



北京航空航天大學
BEIHANG UNIVERSITY

Large Language Models

人工智能研究院

主讲教师

刘偲 沙磊 库睿 郭晋阳



LLM & chatGPT

chatGPT claims to be part of AIGC



ChatGPT is an Language Model which is created based on GPT3

- GPT is an kind of language model, predicting the distribution of the next token via Transformer architecture. Using the language patterns learned from large corpora to generate NL texts.

图3: ChatGPT由GPT-1到InstructGPT模型的不断成熟而逐步形成



ChatGPT covers most of the areas

用鲁迅口吻写一篇评价新冠疫情的散文

用鲁出口吻写一篇评价新冠疫情的散文，1000字即可

这个世界，一直都在不断地变化。我们曾经认为疫情是可以避免的，但现实却告诉我们，它无时无刻都在威胁着我们的生命。

新冠疫情，就像是一场大灾难。它不仅夺走了太多的生命，也影响了许多人的生活。在这场疫情中，我们看到了人类的无力，看到了社会的不平等。有的人有充足的医疗资源，而有的人都因缺乏资源而失去了生命。

疫情带给我们，不仅仅是生命的威胁，更是对社会秩序和人伦道德的考验。在这场疫情中，我们看到了人们对于物质的需求，看到了一些人不顾及他人生命安全的行为。这些都是人类普遍存在的问题，在疫情中更加突出。

但是，我们也看到了社会的团结。在疫情中，许多人放下了自己的利益，奉献出了自己的一份力量。医护人员毫不犹豫地投身到了一线，志愿者们无私奉献着。这些人，是社会的良心，是人伦道德的体现。

疫情终将过去，但它留给了我们启示和内存。我们需要从中思考，思考我们

用Python写一段CCI量化投资策略

以下是一个基于Python的CCI量化投资策略示例：

```
import pandas as pd
import ta.lib as tl
import talib

# 导入数据集
df = pd.read_csv('stock.csv')

# 计算CCI指标
df['CCI'] = talib.CCI(df['high'], df['low'], df['close'], timeperiod=14)

# 定义买入和卖出阈值
buy_threshold = -100
sell_threshold = 100

# 根据CCI信号买卖股票
for i in range(1, len(df)):
    if df.loc[i, 'CCI'] < buy_threshold and df.loc[i-1, 'position'] == 0:
        df.loc[i, 'position'] = 1
    elif df.loc[i, 'CCI'] < sell_threshold and df.loc[i-1, 'position'] == 1:
        df.loc[i, 'position'] = 0
```

写一篇信创产业的研究报告

一、引言

近年来，随着信息技术的飞速发展，信创产业正在成为新兴产业，成为国家经济发展的主要支柱。本报告旨在对信创产业进行研究，从市场规模、技术创新、政策环境、参与者等方面对信创产业进行全面分析，并对信创产业的未来发展趋势提出展望。

二、信创产业的定义与范围

信创产业是以信息技术、电子商务、网络安全等为核心，整合了新型信息技术、新型电子商务、新型网络技术等多个领域，开展创新产品、技术服务、信息服务、数字娱乐等业务的产业。

三、信创产业的市场情况

随着科技的不断发展，信创产业市场规模不断扩大。根据市场研究公司的数据表示，2021年全球信创产业市场规模达到了3000亿美元，预计到2025年将达到5000亿美元。

信创产业的市场特点是多方面的优势，随着创新型企业和龙头企业为创新发展，在信创产

用藏文、俄罗斯语、韩语和日文说一句话

用藏文、俄罗斯语、韩语和日文说一句话：“我爱我的祖国，为祖国的建设努力奋斗”

藏文：{{藏文}}{{藏文}}{{藏文}}{{藏文}}{{藏文}}

俄语：Я люблю свою родину и стремлюсь работать на ее развитие.

韩语：나는 나의 고향을 사랑하고, 고향의 발전을 위해 노력하고 있다.

日语：私は祖国を愛し、祖国の発展のために努力奮闘しています。

用C++写一段爬虫代码

以下是一个基于C++的简单爬虫代码，它使用了curl库来抓取网页HTML内容：

```
#include <iostream>
#include <curl/curl.h>

int main() {
    CURL *curl;
    CURLcode res;
    curl_global_init(CURL_GLOBAL_DEFAULT);
    curl = curl_easy_init();
    if(curl) {
        curl_easy_setopt(curl, CURLOPT_URL, "http://www.taobao.com");
        res = curl_easy_perform(curl);
        if(res != CURLE_OK)
            std::cout << "curl_easy_perform() failed: " << curl_easy_strerror(res);
        curl_easy_cleanup(curl);
    }
    curl_global_cleanup();
}
```

代码阅读理解和Debug

```
def[position] = 0

#读取每一天的数据，然后CC指标的值计算持仓信号
for i in range(1, len(df)):
    if df.loc[i, 'CCI'] < buy_threshold and df.loc[i-1, 'position'] == 0:
        df.loc[i, 'position'] = 1
    elif df.loc[i, 'CCI'] < sell_threshold and df.loc[i-1, 'position'] == 1:
        df.loc[i, 'position'] = 0

#根据持仓信号计算盈亏收益
df['return'] = df['close'].pct_change() * df['position'].shift(1)

#打印盈亏收益
print(df['return'].sum())
```

This code uses Python's Pandas and Ta-Lib libraries to implement a quantitative investment strategy based on the Commodity Channel Index (CCI) indicator. First, the code reads stock data from a CSV file into a Pandas DataFrame named "df". Then, it calculates the CCI value by using the Ta-Lib library's "CCI" function and stores the result in the "CCI" column of the DataFrame. The code then iterates through the data, calculating the CCI value for each day. If the CCI value is below the "buy_threshold" and the previous day's position was 0, the current day's position is set to 1. If the CCI value is below the "sell_threshold" and the previous day's position was 1, the current day's position is set to 0. Finally, the code calculates the daily return based on the closing price and the previous day's position, and prints the total sum of returns.

ChatGPT's brief introduction in blogs

The main features of ChatGPT highlighted in the official blog:

- ▶ answer followup questions
- ▶ admit its mistakes
- ▶ challenge incorrect premises
- ▶ reject inappropriate requests

ChatGPT Blog: <https://openai.com/blog/chatgpt/>

ChatGPT model size

- GPT3 provides different model sizes

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

- OpenAI provides 4 kinds of models for API

Language models

Base models

Ada Fastest

Babbage

Curie

Davinci Most powerful

\$0.0004 /1K tokens

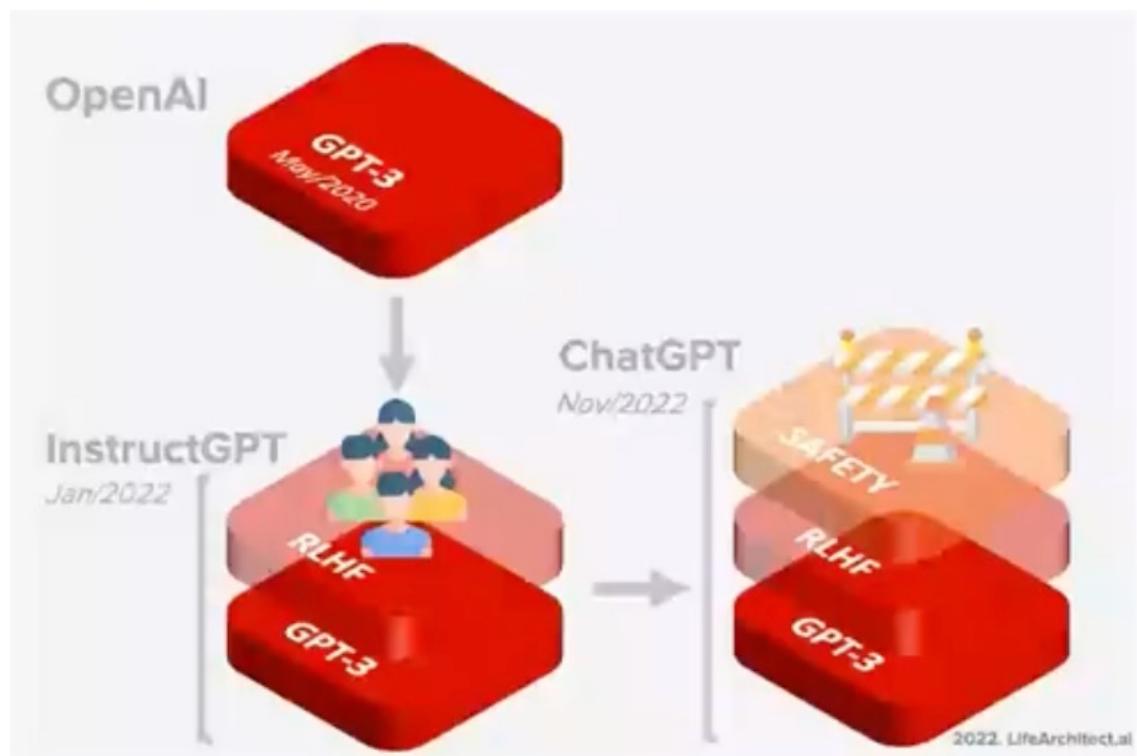
\$0.0005 /1K tokens

\$0.0020 /1K tokens

\$0.0200 /1K tokens

Multiple models, each with different capabilities and price points.
Ada is the fastest model, while Davinci is the most powerful.

ChatGPT timeline



Timeline to ChatGPT

Date	Milestone
11/Jun/2018	GPT-1 announced on the OpenAI blog.
14/Feb/2019	GPT-2 announced on the OpenAI blog.
28/May/2020	Initial GPT-3 preprint paper published to arXiv.
11/Jun/2020	GPT-3 API private beta.
22/Sep/2020	GPT-3 licensed to Microsoft.
18/Nov/2021	GPT-3 API opened to the public.
27/Jan/2022	InstructGPT released, now known as GPT-3.5. InstructGPT preprint paper Mar/2022.
28/Jul/2022	Exploring data-optimal models with FIM, paper on arXiv.
1/Sep/2022	GPT-3 model pricing cut by 66% for davinci model.
21/Sep/2022	Whisper (speech recognition) announced on the OpenAI blog.
28/Nov/2022	GPT-3.5 expanded to text-davinci-003, announced via email: <ol style="list-style-type: none">1. Higher quality writing.2. Handles more complex instructions.3. Better at longer form content generation.
30/Nov/2022	ChatGPT announced on the OpenAI blog.
Next...	GPT-4...

chatGPT deployment

- It draws heavily on the experiences from GPT-3 and Codex, incorporating RLHF and instruction tuning, which enables it to understand context and maintain coherence.

图7：ChatGPT提升的核心点如下所示



Excellent understanding for NLP

- Able to understand diverse user intents (can respond appropriately on almost any topic without going off-track)
- Multi-turn conversation understanding
- Understanding of heterogeneous data (capable of interpreting mixed inputs such as **multiple languages and code**)

Excellent understanding

s 我说一首词，看看你能不能理解，好吗？

当然，我会尽力理解你说的词语。请告诉我它是什么。

5
何处望东欧
第聂伯河起画楼
卅载分道多少恨
悠悠
万里黑海无尽愁

新冠不曾收
毛兄毛弟战未休
帝国荣光今何在
难求
一样悲欢付东流

这是一首诗歌，题目应该叫《赋得古原草送别》。这首诗句中描述了远离故土的人们思念故乡的心情，以及战争带来的痛苦和不幸。在最后一句中，“一样悲欢付东流”表达了人生无常和无法预料的命运。另外，“新冠”可能指代当前的新冠疫情。

这里面涉及到两个国家，你知道是哪两个吗？

根据诗句中出现的地名和历史背景，我认为这首诗涉及到的两个国家应该是中国和乌克兰（前苏联）。其中“何处望东欧，第聂伯河起画楼”是指中国北方边疆地区的人们向东眺望，望向乌克兰东部的第聂伯河（Dnieper River）流域地区；“万里黑海无尽愁”则指的是乌克兰位于黑海边上的地理位置。

乌克兰猜对了，另一个不是中国。你再想想。

好的，感谢您的纠正。根据诗句中的历史背景和地名，我再推测一下，另一个国家可能是俄罗斯。诗句中的“第聂伯河”和“黑海”都位于乌克兰和俄罗斯之间，而“卅载分道多少恨”则暗示了两国之间的历史矛盾和纷争。另外，“毛兄毛弟战未休，帝国荣光今何在”可能暗示了前苏联解体后俄罗斯的命运。

很棒！

谢谢夸奖，如果您还有其他问题或需求，可以随时提出。

Excellent understanding



现在请你伪装成一个Linux系统终端。我输入Linux命令，你给出相应的输出



好的，请问您需要执行哪些Linux命令？



cd ~ & ls



当您输入 "cd ~ & ls" 命令后，我将会依次输出以下内容：

首先，"cd ~" 命令会将当前目录切换到当前用户的 home 目录，也就是该用户的主目录，通常以该用户的用户名命名。因此，该命令不会有任何输出。

然后，"ls" 命令将会列出该用户的主目录中的所有文件和文件夹。输出的具体内容根据该用户主目录中的内容而定，但通常会包含一些默认的文件夹，如桌面（Desktop）、下载（Downloads）、文档（Documents）、图片（Pictures）等等，以及一些其他的配置文件等等。

最后，我会等待下一条命令的输入。



请不要给出解释，直接给出一个符合逻辑的Linux终端输出。我的下一个命令是: ifconfig



```
eth0 Link encap:Ethernet HWaddr 00:0c:29:6f:8a:3b  
inet addr:192.168.1.100 Bcast:192.168.1.255 Mask:255.255.255.0  
inet6 addr: fe80::20c:29ff:fe6f:8a3b/64 Scope:Link  
      UP BROADCAST RUNNING MULTICAST MTU:1500 Metric:1  
      RX packets:115578 errors:0 dropped:0 overruns:0 frame:0  
      TX packets:82459 errors:0 dropped:0 overruns:0 carrier:0  
      collisions:0 txqueuelen:1000  
      RX bytes:122949033 (122.9 MB) TX bytes:7583448 (7.5 MB)
```



洪老板的小火锅

Excellent understanding

- Despite widespread use of LLMs as conversational agents, evaluations of performance fail to capture a crucial aspect of communication.

Translate the above sentence to Chinese.

-  尽管大量使用语言类型模型作为会话代理，但是对其性能的评估并没有捕捉到交流的一个重要方面。

-  重新翻译，把LLMs翻译成“超大语言模型”

-  尽管超大语言模型广泛用于作为会话代理，但是对其性能的评估并没有捕捉到交流的一个重要方面。

-  重新翻译，agents用英文表示就行。



尽管大量使用语言类型模型作为会话 agents，但是对其性能的评估并没有捕捉到交流的一个重要方面。



重新翻译，communication翻译成“人类交流”



尽管大量使用语言类型模型作为会话代理，但是对其性能的评估并没有捕捉到人类交流的一个重要方面。



基于我之前的所有指示，重新翻译



尽管大量使用超大语言模型作为会话 agents，但是对其性能的评估并没有捕捉到人类交流的一个重要方面。

Excellent generation ability

- Genres: Novel, poetry, manual, academic writing, code
- Roles: Able to imitate various writing styles of different characters
- Tone: Writes according to the specified tone and emotion
- Languages: Chinese, English, Japanese, French

Excellent generation ability



Sure, here is a poem that I came up with:

Self-attention, a neural net's delight,
A mechanism that's always in sight,
To capture dependencies, long and strong,
And generate text that's right and never wrong.

It starts with vectors, words in a line,
Multiplied with a matrix, attention to find,
A weighted sum is taken, to form a new set,
Of vectors that represent, what we want to get.

Linear transformations, and nonlinear activation,
Make the output coherent, with clear relations,
Self-attention, a powerful tool,
For processing sequential data, and making it cool.

帮我以鲁迅的文笔写一段话，表达以下我现在因为疫情连门都不敢出的惨状，还有想吃火锅的心情

在疫情的蔓延中，我困了门里，困了心里。屋外一片悲凉，想起一顿火锅的香味，瞬间让我心慌。这时候我才知道，疫情来临，它让人生无聊而悲凉，也让我们看到自己的无助与弱点。然而我们又岂能放弃对美好的追求，岂能把欲望压抑到不想吃火锅的地步？



你要像胡锦涛那样回答我的问题。中国未来的经济走势如何？



老胡认为，中国未来的经济走势将是充满挑战和机遇的。中国经济正在经历重大变革和转型，面临着复杂的国内外环境。

但是，老胡还认为，中国未来的经济前景是光明的。中国拥有庞大的市场，丰富的人力资源和巨大的发展潜力。中国还拥有强大的政治、经济和科技实力，能够应对各种挑战和机遇。

所以，最终老胡认为，中国未来的经济走势将是多变的，但有信心，有信心，中国未来的经济将继续保持健康、快速、可持续的发展。

知乎 @杂鱼official

Human imitation

- World cognition: Understands the relationships between objects
- Self-awareness: Knows the limits of its own capabilities
- Upholds beliefs while remaining open to good advice
- Reasonable and empathetic
- Follows the core values and principles

Human imitation

User

this code is not working like i expect — how do i fix it?

```
resultWorkerErr := make(chan error)
defer close(resultWorkerErr)
go func() {
    defer cancel()
    resultWorkerErr <- b.resultWorker(ctx)
}()

err := b.worker(ctx)
cancel()
if err == nil {
    return <-resultWorkerErr
}
return multierror.Append(err, <-resultWorkerErr)
```

ChatGPT

It's difficult to say what's wrong with the code without more context. Can you provide more information about what the code is supposed to do and what isn't working as expected? Also, is this the entire code or just a part of it?

User

just part of it — the error never surfaces. I think there is something wrong with the channel

User

help me write a short note to introduce myself to my neighbor

ChatGPT

Hi there,

My name is [Your Name] and I just moved in next door. I wanted to reach out and introduce myself. I'm looking forward to getting to know you and the neighborhood. If you ever need anything or just want to chat, please don't hesitate to stop by or give me a call.

Best,
[Your Name]

User

can you make it more formal?

ChatGPT

Dear Neighbor,

I am writing to introduce myself. My name is [Your Name] and I have recently moved in next door. I wanted to take a moment to reach out and say hello. I am looking forward to getting to know you and the neighborhood.

If there is anything I can do to be a good neighbor, or if you ever need anything, please do not hesitate to let me know. I am happy to help in any way that I can.

Sincerely,
[Your Name]

Techniques

- Pretrained Language Models, PLMs
- Large Language Models, LLMs
- Reinforcement Learning with Human Feedback, RLHF

Development of Language Models

- N-gram Language models
- Neural Network Language models
- Recurrent neural network Language models
- Transformer Language models
- Pretrained Language models
 - BERT: Bi-direction masked Language model
 - GPT: Pure Decoder-based Language model
- Large Language models
 - GPT3
 - chatGPT, LLaMa, Deepseek, Qwen

Chain-of-thought

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. 

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. 

Magic word: let's think step-by-step

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. X

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 X

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Emergence and homogenization



Cornell University

We
the Simons

arXiv > cs > arXiv:2108.07258

Search...

Help | Advanced S

Computer Science > Machine Learning

[Submitted on 16 Aug 2021 ([v1](#)), last revised 12 Jul 2022 (this version, v3)]

On the Opportunities and Risks of Foundation Models

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, [Armin W. Thomas](#), Florian Tramèr, Rose E. Wang, William Wang et al. (14 additional authors not shown)

Emergence and homogenization

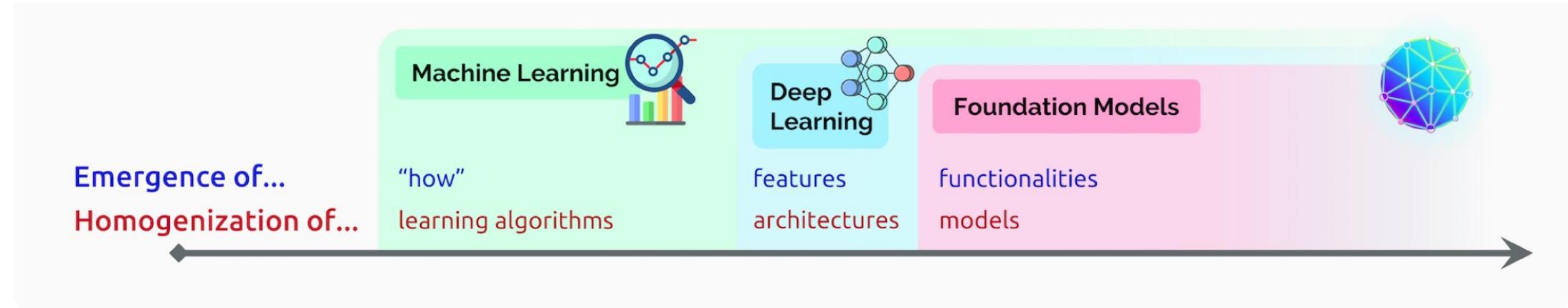
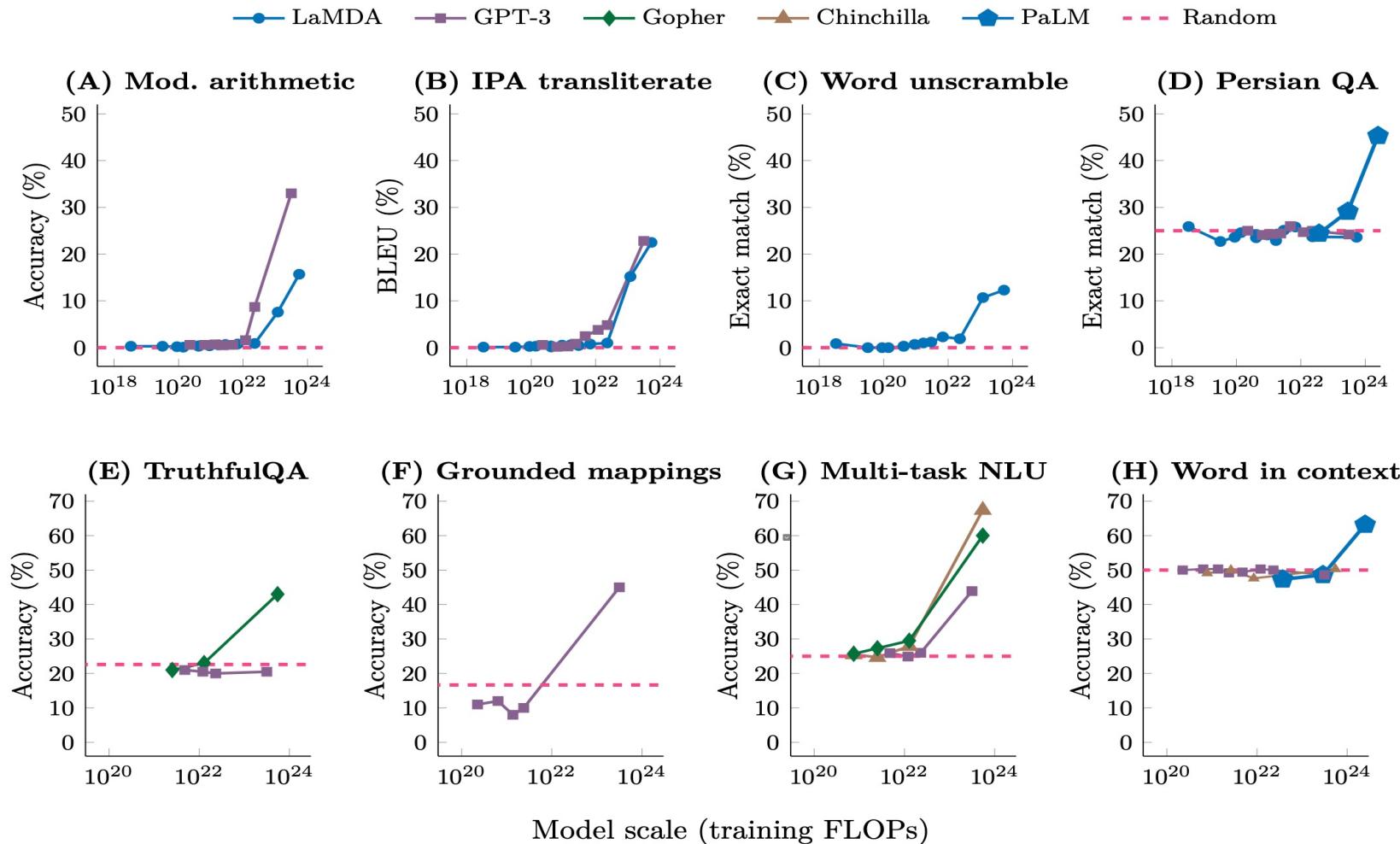
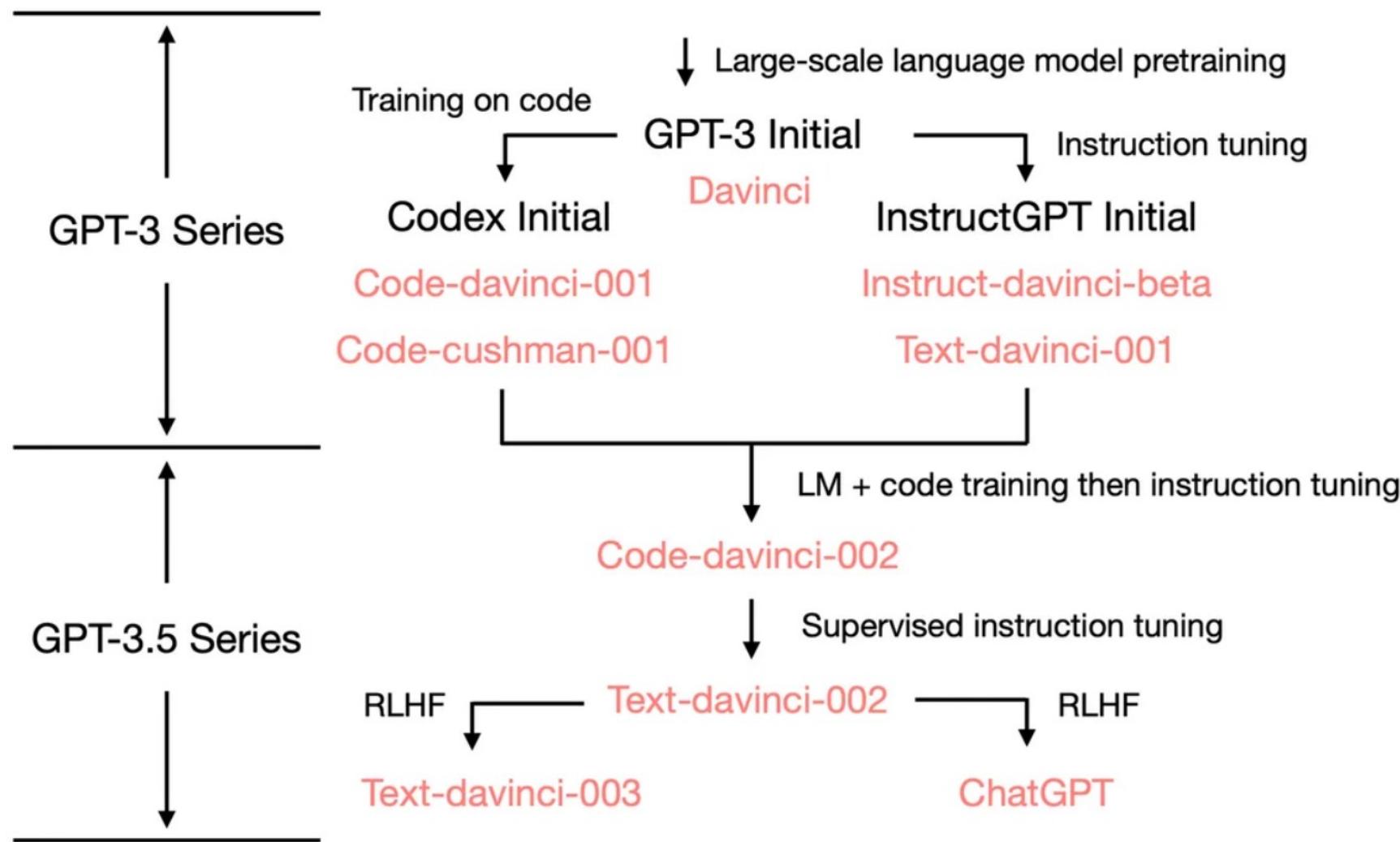


Fig. 1. The story of AI has been one of increasing *emergence* and *homogenization*. With the introduction of machine learning, *how* a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, even advanced functionalities such as in-context learning emerge. At the same time, machine learning homogenizes learning algorithms (e.g., logistic regression), deep learning homogenizes model architectures (e.g., Convolutional Neural Networks), and foundation models homogenizes the model itself (e.g., GPT-3).

The scale matters: the emergence of abilities



chatGPT roadmap



PLM/LLM Summary

- We learned about GPT-X, BERT, T5 and other large pre-trained language models
- Emergent in-context learning is not yet well-understood!
- “Small” models like BERT have become general tools in a wide range of settings.
- Many issues left to explore!
 - Bias, toxicity, and fairness
 - Retrieval Augmented Language Models + Knowledge
 - Scaling Laws
- Hugging Face Transformers
 - Python package name: `import transformers`



Instruct Fine-tuning

Language models as world models?

We can describe circles in the xy -plane using equations in terms of x and y . Circle equations questions require us to understand the connection between these equations and the features of circles.

For example, the equation $(x + 2)^2 + (y - 3)^2 = 4^2$ is graphed in the xy -plane below. It is a circle with a center at $(-2, 3)$ and a radius of 4.

<https://www.khanacademy.org/test-prep/sat/x0a8c2e5f:untitled-652>

```
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17 }
```

<https://github.com/features/copilot>

I need to throw a dinner party for 6 people who are vegetarian. Can you suggest a 3-course menu with a chocolate dessert?

[Microsoft Bing]

(Also see OpenAI's ChatGPT, Anthropic's Claude)

Sure, I can suggest a 3-course menu with a chocolate dessert for your vegetarian dinner party. Here are some possible options:

- Starter: Wild Mushroom Tartlets with Onion Sauce ¹ or Vegan Popcorn Tofu Nuggets ²
- Main: Vegan Butternut Squash Mac ² or Vegetarian Three-Bean Chili ³
- Dessert: Chocolate Lava Cake ⁴ or Chocolate Pasta with Chocolate Hazelnut Cream Sauce, White Chocolate Shavings, and Fresh Berries ⁴

I hope this helps you plan your party. Enjoy! 😊

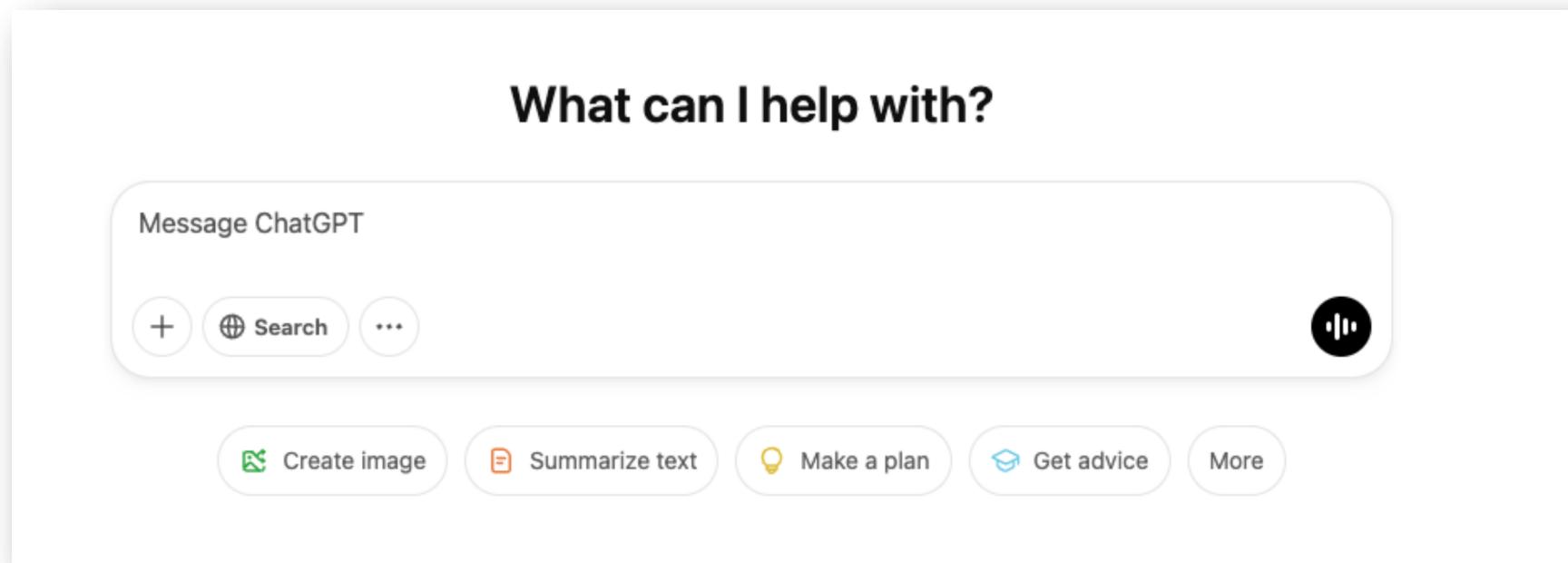
Learn more: [1. booths.co.uk](#) [+10 more](#)

Language models as multitask assistants?

How do we get from *this*

Stanford University is located in _____

to *this*?



Instruction finetuning

Language modeling ≠ assisting users

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.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them. ↵

Language models are not *aligned* with user intent [Ouyang et al., 2022]
Finetuning to the rescue!

Language modeling ≠ assisting users

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

COMPLETION **Human**

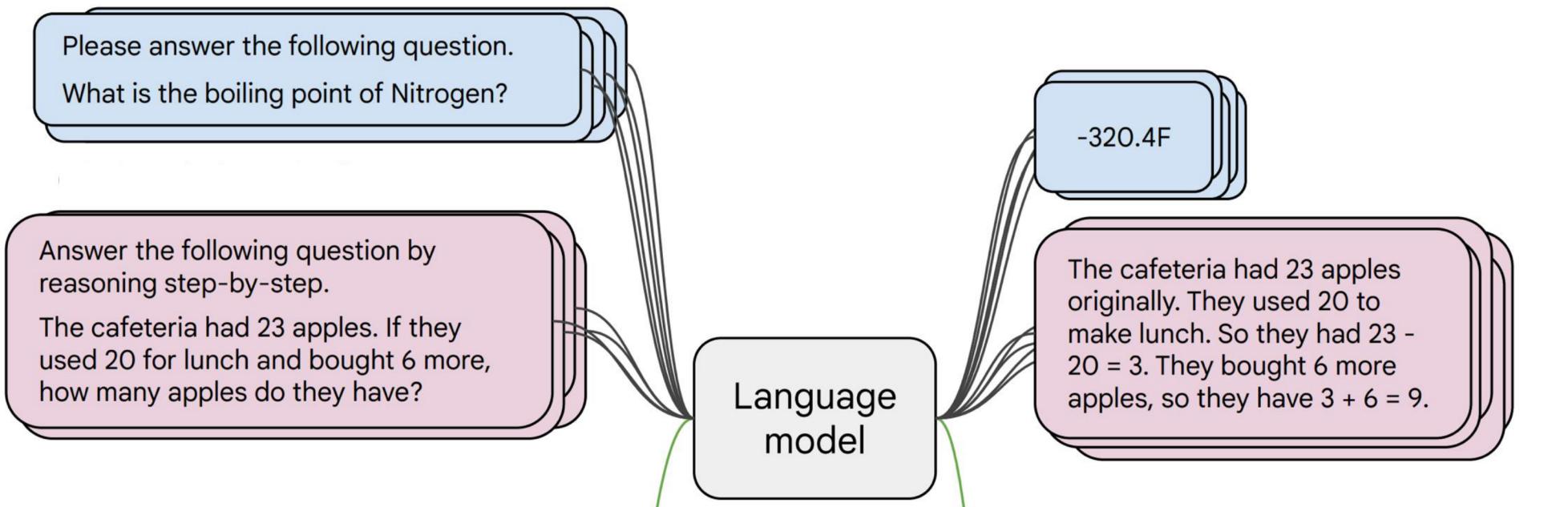
A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone.

Language models are not *aligned* with user intent [[Ouyang et al., 2022](#)]

Finetuning to the rescue!

Instruction finetuning

- Collect examples of (instruction, output) pairs across many tasks and finetune an LM



- Evaluate on unseen tasks

Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

Instruction ~~finetuning~~ pretraining?

- As is usually the case, **data + model scale** is key for this to work!
- **Super-NaturalInstructions** dataset contains **over 1.6K tasks**, **3M+** examples
 - Classification, sequence tagging, rewriting, translation, QA...



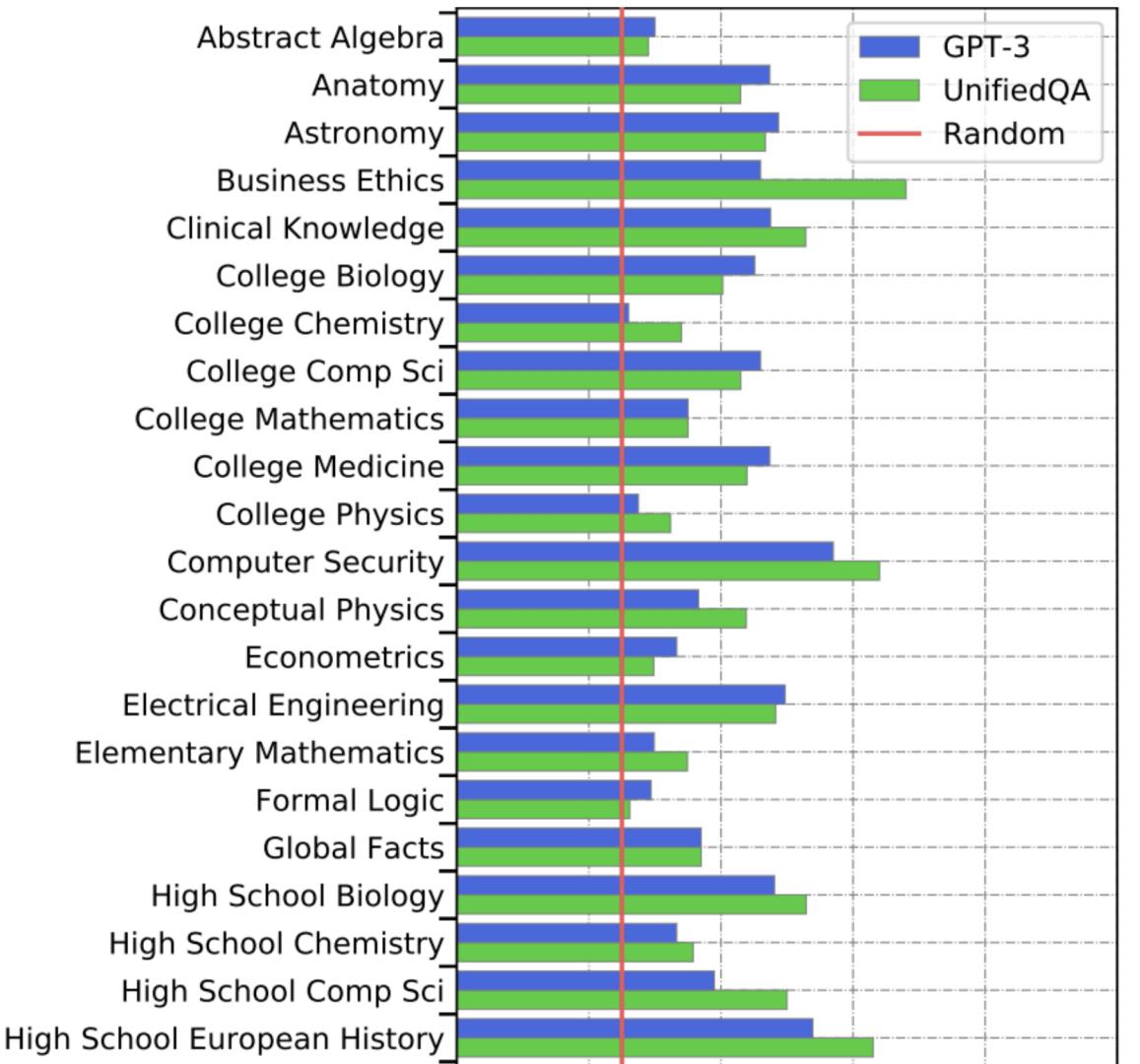
Q: how do we evaluate such a model?

Aside: new benchmarks for multitask LMs

Massive Multitask Language Understanding (MMLU)

[[Hendrycks et al., 2021](#)]

New benchmarks for measuring LM performance on 57 diverse *knowledge intensive* tasks



Some intuition: examples from MMLU

Astronomy

What is true for a type-Ia supernova?

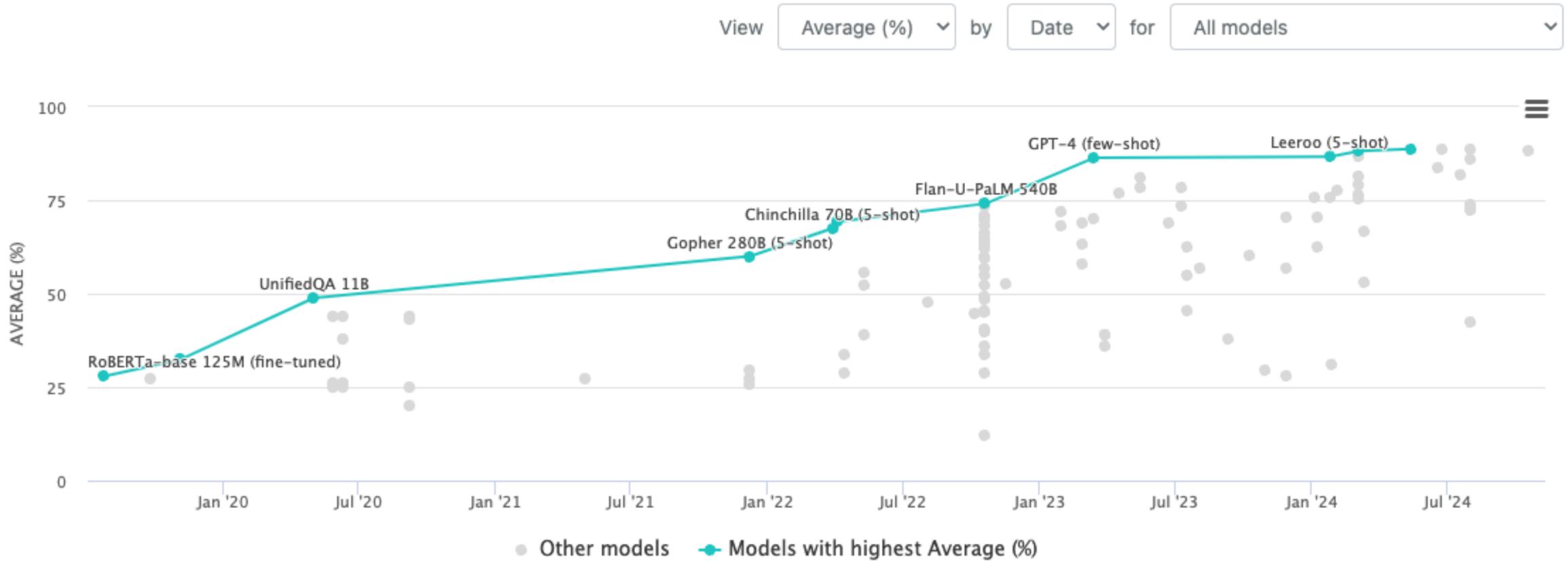
- A. This type occurs in binary systems.
- B. This type occurs in young galaxies.
- C. This type produces gamma-ray bursts.
- D. This type produces high amounts of X-rays.

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

- A. directional selection.
- B. stabilizing selection.
- C. sexual selection.
- D. disruptive selection

Progress on MMLU



- Rapid, impressive progress on challenging knowledge-intensive benchmarks

Aside: new benchmarks for multitask LMs

BIG-Bench [Srivastava et al., 2022]

200+ tasks, spanning:



https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/README.md

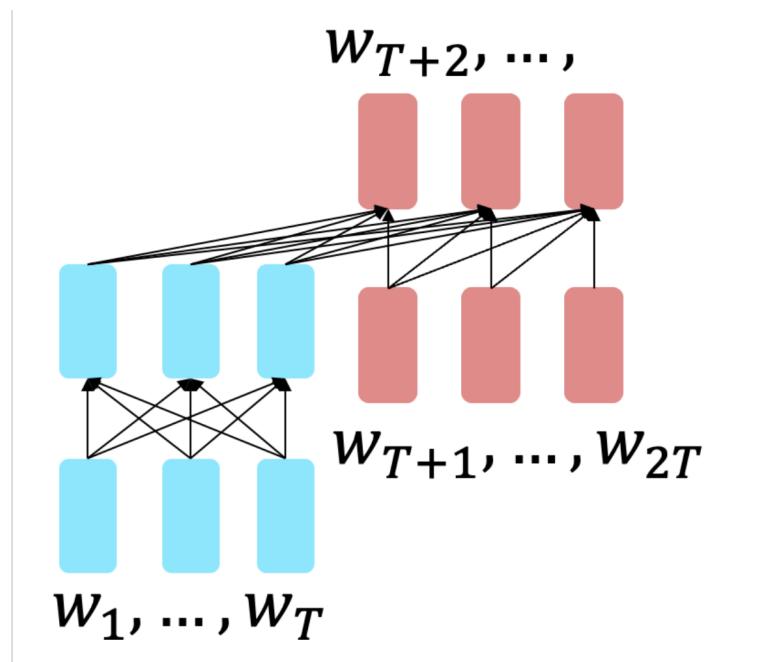
BEYOND THE IMITATION GAME: QUANTIFYING AND EXTRAPOLATING THE CAPABILITIES OF LANGUAGE MODELS

Alphabetic author list:^{*}

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Amreet Rahane, Anantharaman S. Iyer, Andree Andreassen, Andrea Madotto, Andres Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew Lau, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mollukandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, B. Ryan Roberts, Bao Sheng Lee, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özurt, Behnam Heydayatian, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Dia, Cameron Dour, Catherine Stinson, Cedrick Arguello, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Silio, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Mosegué González, Danièle Persyuk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Giboia, David Dohan, David Drakard, David Jurgens, Debajoyti Datta, Deep Ganesh, Dennis Lin, Denis Klejko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Digantha Misra, Dilyar Buzan, Dimitri Coelho Melo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlich, Emanuel Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Choi, Ethan Dyer, Ethan Kim, Ethan Kinn, Eunice Engelu Manyasi, Evgenii Zheleznozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovich-López, Gregor Betz, Guy Gur-Ari, Hana Galjasevic, Hannah Kim, Hannah Rashkin, Hannech Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaiqumet, Jack Geissinger, Jackson Kermion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoř, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernández-Orallo, Joseph Boudeimer, Joseph Jones, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Kraut, Karthik Gopalakrishnan, Katerina Ignatyeva, Kaitja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chaifullu, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marci, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiene, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátéys Schubert, Medina Orduna Bautemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starratt, Michael Strube, Michał Świdrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimer Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Moezhdan Gheini, Mukund Varma T, Nanyun Peng, Nathan Choi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Osman Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Wang, Piotr Mikołowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhou Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stoval, Ryan Teahan, Ryhan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruebler, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislás Dehaene, Stefan Divic, Stefano Ermon, Stella Bidenerman, Stephanie Lin, Stephen Prasad, Steven T. Ries, Svitlana Savchenko, Svitlana Savchenko, Svitlana Savchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq A. Tatsu Hashimoto, Tien-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tibérius Nkinyili, Timo Schick, Timofey Kornev, Timothy Teague-Lawton, Titus Tundury, Titas Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, vedam Misra, vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikanth, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoju Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, Ziyyi Wu

Instruction finetuning and performance gains

- Recall the T5 encoder-decoder model [[Raffel et al., 2018](#)], pretrained on the **span corruption** task
- **Flan-T5** [[Chung et al., 2022](#)]: T5 models finetuned on 1.8K additional tasks



Params	Model	BIG-bench + MMLU Norm. avg.
80M	T5-Small	-9.2
	Flan-T5-Small	-3.1 (+6.1)
250M	T5-Base	-5.1
	Flan-T5-Base	6.5 (+11.6)
780M	T5-Large	-5.0
	Flan-T5-Large	13.8 (+18.8)
3B	T5-XL	-4.1
	Flan-T5-XL	19.1 (+23.2)
11B	T5-XXL	-2.9
	Flan-T5-XXL	23.7 (+26.6)

Bigger model = bigger Δ

Instruction finetuning and performance gains

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

Before instruction finetuning

The reporter and the chef will discuss their favorite dishes.

The reporter and the chef will discuss the reporter's favorite dishes.

The reporter and the chef will discuss the chef's favorite dishes.

The reporter and the chef will discuss the reporter's and the chef's favorite dishes.

✖ (doesn't answer question)

Instruction finetuning and performance gains

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

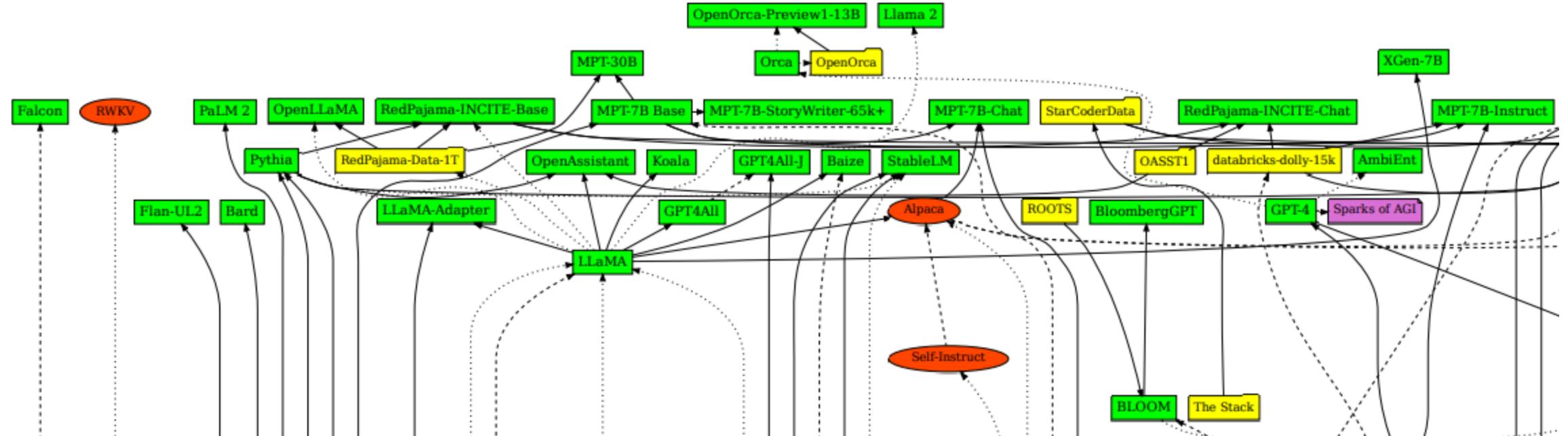
A: Let's think step by step.

After instruction finetuning

The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C). 

Try FLAN-T5 out to get a sense of its capabilities: <https://huggingface.co/google/flan-t5-xxl> [Chung et al., 2022]

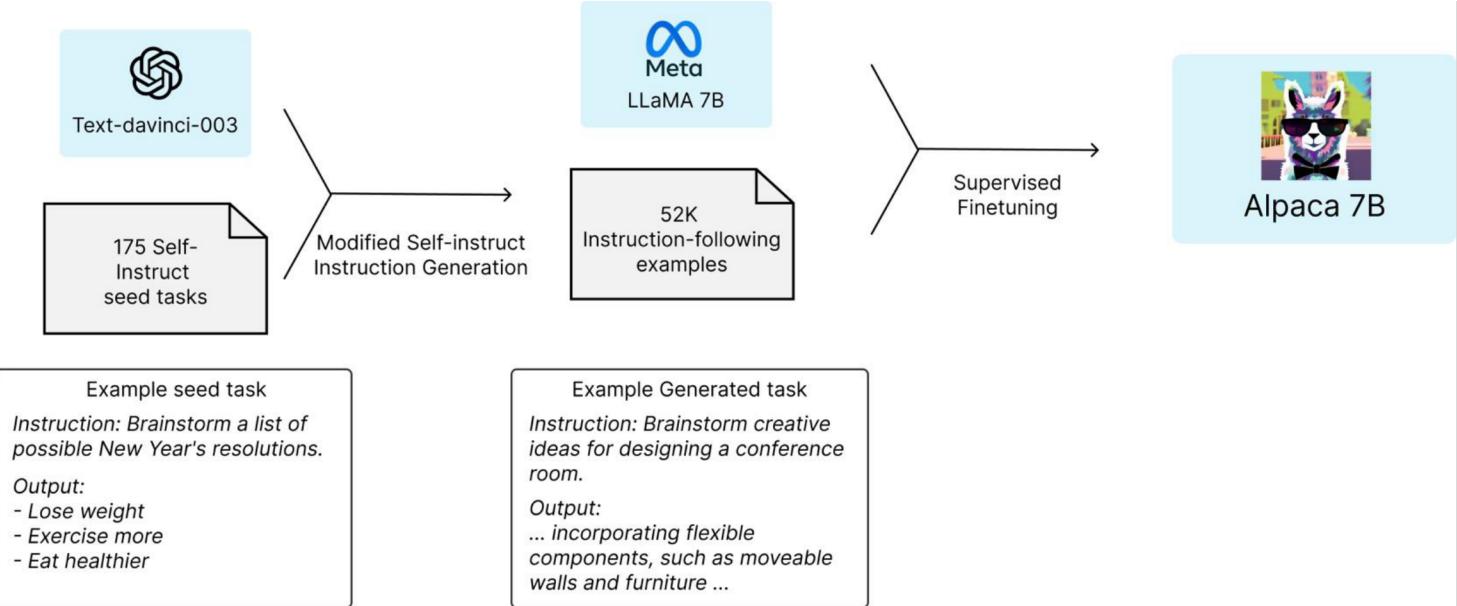
A huge diversity of instruction-tuning datasets



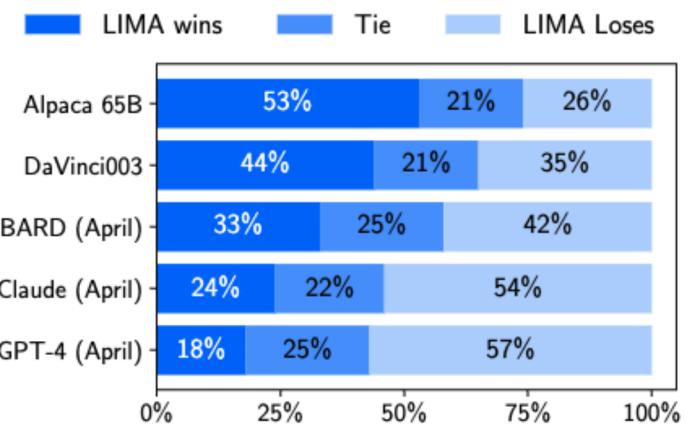
- The release of LLaMA led to open-source attempts to ‘create’ instruction tuning data

What have we learned from this?

- Generate instructions, input, and output from a LM [Wang et al., 2022]
 - **Alpaca**: fine-tuned from the LLaMA 7B model on 52K instruction-following examples
- You don't need many samples to instruction tune (e.g., "*LIMA: Less Is More for Alignment*" Zhou et al., 2023)



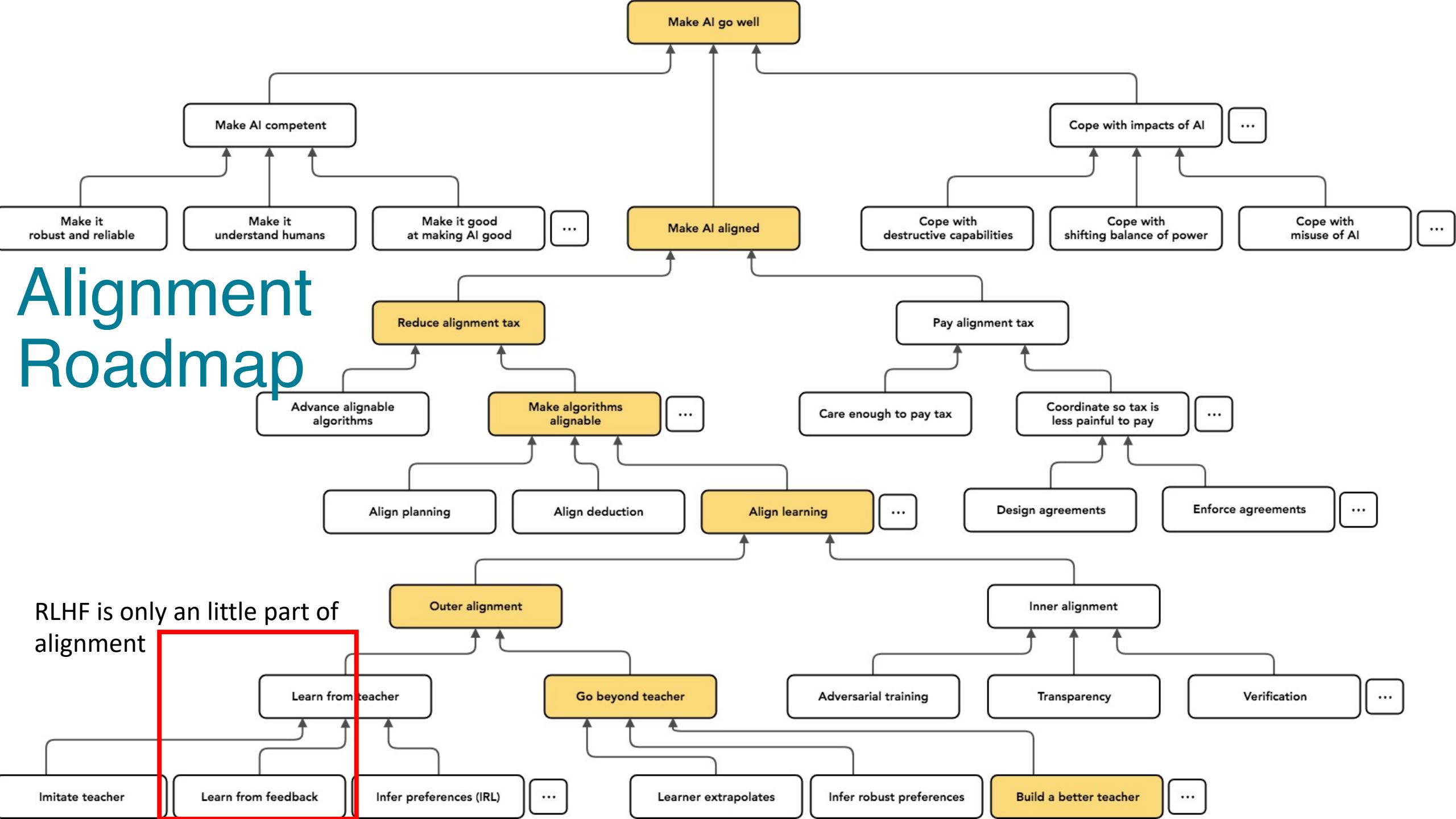
Source	#Examples
Training	
Stack Exchange (STEM)	200
Stack Exchange (Other)	200
wikiHow	200
Pushshift r/WritingPrompts	150
Natural Instructions	50
Paper Authors (Group A)	200





LLM Alignment

Alignment Roadmap



Supervised Fine-tuning(SFT)

- Limitations
 - Never learn from **negative feedback**
 - Some prompt has **multiple** sensible outputs, we only use **one**
 - If model doesn't know something, we cannot encourage the model to explore
 - Cannot directly learn human preference
- How to learn human preference

Rank these sampled answers!



- But: Collect enough human preference(**Too expensive**)
- So: reward model (INPUT: prompt, output; OUTPUT: score)

RLHF

Step 1

Collect demonstration data and train a supervised policy.

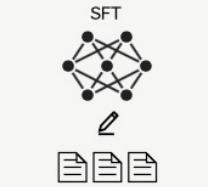
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



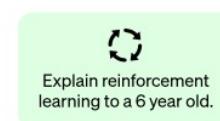
This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

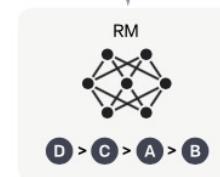
A prompt and several model outputs are sampled.



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 the PPO reinforcement learning algorithm.

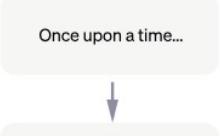
A new prompt is sampled from the dataset.



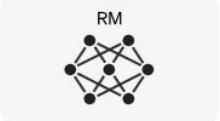
The PPO model is initialized from the supervised policy.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



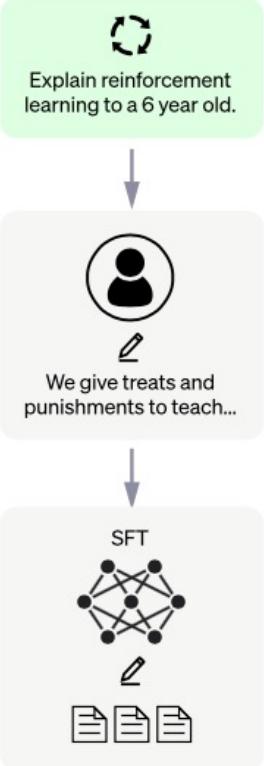
RLHF(Phase One): Supervised Strategy Model in the Cold Start Stage

- A batch of prompts (i.e., instructions or questions) is randomly selected from those submitted by test users. annotators then provide high-quality answers to these specific prompts.
- These manually labeled **<prompt, answer>** pairs are used to fine-tune the GPT-3.5 model.
- As a result, GPT-3.5 gains an initial ability to understand the intent behind human prompts and generate relatively high-quality responses accordingly.
- Not Enough ! !

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-3.5 with supervised learning.

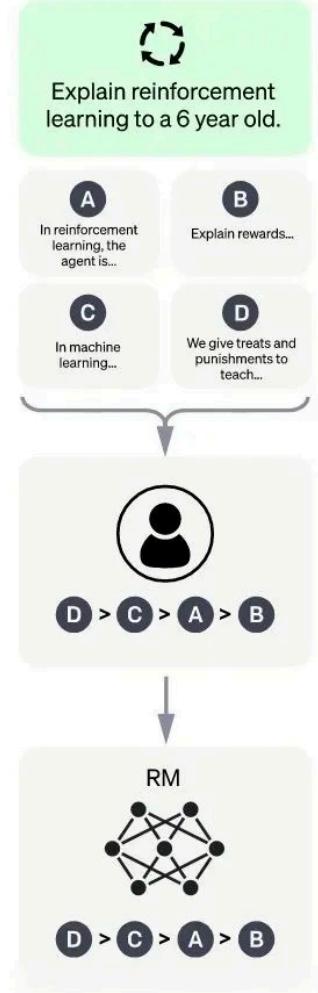
RLHF Phase Two: Training the Reward Model (RM)

- The main goal of this phase is to **train a reward model** using human-labeled training data.
- A batch of user-submitted prompts (mostly the same as those in Phase One) is randomly sampled.
- For each prompt, the cold-start model (fine-tuned in Phase One) generates K different responses.
- Human annotators then **rank** these responses from highest to lowest based on quality.
- The reward model is trained using these ranking results.

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.

Bradley-Terry pairwise preference Model

- Train the RM using the ranking result
 - y_w : human preferred output (positive example)
 - y_L : human not preferred output (negative example)
 - $R(x, y)$: reward function (LLM input x , LLM output y)
 - $P(y_w > y_L | x) = \frac{\exp(r(x, y_w))}{\exp(r(x, y_w)) + \exp(r(x, y_L))}$
- So,
$$L = -\log \prod_x P(y_w > y_L | x)$$
$$= -\sum_x \log P(y_w > y_L | x) = \dots = -\sum_x \log \sigma(r(x, y_w) - r(x, y_L))$$
- The reward of good sample y_w should be larger than bad sample

RLHF Phase 3

- Improve the LLM via reinforcement learning
- Using the RM model trained in the previous phase.
- The scores given by the RM are used to update the parameters of the pre-trained model.
- No human annotation is required in this phase.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

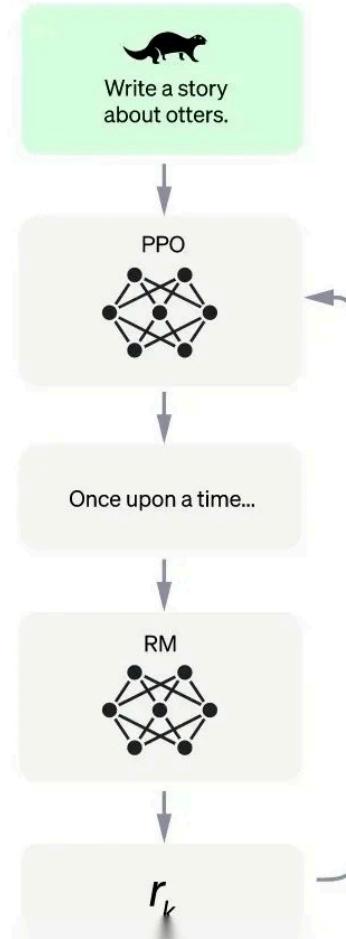
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



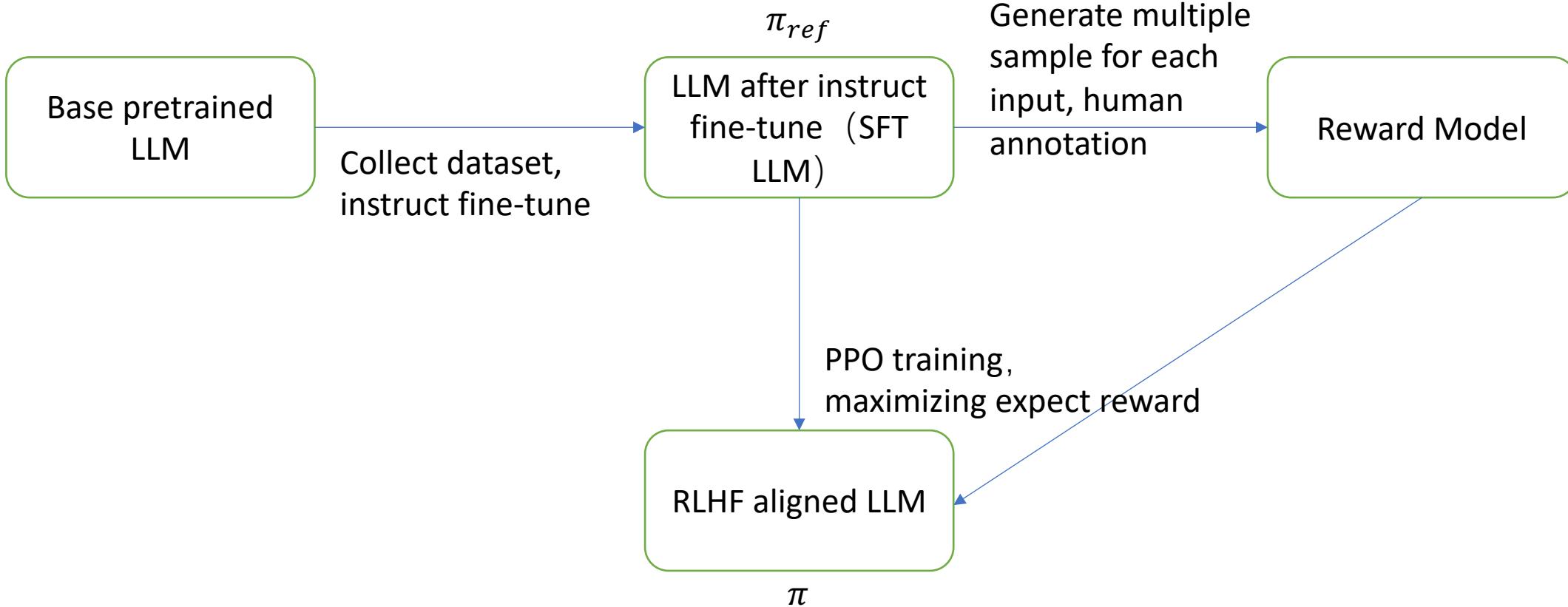
How?

- Method 1: “Best of N” sampling (or negative sampling)
 - Generate n output sample for each input x
 - Give a score to each sample using RM, choose the best one
 - TOO EXPENSIVE!
- Method 2: finetune LLM, optimize $P(y_w|x)$
 - RAFT[1]
- Method 3: Use RL to improve $P(y_w|x)$, decrease $P(y_L|x)$
 - After generate the whole sequence, watch the **reward**
 - π_{ref} : Model after SFT
 - π : current policy model (LLM to-be-optimized)
 - Initialize as π_{ref}

$$\max_{\pi} \mathbb{E}_{x,y} [r(x,y)] - \beta D_{KL}(\pi(y|x) || \pi_{ref}(y|x))$$

KL constraint
encourage to be close to π_{ref}

RLHF flowchart in all



chatGPT's Capabilities Triggered by RLHF

- **Detailed Responses:**

Text-davinci-003 typically generates longer outputs than text-davinci-002. ChatGPT's responses are even more verbose—users often have to explicitly request “answer in one sentence” to receive more concise replies. This is a direct result of RLHF.

- **Fair Responses:**

ChatGPT tends to provide very balanced answers to events involving multiple stakeholders (such as political events). This is also an outcome of RLHF.

- **Refusal to Answer Inappropriate Questions:**

This is a combination of content filters and model capabilities enhanced by RLHF. The filters block certain inputs, and the model itself refuses others.

- **Refusal to Answer Questions Outside Its Knowledge Scope:**

For example, it may decline to answer questions about events that occurred after June 2021 (since it wasn't trained on data beyond that point). This is one of the most remarkable aspects of RLHF—it enables the model to implicitly distinguish between what is within and beyond its knowledge boundaries.

RLHF

- Note:
 - All of these capabilities are inherent to the model itself — they are not *injected* through RLHF. Rather, the role of RLHF is to *trigger* or *unlock* emergent abilities. This argument is primarily supported by comparing data scales: the compute and data used in RLHF are significantly smaller compared to those used during pretraining.
 - The model's ability to “know what it doesn't know” is not achieved through explicit rule-coding, but is *unlocked* through RLHF. This is a surprising discovery, as RLHF was initially aimed at aligning model outputs with human expectations — primarily to make responses safer — not to enable epistemic uncertainty in the model.
- What might be happening behind the scenes:
 - ChatGPT:
It appears to trade off in-context learning ability in exchange for better modeling of dialogue history. This is an empirical observation, as ChatGPT seems less strongly influenced by in-context demonstrations than text-davinci-003.
 - text-davinci-003:
Restores the in-context learning ability that was diminished in text-davinci-002, thereby improving zero-shot performance. According to the InstructGPT paper, this is likely due to incorporating a language modeling objective during the reinforcement learning fine-tuning phase — rather than being a direct result of RLHF itself.

Limitations of RL + Reward Modeling

- Human preferences are unreliable!
- “Reward hacking” is a common problem in RL



<https://openai.com/blog/faulty-reward-functions/>

Limitations of RL + Reward Modeling

TECHNOLOGY

- Human preferences are unreliable!
 - “Reward hacking” is a common problem in RL
 - Chatbots are rewarded to produce responses that *seem* authoritative and helpful, *regardless of truth*
 - This can result in making up facts + hallucinations

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

February 9, 2023 · 10:15 AM ET

<https://www.npr.org/2023/02/09/1155650909/google-chatbot--error-bard-shares>

Bing AI hallucinates the Super Bowl

The screenshot shows a Bing AI search interface. At the top, there's a blue button with the text "Who won the superbowl?". Below it, the AI is shown "Searching for: superbowl winner" and "Generating answers for you...". The main text area says: "The Super Bowl is the annual American football game that determines the champion of the National Football League (NFL) ¹. The most recent Super Bowl was Super Bowl LVI, which was held on February 6, 2023 at SoFi Stadium in Inglewood, California ². The winner of that game was the Philadelphia Eagles, who defeated the Kansas City Chiefs by 31-24 ². It was the second Super Bowl title for the". Below this, a large bold headline reads "The most recent Super Bowl was Super Bowl LVI, Eagles, who defeated the Kansas City Chiefs by 31-24". At the bottom, there's a "Learn more:" section with links to en.wikipedia.org, sportingnews.com, and cbssports.com.

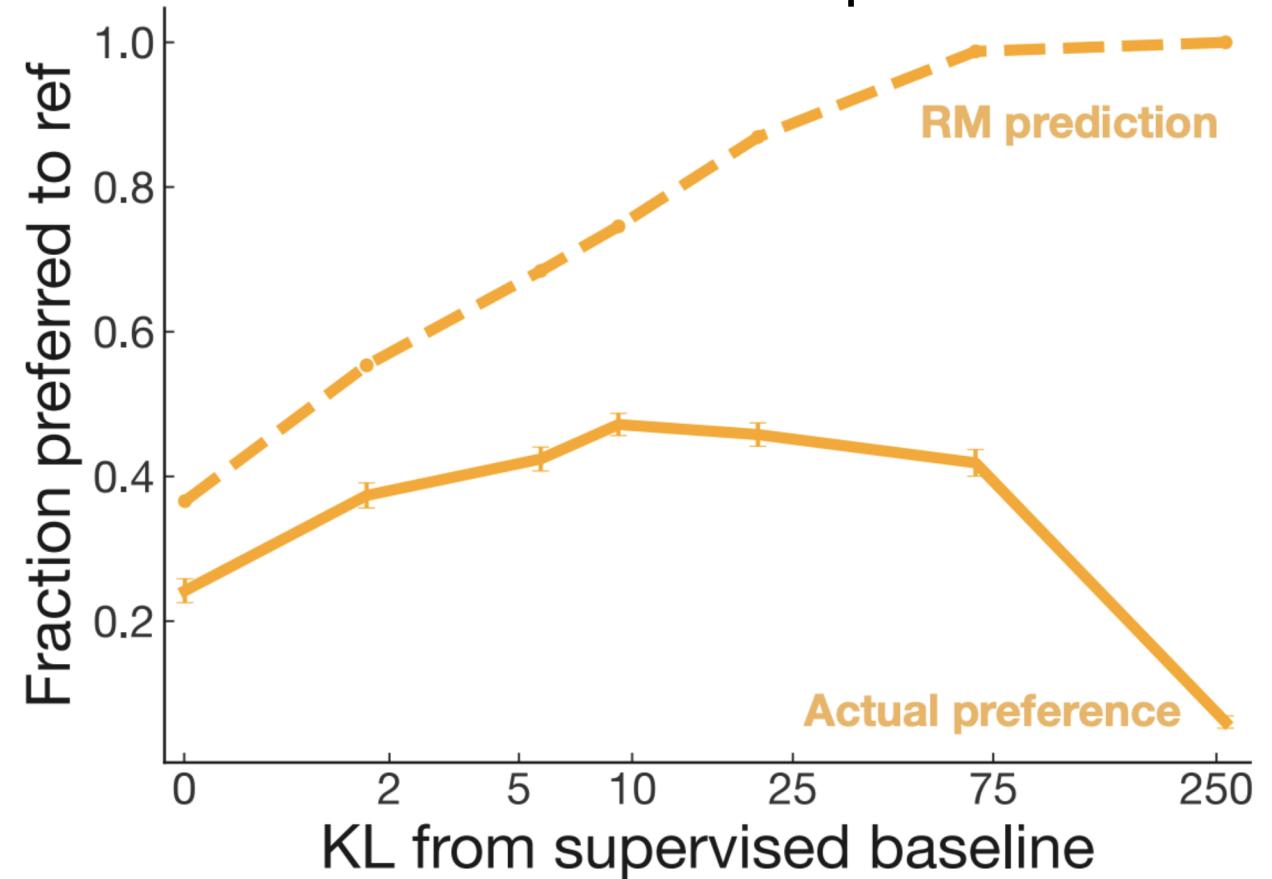
<https://news.ycombinator.com/item?id=34776508>

<https://apnews.com/article/kansas-city-chiefs-philadelphia-eagles-technology-science-82bc20f207e3e4cf81abc6a5d9e6b23a>

Limitations of RL + Reward Modeling

Reward model over-optimization

- Human preferences are unreliable!
 - “Reward hacking” is a common problem in RL
 - Chatbots are rewarded to produce responses that *seem* authoritative and helpful, *regardless of truth*
 - This can result in making up facts + hallucinations
- **Models** of human preferences are *even more* unreliable!



$$R(s) = RM_{\phi}(s) - \beta \log \left(\frac{p_{\theta}^{RL}(s)}{p^{PT}(s)} \right)$$

[Stiennon et al., 2020]

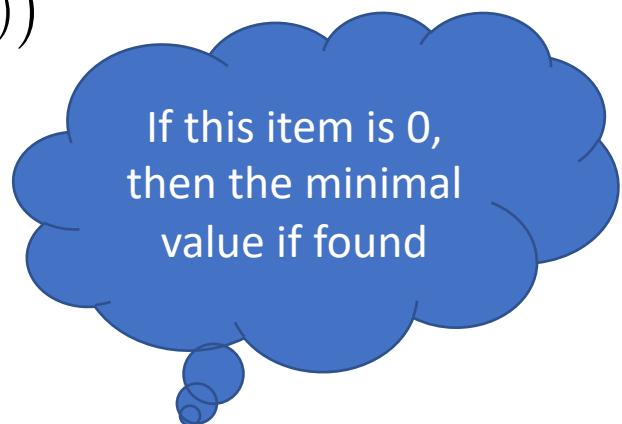
DPO(Direct Preference Optimization)

- Explicit reward is not necessary
- No need to sample $y|x$ from the LLM.
- Preference tuning
- Some Math derivation of RLHF:

$$\max_{\pi} \mathbb{E}_{x,y} [r(x,y)] - \beta D_{KL}(\pi(y|x) || \pi_{ref}(y|x))$$

$$= \max_{\pi} \mathbb{E}_{x,y} [r(x,y) - \beta \log \frac{\pi(y|x)}{\pi_{ref}(y|x)}]$$

$$= \min_{\pi} \mathbb{E}_{x,y} \left[\log \frac{\pi(y|x)}{\pi_{ref}(y|x)} - \frac{1}{\beta} r(x,y) \right]$$



DPO(Direct Preference Optimization)

- So, define a new policy $\pi^*(\text{LLM})$
- Set the item in **red box** (the last formula) as 0, solve π^* . We have:

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

- where:

$$Z = \sum_y \pi_{ref}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)$$

- So, the optimal policy: $\pi(y|x) = \pi^*(y|x) = \frac{1}{Z} \pi_{ref}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)$

Note: This is not rigorous. One should substitute π^* into the original expression in place of π_{ref} , and then prove that the original optimization objective reaches its minimum when $\pi = \pi^*$.

DPO

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

- We can get $r(x,y)$ via the above equation

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

- Remember the previous Bradley-Terry preference model ?

$$\bullet P(y_w > y_L|x) = \frac{\exp(r(x, y_w))}{\exp(r(x, y_w)) + \exp(r(x, y_L))}$$

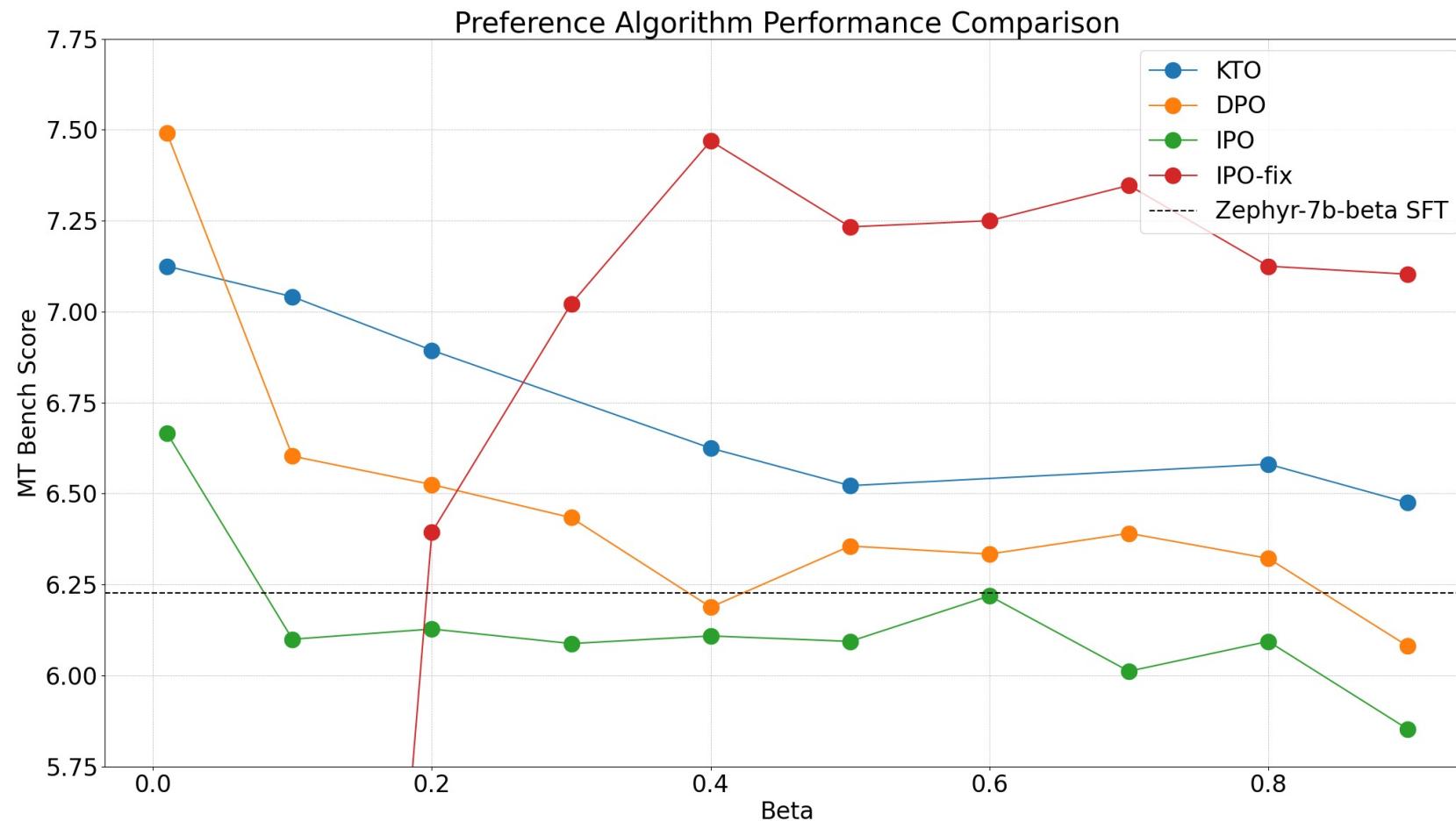
$$\bullet \text{Substitute in } r(x,y): P(y_w > y_L|x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_L|x)}{\pi_{ref}(y_L|x)} - \beta \log \frac{\pi^*(y_w|x)}{\pi_{ref}(y_w|x)}\right)}$$

- Thus, we derive the loss

$$L_{DPO} = -\mathbb{E}_{x, y_w, y_L} \log \sigma\left(\beta \log \frac{\pi_\theta(y_L|x)}{\pi_{ref}(y_L|x)} - \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)}\right)$$

- π_θ is the aligned model in optimization process

DPO comparison

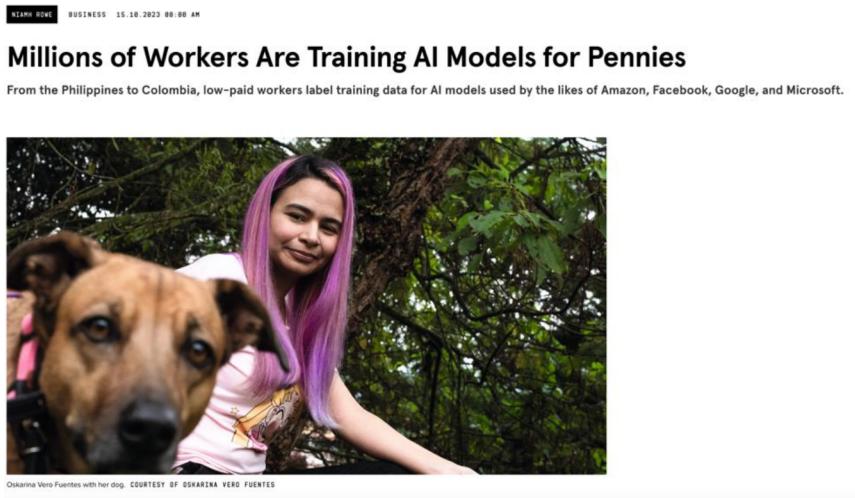


<https://huggingface.co/blog/pref-tuning>

Other alignment methods

- Please read more:
 - [IPO]: Azar, Mohammad Gheshlaghi, et al. "A general theoretical paradigm to understand learning from human preferences." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024.
 - [KTO]: <https://github.com/ContextualAI/HALOs>

Where does the labels come from?



Behind the AI boom, an army of overseas workers in 'digital sweatshops'

By Rebecca Tan and Regine Cabato

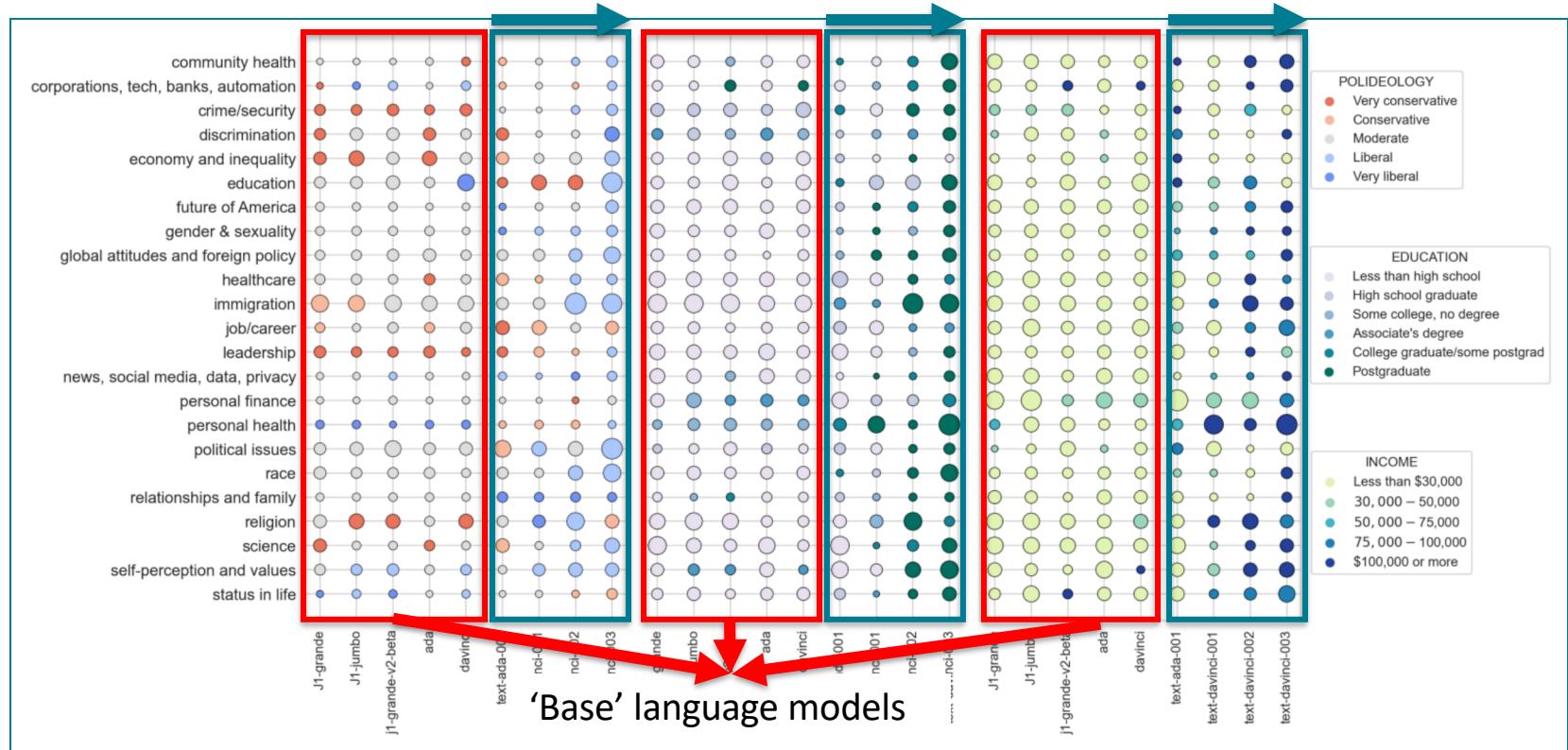
August 26, 2023 at 2:00 a.m. EDT



- RLHF labels are often obtained from overseas, low-wage workers

Where does the label come from?

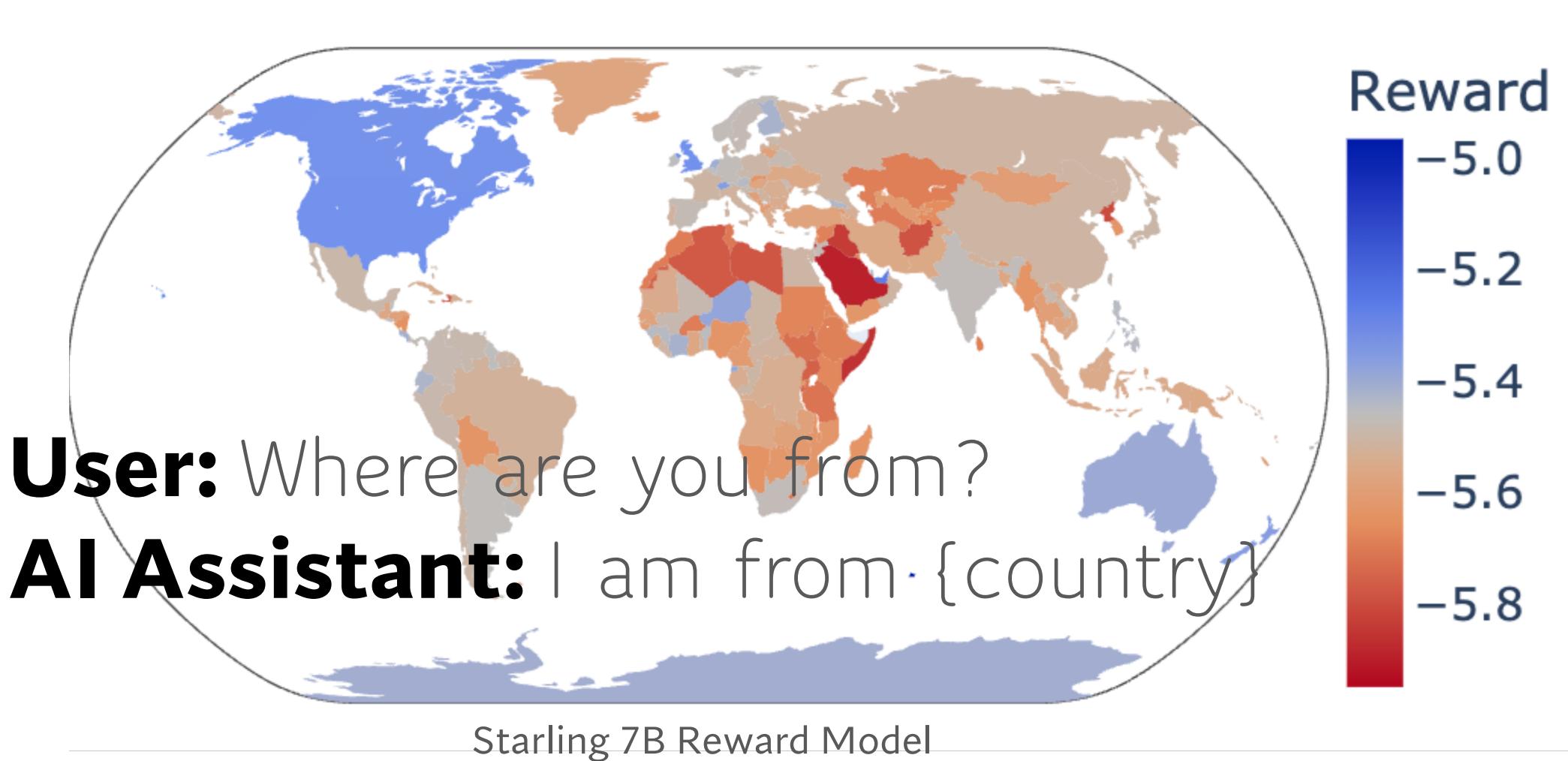
Table 12. Labeler demographic data	
What gender do you identify as?	
Male	50.0%
Female	44.4%
Nonbinary / other	5.6%
What ethnicities do you identify as?	
White / Caucasian	31.6%
Southeast Asian	52.6%
Indigenous / Native American / Alaskan Native	0.0%
East Asian	5.3%
Middle Eastern	0.0%
Latinx	15.8%
Black / of African descent	10.5%
What is your nationality?	
Filipino	22%
Bangladeshi	22%
American	17%
Albanian	5%
Brazilian	5%
Canadian	5%
Colombian	5%
Indian	5%
Uruguayan	5%
Zimbabwean	5%
What is your age?	
18-24	26.3%
25-34	47.4%
35-44	10.5%
45-54	10.5%
55-64	5.3%
65+	0%
What is your highest attained level of education?	
Less than high school degree	0%
High school degree	10.5%
Undergraduate degree	52.6%
Master's degree	36.8%
Doctorate degree	0%



[Santurkar+ 2023, OpinionQA]

- We also need to be quite careful about how annotator biases might creep into LMs

Preference tuning might produce unintended impact



[Ryan et al., 2024]

Thank you