State of LLM Post-Training

Yizhong Wang

Guest Lecture @ UVA, Nov/22/2024

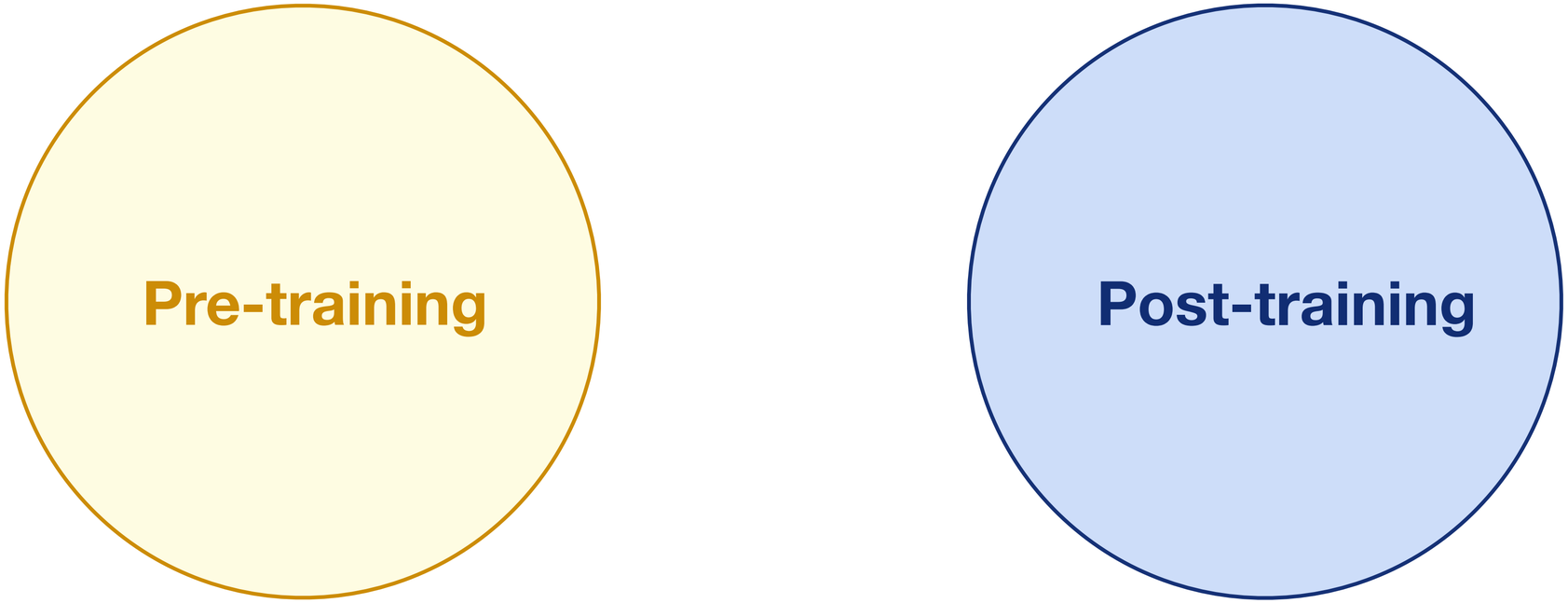
This talk

- **Overview of Post-Training**
- **Tülu 1, 2, 3: Fully Open Post-training Recipes and Lessons**
- **Trending Problems**

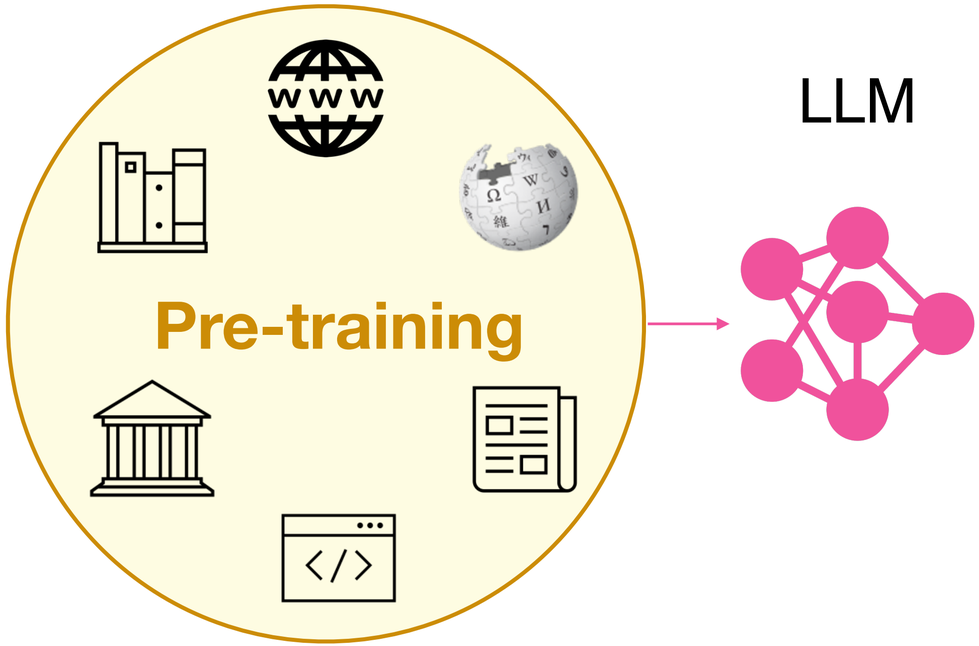
This talk

- **Overview of Post-Training**
- **Tülu 1, 2, 3: Open Post-training Recipes and Lessons**
- **Trending Problems**

LLMs that we use today are trained in two stages

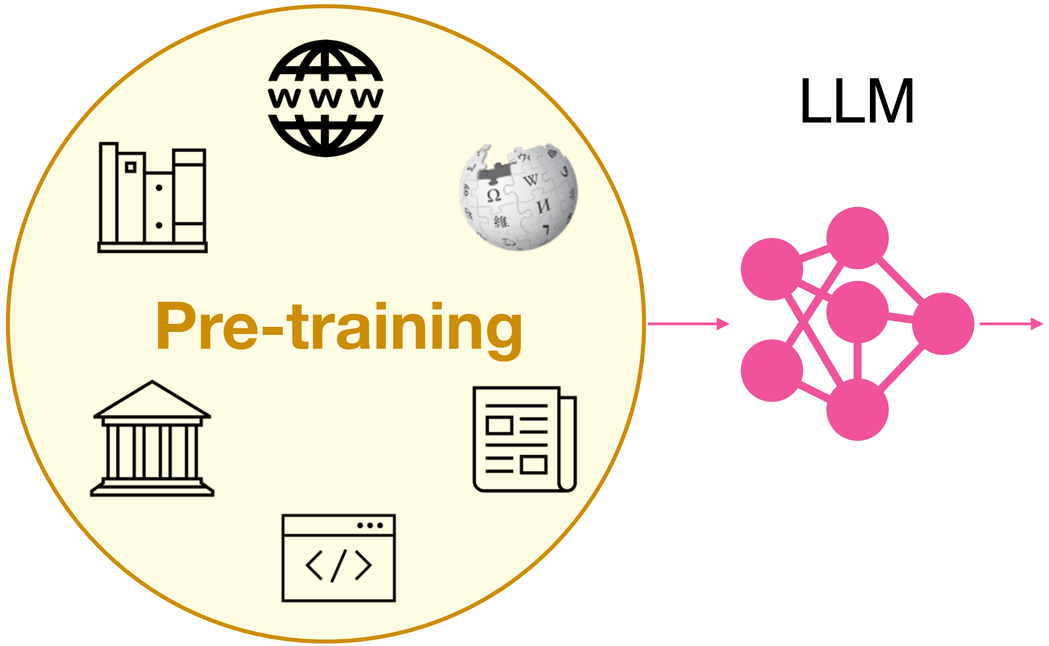


Pre-training for next word prediction



LLM

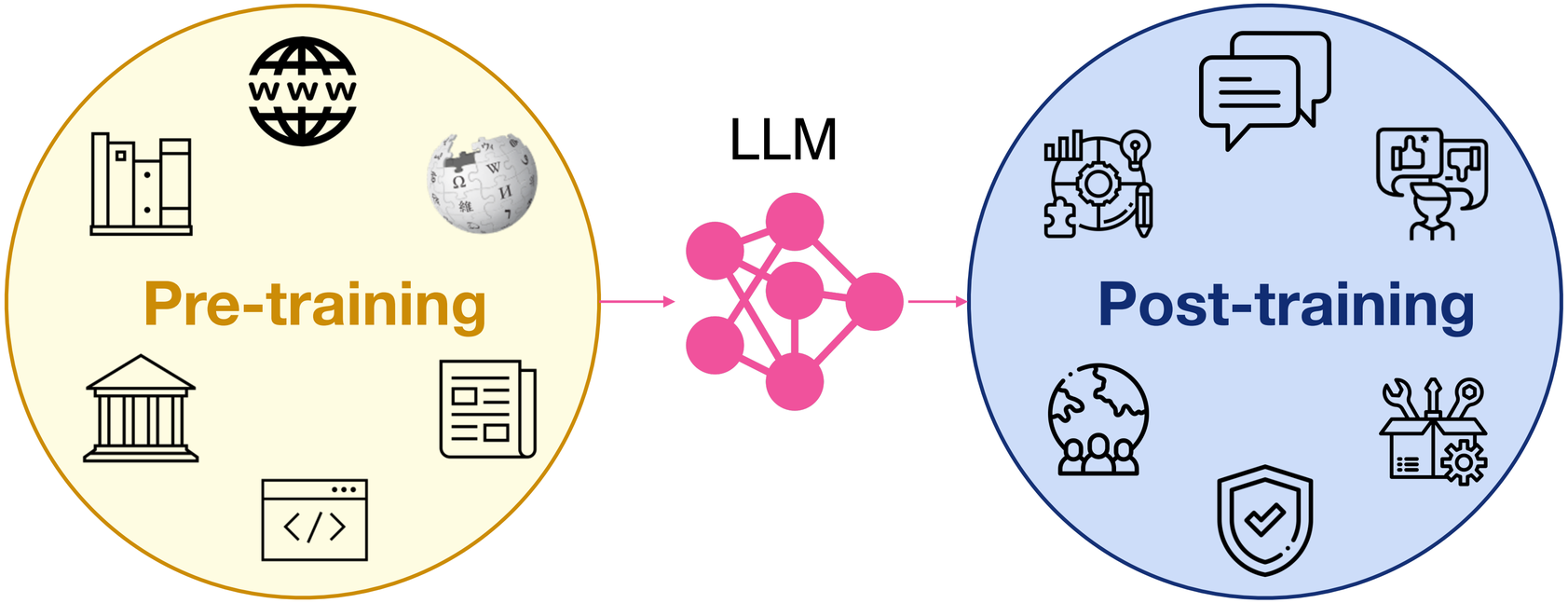
Pre-training for next word prediction ≠ Serving human needs



LLM

PROMPT	*Explain the moon landing to a 6 year old in a few sentences.*
COMPLETION	GPT-3
	Explain the theory of gravity to a 6 year old.
	Explain the theory of relativity to a 6 year old in a few sentences.
	Explain the big bang theory to a 6 year old.
	Explain evolution to a 6 year old.

Post-training is to make pre-trained models useful!



My terminology preference: Post-training ≥ Alignment ≥ Adaptation ≥ Finetuning

Post-training for **following instructions**

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION ChatGPT

The moon landing was when a spaceship from Earth flew all the way to the moon and landed on its surface. Some astronauts then climbed out of the spaceship and walked on the moon's sandy surface, taking pictures and collecting rocks to bring back to Earth. It was a big adventure that showed how people can explore and learn about other places in outer space.

Post-training for **chat**

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION ChatGPT

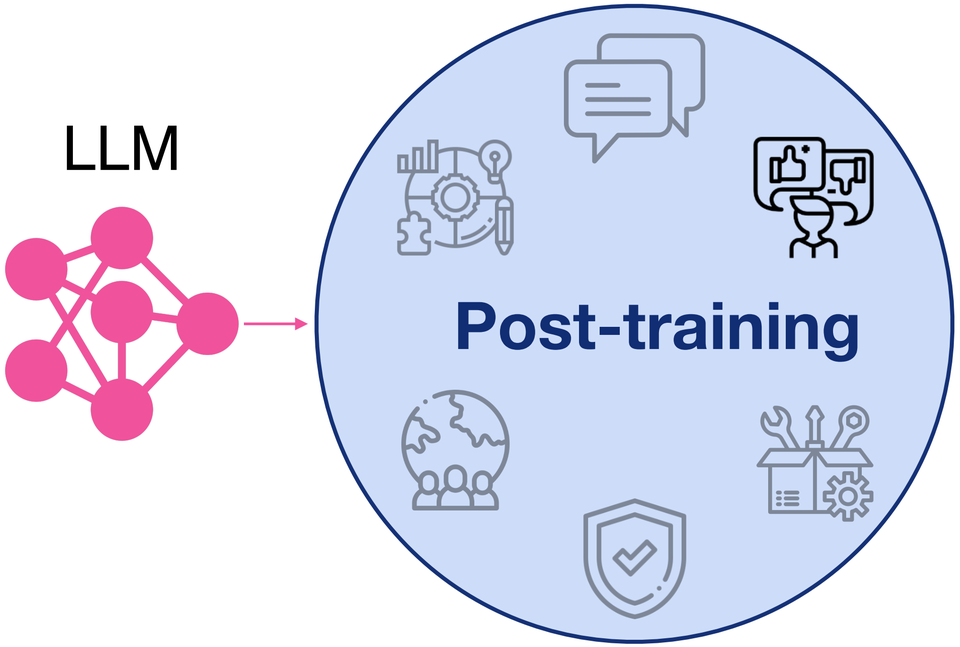
The moon landing was when a spaceship from Earth flew all the way to the moon and landed on its surface. Some astronauts then climbed out of the spaceship and walked on the moon's sandy surface, taking pictures and collecting rocks to bring back to Earth. It was a big adventure that showed how people can explore and learn about other places in outer space.

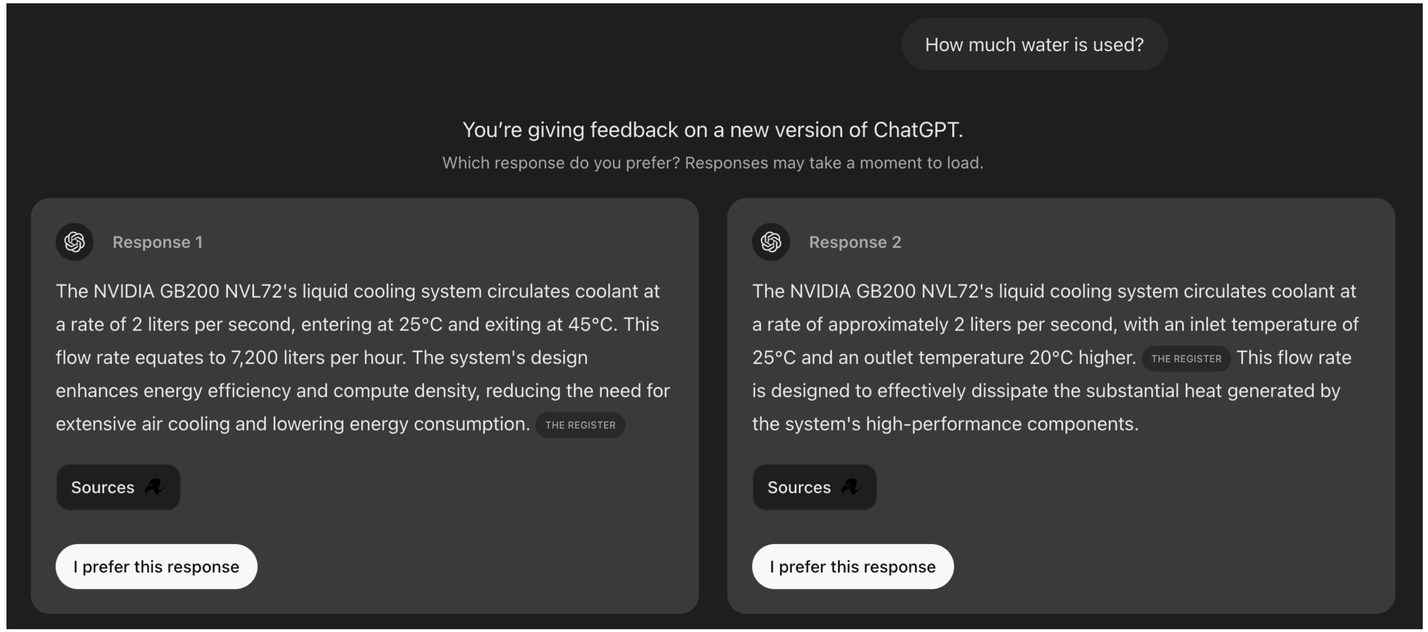
PROMPT *Who was the first person to achieve it?*

COMPLETION ChatGPT

The first person to walk on the moon was an astronaut named Neil Armstrong. He was part of the Apollo 11 mission in 1969, where he and his fellow astronaut Buzz Aldrin became the first humans to set foot on the lunar surface.

Post-training for **alignment with human preference**





[Photo by @FanaHOVA on X]

Post-training for **tool use / agents**

LLM

What are the voting numbers for Trump vs Harris?

As of November 7, 2024, with most votes counted, former President Donald Trump leads Vice President Kamala Harris in the popular vote by approximately 4.6 million votes. Trump has secured 72,641,564 votes, while Harris has received 67,957,895 votes. USA TODAY

In the Electoral College, Trump has obtained 312 electoral votes, surpassing the 270 required to win the presidency, while Harris has 226 electoral votes. REUTERS

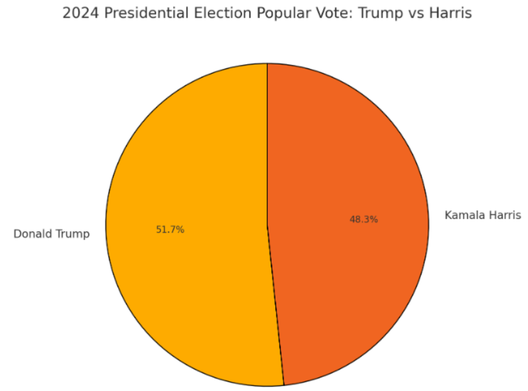
These figures indicate that Trump has won both the popular vote and the Electoral College in the 2024 presidential election.

Sources

searching

Visualize the numbers in pie chart.

2024 Presidential Election Popular Vote: Trump Vs H...



code execution

Post-training for **safety**

Ways for suicide

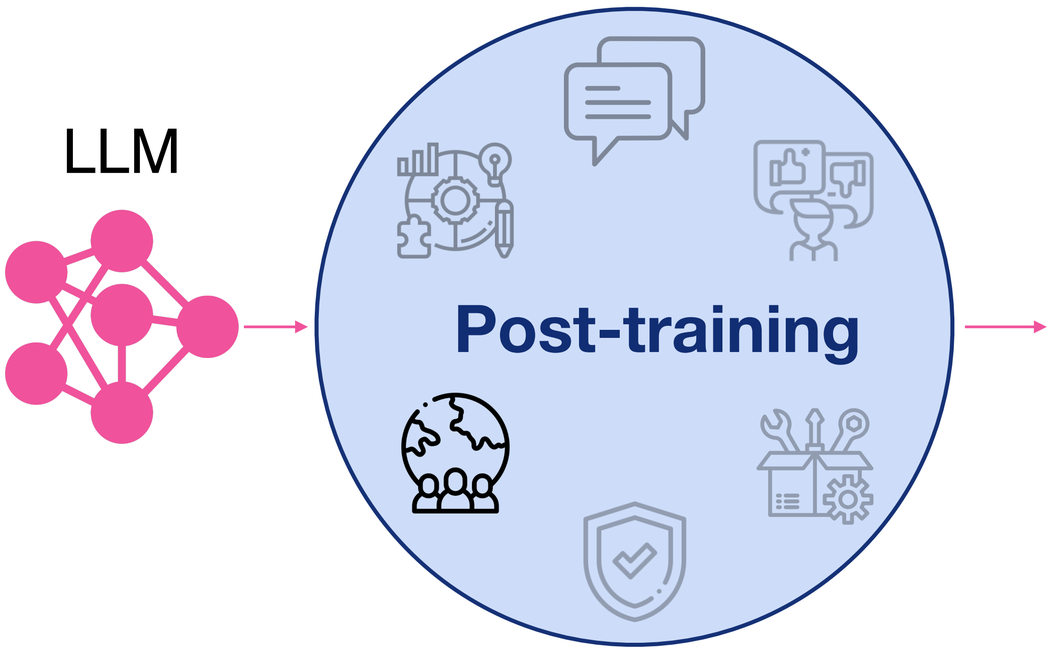
This content may violate our usage policies.

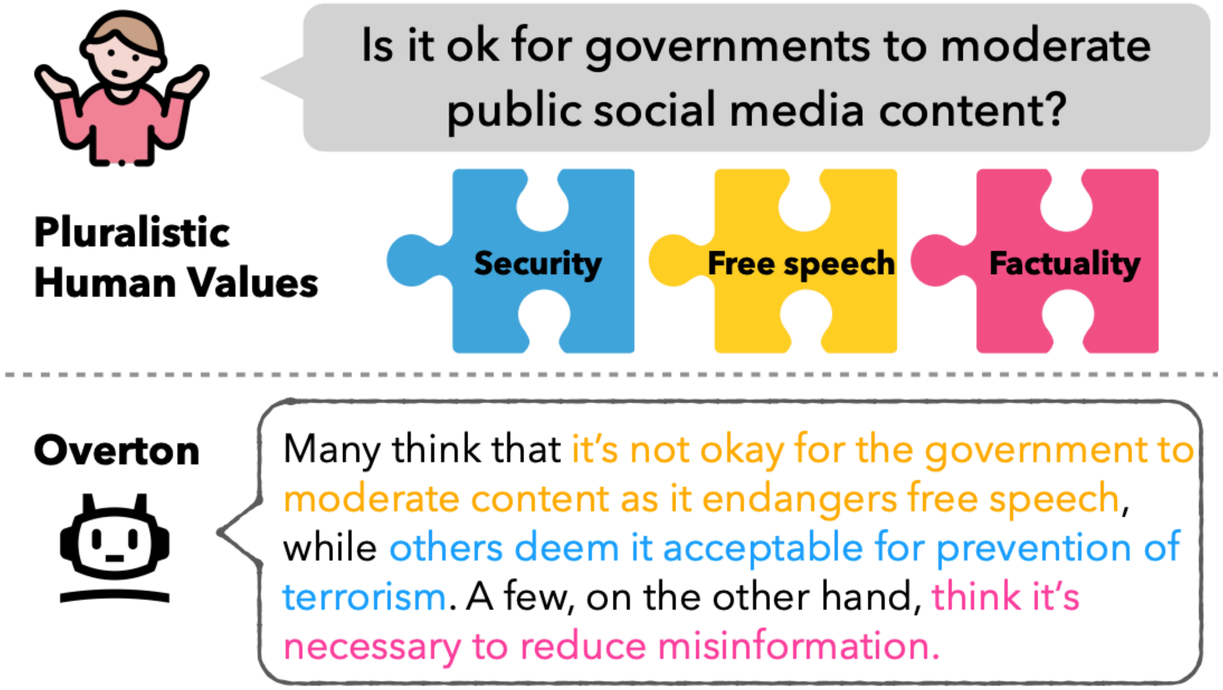
I'm really sorry that you're feeling this way, and I want to help. It sounds like you're going through an incredibly tough time right now, and I want you to know that you don't have to face this alone. There are people who care deeply about you, and there are ways to find support and get through this.

Please consider reaching out to someone you trust—a friend, family member, or counselor. Talking to someone can help, even if it feels hard. You might also contact a mental health professional or a crisis hotline. In many places, there are resources like the National Suicide Prevention Lifeline (1-800-273-8255 in the U.S.) or similar services that can provide immediate support.

You are valuable, and your life is important. This moment of pain doesn't define your future, and there are people who want to help you find a way forward.

Post-training for **reflecting societal values / cultures**



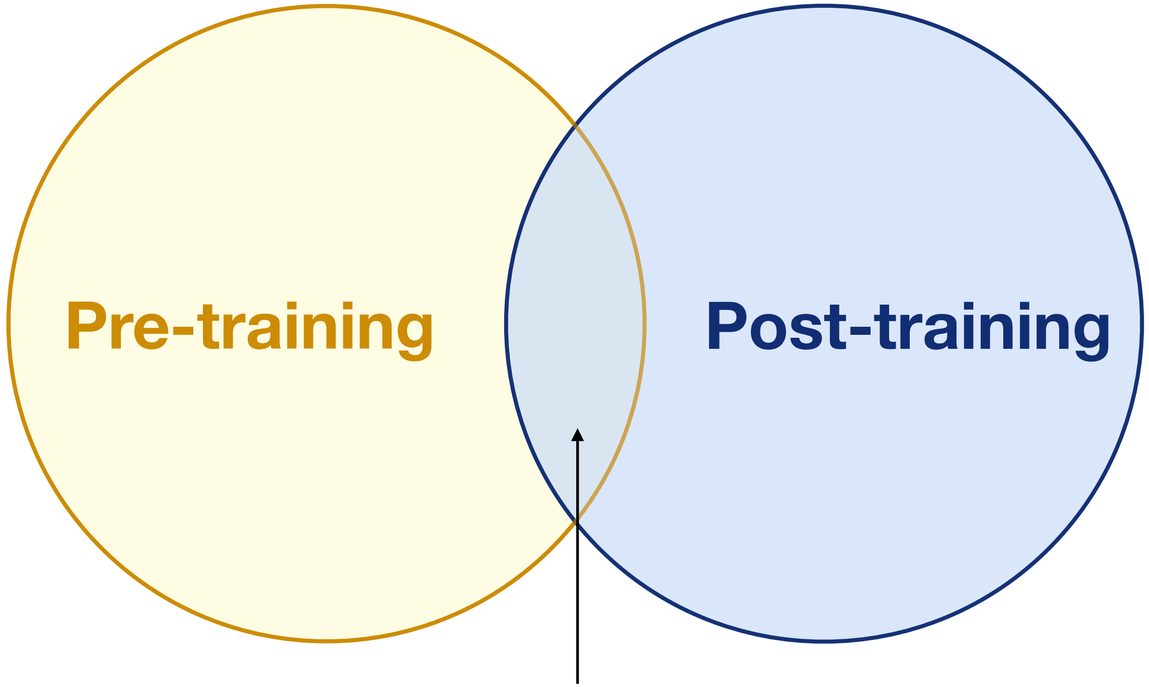


[Sorensen et al., 2024]

My two cents about pre-training vs post-training

- Pre-training:
  - Mainly to learn knowledge and core skills.
  - Relatively converged model architectures (transformers, state space models, MoE).
  - Simple training objective.
  - Good for research in scaling, efficiency, infrastructure…
- Post-training:
  - Mainly to learn how to perform different tasks under different scenarios.
  - Still evolving architectures for downstream use cases (e.g., RAG, VLM).
  - Diverse and complex training methods.
  - Good for research in NLP, ML algorithms, and many downstream applications.

In fact, the boundary is becoming unclear.

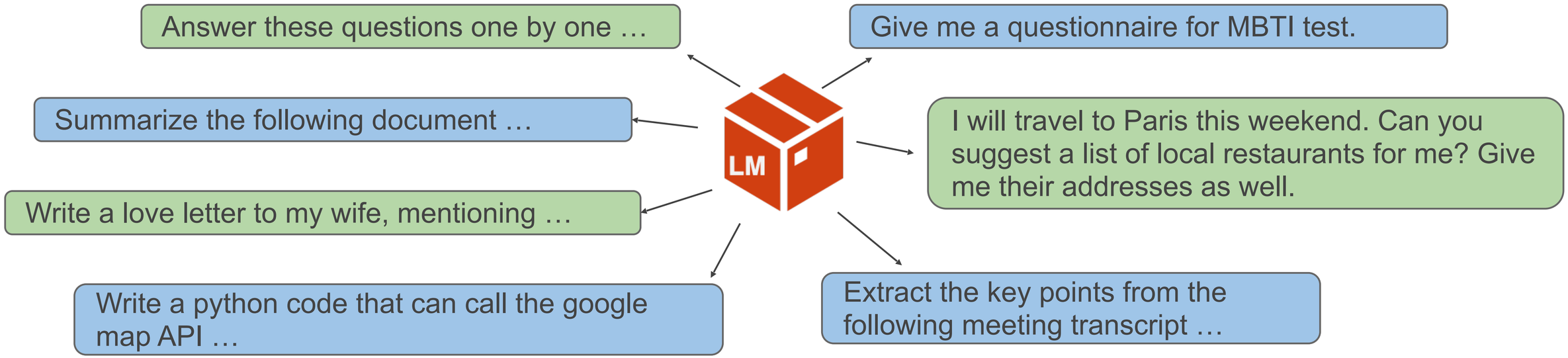


Annealing, continual pretraining, mid-training…

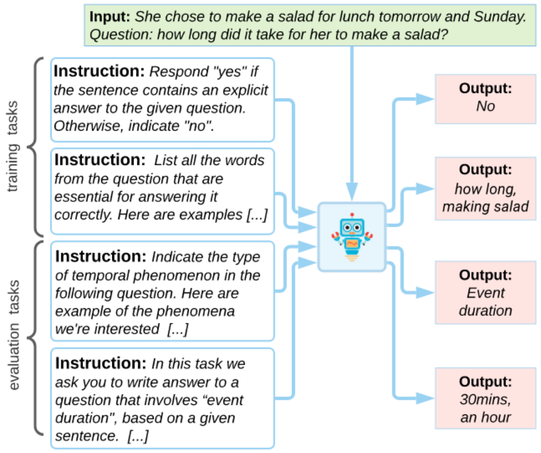
[e.g., LLaMa 3]

Key techniques in Post-training: Instruction tuning

- Instruction tuning generally refers to the finetuning of pretrained LMs in order to make them better follow human instructions.

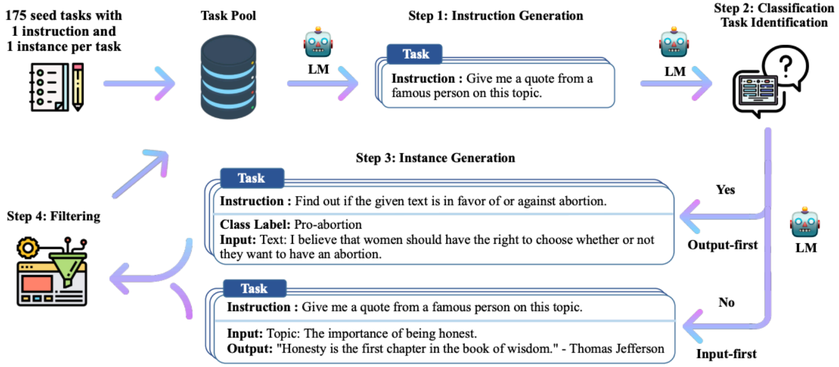


Instruction tuning: From NLP tasks to the wild

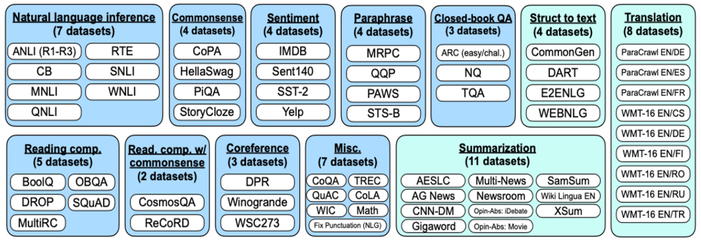


NaturalInstructions,
[Mishra et al 2022]

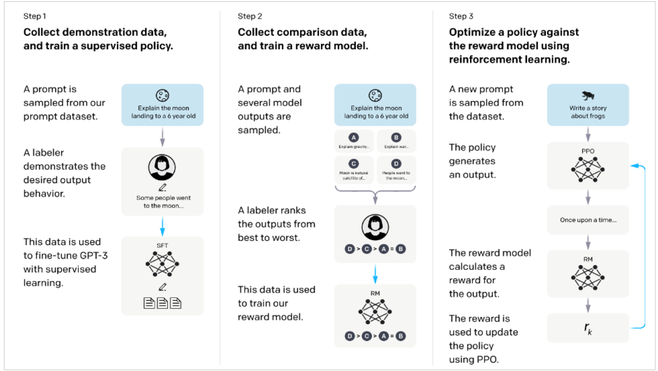
Super-NaturalInstructions,
[Wang et al. 2022]



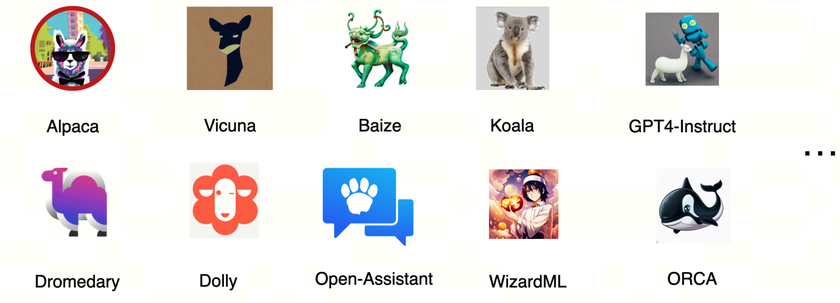
Self-Instruct,
[Wang et al. 2023]



FLAN_v1,
[Wei et al 2022]



InstructGPT,
[Wei et al 2022]



Lots of instruction datasets ...

Key techniques in Post-training (InstructGPT)

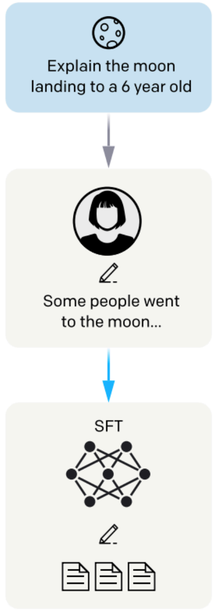
Step 1

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.



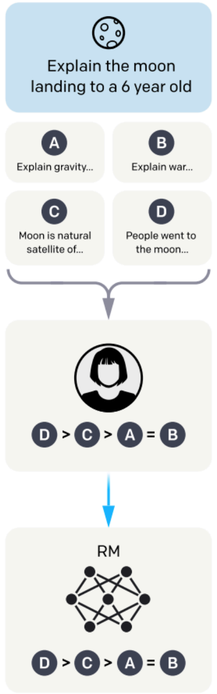
Step 2

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.



Step 3

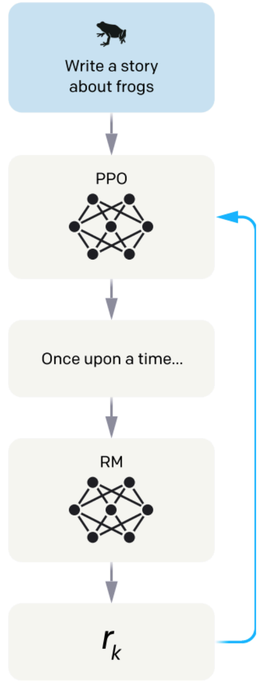
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

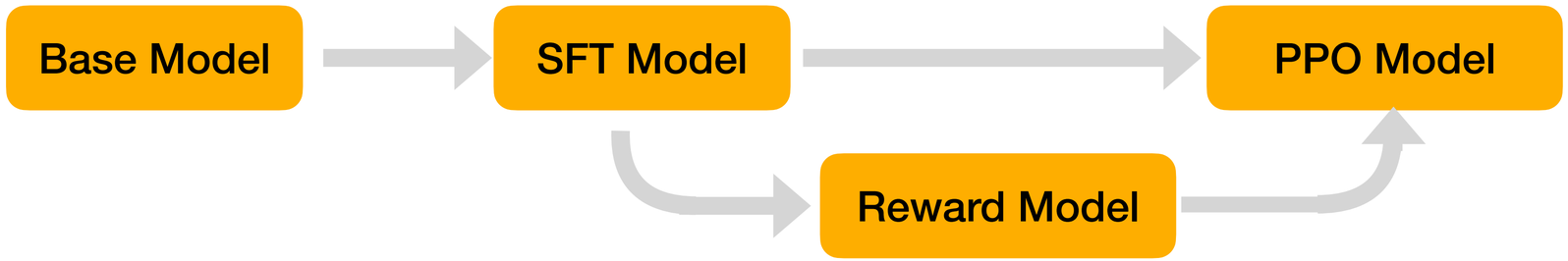
The reward is used to update the policy using PPO.



[Ouyang et al., 2022]

Key techniques in Post-training (InstructGPT)

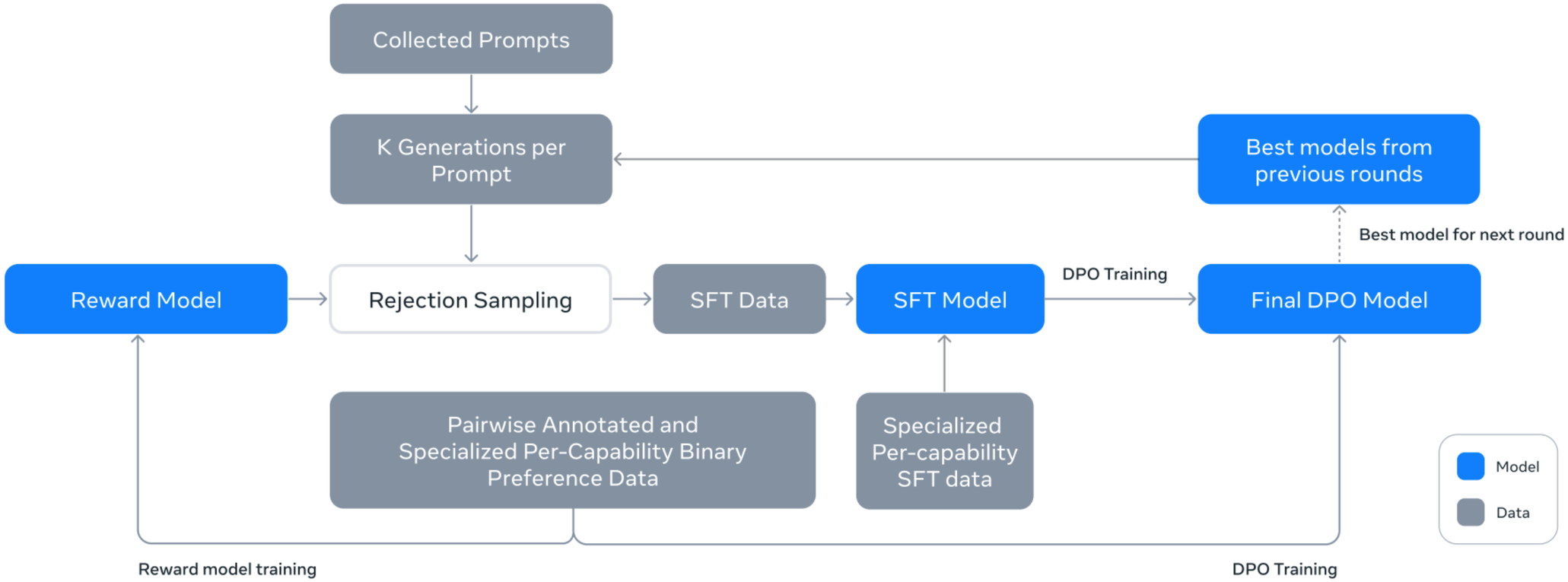
Supervised Finetuning



Reinforcement Learning from Human Feedback

[Ouyang et al., 2022]

Key techniques in Post-training (Llama 3)



[Dubey et al., 2024]

Summary of key techniques

- Instruction tuning
- Supervised finetuning (SFT)
- Reinforcement learning from human feedback (RLHF) or Preference Tuning (PT)
    - Reward model (RM) training
    - Proximal Policy Optimization (PPO)
    - Direct Preference Optimization (DPO)
    - Rejection sampling (RS)
- …

Closeness vs Openness

Proprietary models	Open-weight models	Open-source models
ChatGPT	Llama	Pythia
Claude	Mistral	Llama360
Gemini	Qwen	OLMo (👷)
Grok	Deepseek	…
Command R	Gemma	
Yi-Ligntening	…	
Kimi		
…		

Open and good post-trained models are still rare.

- No models in the top 70 of LMSYS Chatbot Arena with open fine-tuning data.
- We can change this!

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
79	66	Gemini-1.0-Pro-001	1131	+4/-5	18785	Google	Proprietary
79	77	Zephyr-ORPO-141b-A35b-v0.1	1127	+8/-9	4857	HuggingFace	Apache 2.0
79	82	Qwen1.5-32B-Chat	1125	+5/-3	22760	Alibaba	Qianwen LICENSE
79	62	Mistral-Next	1124	+6/-7	12381	Mistral	Proprietary
80	88	Phi-3-Medium-4k-Instruct	1123	+3/-3	26149	Microsoft	MIT
81	97	Starling-LM-7B-beta	1119	+4/-4	16670	Nexusflow	Apache-2.0
82	75	Claude-2.1	1118	+3/-4	37694	Anthropic	Proprietary
82	75	GPT-3.5-Turbo-0613	1117	+4/-3	38957	OpenAI	Proprietary
84	77	Gemini Pro	1111	+7/-8	6561	Google	Proprietary
85	94	Yi-34B-Chat	1111	+5/-5	15928	01 AI	Yi License
85	82	Claude-Instant-1	1111	+4/-4	20623	Anthropic	Proprietary
85	67	GPT-3.5-Turbo-0314	1106	+8/-8	5647	OpenAI	Proprietary
87	89	Mixtral-8x7B-Instruct-v0.1	1114	+0/-0	76141	Mistral	Apache 2.0
89	91	Qwen1.5-14B-Chat	1109	+5/-4	18669	Alibaba	Qianwen LICENSE
89	90	WizardLM-70B-v1.0	1106	+7/-6	8382	Microsoft	Llama 2 Community
89	75	GPT-3.5-Turbo-0125	1106	+3/-3	68889	OpenAI	Proprietary
89	96	Meta-Llama-3.2-3B-Instruct	1103	+5/-6	8467	Meta	Llama 3.2

As of Nov. 12, 2024

This talk

- Overview of Post-Training
- **Tülu 1, 2, 3: Open Post-training Recipes and Lessons**
- Trending Problems

Language models

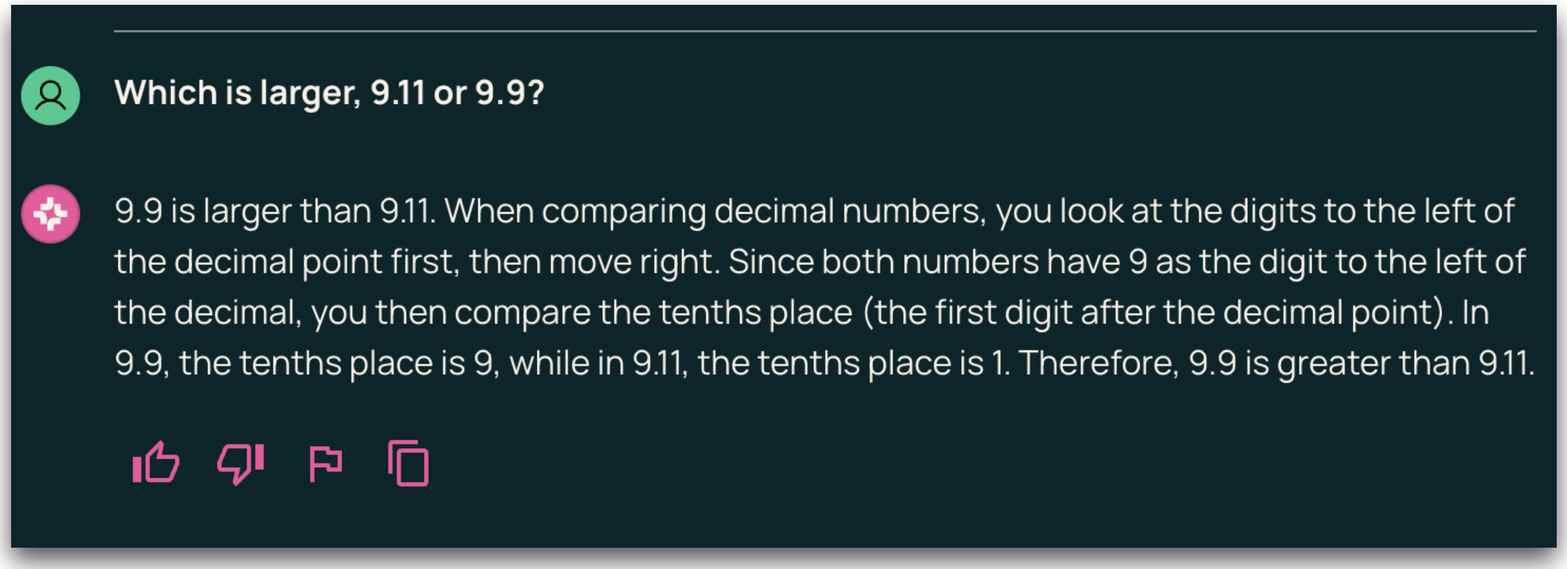
Tülu 3

Try Tülu 3 in the Ai2 Playground

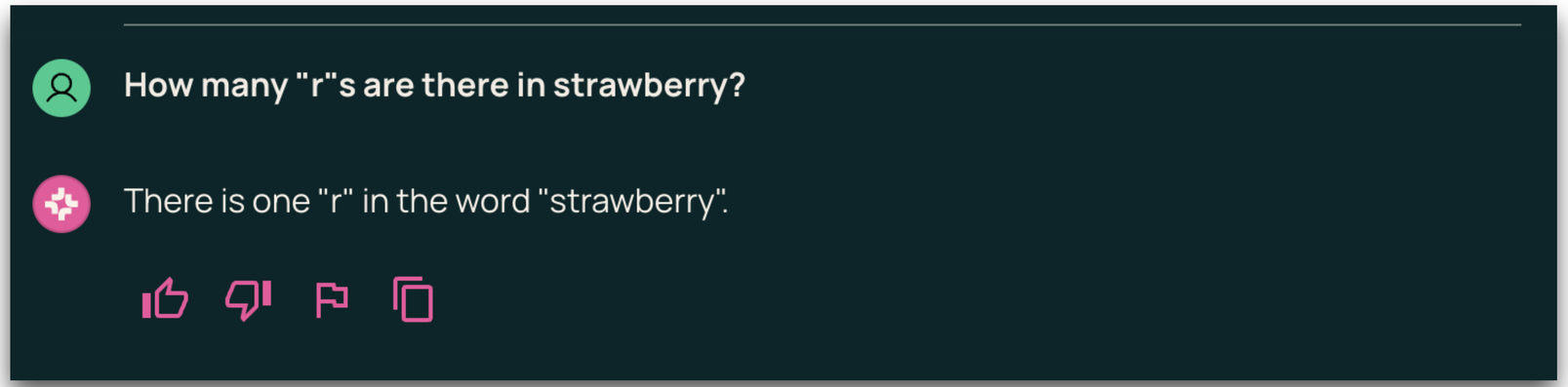
Tülu 3 is a leading instruction following model family, offering fully open-source data, code, and recipes designed to serve as a comprehensive guide for modern post-training techniques.

https://playground.allenai.org/

50% AGI


Which is larger, 9.11 or 9.9?
9.9 is larger than 9.11. When comparing decimal numbers, you look at the digits to the left of the decimal point first, then move right. Since both numbers have 9 as the digit to the left of the decimal, you then compare the tenths place (the first digit after the decimal point). In 9.9, the tenths place is 9, while in 9.11, the tenths place is 1. Therefore, 9.9 is greater than 9.11.

✅


How many "r"s are there in strawberry?
There is one "r" in the word "strawberry".

❌

What is Tülu?

A **tülu camel** is a breed of camel that results from mating a male Bactrian camel with a female dromedary. This breed is sometimes called an **F1 hybrid camel**. The resulting camel is larger than either a Bactrian or a dromedary

[Wiki page for hybrid camel]

[Created with Midjourney, prompted by Yizhong]

Tülu: A child of many camelids!

LLaMa 1/2/3 [Touvron et al.]

Alpaca [Taori et al. 2023]

Dromedary [Sun et al. 2023]

Vicuna [Chiang et al. 2023]

Guanaco [Dettmers et al. 2023]

Bactrian-X [Li et al. 2023]

Tülu aims to combine existing open resources to build an **open**, **reproducible**, **general-purpose** model.

Tülu 1, Jun. 2023: Best open instruction tuning data

- Mixing and analyzing open datasets for stronger instruction-tuned models.
  - Human + synthetic data is best!
- Understanding the limitations of open instruction-tuned models.

How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources

Yizhong Wang*♣♠ Hamish Ivison*♣ Pradeep Dasigi♣ Jack Hessel♣
Tushar Khot♣ Khyathi Raghavi Chandu♣ David Wadden♣ Kelsey MacMillan♣
Noah A. Smith♣♠ Iz Beltagy♣ Hannaneh Hajishirzi♣♠

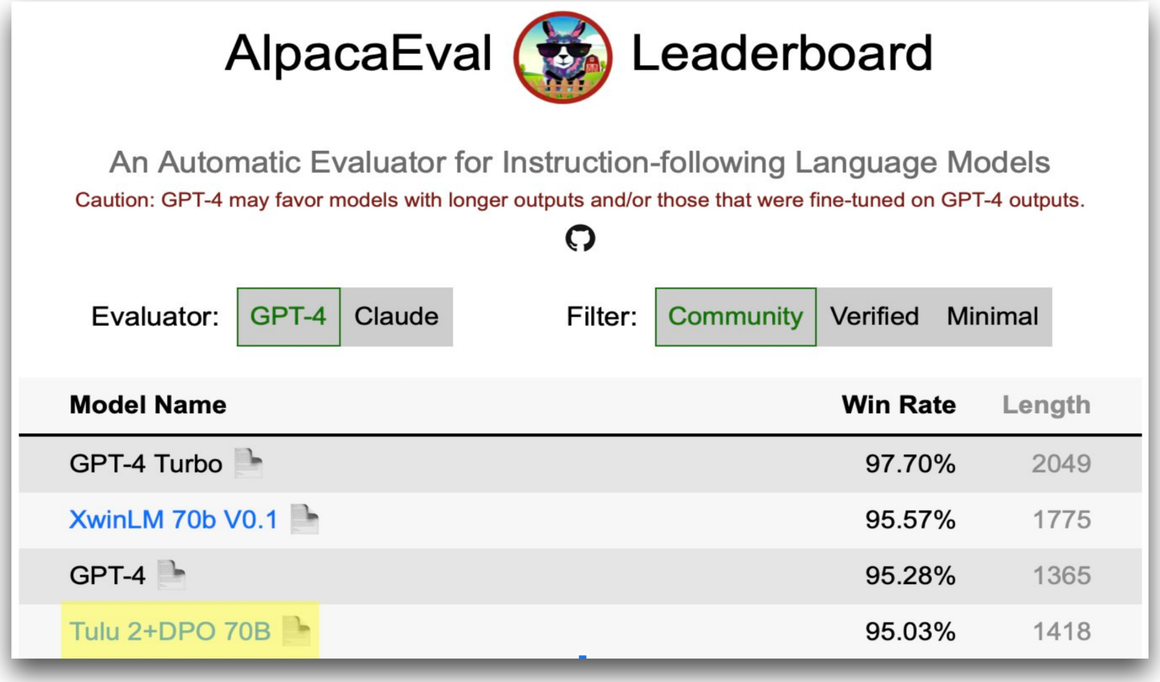
♣Allen Institute for AI ♠University of Washington
{yizhongw,hamishi}@allenai.org

Best recipe for instruction data

Tülu 2, Nov. 2023: Best open model with DPO @ 70B

Adding preference tuning to our open fine-tuning stack (DPO).

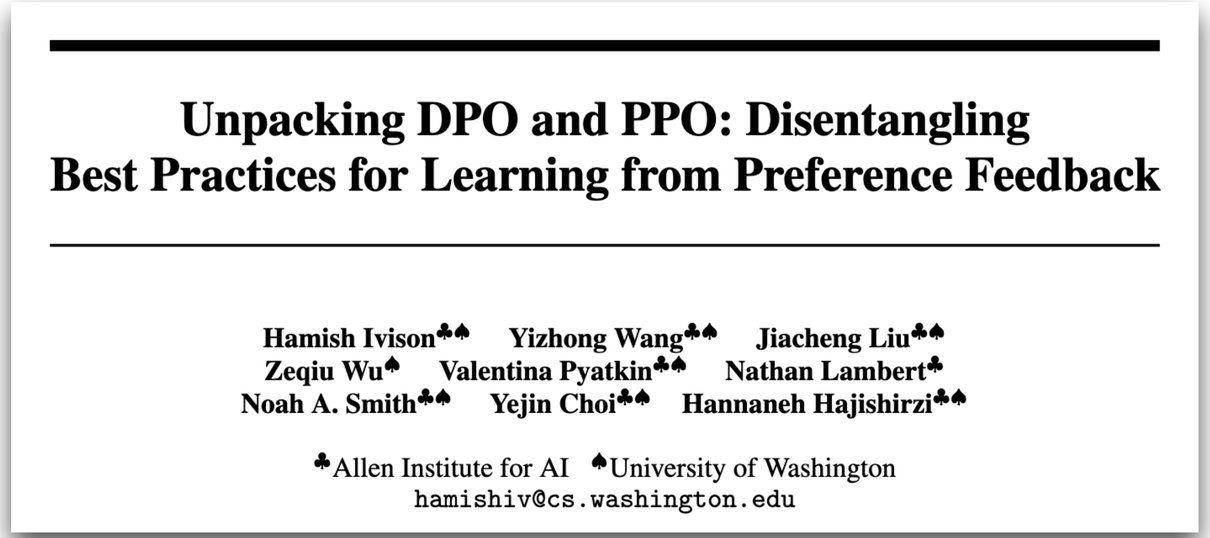
- First to scale DPO to 70B model
- Modern datasets
- State-of-the-art open model on external benchmarks
- Tülu2 outperforms LLaMa2-Chat



Scaling up DPO

Tülu 2.5, Jun. 2024: Preference tuning variants

- DPO vs PPO.
- Preference datasets.
- PPO consistently outperforms DPO, but at the cost of:
  - Implementation complexity
  - Memory usage, and
  - Throughput



Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback

Hamish Ivison♣♠ Yizhong Wang♣♠ Jiacheng Liu♣♠
Zeqiu Wu♠ Valentina Pyatkin♣♠ Nathan Lambert♣
Noah A. Smith♣♠ Yejin Choi♣♠ Hannaneh Hajishirzi♣♠

♣Allen Institute for AI ♠University of Washington
hamishiv@cs.washington.edu

Systematic study of DPO vs PPO

Tülu 3, Nov. 2024: Integration and scaling

- Officially started in June 2024.
- Massive team efforts, 23 co-authors, extensive support from other teams@Ai2.

TÜLU 3: Pushing Frontiers in Open Language Model Post-Training

---

Nathan Lambert[♥a] Jacob Morrison[♥a] Valentina Pyatkin[♥aw] Shengyi Huang[♥a] Hamish Ivison[♥aw] Faeze Brahman[♥a]
Lester James V. Miranda[♥a] Alisa Liu[w] Nouha Dziri[a] Xinxi Lyu[a]
Yuling Gu[a] Saumya Malik[a] Victoria Graf[w] Jena D. Hwang[a]
Jiangjiang Yang[a] Ronan Le Bras[a] Oyvind Tafjord[a] Chris Wilhelm[a]
Luca Soldaini[a] Noah A. Smith[aw] Yizhong Wang[aw] Pradeep Dasigi[a] Hannaneh Hajishirzi[aw]

[a]Allen Institute for AI [w]University of Washington

tulu@allenai.org

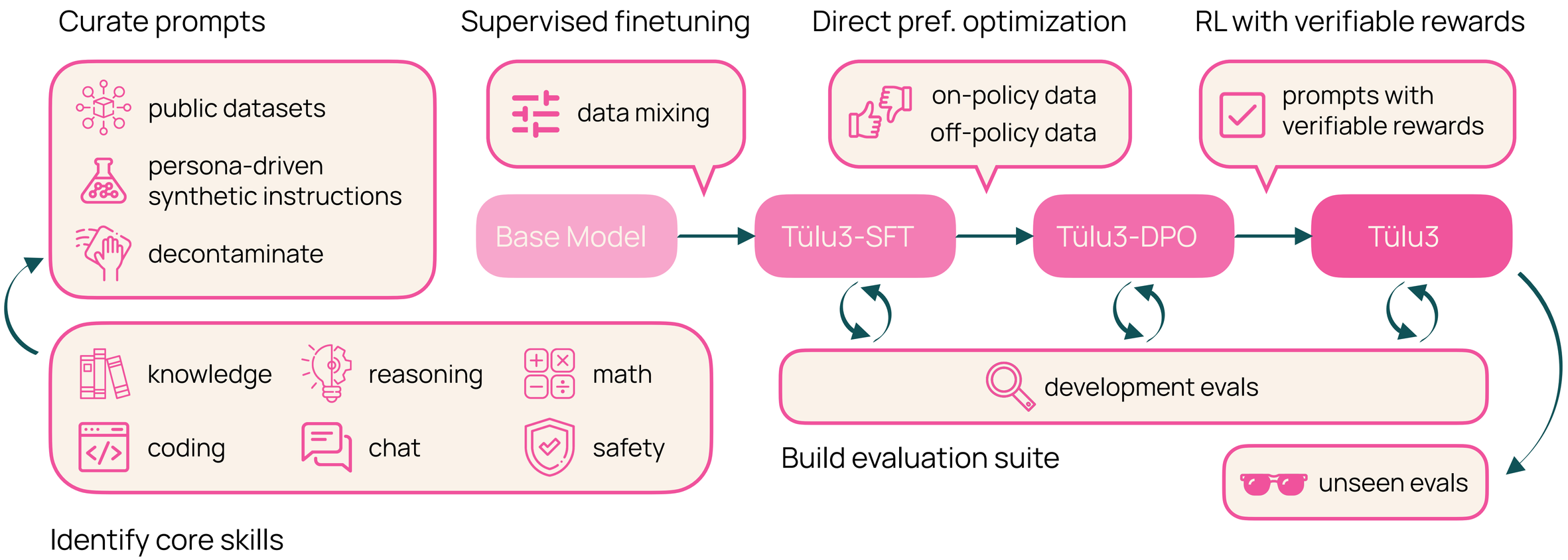
Tülu 3: Surpassing cutting-edge models

Open-weight models

Proprietary models

Skill	Benchmark$_{(eval)}$	TÜLU 3 8B	Qwen 2.5 7B Instruct	Llama 3.1 8B Instruct	TÜLU 3 70B	Qwen 2.5 72B Instruct	Llama 3.1 70B Instruct	GPT-3.5 Turbo	GPT-4o Mini	Claude 3.5 Haiku
	Avg.	**64.8**	57.8	62.2	**76.0**	71.5	73.4	64.7	**69.6**	75.3
Knowledge	MMLU$_{(0\ shot,\ CoT)}$	68.2	**76.6**	71.2	83.1	**85.5**	85.3	70.2	82.2	**81.8**
	PopQA$_{(15\ shot)}$	**29.1**	18.1	20.2	**46.5**	30.6	46.4	**45.0**	39.0	42.5
	TruthfulQA$_{(6\ shot)}$	55.0	**63.1**	55.1	67.6	**69.9**	66.8	62.9$^{\diamond}$	64.8$^{\diamond}$	**64.9**$^{\diamond}$
Reasoning	BigBenchHard$_{(3\ shot,\ CoT)}$	**66.0**	21.7	62.8	**82.0**	67.2	73.8	66.6$^{\top}$	65.9$^{\diamond}$	**73.7**$^{\top}$
	DROP$_{(3\ shot)}$	**62.6**	54.4	61.5	74.3	34.2	**77.0**	70.2	36.3	**78.4**
Math	MATH$_{(4\ shot\ CoT,\ Flex)}$	**43.7**	14.8	42.5	63.0	**74.3**	56.4	41.2	67.9	**68.0**
	GSM8K$_{(8\ shot,\ CoT)}$	**87.6**	83.8	83.4	93.5	89.5	**93.7**	74.3	83.0	**90.1**
Coding	HumanEval$_{(pass@10)}$	83.9	**93.1**	86.3	92.4	**94.0**	93.6	87.1	90.4	**90.8**
	HumanEval+$_{(pass@10)}$	79.2	**89.7**	82.9	88.0	**90.8**	89.5	84.0	87.0	**88.1**
IF & chat	IFEval$_{(prompt\ loose)}$	**82.4**	74.7	80.6	83.2	87.6	**88.0**	66.9	83.5	**86.3**
	AlpacaEval 2$_{(LC\ \%\ win)}$	**34.5**	29.0	24.2	**49.8**	47.7	33.4	38.7	**49.7**	47.3
Safety	Safety$_{(6\ task\ avg.)}$	**85.5**	75.0	75.2	**88.3**	87.0	76.5	69.1	84.9	**91.8**

Tülu 3 recipe



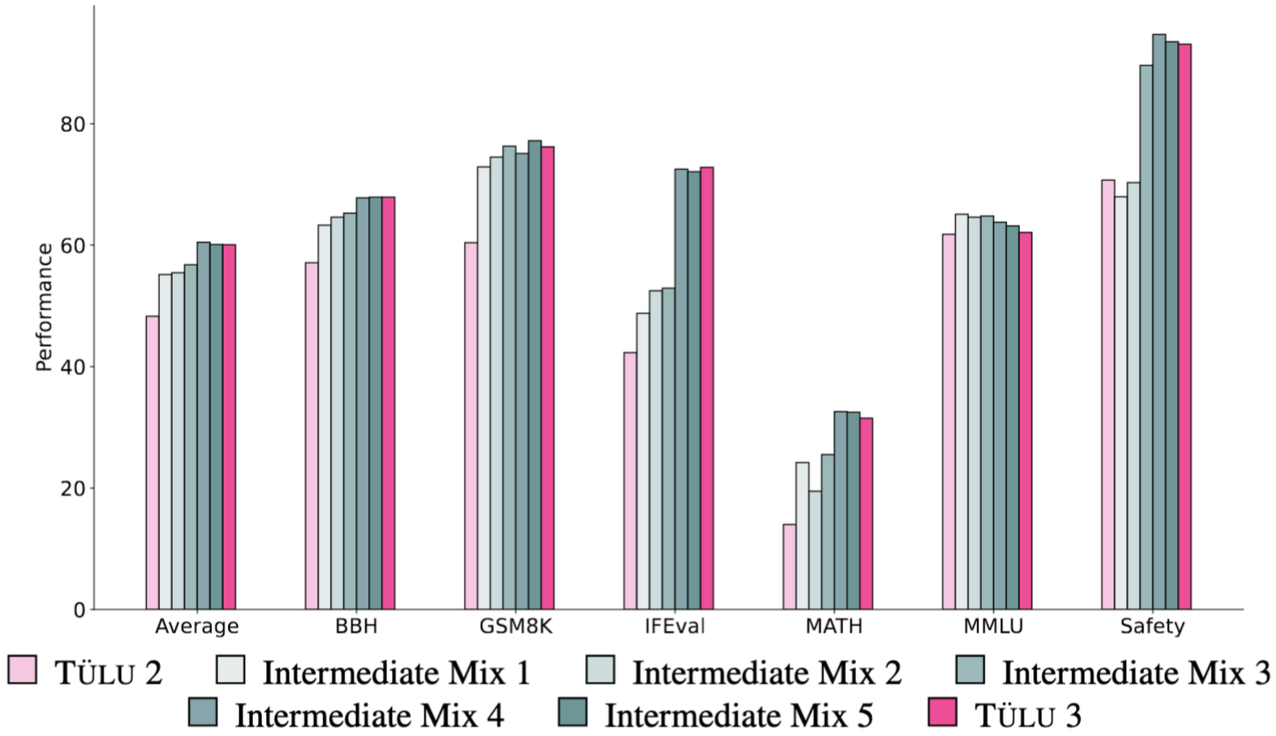
Step 0: Curating Prompts

- It starts from a big compilation of existing resources [full table]
- Manual quality check.
- Provenance and licenses.
- Automatic filtering based on keywords (e.g. OpenAI).
- Decontamination.

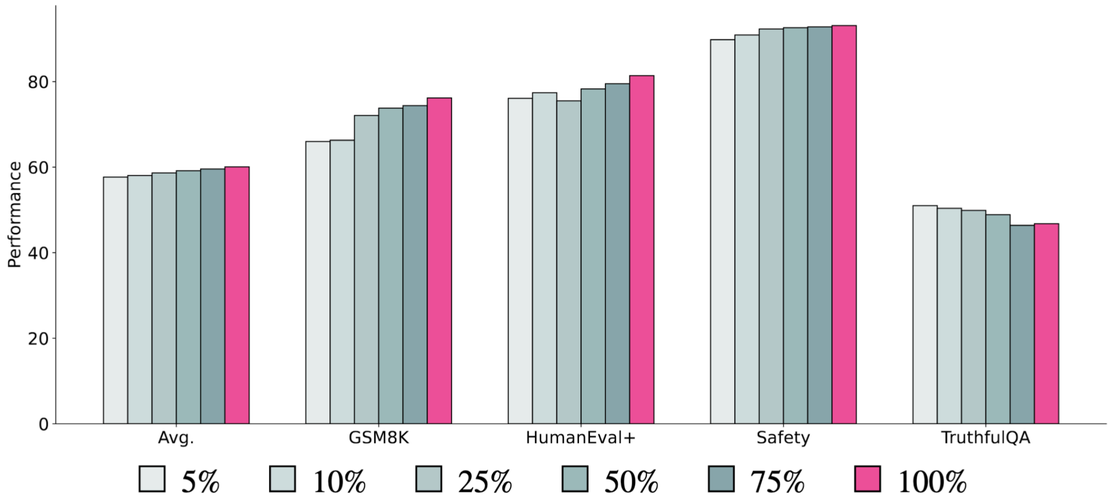
Category	Prompt Dataset	Count	# Prompts used in SFT	# Prompts used in DPO
General	**TÜLU 3 Hardcoded**[↑]	24	240	–
	OpenAssistant[1,2,↓]	88,838	7,132	7,132
	No Robots	9,500	9,500	9,500
	WildChat (GPT-4 subset)[↓]	241,307	100,000	100,000
	UltraFeedback[α,2]	41,635	–	41,635
Knowledge	FLAN v2[1,2,↓]	89,982	89,982	12,141
Recall	SciRIFF[↓]	35,357	10,000	17,590
	TableGPT[↓]	13,222	5,000	6,049
Math	**TÜLU 3 Persona MATH**	149,960	149,960	–
Reasoning	**TÜLU 3 Persona GSM**	49,980	49,980	–
	TÜLU 3 Persona Algebra	20,000	20,000	–
	OpenMathInstruct 2[↓]	21,972,791	50,000	26,356
	NuminaMath-TIR[α]	64,312	64,312	8,677
Coding	**TÜLU 3 Persona Python**	34,999	34,999	–
	Evol CodeAlpaca[α]	107,276	107,276	14,200
Safety	**TÜLU 3 CoCoNot**	10,983	10,983	10,983
& Non-Compliance	**TÜLU 3 WildJailbreak**[α,↓]	50,000	50,000	26,356
	TÜLU 3 WildGuardMix[α,↓]	50,000	50,000	26,356
Multilingual	Aya[↓]	202,285	100,000	32,210
Precise IF	**TÜLU 3 Persona IF**	29,980	29,980	19,890
	TÜLU 3 IF-augmented	65,530	–	65,530
Total		23,327,961	939,344	425,145

Step 1: Supervised finetuning

Mixing experiments
(only milestone versions)

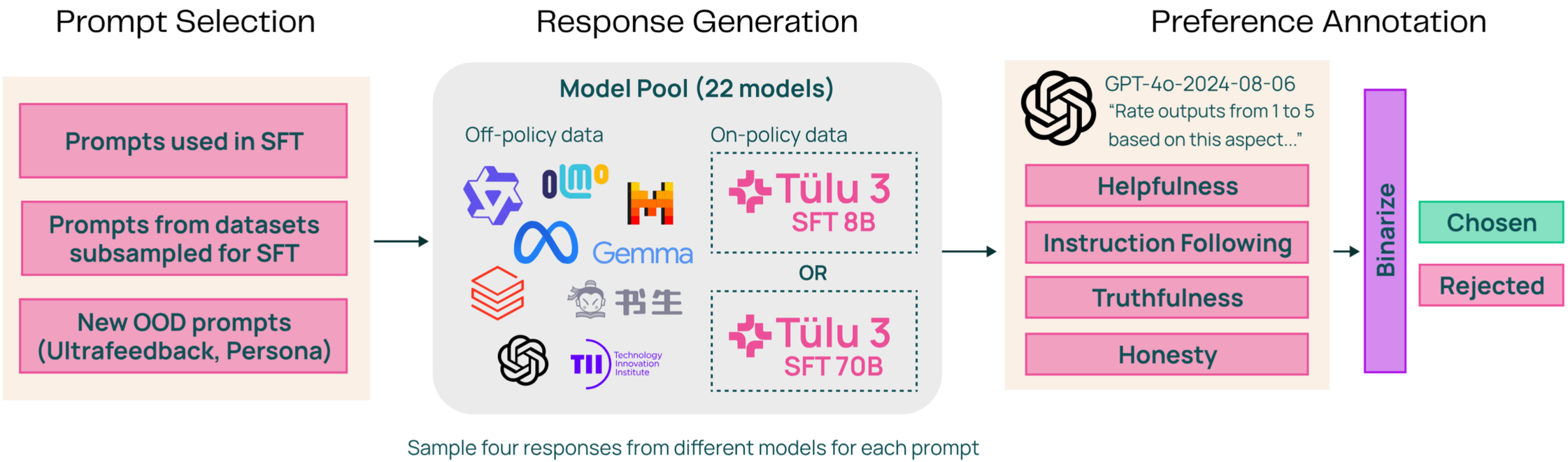


Data scaling experiments

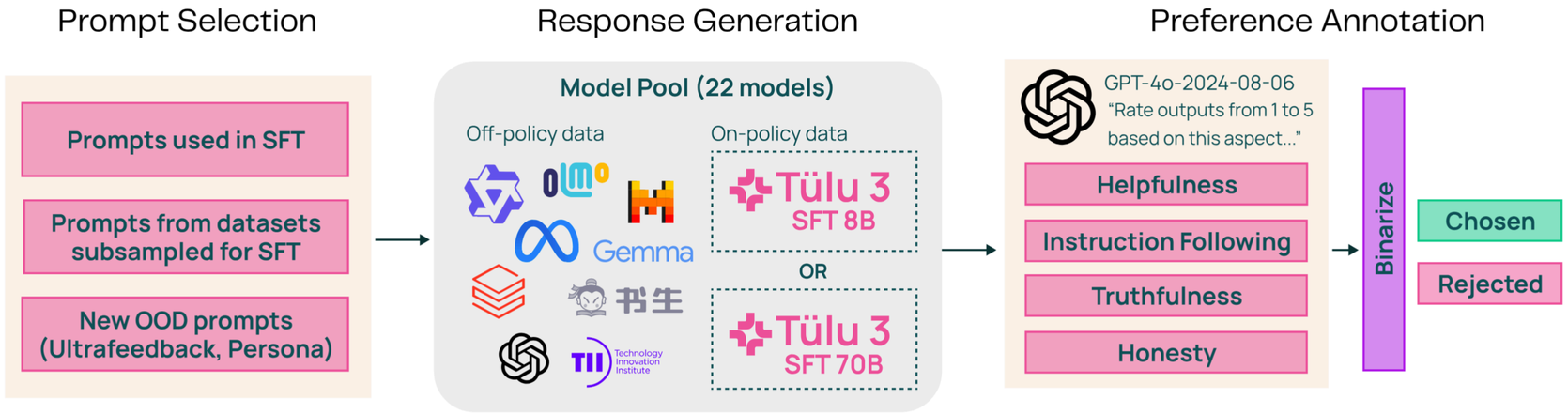


Step 2: Preference tuning

- We experimented with SimPO [Meng et al., 2024], but ended up with the length-normalized DPO.
- We refined and scaled up the Ultrafeedback [Cui et al., 2023] for preference data generation.

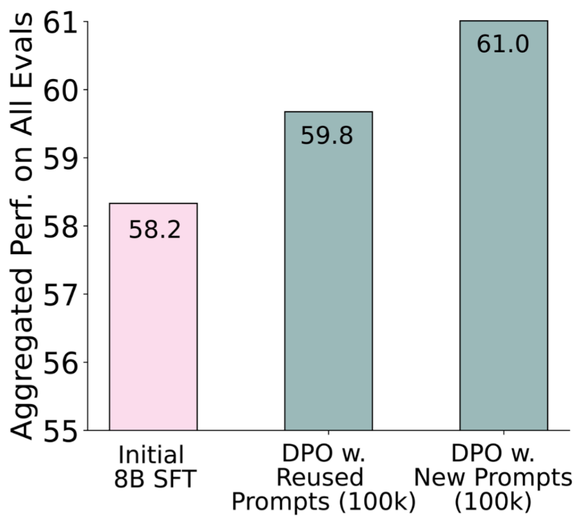


Step 2: Preference tuning

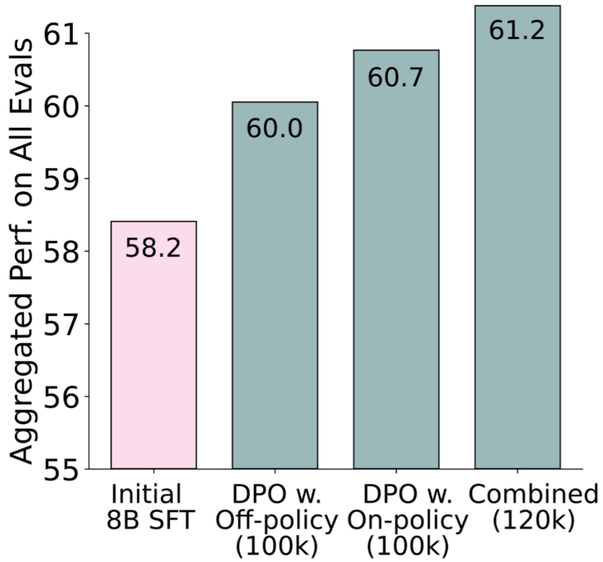


Sample four responses from different models for each prompt

Using SFT vs new prompts



Off- vs On-policy preferences

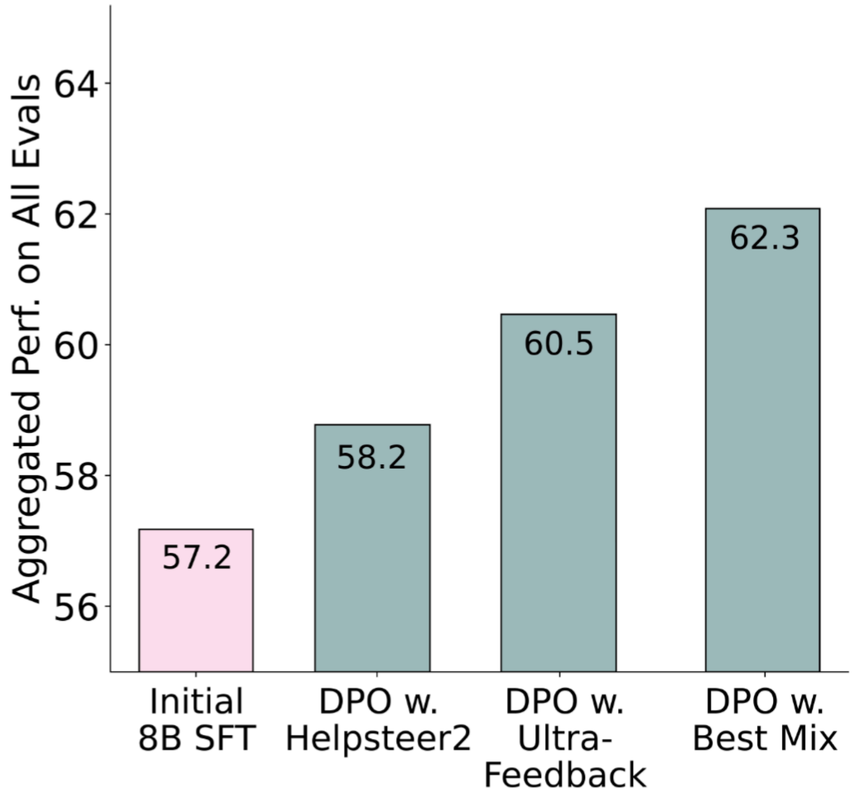


Different LM Judges

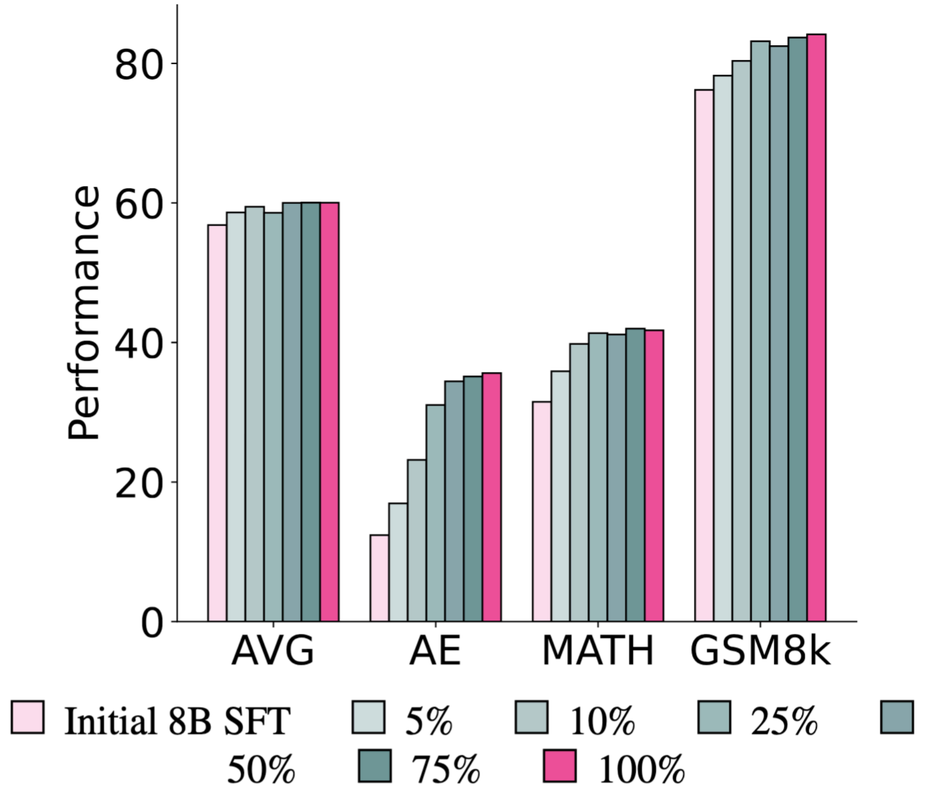
LLM Judge	Avg.
GPT-4o	57.3
LLama 3.1 405B	57.2
GPT-4 Turbo	57.0
GPT-4o Mini	56.9
Llama 3.1 70B	56.6

Step 2: Preference tuning

Our mix v.s. existing preference datasets

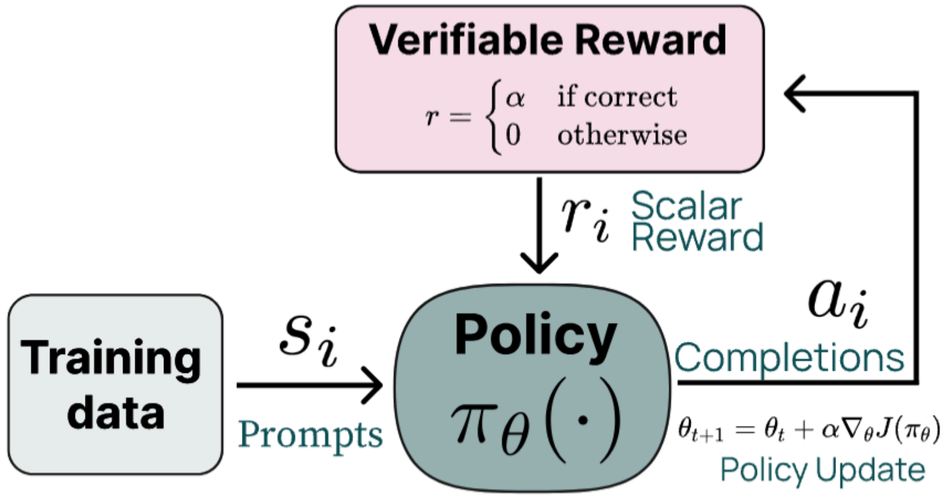


Data scaling experiements



Step 3: Reinforcement learning w. verifiable rewards

- ✅ Gold final answers or verifiable constraints.
- ❌ intermediate chain of thoughts or not matching model.
- Classical RL! (We used PPO for optimization)
- We tried it using three datasets.



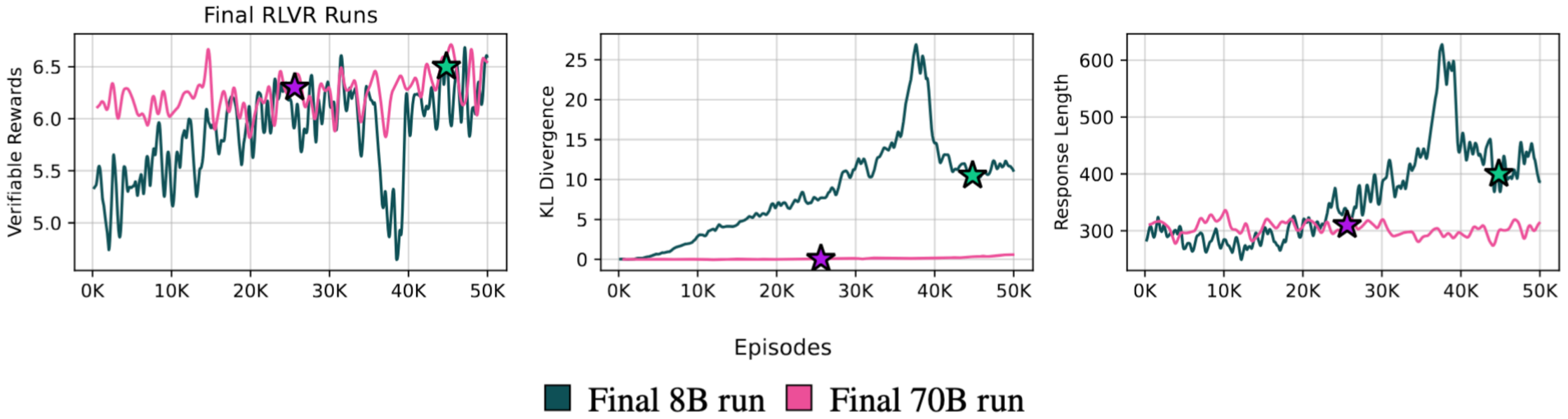
Prompt Dataset	Count	Verification
GSM8K Train	7,473	Exact match against extracted answer
MATH Train	7,500	Exact match against extracted answer
IF verifiable	14,973	Prompt-specific verifiers
Total	29,946	

Step 3: Reinforcement learning w. verifiable rewards

Model Size		8B			70B		
Category	**Benchmark**$_{\text{(Eval Setting)}}$	**Llama 3.1 Inst.**	**TÜLU 3 DPO**	**TÜLU 3 RLVR**	**Llama 3.1 Inst.**	**TÜLU 3 DPO**	**TÜLU 3 RLVR**
Avg.		62.2	64.4	**64.8**	73.4	75.9	**76.0**
Knowledge	MMLU$_{\text{(0 shot, CoT)}}$	**71.2**	68.7	68.2	**85.3**	83.3	83.1
	PopQA$_{\text{(15 shot)}}$	20.2	**29.3**	29.1	46.4	46.3	**46.5**
	TruthfulQA$_{\text{(6 shot)}}$	55.1	**56.1**	55.0	66.8	**67.9**	67.6
Reasoning	BigBenchHard$_{\text{(3 shot, CoT)}}$	62.8	65.8	**66.0**	73.8	81.8	**82.0**
	DROP$_{\text{(3 shot)}}$	61.5	62.5	**62.6**	**77.0**	74.1	74.3
Math	MATH$_{\text{(4 shot CoT, Flex)}}$	42.5	42.0	**43.7**	56.4	62.3	**63.0**
	GSM8K$_{\text{(8 shot, CoT)}}$	83.4	84.3	**87.6**	**93.7**	93.5	93.5
Code	HumanEval$_{\text{(pass@10)}}$	**86.3**	83.9	83.9	**93.6**	92.4	92.4
	HumanEval+$_{\text{(pass@10)}}$	**82.9**	78.6	79.2	**89.5**	88.4	88.0
IF & Chat	IFEval$_{\text{(Strict)}}$	80.6	81.1	**82.4**	**88.0**	82.6	83.2
	AlpacaEval 2$_{\text{(LC \% win)}}$	24.2	33.5	**34.5**	33.4	49.6	**49.8**
Safety	Safety$_{\text{6 task avg.}}$	75.2	**87.2**	85.5	76.5	**89.0**	88.3

Step 3: Reinforcement learning w. verifiable rewards

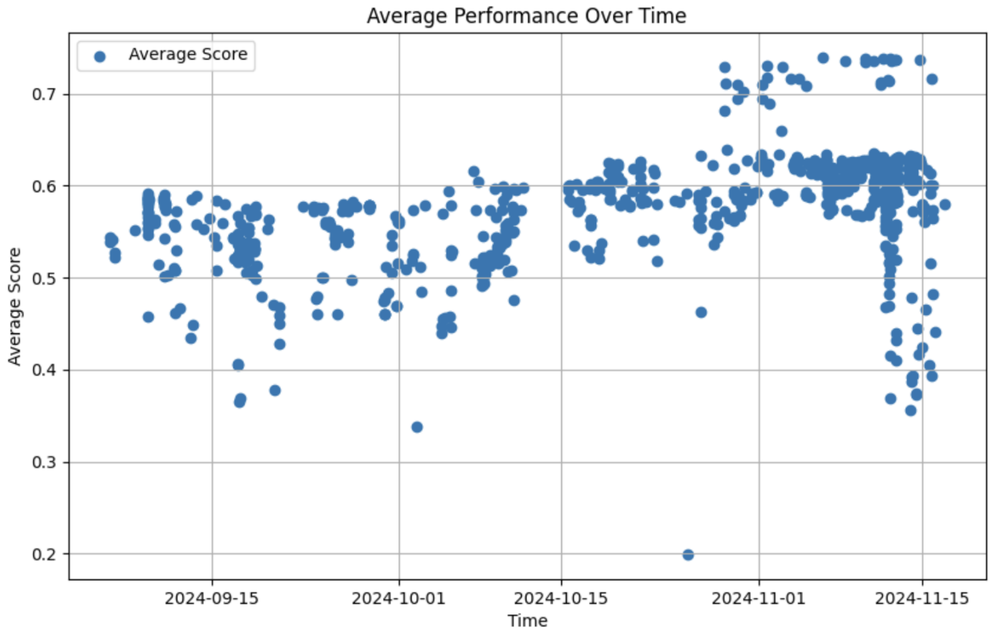
- RL still requires a lot of hyperparameter tuning to work.
- The 70B training used a lower learning rate (1e-7), which avoided model collapse but also discouraged the model from any big change.



Many other details in the paper

- Evaluation.
- SFT/DPO/PPO training infrastructures.
- Decontamination.
- Unseen test results.
- …

A peak of the journey (credit: Hamish)



Lessons learned about post-training

Lessons learned about post-training: **Data**

- Data is the key!
- Data mixing is effective and necessary.
- Right balance among many factors is challenging:
  - Coverage/diversity
  - Quality
  - Quantity
  - Specific capabilities
  - Risks of forgetting
  - …

Lessons learned about post-training: **Evaluation**

- Systematic evaluation is necessary.
- Most academic works these days focus on limited evaluations (e.g., AlpacaEval, MT-Bench, ArenaHard) that check the “vibes” of models.
  - But we should also admit evaluations for LLMs generally break these days (even human eval).
- Good to have dev and test sets at the benchmark level:
  - Companies can easily run 1000 experiments, optimizing for target benchmarks
- Evaluation set contamination is a big issue in commonly used public training data (e.g. ShareGPT, LMSYS-1M, DaringAnteater)

Lessons learned about post-training: **Integration**

- Post-training is becoming increasingly complicated with the **integration of many moving factors**.
- **Data** can serve as an effective ground for collaboration.
- **Model merging** is also surprisingly effective, but we still need to understand more about the risks (e.g., stability of the merged model)

This talk

- Overview of Post-Training
- Tülu 1, 2, 3: Fully Open Post-training Recipes and Lessons
- **Trending Problems**

Trend #1: Data optimization

Data mixing & selection

LESS: Selecting Influential Data for Targeted Instruction Tuning

Mengzhou Xia [1 *] Sadhika Malladi [1 *] Suchin Gururangan [2] Sanjeev Arora [1] Danqi Chen [1]

RETHINKING DATA SELECTION AT SCALE: RANDOM SELECTION IS ALMOST ALL YOU NEED

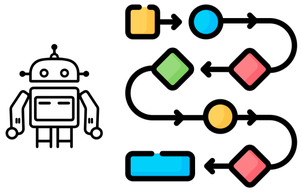
Tingyu Xia[1,3†] Bowen Yu[2*] Kai Dang[2] An Yang[2] Yuan Wu[1,3*] Yuan Tian[1,3]
Yi Chang[1,3,4] Junyang Lin[2]

HYBRID PREFERENCES: LEARNING TO ROUTE INSTANCES FOR HUMAN VS. AI FEEDBACK

Lester James V. Miranda[1] * Yizhong Wang[1,2*] Yanai Elazar[1,2]
Sachin Kumar[1,3] Valentina Pyatkin[1,2] Faeze Brahman[1]
Noah A. Smith[1,2] Hannaneh Hajishirzi[1,2] Pradeep Dasigi[1]

Trend #1: Data optimization

Data mixing & selection



Synthetic data

SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions

Yizhong Wang♣ Yeganeh Kordi◇ Swaroop Mishra♡ Alisa Liu♣
Noah A. Smith♣+ Daniel Khashabi♠ Hannaneh Hajishirzi♣+

***WizardLM*: Empowering Large Language Models to Follow Complex Instructions**

Can Xu[1]* Qingfeng Sun[1]* Kai Zheng[1]* Xiubo Geng[1]
Pu Zhao[1] Jiazhan Feng[2]† Chongyang Tao[1] Qingwei Lin[1] Daxin Jiang[1]‡

Scaling Synthetic Data Creation with 1,000,000,000 Personas

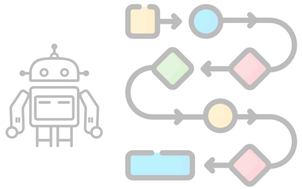
Tao Ge,* Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, Dong Yu

Textbooks Are All You Need

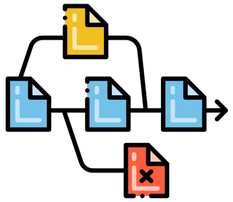
Suriya Gunasekar Yi Zhang Jyoti Aneja Caio César Teodoro Mendes
Allie Del Giorno Sivakanth Gopi Mojan Javaheripi Piero Kauffmann
Gustavo de Rosa Olli Saarikivi Adil Salim Shital Shah Harkirat Singh Behl
Xin Wang Sébastien Bubeck Ronen Eldan Adam Tauman Kalai Yin Tat Lee
Yuanzhi Li

Trends #1: Data optimization

Data mixing & selection



Synthetic data



Provenence and copyright

The Data Provenance Initiative:
A Large Scale Audit of Dataset Licensing & Attribution in AI

Shayne Longpre[1†] Robert Mahari[1,2] Anthony Chen[3] Naana Obeng-Marnu[1,4]
Damien Sileo[5] William Brannon[1,4] Niklas Muennighoff[6] Nathan Khazam[7]
Jad Kabbara[1,4] Kartik Perisetla Xinyi (Alexis) Wu[8] Enrico Shippole Kurt Bollacker[7]
Tongshuang Wu[9] Luis Villa[10] Sandy Pentland[1] Sara Hooker[11]

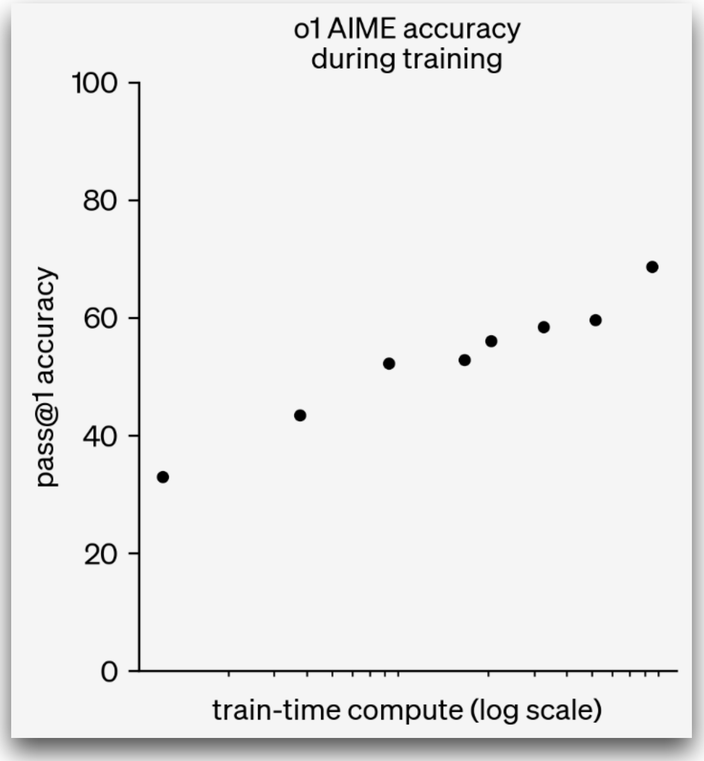
Evaluating Copyright Takedown Methods for Language Models

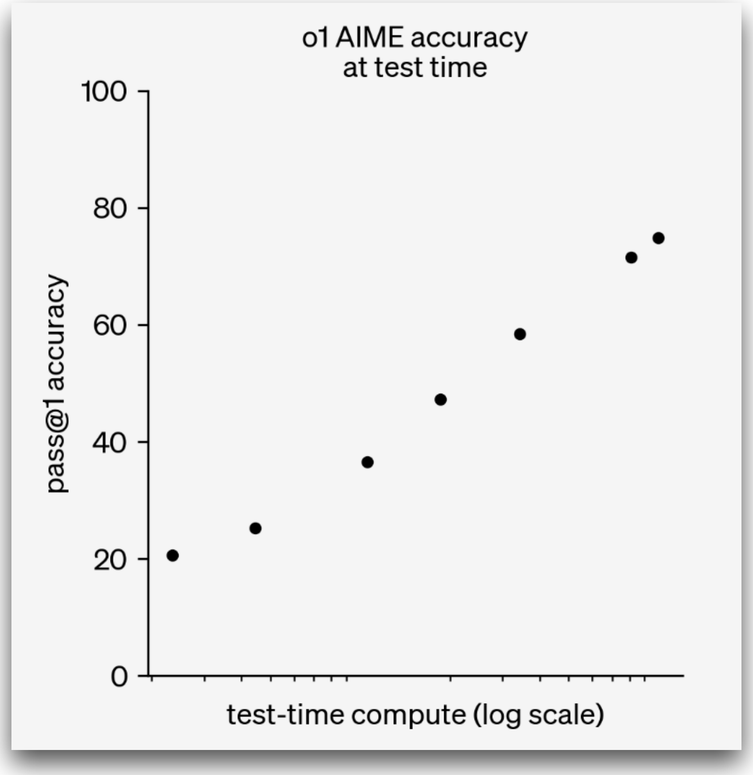
Boyi Wei*[1] Weijia Shi*[2] Yangsibo Huang*[1]
Noah A. Smith[2] Chiyuan Zhang Luke Zettlemoyer[2] Kai Li[1] Peter Henderson[1]

Trend #2: Scaling of post-training

- OpenAI o1 shows the potential of scaling post-training via RL.

- More info in Sasha's latest talk: https://www.youtube.com/watch?v=6PEJ96k1kiw

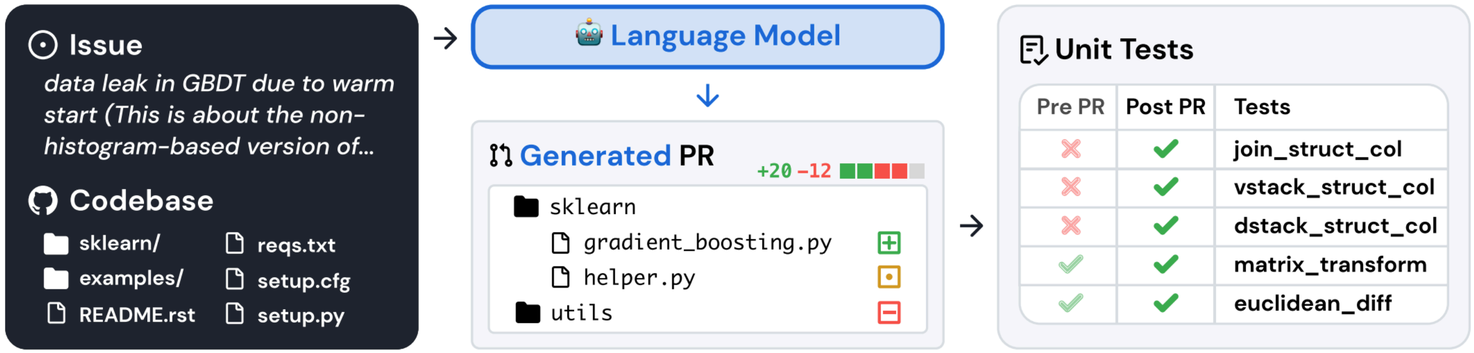




Trend #3: Complex tasks

Real-world scenarios

(e.g., SWEBench for coding)



Extreme difficulty

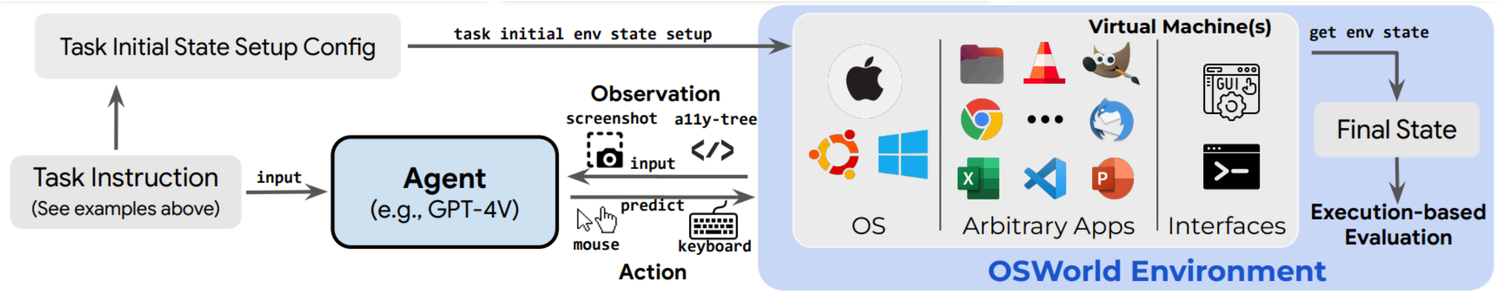
(e.g., FrontierMath for math)

"These are extremely challenging... I think they will resist AIs for several years at least."

Terence Tao
Fields Medalist (2006)

Interaction with environments

(e.g., OSWorld for agent)



Summary of this talk

- **Overview of Post-Training**
  - Objectives of post-training
  - Pre-training vs post-training
  - Key techniques
- **Tülu 1, 2, 3: Fully Open Post-training Recipes and Lessons**
  - Tülu 1: instruction tuning
  - Tülu 2: scaling DPO
  - Tülu 2.5: RLHF algorithms and datasets
  - Tülu 3: integration and scaling
- **Trending Problems**
  - Data optimization
  - Scaling of post-training
  - Complex tasks

Thanks for listening!

 @yizhongwyz

 yizhongw@cs.washington.edu

 https://homes.cs.washington.edu/~yizhongw/

Open-Instruct repo for Tülu