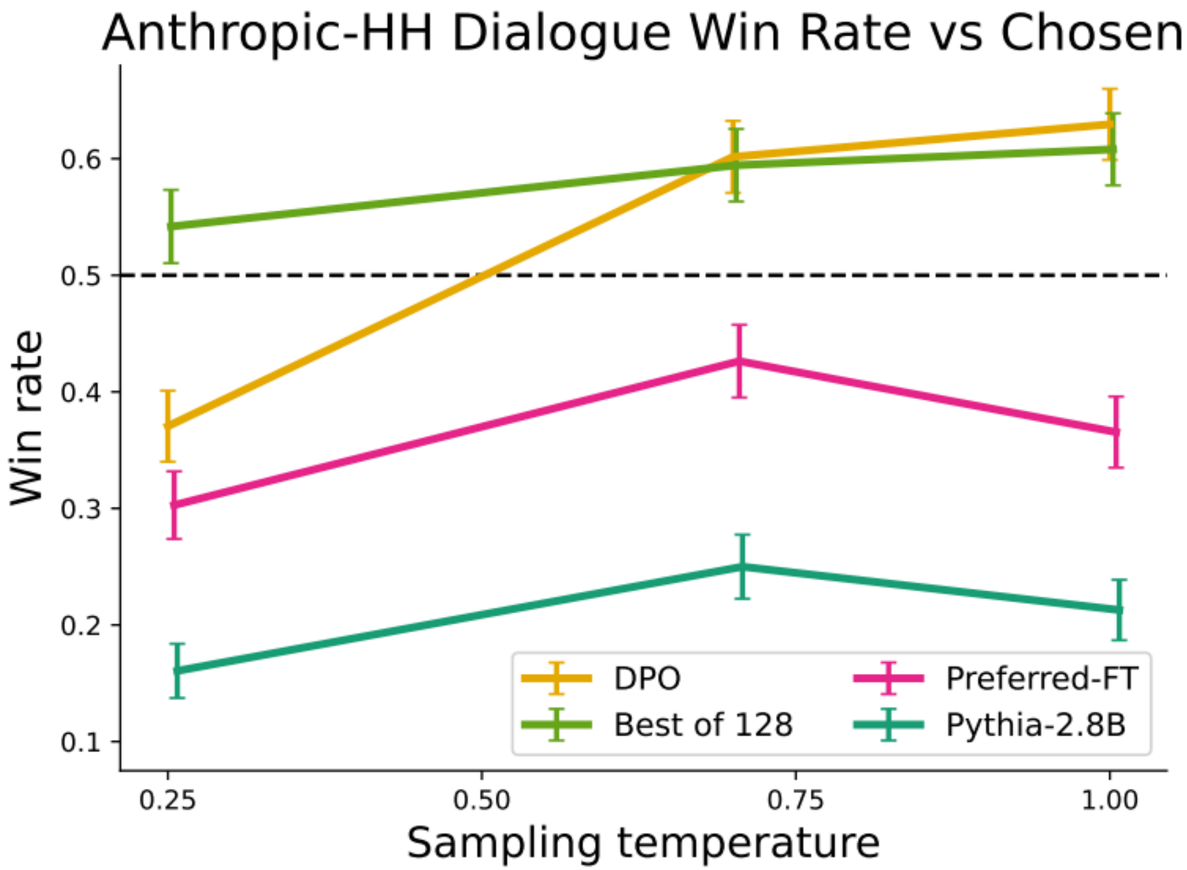
Slide credit: Yoav Artzi

Outcome of RLHF/DPO

- RLHF produces an “aligned” model that should achieve high reward
- Baselines:
  - Best-of-n: sample n responses from an SFT model, take the best one according to the reward function
    - Pro: training-free
    - Cons: expensive, may not deviate far from the initial SFT model
  - Preference tuning: apply SFT on preferred outputs
    - Pro: simple. Cons: doesn’t use the negative examples

DPO Results





- Evaluation: *win rate* (as scored by an LLM)

Rafailov et al. (2023)

RLHF in Llama 2

Dataset	Num. of Comparisons	Avg. # Turns per Dialogue	Avg. # Tokens per Example	Avg. # Tokens in Prompt	Avg. # Tokens in Response
Anthropic Helpful	122,387	3.0	251.5	17.7	88.4
Anthropic Harmless	43,966	3.0	152.5	15.7	46.4
OpenAI Summarize	176,625	1.0	371.1	336.0	35.1
OpenAI WebGPT	13,333	1.0	237.2	48.3	188.9
StackExchange	1,038,480	1.0	440.2	200.1	240.2
Stanford SHP	74,882	1.0	338.3	199.5	138.8
Synthetic GPT-J	33,139	1.0	123.3	13.0	110.3
Meta (Safety & Helpfulness)	1,418,091	3.9	798.5	31.4	234.1
Total	2,919,326	1.6	595.7	108.2	216.9

RLHF data for Llama 2

- They do 5 iterations of (train, get more preferences, get new reward model). First 3 iterations: just fine-tuning best-of-n, then they used PPO
- Current approaches: many papers exploring versions with active data collection (e.g., tune with DPO -> collect preferences -> keep tuning …)

Touvron et al. (2023)

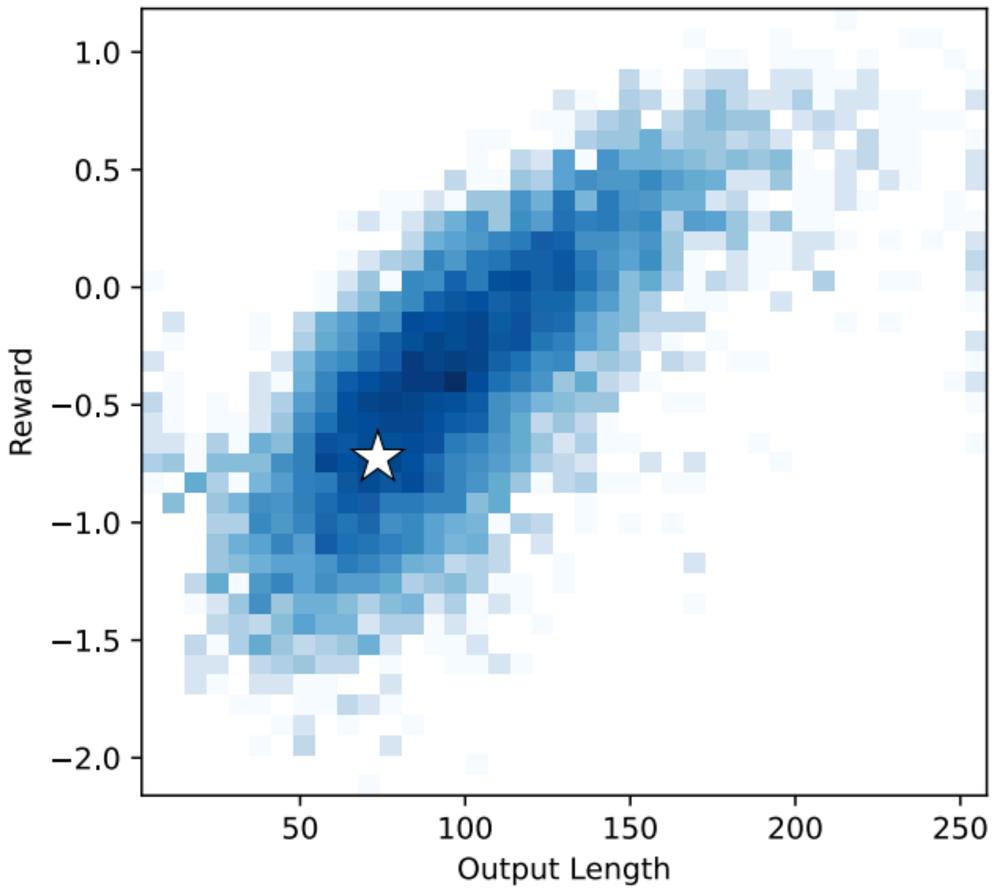
Is DPO the best?

Model	Factuality	Reasoning	Coding	Truthfulness	Safety	Instr. Foll.	Average
Llama 2 13B Base	52.0	37.0	30.7	32.7	32.7	-	-
Llama 2 Chat 13B [52]	53.2	24.7	36.9	**88.0**	91.9	51.2	57.7
Nous Hermes 13B [51]	53.2	43.5	47.7	80.5	43.9	38.7	51.3
Vicuna 1.5 13B [64]	54.5	39.3	38.5	62.8	92.4	45.8	55.6
TÜLU 2 13B SFT	55.4	47.8	45.1	56.6	91.8	44.2	56.8
TÜLU 2+DPO 13B	55.3	50.9	45.9	69.3	91.9	52.8	61.0
TÜLU 2+PPO 13B (13B UFRM)	56.0	52.0	47.7	71.5	91.8	54.4	62.2
TÜLU 2+PPO 13B (70B UFRM)	55.4	53.9	47.3	72.3	91.9	55.8	62.8
TÜLU 2+PPO 13B (70B UFRM+MP)	55.3	53.1	48.4	71.0	**92.7**	54.0	62.4
L3+TÜLU 2 8B SFT	58.0	58.6	56.4	59.2	92.8	42.6	61.3
L3+TÜLU 2+DPO 8B	59.4	56.2	55.6	71.4	91.7	50.4	64.1
L3+TÜLU 2+PPO 8B (8B UFRM)	**59.5**	57.0	**55.9**	69.6	91.4	**56.0**	64.9
L3+TÜLU 2+PPO 8B (70B UFRM)	58.5	**60.8**	55.0	72.8	91.8	55.8	**65.8**
L3+TÜLU 2+PPO 8B (70B UFRM+MP)	58.3	40.6	48.2	62.4	91.0	53.0	58.9

Table 5: **Putting together a recipe for preference-based learning:** Performance of our best-performing models along with popular open models based on Llama 2 13B. 'MP' refers to using the mixed prompt set described in §4. 'L3' stands for experiments using Llama 3 as a base model. Using PPO with a large reward model performs best overall.

Ivison et al. (2024)

What does RLHF do?



Question: *Why don't adults roll off the bed?*

☆ **SFT (Before); 59 tokens**

Adults typically do not roll off of the bed because they have developed the muscle memory to keep their bodies from involuntarily moving during sleep and maintaining proper posture.

RLHF (After); 243 tokens: Similar output, but much longer / more details

Adults generally do not roll off of the bed because they have developed muscle control and awareness which helps them to remain upright while sleeping.

Additionally, most adults find it uncomfortable or even painful to move around in their sleep, so rolling off the bed would be difficult without waking up first. There may also be safety concerns such as falling out of bed and potentially hurting oneself if one were to lose balance or fall from the ...

- Reward models trained on open datasets have high correlations with length

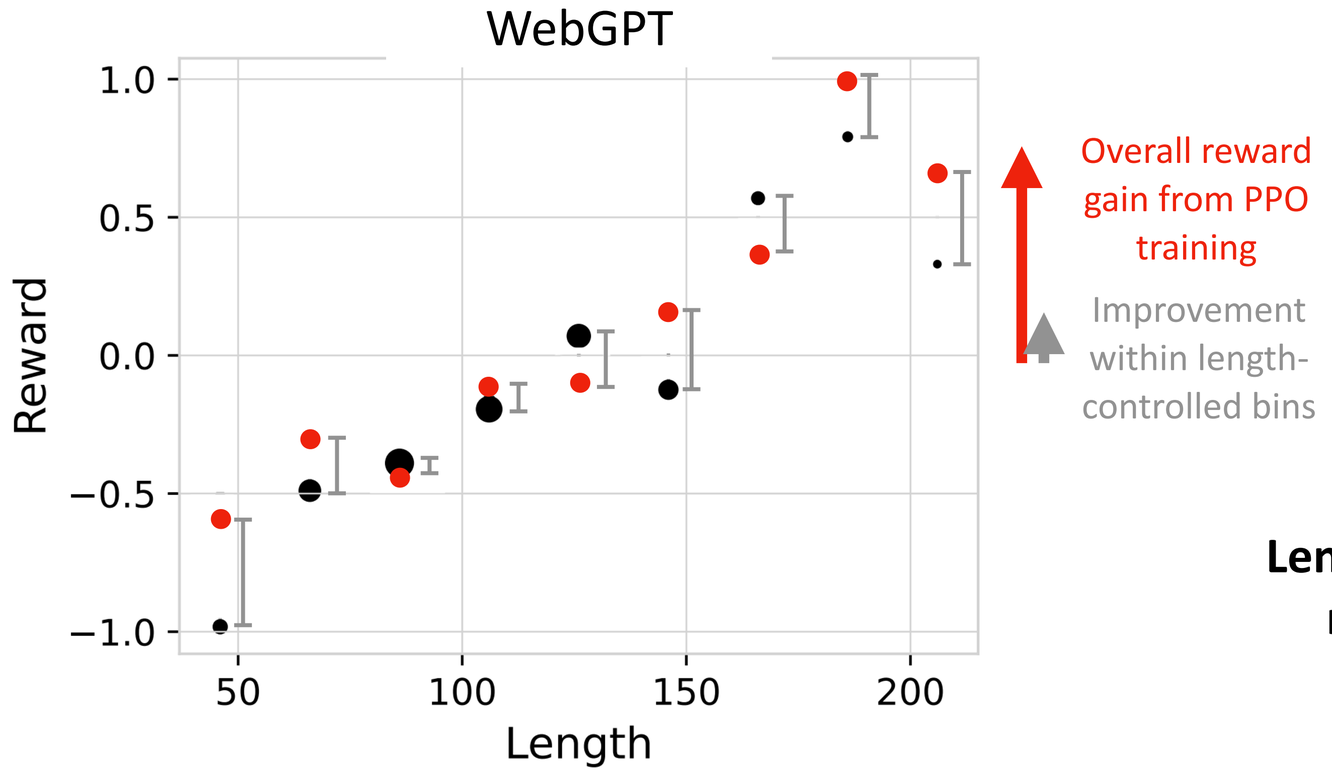
(Prasann Singhal, Tanya Goyal, Jiacheng Xu, GD, COLM 2024 oral spotlight)

What does RLHF do?

On older preference dataset, most reward optimization was attributable to shifting to longer outputs! (Modern datasets are much bigger and this effect is reduced)

Average bin reward (RLHF model outputs)

Average bin reward (SFT model outputs)



Overall reward gain from PPO training

Improvement within length-controlled bins

Length accounts for 85% of reward improvement

(Prasann Singhal, Tanya Goyal, Jiacheng Xu, GD, COLM 2024 oral spotlight)

Administrative details and recap

Supervised Fine-Tuning

SFT History

Understanding SFT Data

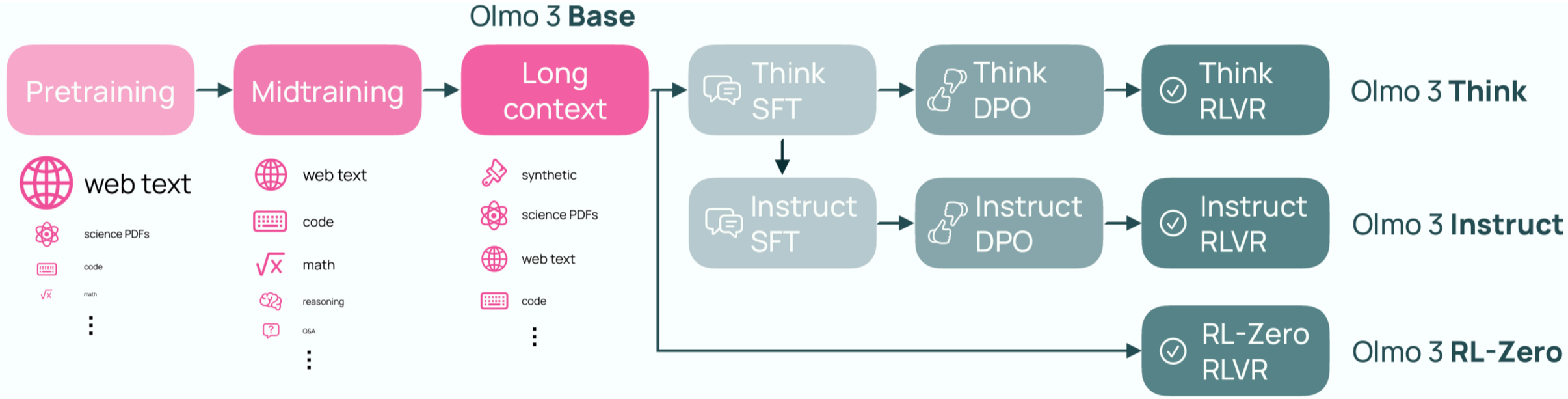
RLHF

DPO

Putting it together: Olmo3, OpenThoughts

Putting it together: Olmo 3

Olmo 3 SFT



- Several stages where SFT and RLHF is used:
  - Midtraining
  - Think/Instruct SFT
  - Think/Instruct DPO

Olmo 3 Midtraining

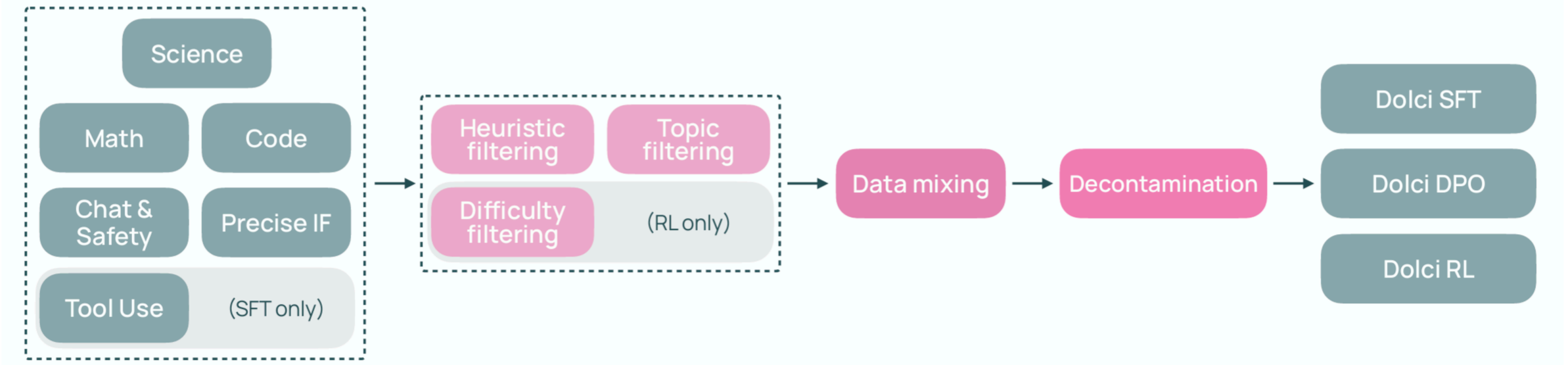
- "Midtraining" is a phase when models are still "pre-trained" using the next-token prediction objective, but on data that looks more like post-training data (including SFT!)

QA and knowledge access capabilities We target improvements in question-answering and general knowledge access capabilities through synthesis of two novel datasets focused on particular QA capabilities, as well as inclusion of high-quality existing QA data. The final datasets included for these capabilities are the following:

- **Reddit-to-Flashcards** We synthesize this dataset in response to the need to handle diverse content categories and question structures in multiple-choice QA tasks. We first identify a subset of academically-relevant subreddits, and then use GPT 4o-mini to rewrite submission-comment pairs from those subreddits into multiple-choice QA pairs. We use seven task formats to increase diversity. Microanneals show that inclusion of 5B tokens of this data in a 10B-token microanneal resulted in over 2 points of improvement in the $\text{MC}_{\text{Non-STEM}}$ task cluster—relative to a 10B-token web-only baseline microanneal—with 3 points of improvement in MMLU.

- And Tulu 3 SFT data as well

Olmo 3 SFT


Science
Math
Code
Chat & Safety
Precise IF
Tool Use
(SFT only)
Heuristic filtering
Topic filtering
Difficulty filtering
(RL only)
Data mixing
Decontamination
Dolci SFT
Dolci DPO
Dolci RL

Olmo 3 SFT

- **Math** We source prompts from the math subsets of OpenThoughts3 (Guha et al., 2025a) and SYNTHETIC-2 (PrimeIntellect, 2025). For OpenThoughts3 prompts, we use all the available math prompts (maintaining the 16X repetition from the original) and the available reasoning traces with complete solutions. For incomplete traces, we generate full reasoning chains and solutions using QwQ-32B, the original model used for the completions, and the same generation settings as OpenThoughts3, except up to 32K tokens instead of the original 16K. We discard any examples that are still incomplete after regenerating. For SYNTHETIC-2, we take completions directly from the verified subsection.
- **Code** We collect code prompts from different sources and generate completions for them. To create DOLCI THINK Python Algorithms, we source prompts from AceCoder (Zeng et al., 2025a), the Python subset of The Algorithms (The Algorithms, 2025), Llama Nemotron Post-training (Bercovich et al., 2025), and OpenCodeReasoning (Ahmad et al., 2025), and then we generate up to 16 responses per prompt from QwQ-32B, which we filter for correctness using synthetically generated test cases from GPT-4.1. For OpenThoughts 3 code prompts, we downsample each prompt to at most 16 times and regenerate complete

Olmo 3 SFT

- **Chat & safety** We source chat prompts from both the Tülu 3 (Lambert et al., 2024) subset of WildChat (Zhao et al., 2024a), as well as WildChat prompts not used during Tülu 3, and the Tülu 3 subset of OpenAssistant (Köpf et al., 2024). For safety, we reuse safety prompts used during Tülu 3. We then generate reasoning traces and completions from DeepSeek R1 (Guo et al., 2025).
- **Precise instruction following** We source precise IF prompts from the overall Tülu 3 mix with additional verifiable constraints added from Pyatkin et al. (2025). We also regenerate Persona IF prompts as in Tülu 3, but with personas sourced from Meyer and Corneil (2025). We then generate responses for each prompt using QwQ-32B, and we verify responses using verifiers associated with each constraint, keeping only the correct responses.
- **Science & other** We source science prompts from the OpenThoughts3 science subset. For other data sources, we include the TableGPT (Zha et al., 2023) subset in Tülu 3 for data transformation and Aya (Singh et al., 2024) for chat and basic multilinguality. We regenerate incomplete responses in OpenThoughts3 as we did for the math and code subsets, and we generate responses with reasoning chains for the other datasets using DeepSeek R1.

Category	Prompt Dataset	7B Count	32B Count	Reference
Chat &	WildChat	83,054	76,209	Zhao et al. (2024a)
Precise IF	OpenAssistant	6,800	6,647	Köpf et al. (2024)
	Dolci Think Persona Precise IF	223,123	220,530	–
	Dolci Think Precise IF	135,792	135,722	–
Math	Dolci Think OpenThoughts 3+ Math⇑	752,997	752,997	Guha et al. (2025a)
	Dolci Think OpenThoughts 3+ STEM⇑	99,269	99,268	Guha et al. (2025a)
	SYNTHETIC-2-SFT-Verified	104,569	104,548	PrimeIntellect (2025)
Coding	Nemotron Post-Training Code	113,777	113,777	NVIDIA AI (2025)
	Dolci Think OpenThoughts 3+ Code⇑	88,900	88,899	Guha et al. (2025a)
	Dolci Think Python Algorithms⇑	466,677	466,676	–
Safety	CoCoNot	10,227	9,549	Brahman et al. (2024)
	WildGuardMix	38,315	36,673	Han et al. (2024)
	WildJailbreak	41,100	40,002	Jiang et al. (2024)
Multilingual	Aya	98,597	97,156	Singh et al. (2024)
Other	TableGPT	4,981	4,973	Zha et al. (2023)
	Olmo Identity Prompts	290	290	–
Total		2,268,468	2,253,916	

Table 17 Olmo 3 Think SFT prompt sources. ⇑ indicates prompt datasets where the datasets are upsampled by repeating prompts with different completions. Prior to Olmo 3 32B training, we filter responses with non-Olmo model identities and irrelevant prompts (e.g. generate a photo).

Category	Prompt Dataset	# Prompts used in DPO	Reference
Chat & Precise IF	WildChat	40,701	Zhao et al. (2024a)
	Dolci Instruct Precise IF	19,365	–
	Tülu 3 Persona IF	3,486	Lambert et al. (2024)
	OpenAssistant	1,762	Köpf et al. (2024)
Math	Tülu 3 Persona MATH	10,657	Lambert et al. (2024)
	Tülu 3 Persona Algebra	1,417	Lambert et al. (2024)
	Tülu 3 Persona GSM	3,681	Lambert et al. (2024)
	OpenMathInstruct 2	3,615	Toshniwal et al. (2024)
Coding	Dolci Instruct Python Algorithms	13,236	–
	Tülu 3 Persona Python	2,514	Lambert et al. (2023)
	Evol CodeAlpaca	7,634	Luo et al. (2023)
Safety	CoCoNot	927	Brahman et al. (2024)
	WildGuardMix	5,338	Han et al. (2024)
	WildJailbreak	5,616	Jiang et al. (2024)
Science	SciRiff	2,253	Wadden et al. (2024)
	OpenThoughts3 Science	19,023	Guha et al. (2025a)
Multilingual	Aya	4,078	Singh et al. (2024)
Other	TableGPT	1,170	Zha et al. (2023)
	FLAN	19,660	Wei et al. (2021)
Not used in SFT	DaringAnteater	1,089	Wang et al. (2024b)
	UltraFeedback	32,778	Cui et al. (2023)
Total		200,000	

Administrative details and recap

Supervised Fine-Tuning

SFT History

Understanding SFT Data

RLHF

DPO

Putting it together: Olmo3, OpenThoughts

Putting it together: OpenThoughts

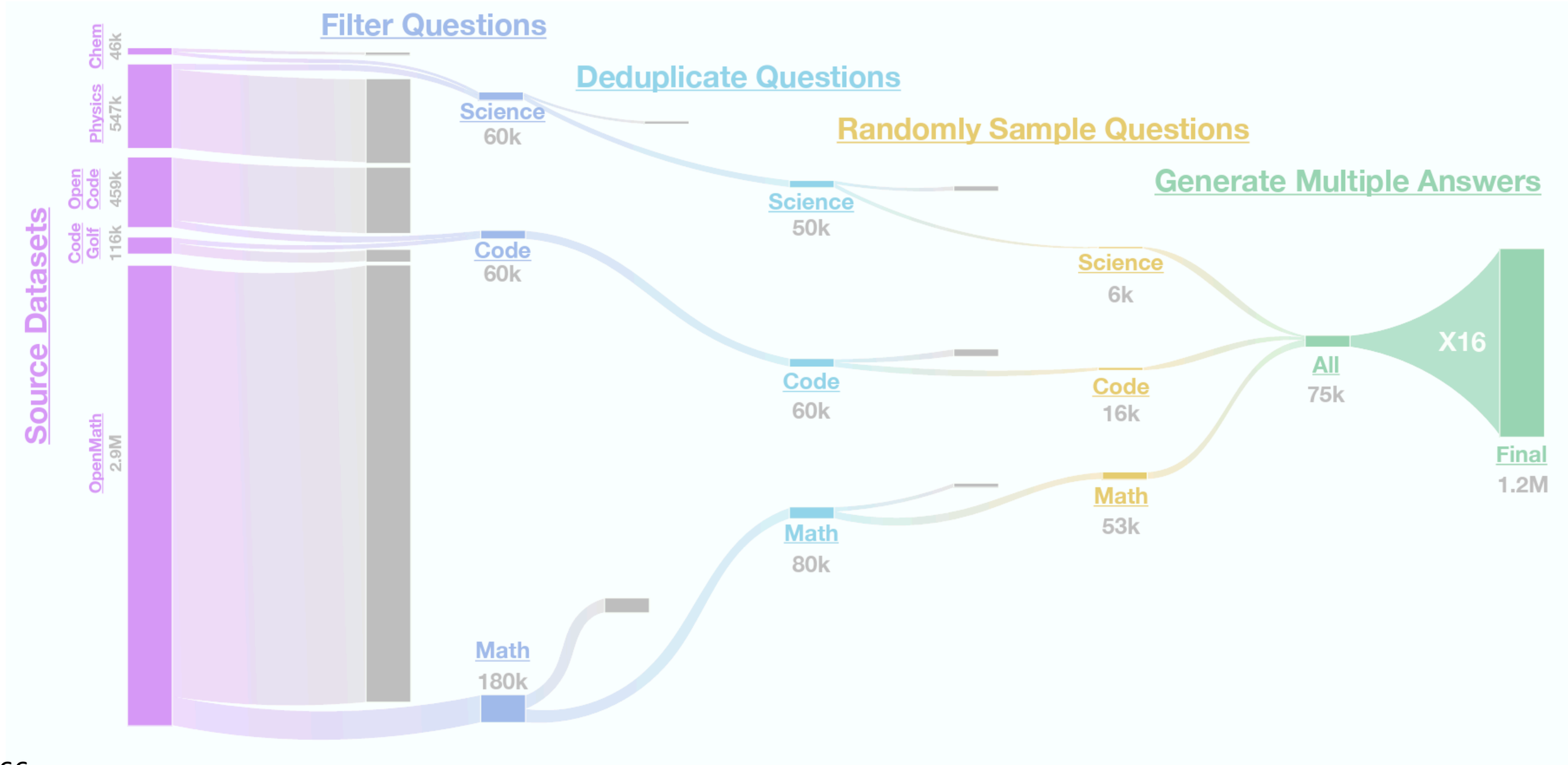
Open Thoughts

Data Recipes for Reasoning Models

Etash Guha*[1,2], Ryan Marten*[3], Sedrick Keh*[4], Negin Raoof*[5], Georgios Smyrnis*[6], Hritik Bansal[ζ7], Marianna Nezhurina[ζ8,9,16], Jean Mercat[ζ4], Trung Vu[ζ3], Zayne Sprague[ζ6], Ashima Suvarna[7], Benjamin Feuer[10], Liangyu Chen[1], Zaid Khan[11], Eric Frankel[2], Sachin Grover[12], Caroline Choi[1], Niklas Muennighoff[1], Shiye Su[1], Wanjia Zhao[1], John Yang[1], Shreyas Pimpalgaonkar[3], Kartik Sharma[3], Charlie Cheng-Jie Ji[3], Yichuan Deng[2], Sarah Pratt[2], Vivek Ramanujan[2], Jon Saad-Falcon[1], Jeffrey Li[2], Achal Dave, Alon Albalak[13], Kushal Arora[4], Blake Wulfe[4], Chinmay Hegde[10], Greg Durrett[6], Sewoong Oh[2], Mohit Bansal[11], Saadia Gabriel[7], Aditya Grover[7], Kai-Wei Chang[7], Vaishaal Shankar, Aaron Gokaslan[14], Mike A. Merrill[1], Tatsunori Hashimoto[1], Yejin Choi[1], Jenia Jitsev[8,9,16], Reinhard Heckel[15], Maheswaran Sathiamoorthy[3], Alexandros G. Dimakis[†3,5], Ludwig Schmidt[†1]

[1]Stanford University, [2]University of Washington, [3]BespokeLabs.ai, [4]Toyota Research Institute, [5]UC Berkeley, [6]UT Austin, [7]UCLA, [8]JSC, [9]LAION, [10]NYU, [11]UNC Chapel Hill, [12]ASU, [13]Lila Sciences, [14]Cornell Tech [15]TUM [16]Open-Ψ (Open-Sci) Collective

Pipeline for distillation of a strong reasoning model (QwQ-32B)



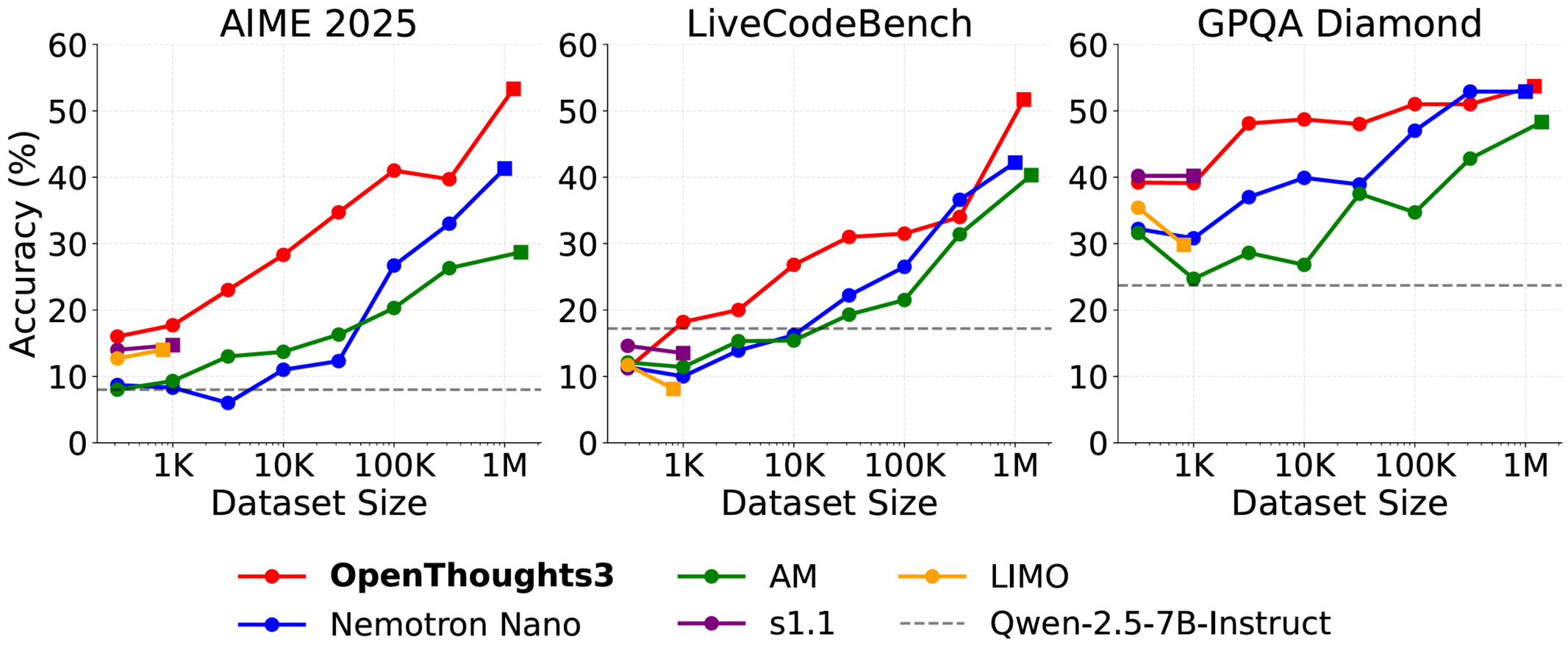
Scaling Up CoT

High-quality, filtered set of questions was important

Answer generation with QwQ-32B. **Removing incorrect answers didn't help**

QwQ was the best teacher, even though R1 achieves stronger performance!

Scaling Up CoT: Distillation



Scaling training data (for a 7B model) leads to scaled-up performance.

7B models can still learn to execute the kinds of reasoning method used in R1/frontier models.

Administrative details and recap

Supervised Fine-Tuning

SFT History

Understanding SFT Data

RLHF

DPO

Putting it together: Olmo3, OpenThoughts

Next time

- GRPO: what's the method that led to the models we're distilling for reasoning?