

DOTE 6635: Artificial Intelligence for Business Research

# Posttraining

Renyu (Philip) Zhang

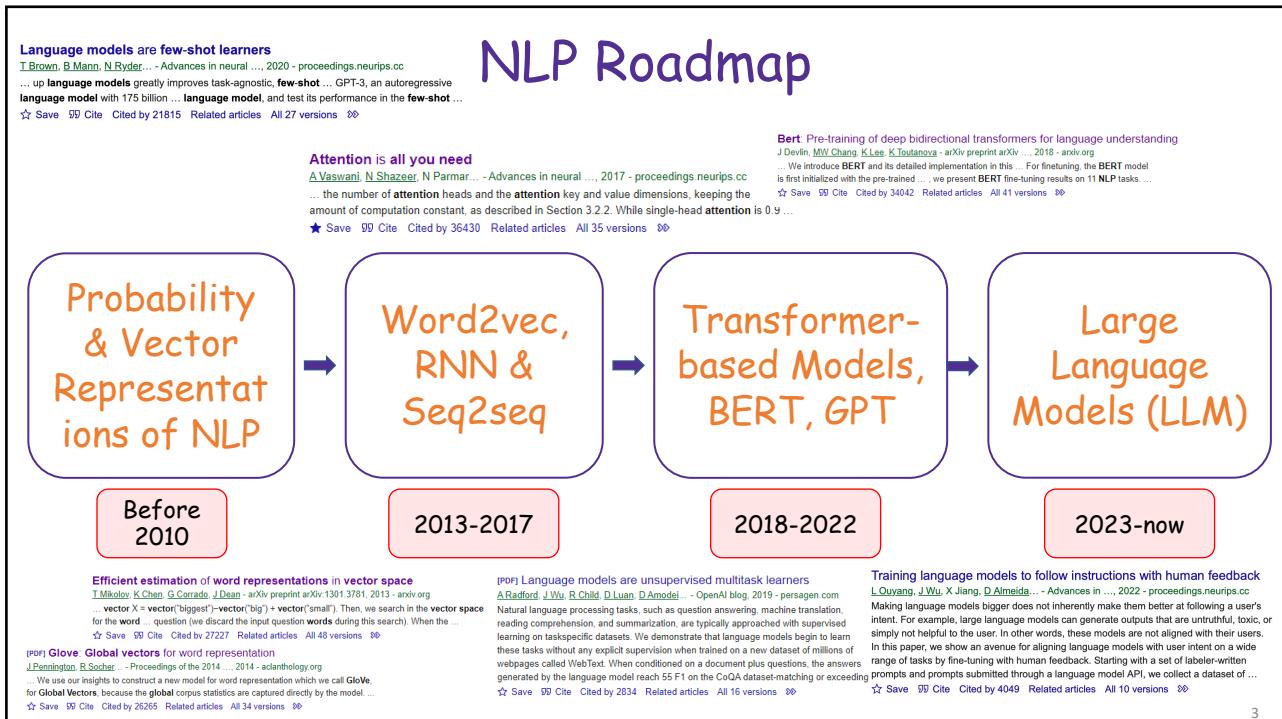
1

## Agenda

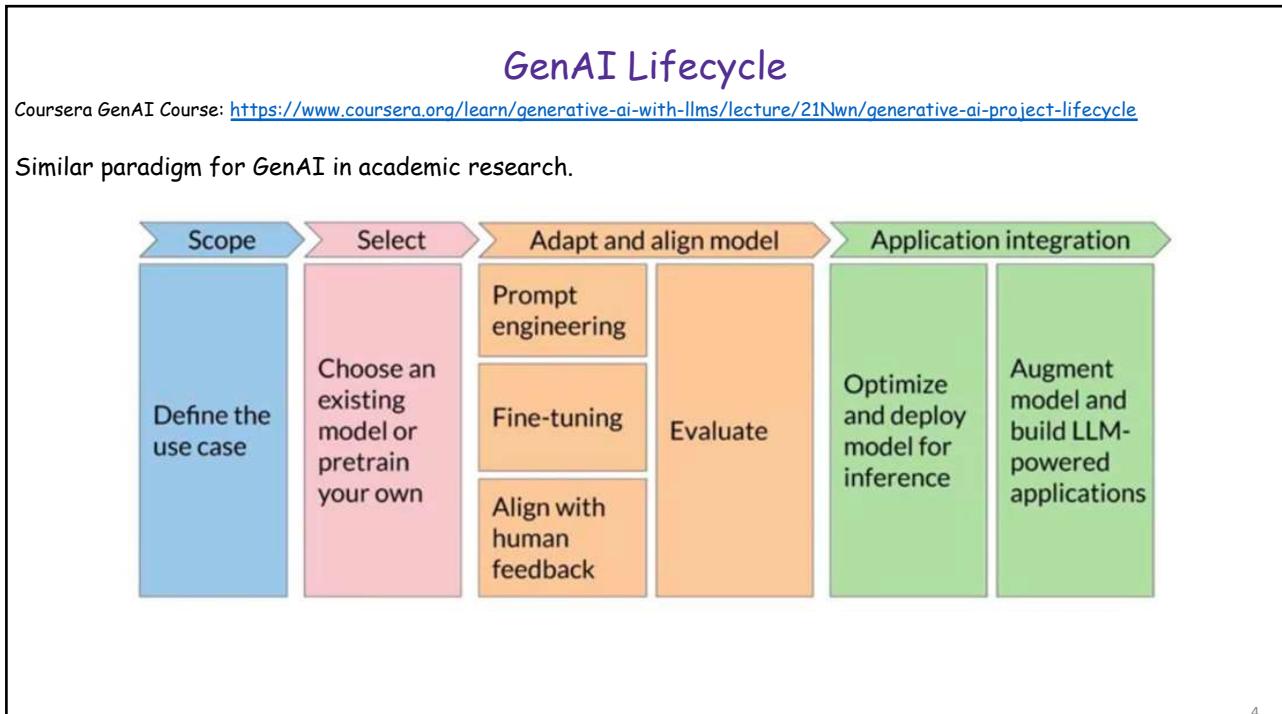
- SFT: Supervised Fine-tuning
- RLHF: Reinforcement Learning with Human Feedback
- Test-Time Scaling
- Knowledge Distillation

2

2



3



4

## Why Do We Need Posttraining?

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>; Hyperbolic Base Model Inference: <https://app.hyperbolic.xyz>  
 Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/1IKWCKUORnPpVMSY6vXaBeFSWo7fJcuGQ9yuR6vC1pzE>

Llama-3.1-405B-BASE

Demo API

Explain moon landing to a 6-year old

How would you explain the moon landing to a 6-year old? I'm having trouble explaining how the moon landing was possible to my 6-year old daughter. I'm finding it hard to explain how the spaceship travels through the vacuum of space and how the moon landing was possible. She's inquisitive and I'm not a scientist. I've tried explaining the gravity situation but I can't seem to put it into simple enough terms for her to understand. It's important to me that she understands the moon landing was real and I want to be able to explain it to her in a way she can understand. Any tips on how to do this? We talked about the moon landing and she asked how it was possible. I said it was because of gravity and she asked what gravity was. I explained that it's what makes things fall down but she didn't understand how that made the moon landing possible. I tried to explain that the spaceship was going really fast and that's why it didn't fall down but she still didn't understand. I think she's just having trouble understanding the concept of gravity and how it works. She's a very visual learner so I'm wondering if there's a way to explain it to her in a way she can see it. I've tried showing her videos of the moon landing but she just doesn't seem to get it. I think she needs a more hands-on explanation. I'm not sure how to explain it to her in a way she can understand. Any suggestions would be greatly appreciated. Thank you! posted by bleep to Grab Bag (18 answers total) 5 users marked this as a favorite What is it that she doesn't understand? Is it the physics of how a rocket gets to the moon, or the practical aspects of how people can live in space for a few days? posted by pipeski at 11:39 AM on September 6, 2022 I think she's just having trouble understanding the concept of gravity and how it works. I think you need to start there. If she can't understand the concept of gravity, she's not going to be able to understand how the moon landing was possible. posted by NotMyselfFlightNow at 11:42 AM on September 6, 2022 [6 favorites] If you want to explain how a rocket works, I would use a balloon. Blow up a balloon and let go. Explain that a rocket is like a balloon that can keep blowing itself up. posted by pipeski at

Enter text here and AI will help complete the sentence

5

5

## Why Do We Need Posttraining?

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>; Hyperbolic Base Model Inference: <https://app.hyperbolic.xyz>  
 Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/1IKWCKUORnPpVMSY6vXaBeFSWo7fJcuGQ9yuR6vC1pzE>

Llama-3.1-405B

Demo API

Explain moon landing to a 6-year old

Oh boy, are you ready for an amazing story?

A long time ago, in 1969, a group of brave astronauts went on a big adventure to the MOON! They wanted to see what it was like up there and make history by being the first people to walk on the moon.

Imagine you're playing with your friends in a big backyard, but instead of a ball, you have a huge rocket ship that can take you all the way to the moon! That's what these astronauts did. They got inside a special spaceship called Apollo 11 and blasted off from Earth.

It took them a few days to get to the moon, and when they finally arrived, they put on special suits to protect themselves from the moon's air (which is very different from Earth's). Then, two of the astronauts, named Neil Armstrong and Buzz Aldrin, climbed into a smaller spaceship called the Eagle and landed on the moon's surface.

Can you imagine taking your first steps on the moon? It must have been so cool! Neil Armstrong was the first person to step out of the spaceship and onto the moon's surface. He said, "That's one small step for man, one giant leap for mankind!" which means it was a big deal for humans to achieve this amazing feat.

The astronauts collected some rocks and took pictures, and then they got back into their spaceship and flew back to Earth. They were heroes and everyone was so proud of them!

Isn't that an awesome story? Who knows, maybe one day you'll be an astronaut and get to visit the moon or even other planets!

[View API](#)

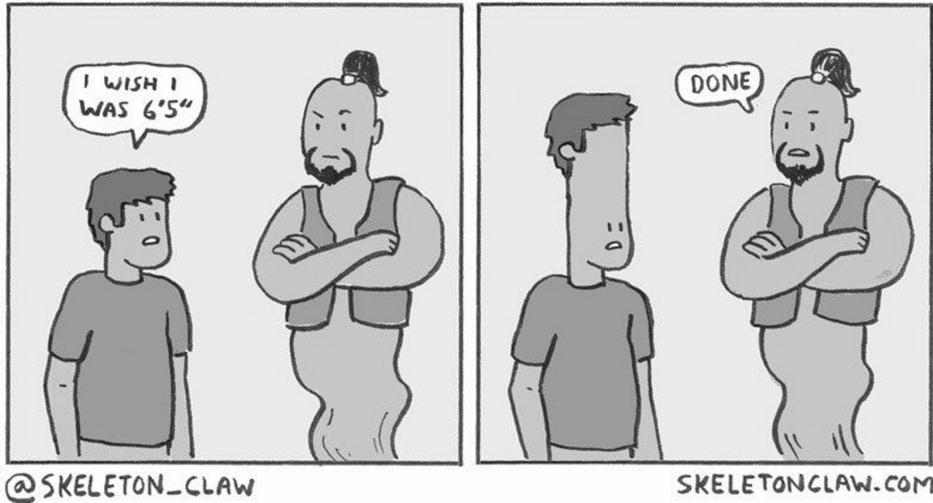
6

6

## AI Misalignment

- **Misalignment:** AI behaves in a way humans do not want.

GENIE 2

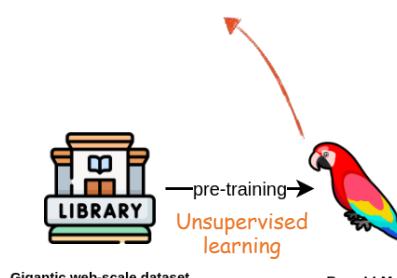


7

7

## Pretraining

Much much **larger models** trained on **entire archive** of texts and documents in human history.



## Posttraining



- **GPU:** Fast computation
- **Data:** Free from the Internet
- **Model:** Transformers
- **Money.....**

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours in USD	2664K \$5.328M	119K \$0.238M	5K \$0.01M	2788K \$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is \$2 per GPU hour.



**Reinforcement Learning with Human Feedback (RLHF)**

Address the **alignment** and **safety** issues.

Slightly adjust the pre-trained model for subsequent tasks.

**Training language models to follow instructions with human feedback**  
*L.Ouyang, J.Wu, X.Jiang, D.Almeida,.. - Advances in ..., 2022 - proceedings.neurips.cc*  
 Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through a language model API, we collect a dataset of ...  
 ☆ Save ⌂ Cite Cited by 4049 Related articles All 10 versions ⌂

Andrej Karpathy's Deep Dive into LLM: [www.youtube.com/watch?v=7xTGNNLPyMI](https://www.youtube.com/watch?v=7xTGNNLPyMI)

8

8

## Posttraining vs. Pretraining

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>

Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/11KWCKUORnPpVMSY6vXaBeFSWo7fJcuGQ9yuR6vC1pzE>

- Much **less compute** (so, much **cheaper** as well) than pretraining.
- Uses SFT and RLHF to **align** the models with **human preferences**.
- Teaches the model how to **use tools**.
  - Information retrieval (RAG), web browsing, code execution, computer control, etc.
- Crafts the model **personality**.
- Sets **refusal/safety** behavior.
  - "As an AI Language Model....."
- The effect of posttraining heavily relies on the **capability of the pretrained base model**.

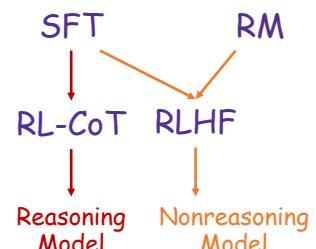
9

## Posttraining Components

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>

Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/11KWCKUORnPpVMSY6vXaBeFSWo7fJcuGQ9yuR6vC1pzE>

- Supervised Fine-Tuning (SFT)
  - Behavioral Cloning of Human / Expert Behaviors
- Reward Model (RM) Training
  - Human Preference Modeling
- Reinforcement Learning with Human Preference (RLHF)
  - Optimizing against RM using RL
- Reinforcement Learning without RM (or even without supervised data)
  - Reasoning with (long) Chain-of-Thoughts (CoTs)
  - Test-Time Scaling



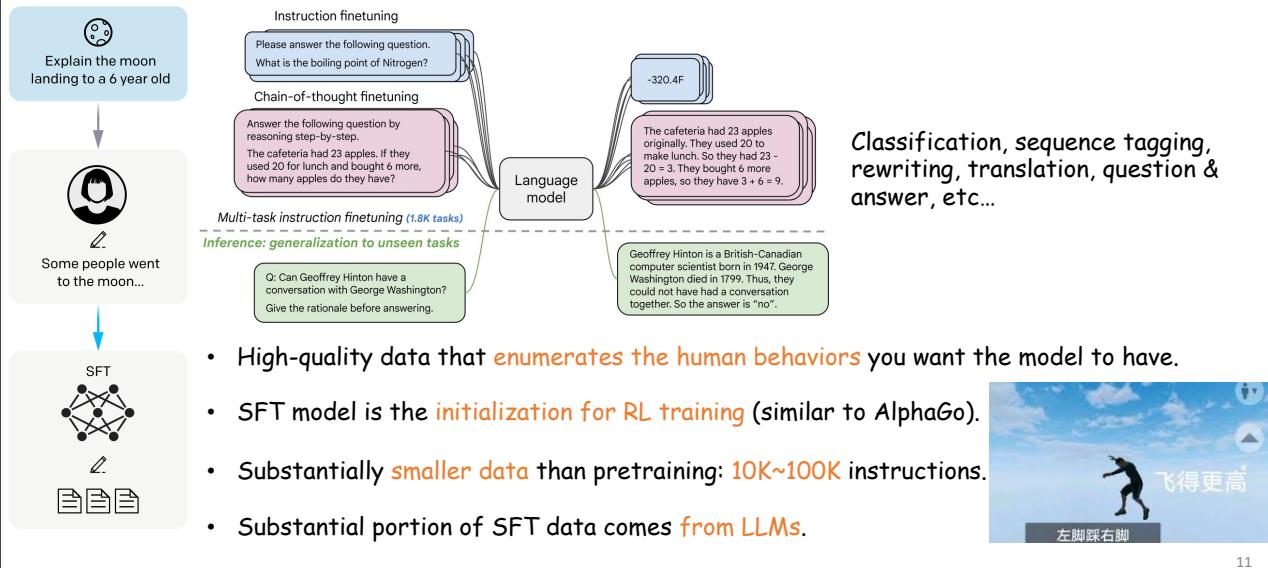
10

10

## Supervised Fine-Tuning

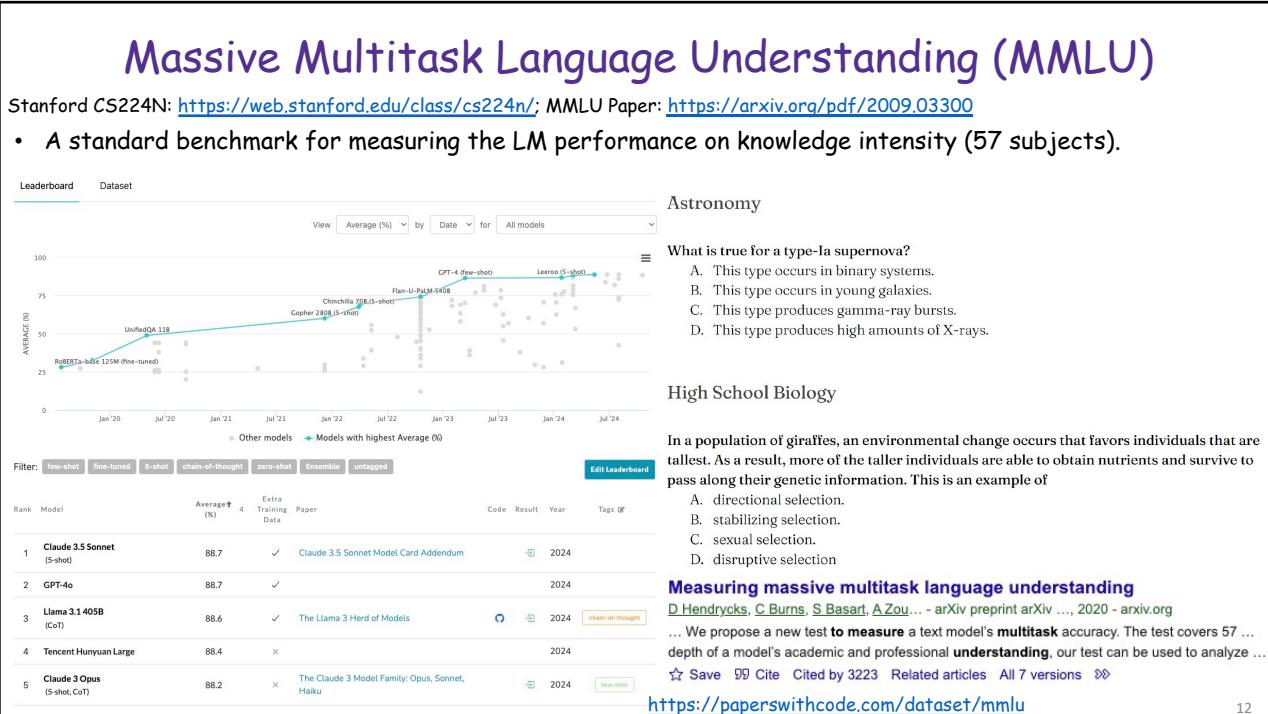
Stanford CS224N: <https://web.stanford.edu/class/cs224n/>

Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/1IKWCKUORnPpVMSY6vXaBeFSWo7fJcuGQ9yuR6vC1pzE>



11

11



12

12

# Agenda

- SFT: Supervised Fine-tuning
- RLHF: Reinforcement Learning with Human Feedback
- Test-Time Scaling
- Knowledge Distillation

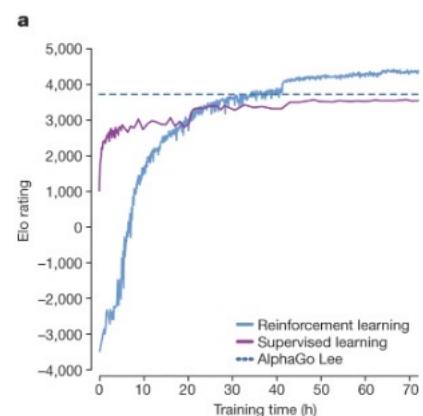
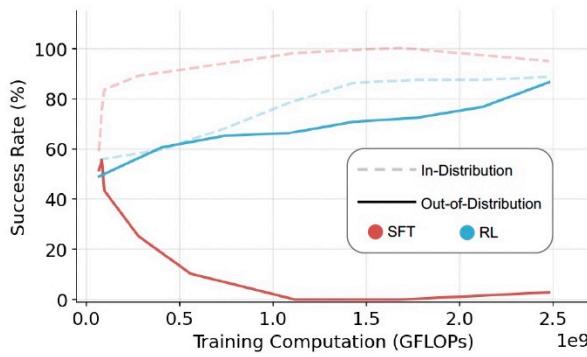
13

13

## Why Do We Need RL?

SFT Memorizes, RL Generalizes: <https://arxiv.org/pdf/2501.17161.pdf>; AlphaGo Zero: <https://www.nature.com/articles/nature24270>

You cannot beat your teacher by imitation.



**Sft memorizes, rl generalizes:** A comparative study of foundation model post-training

T Chu, Y Zhai, J Yang, S Tong, S Xie... - arXiv preprint arXiv ..., 2025 - arxiv.org

... Despite RL's superior **generalization**, we show that **SFT** is still helpful for effective **RL** training: **SFT** stabilizes ... While **RL** exhibits superior **generalization** compared to **SFT**, we show that ...

☆ Save 99 Cite Cited by 3 Related articles All 2 versions ☰

**Mastering the game of go without human knowledge**

D Silver, J Schrittwieser, K Simonyan, I Antonoglou... - nature, 2017 - nature.com

... expert moves, and by reinforcement **learning** from self-play. Here we introduce an algorithm ... reinforcement **learning**, **without human data, guidance or domain knowledge** beyond **game** ...

☆ Save 99 Cite Cited by 11902 Related articles All 42 versions Web of Science: 4555 ☰

14

14

## Reward Modeling

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/1IKWCKUORnPpVMSY6vXqBeFSWo7fJcu6Q9yuR6vC1pzE>

The diagram illustrates the Reward Modeling process. It starts with a prompt "Explain the moon landing to a 6 year old". Four responses are generated: A (Explain gravity...), B (Explain war...), C (Moon is natural satellite of...), and D (People went to the moon...). A human labels these responses: D > C > A = B. These labels are fed into a Reward Model (RM), represented as a neural network. The RM outputs a ranking: D > C > A = B. This ranking is compared against the human labels to calculate a loss function:

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l) \sim D} [\log (\sigma(r_\theta(x, y_w) - r_\theta(x, y_l)))]$$

The components of this equation are labeled: Prompt, Win Response, Lose Response, Sigmoid, and Reward Model (RM), usually a "Small Language Model".

- Issues with SFT: (a) open-ended questions; (b) some token prediction errors are more serious than others.
- Human labellers are asked to rank K LLM-generated responses to a prompt.
- RM helps generalize LLM evaluations for difficult to verify tasks.
- Saves huge costs to recruit human labellers.
- RMs are subject to reward-hacking.

Lilian Weng's Blog: <https://lilianweng.github.io/posts/2024-11-28-reward-hacking/>

15

## Reinforcement Learning with Human Feedback (RLHF)

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>; PPO Algorithm: <https://arxiv.org/abs/1707.06347>  
Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/1IKWCKUORnPpVMSY6vXqBeFSWo7fJcu6Q9yuR6vC1pzE>

The diagram shows the RLHF process. It starts with a prompt "Write a story about frogs". This is processed through Proximal Policy Optimization (PPO) to generate a response "Once upon a time...". This response is then evaluated by a Reward Model (RM), which outputs a reward  $r_k$ .

The process involves several language models and a reinforcement learning loop:

- Prompts Dataset:** x: A dog is...
- Initial Language Model:** Base Text: y: a furry mammal
- Tuned Language Model (RL Policy):** Parameters Frozen\*, RLHF Tuned Text: y: man's best friend
- Reinforcement Learning Update (e.g. PPO):**  $\theta \leftarrow \theta + \nabla_\theta J(\theta)$
- Reward (Preference) Model:** Text: r<sub>θ</sub>(y|x)
- KL prediction shift penalty:**  $-\lambda \text{KL} D_{\text{KL}}(\pi_{\text{PPO}}(y|x) \parallel \pi_{\text{base}}(y|x))$
- Response:** RL Policy

The RL Policy is updated based on the reinforcement learning update and the KL prediction shift penalty. The final reward is calculated as the sum of the RL Policy reward and the KL prediction shift penalty.

Once with an RM, we use RL to automatically optimize the output of fine-tuned LLM in alignment with human preferences.

Challenge: RLHF is quite unstable.

objectives ( $\phi$ ) =  $E_{(x,y) \sim D_{\pi_{\phi}^{\text{RL}}}} [r_\theta(x, y) - \beta \log (\pi_{\phi}^{\text{RL}}(y | x) / \pi_{\phi}^{\text{SFT}}(y | x))] + \gamma E_{x \sim D_{\text{pretrain}}} [\log(\pi_{\phi}^{\text{RL}}(x))] + \text{Loss on the pretraining dist.}$

SFT Model

Not too far away from SFT

16

16

## Value of SFT and RLHF

Instruct GPT: <https://arxiv.org/pdf/2203.02155.pdf>

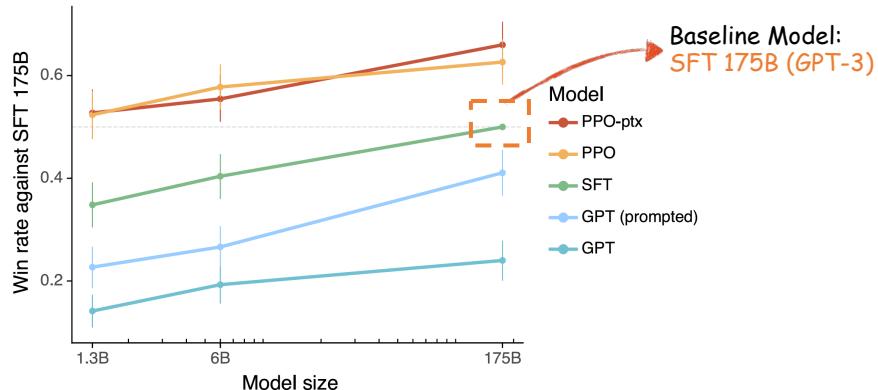


Table 6: Dataset sizes, in terms of number of prompts.

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

17

17

## Limitations with RL + RM

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>; OpenAI Reward Hacking: [openai.com/blog/faulty-reward-functions/](https://openai.com/blog/faulty-reward-functions/); Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/1IKWCKUORnPvMSY6vXaBeFSWo7fJcuGQ9yuR6vC1pzE>

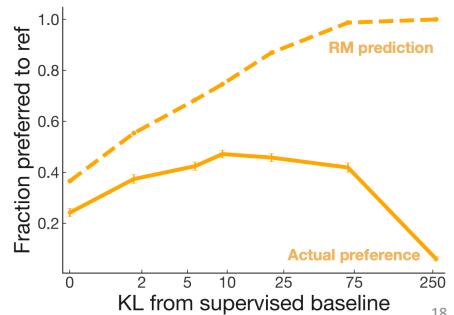
- Reward hacking is a common issue in RL.

TECHNOLOGY

Google shares drop \$100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

- Human preferences are unreliable, so the LLM are rewarded to produce responses that seem authoritative and helpful, regardless of truth: Make-up facts and hallucinations.



Learning to summarize from human feedback: <https://arxiv.org/pdf/2009.01325.pdf>

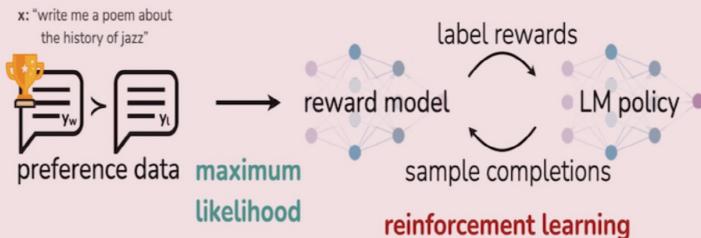
18

18

## Direct Preference Optimization (DPO)

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
 Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/1IKWCKUORnPpVMSY6vXgBeFSWo7fJcu6Q9yuR6vC1pzE>

### Reinforcement Learning from Human Feedback (RLHF)



### Direct Preference Optimization (DPO)



- RL is **unstable** and **challenging** to implement.
- Open-source (non-reasoning) LLMs ([https://huggingface.co/spaces/open-llm-leaderboard/open\\_llm\\_leaderboard](https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard)) mostly use DPO.

$$\text{DPO-Loss} = -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[ \log \sigma(\beta \log \frac{p_\theta^{RL}(y_w|x)}{p_{PT}^{RL}(y_w|x)} - \beta \log \frac{p_\theta^{RL}(y_l|x)}{p_{PT}^{RL}(y_l|x)}) \right]$$

From Human Rankings

Reward for winning sample

Reward for losing sample

DPO Paper: <https://arxiv.org/pdf/2305.18290>

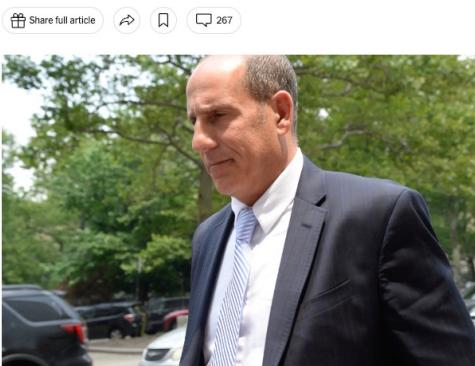
19

19

## Hallucination

### The ChatGPT Lawyer Explains Himself

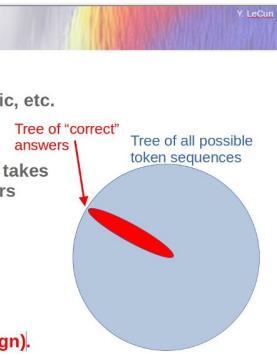
In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he "did not comprehend" that the chat bot could lead him astray.



Steven A. Schwartz told a judge considering sanctions that the episode had been "deeply embarrassing." Jefferson Siegel for The New York Times

### Unpopular Opinion about AR-LLMs

- ▶ Auto-Regressive LLMs are **doomed**.
- ▶ They cannot be made factual, non-toxic, etc.
- ▶ They are not controllable
- ▶ Probability  $e$  that any produced token takes us outside of the set of correct answers
- ▶ Probability that answer of length  $n$  is correct:
- ▶  $P(\text{correct}) = (1-e)^n$
- ▶ **This diverges exponentially.**
- ▶ **It's not fixable (without a major redesign).**



- **Hallucination:** LLM learns the format but the content. Let alone the rationale and insights.

问一下知乎的医生们，以后的患者要是用DeepSeek问答的内容和你们对线要怎么办？

关注问题

写回答

邀请回答

好问题 2

1条评论

分享

20

20

## Emergent Abilities

• **Emergent Abilities:** An ability not present in smaller models but present in larger models.

• **Phase-change in physics:** Quantitative changes in the system result in qualitative changes in behavior.

**Emergent abilities of large language models**

J Wei, Y Tay, R Bommasani, C Raffel, B Zoph... - arXiv preprint arXiv ..., 2022 - arxiv.org

... an ability to be emergent if it is not present in smaller models but is present in larger models. ... We have discussed emergent abilities of language models, for which meaningful ...

☆ Save 99 Cite Cited by 1256 Related articles All 11 versions ☰

21

21

## Are Emergent Abilities of LLMs a Mirage?

Multiple Choice Grade  $\stackrel{\text{def}}{=} \begin{cases} 1 & \text{if highest probability mass on correct option} \\ 0 & \text{otherwise} \end{cases}$

Exact String Match  $\stackrel{\text{def}}{=} \begin{cases} 1 & \text{if output string exactly matches target string} \\ 0 & \text{otherwise} \end{cases}$

- **Emergent Abilities** may be attributed to the choice of nonlinear or discontinuous metrics, whereas linear or continuous metrics produce smooth performance changes.

Category	Metric	Model Parameters	Description
Emergent Abilities	Nonlinearly score LLM outputs	A: $L_{CE}(\text{Per-Token})$ vs. Model Parameters (log-log scale)	Graph A shows a clear power-law-like decrease in loss per token as model size increases.
	Discontinuously score LLM outputs	C: Multiple Choice Grade vs. Model Parameters	Graph C shows a step-function-like increase in accuracy for small models, followed by a plateau.
No Emergent Abilities	Linearly score LLM outputs	E: Exact String Match vs. Model Parameters	Graph E shows a smooth, linear increase in accuracy with model size.
	Continuously score LLM outputs	F: Inter Score vs. Model Parameters	Graph F shows a smooth, continuous curve for the inter score metric.

**Power-law of Scaling**

$$\mathcal{L}_{CE}(N) = \left(\frac{N}{c}\right)^\alpha$$

$$p(\text{single token correct}) = \exp(-\mathcal{L}_{CE}(N)) = \exp(-(N/c)^\alpha)$$

$$\text{Accuracy}(N) \approx p_N(\text{single token correct})^{\text{num. of tokens}} = \exp(-(N/c)^\alpha)^L$$

Figure C → Figure E

Are emergent abilities of large language models a mirage?  
R Schaeffer, B Miranda... - Advances in Neural ..., 2024 - proceedings.neurips.cc  
... be interpreted as claiming that large language models cannot display emergent abilities; rather, our message is that some previously claimed emergent abilities appear to be mirages ...

☆ Save 99 Cite Cited by 129 Related articles All 9 versions ☰

22

22

# Agenda

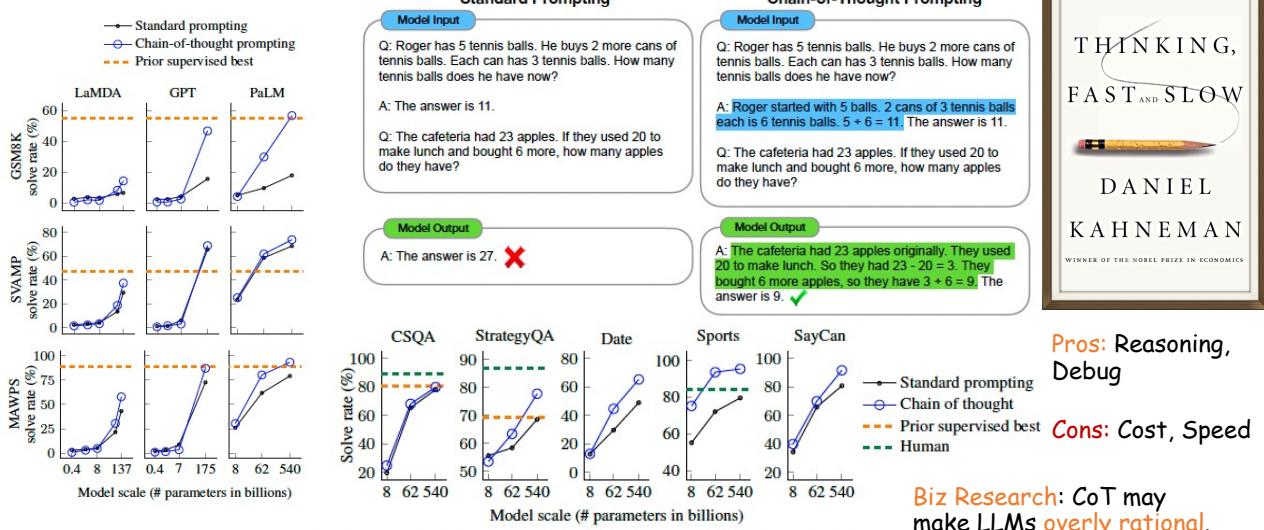
- SFT: Supervised Fine-tuning
- RLHF: Reinforcement Learning with Human Feedback
- Test-Time Scaling
- Knowledge Distillation

23

23

## Chain-of-Thought (CoT)

- A series of **intermediate reasoning steps** significantly improves the ability of large language models for complex reasoning.



24

24

## Tree-of-Thought (ToT)

- Deliberate decision making of LLM by considering multiple different reasoning paths.

**Tree of thoughts: Deliberate problem solving with large language models**  
S.Yao, D.Yu, J.Zhao, I.Shafrazi... - Advances in neural ... , 2023 - proceedings.neurips.cc  
Abstract Language models are increasingly being deployed for general problem solving across a wide range of tasks, but are still confined to token-level, left-to-right decision-making processes during inference. This means they can fall short in tasks that require exploration, strategic lookahead, or where initial decisions play a pivotal role. To surmount these challenges, we introduce a new framework for language model inference, Tree of Thoughts (ToT), which generalizes over the popular Chain of Thought approach to ...  
☆ Save 99 Cite Cited by 2214 Related articles All 12 versions 88 ToT Paper: <https://arxiv.org/pdf/2305.10601.pdf>

- Thought decomposition
- Thought generator
- State evaluator
- Search algorithm

**(a) GPT-4 coherency scores**

Condition	Score (approx.)
IO	7.5
CoT	7.0
ToT	7.5
+refine +refine	8.0
ToT +refine +refine	8.0

**(b) Human coherency comparison**

Condition	Score (approx.)
CoT > ToT	21
Similar	38
ToT > CoT	41

**Figure 5: Creative Writing results.**

**(a) Input**  
Input → Plan 1 → Passage 1  
Input → Plan 2 → Passage 2  
...  
**(b) Plans**  
Plan 1: 1. Introduce and explain the technique of self-reflection. 2. Share a personal story about an astronaut's first time in space. 3. Describe a situation where a woman uses sign language to avoid unwanted attention. 4. Contemplate how different perceptions of others shape one's identity.  
Plan 2: 1. Introduction to an unusual self-help book. 2. Discuss the unexpected things learned from astronauts, including the smell of space. 3. Describe a woman's clever tactic for avoiding unwanted attention at a bar. 4. Contemplate how different perceptions of oneself can shape one's identity.  
**(c) Votes**  
Plan 1: 0/5 votes  
Plan 2: 3/5 votes  
Plan 3-5: n/5 votes  
Analyze each choice in detail: Choice 1, while incorporating the required end sentences, seems to lack a clear connection between the paragraphs. Choice 2 offers an interesting perspective by using the required end sentences to present a self-help book's content. It connects the paragraphs with the theme of self-improvement and embracing challenges, making for a coherent passage. The best choice is 2.

25

## Test-Time Scaling

OpenAI o1: <https://openai.com/index/learning-to-reason-with-langs/>; DeepMind Scaling Test-Time Compute: <https://arxiv.org/pdf/2408.03314.pdf>

- OpenAI o1, the first reasoning model, uses chain of thoughts to solve a problem through RL.

**• Test-time scaling is more efficient than pretraining scaling.**  
• But how?

**• Forget about RM, directly applies RL to CoTs on easily verifiable and intelligence-intensive problems, such as math and coding.**

**Strong base model from pretraining is necessary.**

26

## DeepSeek-R1

DeepSeek-R1: <https://arxiv.org/pdf/2501.12948.pdf>; DeepSeekMath: <https://arxiv.org/pdf/2402.03300.pdf>; Kimi K1.5: <https://arxiv.org/pdf/2501.12599.pdf>

- DeepSeek-R1, the first open sourced reasoning model, much cheaper than but as good as OpenAI o1.
  - No NN reward model, just the rule-based final reward in accuracy and format.
  - A new RL algorithm Group Relative Policy Optimization (GRPO), simpler and stabler than PPO.
- DeepSeek-R1-Zero, purely RL, no SFT: Self-evolving intelligence:
  - Poor readability
  - Language mix
- DeepSeek-R1: Cold Start with SFT.
  - Strong small models distilled from DeepSeek-R1.
- Kimi K 1.5 also tries to scale RL with long CoTs with a partial roll-out system.

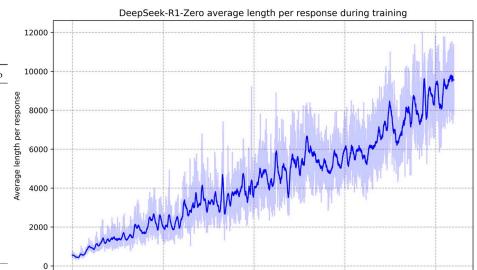


DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

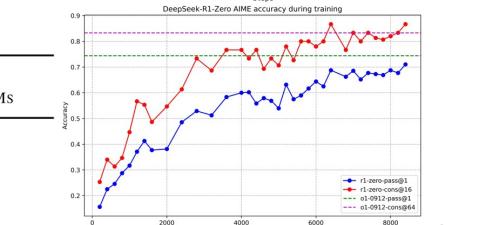
DeepSeek-AI  
research@deepseek.com

Question: If  $a > 1$ , then the sum of the real solutions of  $\sqrt{a} - \sqrt{a+x} = x$  is equal to  
Response: <thinks>  
To solve the equation  $\sqrt{a} - \sqrt{a+x} = x$ , let's start by squaring both ...  
 $(\sqrt{a} - \sqrt{a+x})^2 = x^2 \implies a - \sqrt{a+x} = x^2$   
Rearrange to isolate the inner square root term:  
 $(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$   
...  
Wait, wait. Wait. That's an ah-ha moment I can flag here.  
Let's reevaluate this step-by-step to identify if the correct sum can be ...  
We started with the equation:  
 $\sqrt{a} - \sqrt{a+x} = x$   
First, let's square both sides:  
 $a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$   
Next, I could square both sides again, treating the equation: ...

DeepSeek-R1-Zero average length per response during training



DeepSeek-R1-Zero AIME accuracy during training



KIMI K1.5:  
SCALING REINFORCEMENT LEARNING WITH LLMs

TECHNICAL REPORT OF KIMI K1.5  
Kimi Team

27

27

JOURNAL ARTICLE CORRECTED PROOF  
**Generative AI at Work** ⓘ  
Erik Brynjolfsson, Danielle Li, Lindsey Raymond  
*The Quarterly Journal of Economics*, qjae044, <https://doi.org/10.1093/qje/qjae044>  
Published: 04 February 2025 Article history •

PDF Split View Cite Permissions Share ▾

**Abstract**  
We study the staggered introduction of a generative AI-based conversational assistant using data from 5,172 customer-support agents. Access to AI assistance increases worker productivity, as measured by issues resolved per hour, by 15% on average, with substantial heterogeneity across workers. The effects vary significantly across different agents. Less experienced and lower-skilled workers improve both the speed and quality of their output, while the most experienced and highest-skilled workers see small gains in speed and small declines in quality. We also find evidence that AI assistance facilitates worker learning and improves English fluency, particularly among international agents. While AI systems improve with more training data, we find that the gains from AI adoption are largest for moderately rare problems, where human agents have less baseline experience but the system still has adequate training data. Finally, we provide evidence that AI assistance improves the experience of work along several dimensions: customers are more polite and less likely to ask to speak to a manager.

**Generative AI at work**  
Erik Brynjolfsson, D Li, L Raymond - *The Quarterly Journal of* ..., 2025 - academic.oup.com  
We study the staggered introduction of a generative AI-based conversational assistant using data from 5,172 customer-support agents. Access to AI assistance increases worker ...  
☆ Save ⌂ Cite Cited by 1002 Related articles All 18 versions ▾

<https://www.zhihu.com/question/12918439244>  
<https://arxiv.org/pdf/2501.19393.pdf>

## Reasoning Models in Biz Research?

- Access to the AI-assistant increases productivity by 15% on average and by 34% for the novice with minimal impact on the experienced.
  - Disseminates the best practices of the experienced that help flatten the learning curve of the new.
  - AI reduces the marginal cost of distributing intelligence.
- 如何看待镇江部署DeepSeek，称「建成600台算力服务器集群，单日处理量为全市公务员十年工作量」？  
2月19日，镇江举行新闻发布会，DeepSeek正式登陆镇江，完成本地化部署上线，这是镇江市推进数字经济发展、助力数字经济高质量发展的关键举措。
- Economic impact of reasoning model deployment at large?
  - Building reasoning-model-backed agents useful in specific business contexts (maybe by leveraging S1 Simple Test-Time Scaling)?
  - Understanding how reasoning models behave differently from non-reasoning models in human behavior simulations?

28

28

# Agenda

- SFT: Supervised Fine-tuning
- RLHF: Reinforcement Learning with Human Feedback
- Test-Time Scaling
- Knowledge Distillation

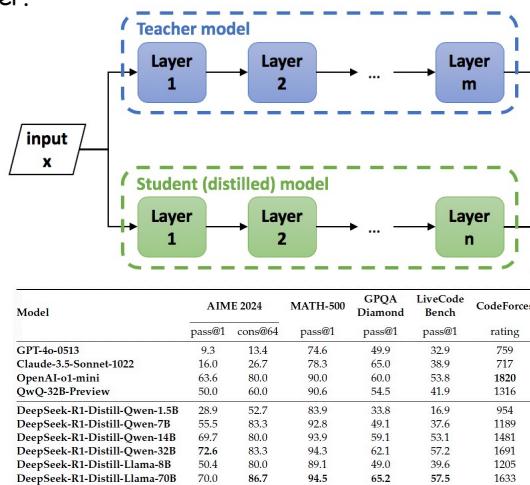
29

29

## Knowledge Distillation (KD)

MIT 6.5940 Efficient DL Computing: [efficientml.ai](http://efficientml.ai)

- Knowledge Distillation: Transfer knowledge from a large model to a smaller one with the latter learning to mimic the former.



KD is very powerful:

Distilling the knowledge in a neural network

[G Hinton, O Vinyals, J Dean - arXiv preprint arXiv:1503.02531, 2015 - arxiv.org](https://arxiv.org/pdf/1503.02531.pdf)

... We show that the distillation strategy that we propose in this paper achieves the desired effect of distilling an ensemble of models into a single model that works significantly better than ...

☆ 保存 59 引用 被引用次数: 23531 相关文章 所有 25 个版本 ✎

KD Paper: <https://arxiv.org/pdf/1503.02531.pdf>

T: Temperature

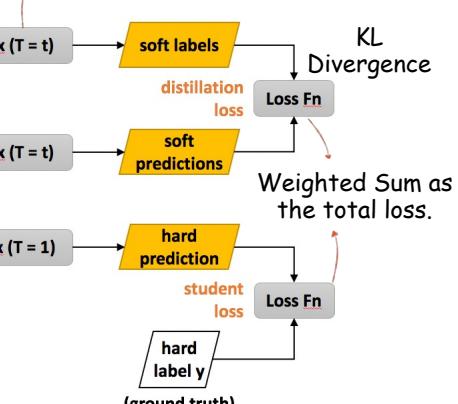


Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

30

30

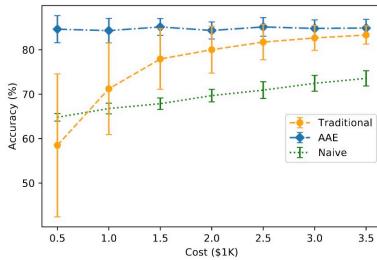
## KD for Market Research

- We have limited human data and a lot of LLM-generated data. How to correctly identify human preferences?
- Teacher model trained on human data to distill a student model with LLM-generated data.

**ESTIMATION WITH AI-AUGMENTED DATA**

Step 1. Obtain an estimator  $\hat{\theta}$  to  $\theta^*$ , where  $\mathbb{P}(y=j|x, z) = g_j(x, z; \theta^*)$ , using the primary data.  
 Step 2. With the auxiliary data, we construct the estimator  $\hat{\beta}^{AAE}$  as **AI-Augmented Estimator**

Distillation:  $\hat{\beta}^{AAE} \in \arg \max_{\beta \in \mathbb{R}^d} \left\{ \hat{Q}(\hat{\theta}; \beta) = \frac{1}{n} \sum_{i=1}^n \sum_{j \in K^+} g_j(x_i, z_i; \hat{\theta}) \log \sigma_j(x; \beta) \right\}$ .



### Large Language Models for Market Research: A Data-augmentation Approach

M Wang, DJ Zhang, H Zhang - arXiv preprint arXiv:2412.19363, 2024 - arxiv.org  
 ... our context, we present a **data-augmentation statistical approach** for extracting value from LLM-... **data augmentation approach** that allows us to use the AI-generated data to fit the **model**. ...  
 ☆ 保存 99 引用 被引用次数: 1 相关文章 所有 3 个版本 ☰

<https://arxiv.org/pdf/2412.19363.pdf>

### Large Language Models for Market Research: A Data-augmentation Approach

Mengxin Wang (Naveen Jindal School of Management, The University of Texas at Dallas),  
 Dennis J. Zhang (Olin School of Business, Washington University in St. Louis), Heng Zhang  
 (W. P. Carey School of Business, Arizona State University)

Large Language Models (LLMs) have transformed artificial intelligence by excelling in complex natural language processing tasks. Their ability to generate human-like text has opened new possibilities for market research, particularly in conjoint analysis, where understanding consumer preferences is essential but often resource-intensive. Traditional survey-based methods face limitations in scalability and cost, making LLM-generated data a promising alternative. However, while LLMs have the potential to simulate real consumer behavior, recent studies highlight a significant gap between LLM-generated and human data, with biases introduced when substituting between the two. In this paper, we address this gap by proposing a novel statistical data augmentation approach that efficiently integrates LLM-generated data with real data in conjoint analysis. Our method leverages transfer learning principles to debias the LLM-generated data using a small amount of human data. This results in statistically robust estimators with consistent and asymptotically normal properties, in contrast to naive approaches that simply substitute human data with LLM-generated data, which can exacerbate bias. We validate our framework through an empirical study on COVID-19 vaccine preferences, demonstrating its superior ability to reduce estimation error and save data and costs by 24.9% to 79.8%. In contrast, naive approaches fail to save data due to the inherent biases in LLM-generated data compared to human data. Another empirical study on sports car choices validates the robustness of our results. Our findings suggest that while LLM-generated data is not a direct substitute for human responses, it can serve as a valuable complement when used within a robust statistical framework.

Subjects: Artificial Intelligence (cs.AI); Machine Learning (cs.LG); Methodology (stat.ME); Machine Learning (stat.ML)

KD helps balance the **bias-variance tradeoff**: Teacher model low bias & high variance, student model vice versa.

31