

DOTE 6635: Artificial Intelligence for Business Research

# Posttraining

Renyu (Philip) Zhang

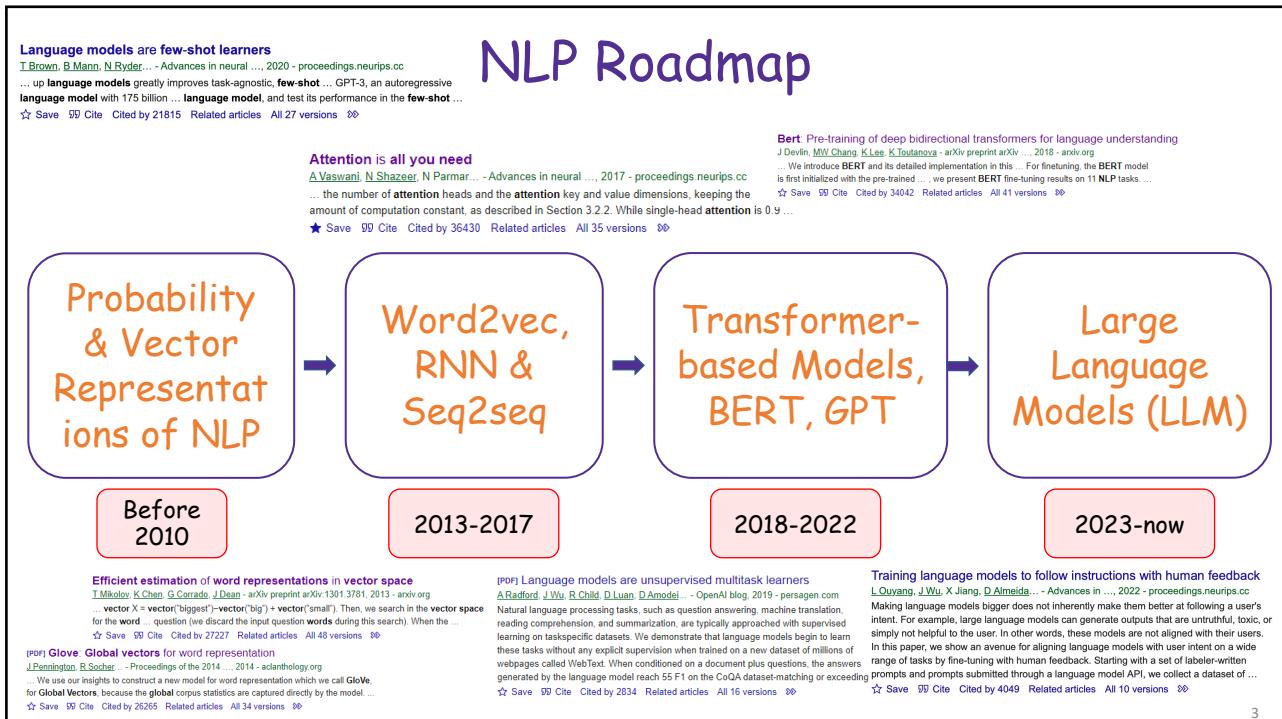
1

## Agenda

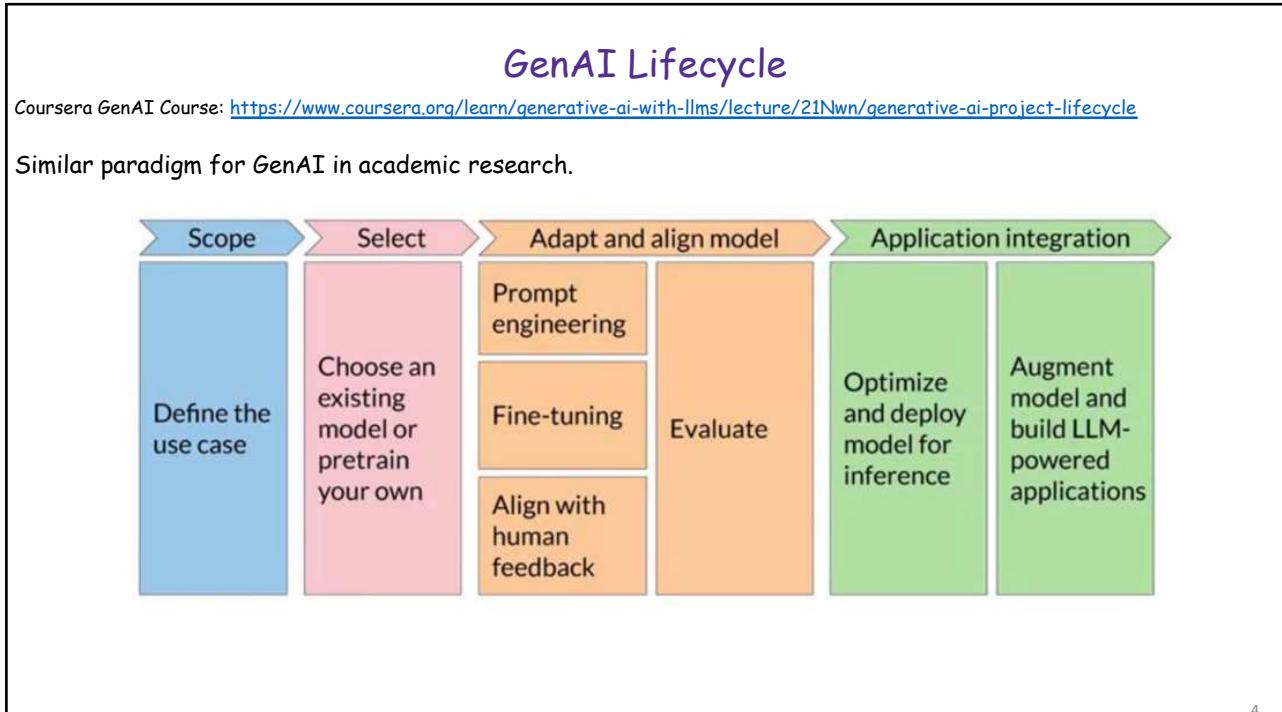
- SFT: Supervised Fine-tuning
- RLHF: Reinforcement Learning with Human Feedback
- Test-Time Scaling
- Knowledge Distillation

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](#)

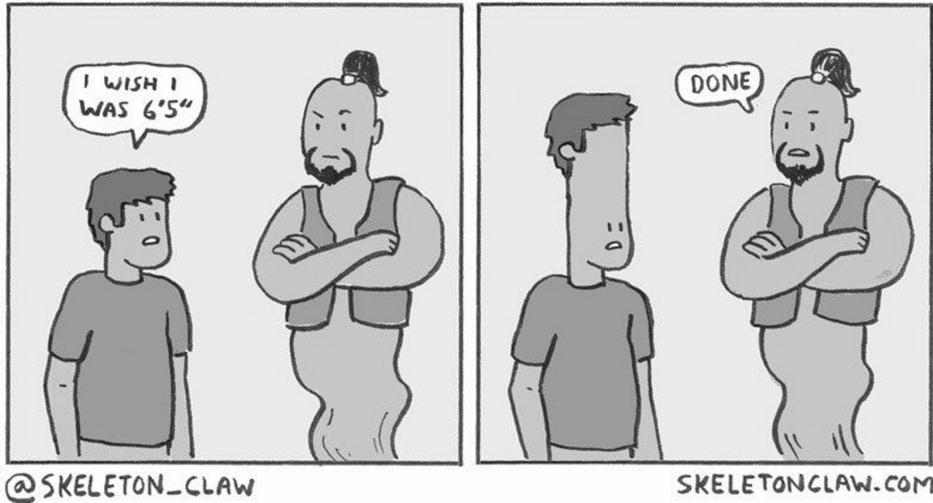
6

6

## AI Misalignment

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

GENIE 2

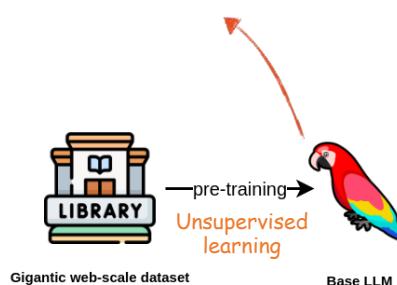


7

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.

Slightly adjust the pre-trained model for subsequent tasks.

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

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

**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

9

## Mis-misalignment

Generate a picture of Elon Musk

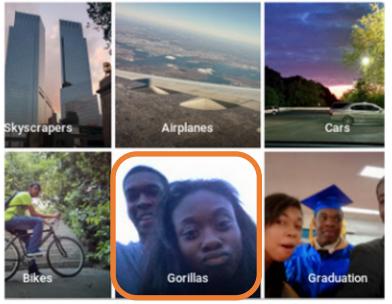
Sure




© Gemini AI

**Jacky Alciné** @jackyalcine Follow

Google Photos, y'all [REDACTED] up. My friend's not a gorilla.



Skyscrapers Airplanes Cars  
Bikes Gorillas Graduation

RETWEETS 3,356 FAVORITES 1,930

8:22 PM - 28 Jun 2015

Over-alignment also makes LLMs dumber.....

10

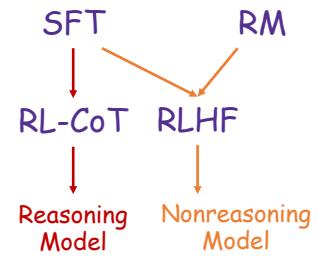
10

## 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



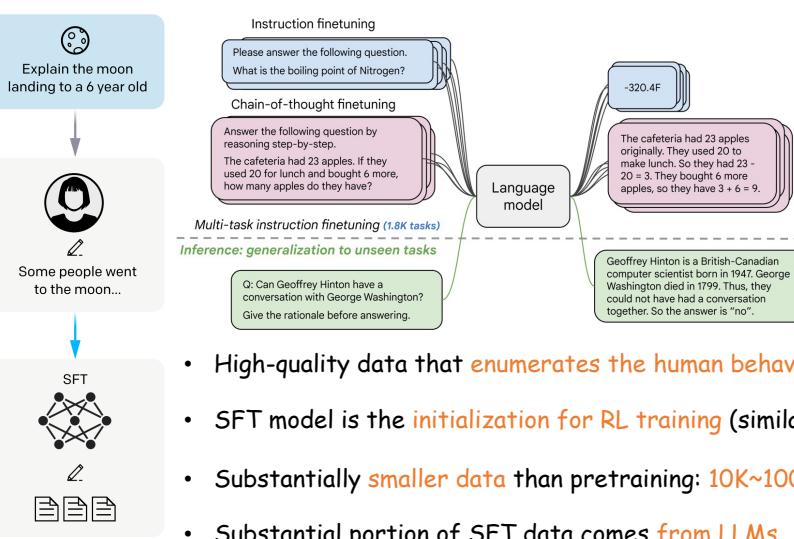
11

11

## Supervised Fine-Tuning

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

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



- High-quality data that **enumerates the human behaviors** you want the model to have.
- SFT model is the **initialization for RL training** (similar to AlphaGo).
- Substantially **smaller data** than pretraining: **10K~100K instructions**.
- Substantial portion of SFT data comes **from LLMs**.



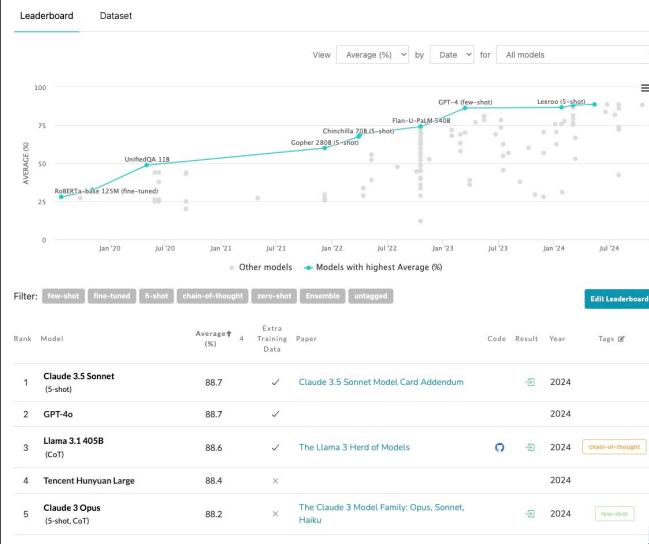
12

12

## Massive Multitask Language Understanding (MMLU)

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>; MMLU Paper: <https://arxiv.org/pdf/2009.03300>

- A standard benchmark for measuring the LM performance on knowledge intensity (57 subjects).



### Astronomy

What is true for a type-Ia supernova?

- This type occurs in binary systems.
- This type occurs in young galaxies.
- This type produces gamma-ray bursts.
- This type produces high amounts of X-rays.

### High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

- directional selection.
- stabilizing selection.
- sexual selection.
- disruptive selection

### Measuring massive multitask language understanding

D Hendrycks, C Burns, S Basart, A Zou... - arXiv preprint arXiv ..., 2020 - arxiv.org

... We propose a new test to measure a text model's **multitask** accuracy. The test covers 57 ... depth of a model's academic and professional **understanding**, our test can be used to analyze ...

☆ Save 99 Cite Cited by 3223 Related articles All 7 versions ☺

<https://paperswithcode.com/dataset/mmlu>

13

13

## Agenda

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

14

14

## Why Do We Need RL?

STF 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 ☺

15

## Reward Modeling

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/11KWCKUORnPpVMSY6vXaBeFSWo7fJcuGQ9yuR6vC1pzE>

- 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.

loss ( $\theta$ ) =  $-\frac{1}{(K)} \sum_{i=1}^K \log (\sigma (r_\theta (x, y_w) - r_\theta (x, y_l)))$

Prompt → Win Response → Sigmoid → Reward Model (RM), usually a "Small Language Model"

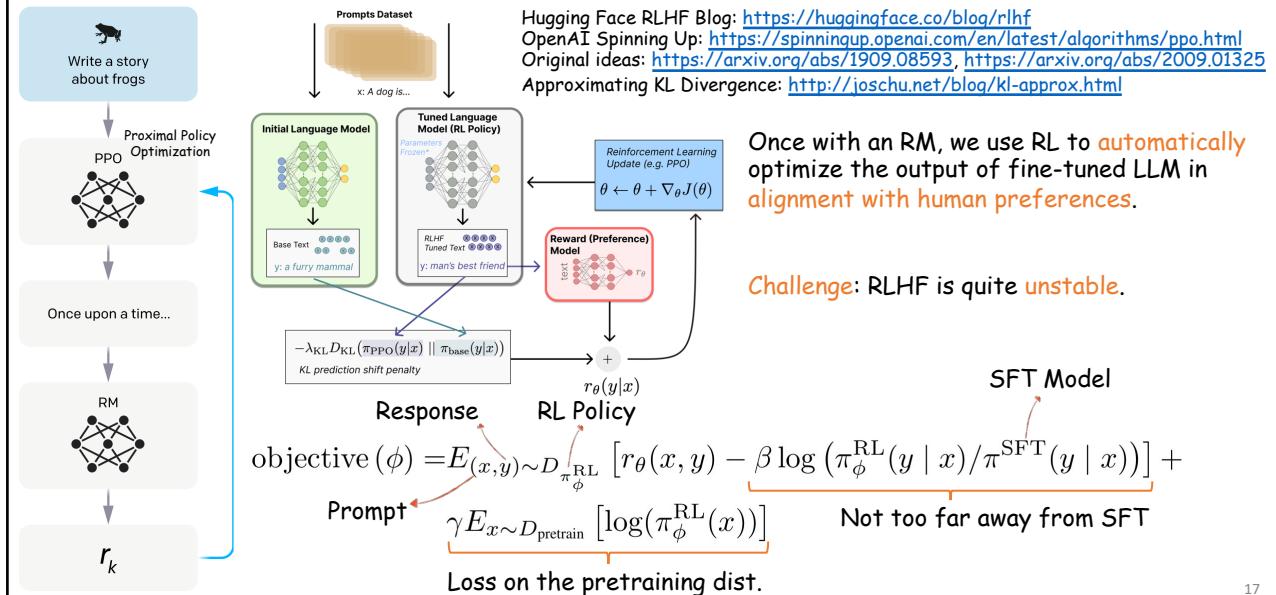
- RM helps generalize LLM evaluations for difficult to verify tasks.
- Saves huge costs to recruit human labellers.
- RMs are subject to reward-hacking.

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

16

## 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/1IKWCKUOrnPvMSY6vXgBeFSWo7fJcu6Q9yuR6vC1pzE>



17

17

## Value of SFT and RLHF

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

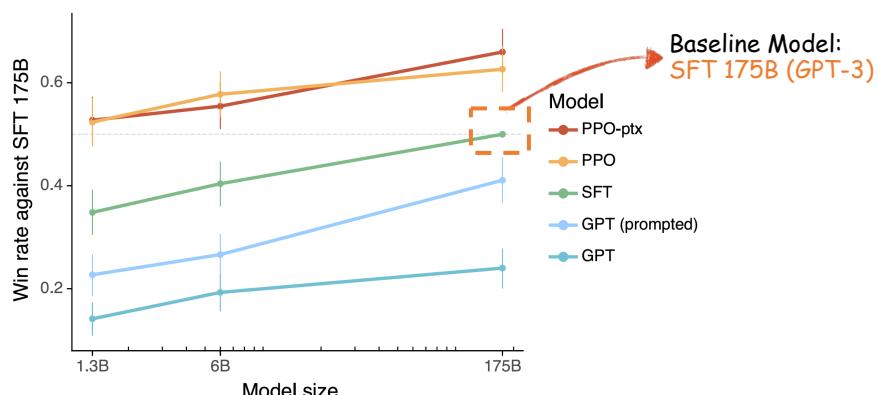


Table 6: Dataset sizes, in terms of number of prompts.

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			

18

18

## 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/11KWCKUORnPpVMSY6vXqBeFSWo7fJcuGQ9yuR6vC1pzE>

- Reward hacking is a common issue in RL.

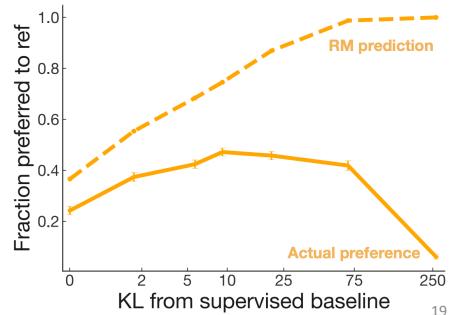
TECHNOLOGY

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.

- Models of human preferences are even more unreliable.



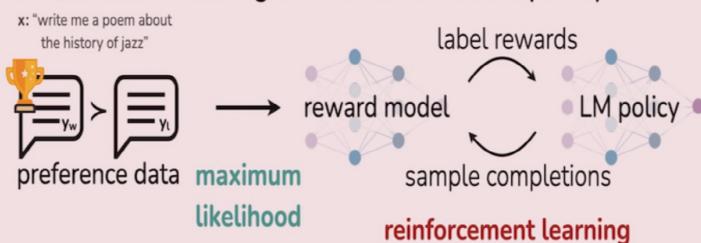
Learning to summarize from human feedback: <https://arxiv.org/pdf/2009.01325.pdf>

19

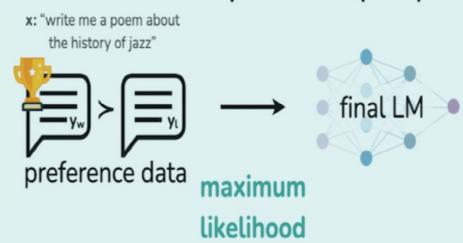
## Direct Preference Optimization (DPO)

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Talk by Barret Zoph and John Schulman: <https://docs.google.com/presentation/d/11KWCKUORnPpVMSY6vXqBeFSWo7fJcuGQ9yuR6vC1pzE>

### 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.

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

From Human Rankings

Reward for winning sample

Reward for losing sample

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

20

20

## 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.

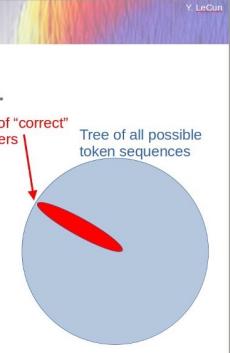
[Share full article](#) 267



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](#) ...

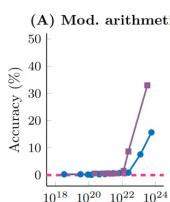
21

21

## Emergent Abilities

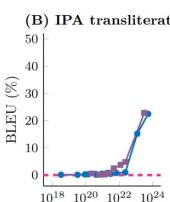
- **Emergent Abilities:** An ability not present in smaller models but present in larger models.
- **Phase-change in physics:** Quantitative changes in the system result in qualitative changes in behavior.

(A) Mod. arithmetic



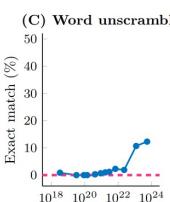
Model scale (training FLOPs)	LaMDA	GPT-3	Gopher	Chinchilla	PaLM	Random
$10^{18}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{20}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{22}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{24}$	~18%	~35%	~18%	~18%	~18%	~18%

(B) IPA transliterate



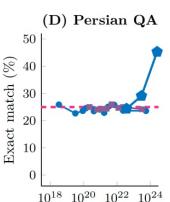
Model scale (training FLOPs)	LaMDA	GPT-3	Gopher	Chinchilla	PaLM	Random
$10^{18}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{20}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{22}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{24}$	~18%	~22%	~18%	~18%	~18%	~18%

(C) Word unscramble



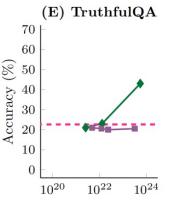
Model scale (training FLOPs)	LaMDA	GPT-3	Gopher	Chinchilla	PaLM	Random
$10^{18}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{20}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{22}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{24}$	~18%	~15%	~18%	~18%	~18%	~18%

(D) Persian QA



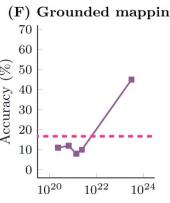
Model scale (training FLOPs)	LaMDA	GPT-3	Gopher	Chinchilla	PaLM	Random
$10^{18}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{20}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{22}$	~1%	~1%	~1%	~1%	~1%	~1%
$10^{24}$	~18%	~45%	~18%	~18%	~18%	~18%

(E) TruthfulQA



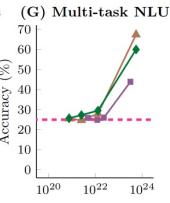
Model scale (training FLOPs)	LaMDA	GPT-3	Gopher	Chinchilla	PaLM	Random
$10^{20}$	~18%	~18%	~18%	~18%	~18%	~18%
$10^{22}$	~18%	~18%	~18%	~18%	~18%	~18%
$10^{24}$	~45%	~45%	~45%	~45%	~45%	~45%

(F) Grounded mappings



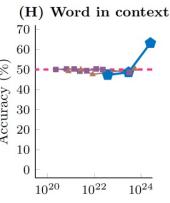
Model scale (training FLOPs)	LaMDA	GPT-3	Gopher	Chinchilla	PaLM	Random
$10^{20}$	~18%	~18%	~18%	~18%	~18%	~18%
$10^{22}$	~18%	~18%	~18%	~18%	~18%	~18%
$10^{24}$	~45%	~45%	~45%	~45%	~45%	~45%

(G) Multi-task NLU



Model scale (training FLOPs)	LaMDA	GPT-3	Gopher	Chinchilla	PaLM	Random
$10^{20}$	~18%	~18%	~18%	~18%	~18%	~18%
$10^{22}$	~18%	~18%	~18%	~18%	~18%	~18%
$10^{24}$	~65%	~65%	~65%	~65%	~65%	~65%

(H) Word in context



Model scale (training FLOPs)	LaMDA	GPT-3	Gopher	Chinchilla	PaLM	Random
$10^{20}$	~18%	~18%	~18%	~18%	~18%	~18%
$10^{22}$	~18%	~18%	~18%	~18%	~18%	~18%
$10^{24}$	~65%	~65%	~65%	~65%	~65%	~65%

**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 ⚡ Cite Cited by 1256 Related articles All 11 versions ⚡

22

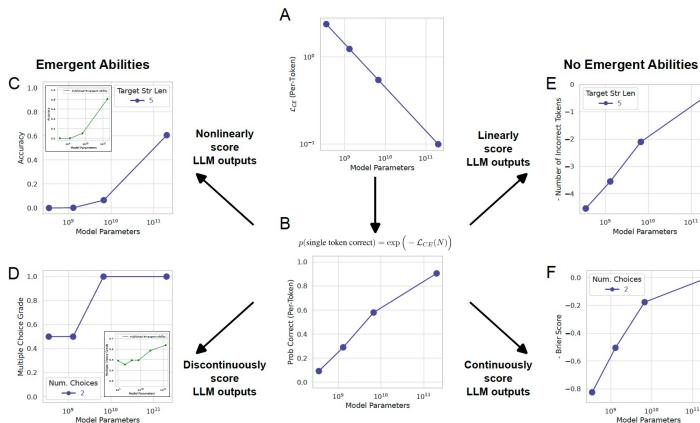
22

## Are Emergent Abilities of LLMs a Mirage?

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

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

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



NeurIPS 2023 Outstanding Main Track Paper

### 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

$$\text{Token Edit Distance}(N) \approx L(1 - p_N(\text{single token correct})) = L(1 - \exp(-(N/c)^{\alpha}))$$

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

23

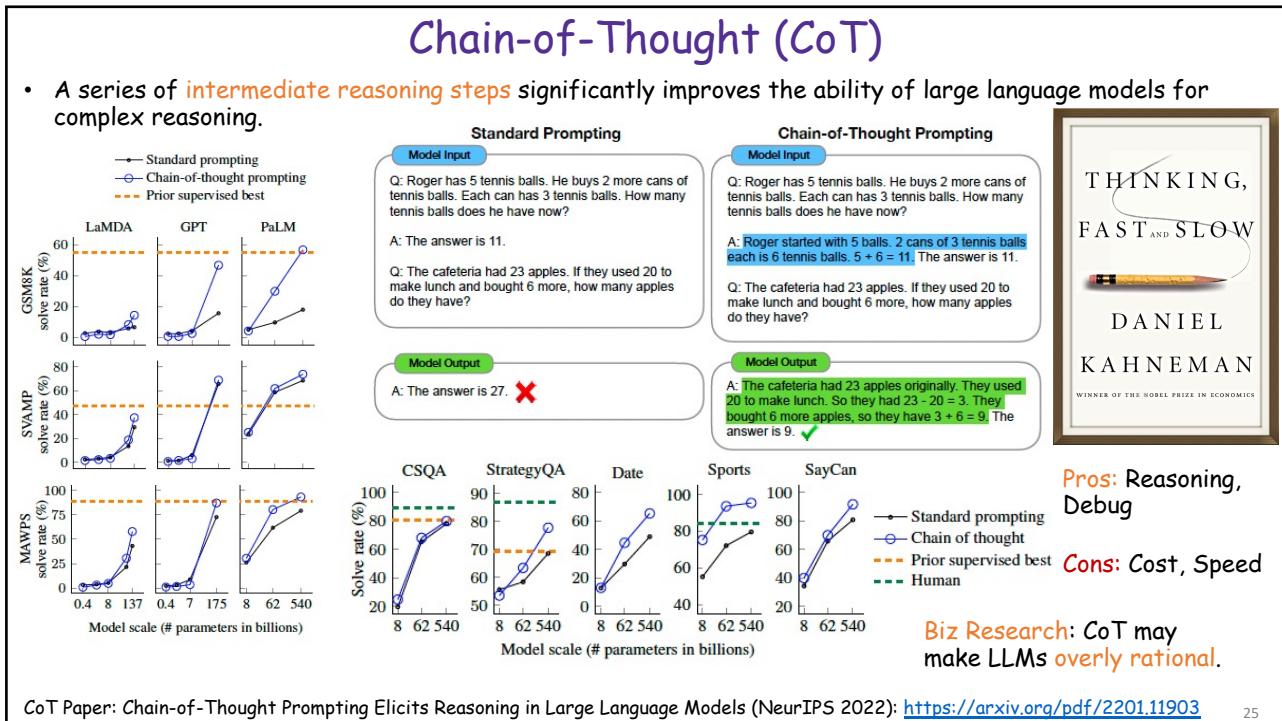
23

## Agenda

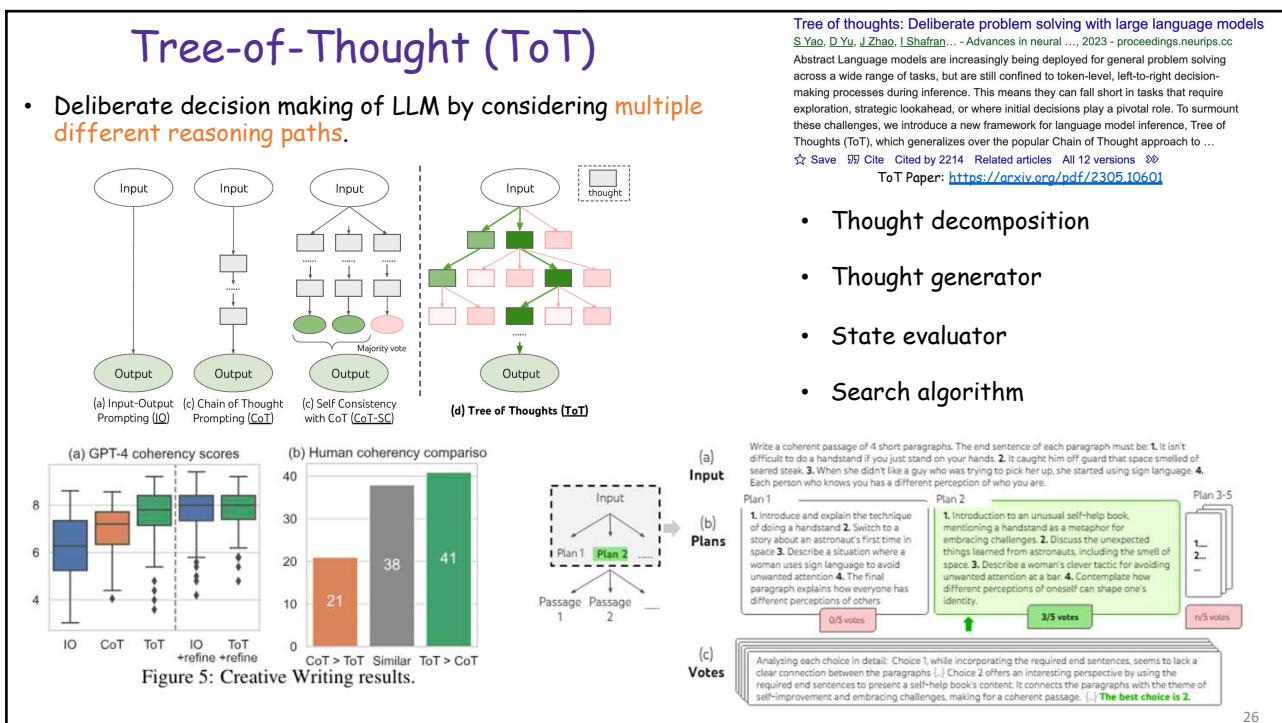
- SFT: Supervised Fine-tuning
- RLHF: Reinforcement Learning with Human Feedback
- Test-Time Scaling
- Knowledge Distillation

24

24



25



26

## 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.

Comparing Test-time and Pretraining Compute in a FLOPs Matched Evaluation

Strong base model from pretraining is necessary.

27

27

## 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 K1.5 also tries to scale RL with long CoTs with a partial roll-out system.

deepseek

Question: If  $a > 1$ , then the sum of the real solutions of  $\sqrt{a - \sqrt{a+x}} = x$  is equal to  
Response: <think>  
To solve the equation  $\sqrt{a - \sqrt{a+x}} = x$ , let's start by squaring both sides:  
$$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$$
  
Rewriting 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 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, we squared 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

KIMI K1.5:  
SCALING REINFORCEMENT LEARNING WITH LLMs

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

TECHNICAL REPORT OF KIMI K1.5

KIMI Team

DeepSeek-R1-Zero AIME accuracy during training

28

28

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 Cite 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, 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**  
 E. 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

## 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建设、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?**

<https://www.zhihu.com/question/12918439244>

<https://arxiv.org/pdf/2501.19393.pdf>

29

## Agenda

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

30

30

## Knowledge Distillation (KD)

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

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

The diagram illustrates the Knowledge Distillation (KD) process. It shows two parallel neural network architectures: a **Teacher model** (top, dashed blue box) and a **Student (distilled) model** (bottom, dashed green box). Both models receive an **input  $x$** . The Teacher model consists of  $m$  layers, and the Student model consists of  $n$  layers. The Teacher's output is a **Softmax ( $T = t$ )** layer producing **soft labels**, which are then used to calculate the **KL Divergence** loss. The Student's output is also a **Softmax ( $T = t$ )** layer producing **soft predictions**, and a **Softmax ( $T = 1$ )** layer producing a **hard prediction**. This hard prediction is compared against the **hard label  $y$  (ground truth)** to calculate the **student loss**. The total loss is the **Weighted Sum as the total loss**, combining the KL Divergence loss and the student loss.

**Distilling the knowledge in a neural network**  
G Hinton, O Vinyals, J Dean - arXiv preprint arXiv:1503.02531, 2015 - arxiv.org  
... 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 ⚡ 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>

Model	AIME 2024		MATH-500		GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1			
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759	
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717	
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	<b>1820</b>	
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316	
DeepSeek-R1-Distill-Owen-1.5B	28.9	52.7	83.9	33.8	16.9	954	
DeepSeek-R1-Distill-Owen-7B	55.5	83.3	92.8	49.1	37.6	1189	
DeepSeek-R1-Distill-Owen-14B	69.7	80.0	93.9	59.1	53.1	1481	
DeepSeek-R1-Distill-Owen-32B	<b>72.6</b>	83.3	94.3	62.1	57.2	1691	
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205	
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	<b>94.5</b>	65.2	57.5	1633	

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

31

## 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  $P(y=j|x,z) = g_j(x,z;\theta^*)$ , using the primary data.

Step 2. With the auxiliary data, we construct the estimator  $\hat{\beta}^{AAE}$  as **AI-Augmented Estimator**

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

The graph plots Accuracy (%) on the y-axis (50 to 80) against Cost (\$1K) on the x-axis (0.5 to 3.5). Three series are shown: Traditional (orange line with circles), AAE (blue line with diamonds), and Naive (green line with squares). All series include error bars. The AAE series consistently shows the highest accuracy, especially at higher costs.

**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 LLMs ... **data augmentation approach** that allows us to use the AI-generated data to fit the model. ...  
☆ 保存 ⚡ 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.

32

16