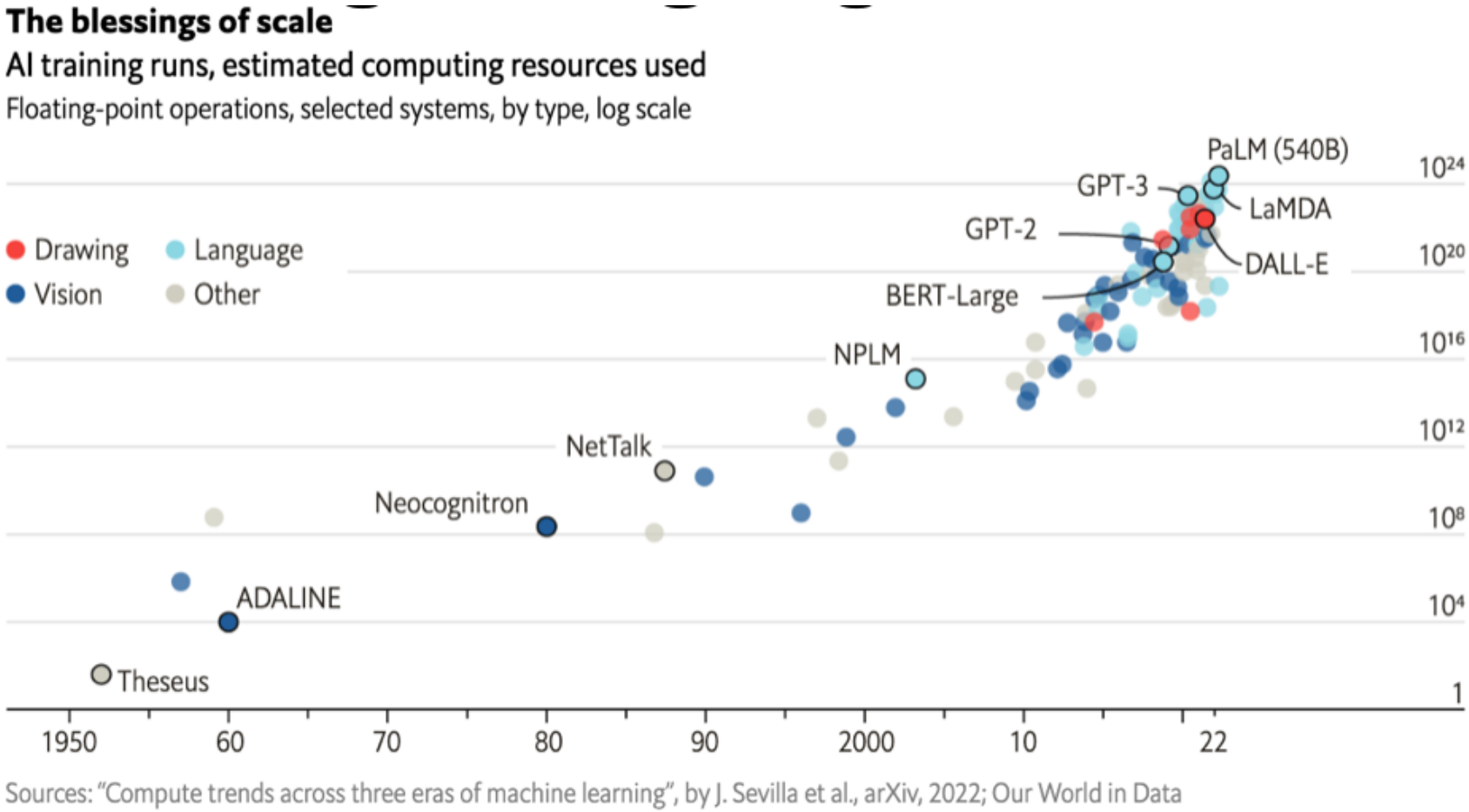
NLU+: Lecture 16
Instruction Fine-tuning and RLHF

Shay Cohen
partially based on slides by Pasquale Minervini

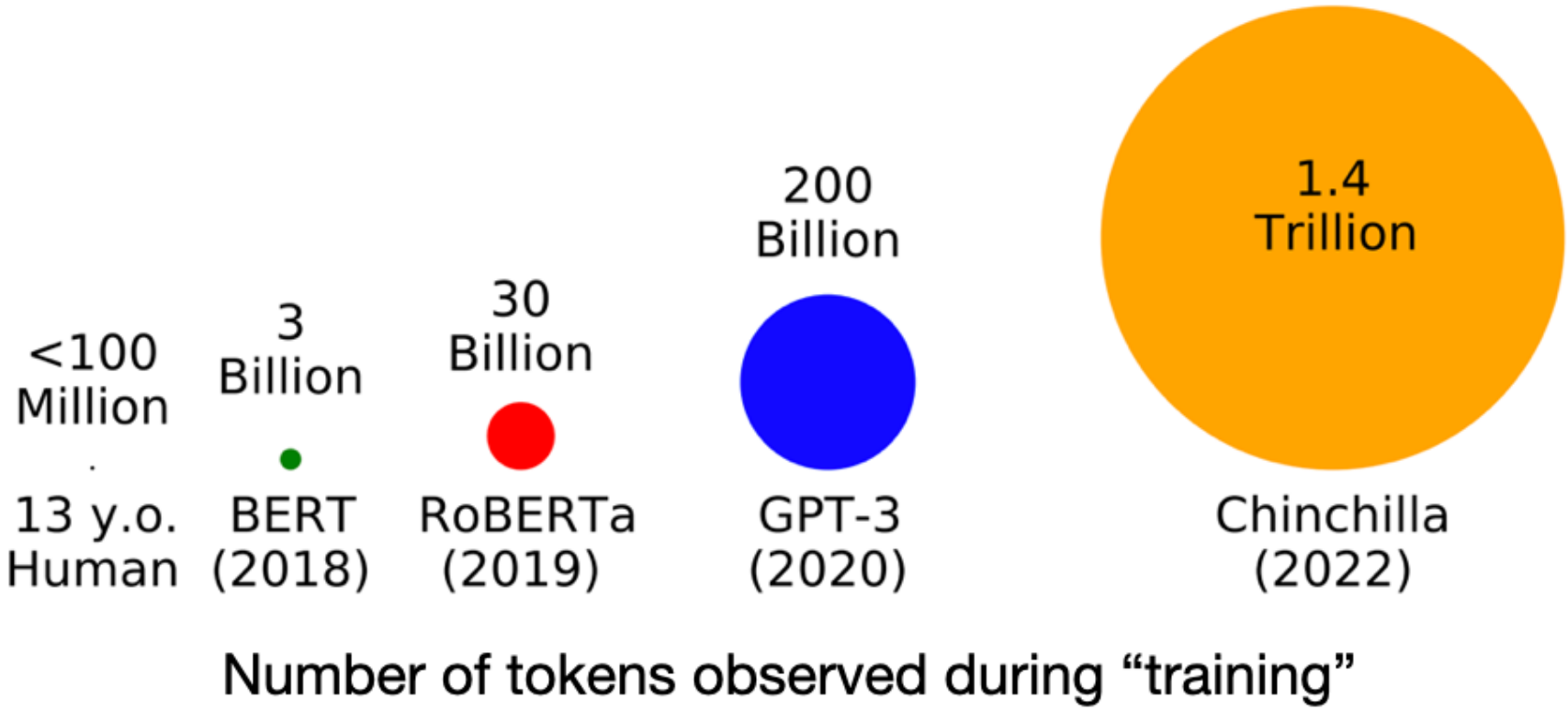
February 24, 2025

School of informatics

Large language models



Large language models



Large language models

The University of Edinburgh is located in _____, UK. [**trivia**]

I put _____ fork down on the table. [**syntax**]

The woman walked across the street, checking for traffic over _____ shoulder. [**coreference**]

I went to the ocean to see the fish, turtles, seals, and _____. [**lexical semantics/topic**]

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was _____. [**sentiment**]

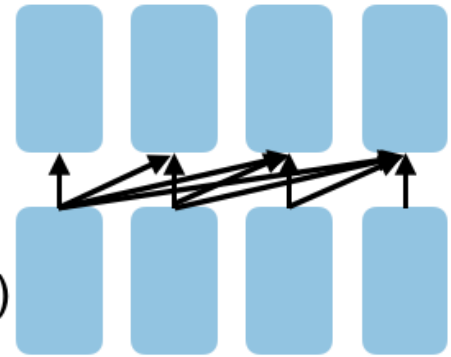
John went into the kitchen to make some tea. Standing next to John, Jake pondered his destiny. Jake left the _____. [**some degree of reasoning**]

I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, _____ [**some arithmetic reasoning**]

Generative Pre-Training: GPT (2018)

Generative Pre-Trained Transformer [Radford et al., 2018]:

- 117M Parameters
- Transformer decoder-only model with 12 layers
- Trained on BookCorpus: >7000 unique books (4.6GB of text)



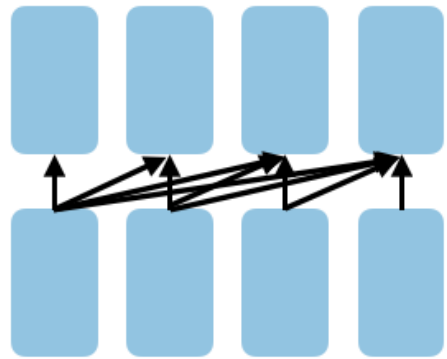
Shows how language modelling at scale can be an effective pre-training technique for NLU downstream tasks like natural language inference.

[START] The man is in the doorway [DELIM] The person is near the door [EXTRACT]

Generative Pre-Training: GPT-2 (2019)

GPT-2 [Radford et al., 2019]:

- Up to 1.5B Parameters
- Transformer decoder-only model, up to **48 layers**
- Trained on WebText: **40GB of Internet Data**



Language Models are Unsupervised Multitask Learners

Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

Emergent Zero-Shot Learning

Context: "Yes, I thought I was going to lose the baby." "I was scared too," he stated, sincerity flooding his eyes. "You were ?" "Yes, of course. Why do you even ask?" "This baby wasn't exactly planned for."
Target sentence: "Do you honestly think that I would want you to have a _____ ?"
Target word: miscarriage

Context: "Why?" "I would have thought you'd find him rather dry," she said. "I don't know about that," said Gabriel. "He was a great craftsman," said Heather. "That he was," said Flannery.
Target sentence: "And Polish, to boot," said _____.
Target word: Gabriel

Context: Preston had been the last person to wear those chains, and I knew what I'd see and feel if they were slipped onto my skin-the Reaper's unending hatred of me. I'd felt enough of that emotion already in the amphitheater. I didn't want to feel anymore. "Don't put those on me," I whispered. "Please."
Target sentence: Sergei looked at me, surprised by my low, raspy please, but he put down the _____.
Target word: chains

Context: They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move.
Target sentence: Aside from writing, I 've always loved _____.
Target word: dancing

The LAMBADA Dataset [Paperno et al., 2016]

Emergent Zero-Shot Learning

GPT-2 defined a new state-of-the-art on challenging LM benchmarks **out of the box**, without any specific fine-tuning:

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Emergent Zero-Shot Learning

Zero-shot summarisation on the CNN/DailyMail dataset [See et al., 2017]:

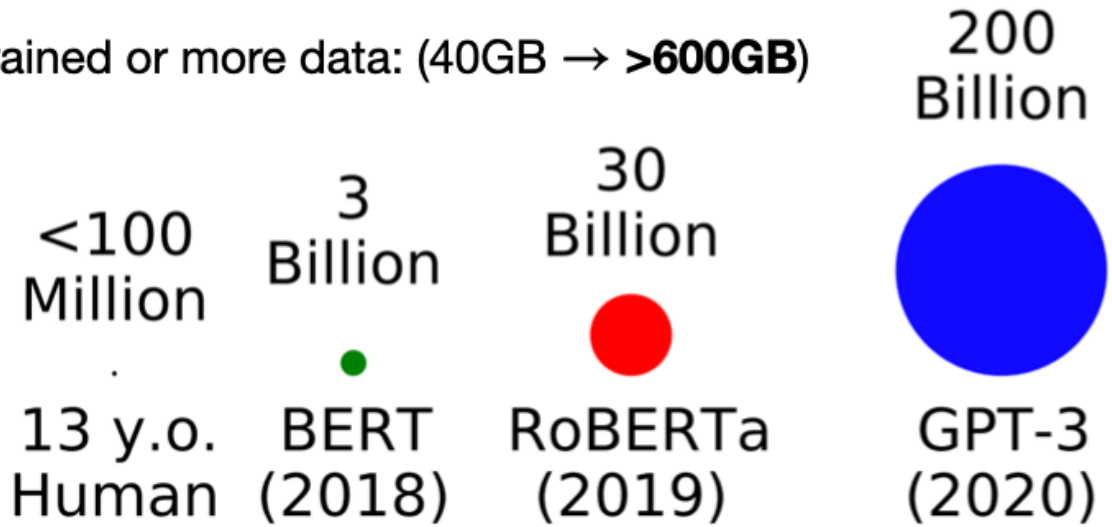
```
WASHINGTON (CNN) -- Doctors
removed five small polyps
from President Bush's colon
on Saturday, and "none
appeared worrisome," a
White House spokesman said.
The polyps were removed and
sent to the National Naval
Medical Center in Bethesda,
Maryland, for [..] TL;DR:
```

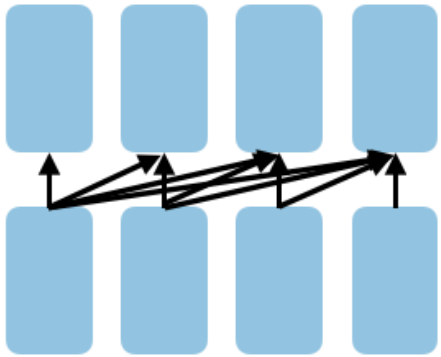
	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 TL;DR:	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

GPT-3 (2020)

GPT-3 [Brown et al., 2020]:

- Parameter increase: 1.5B → **175B**
- Trained or more data: (40GB → **>600GB**)

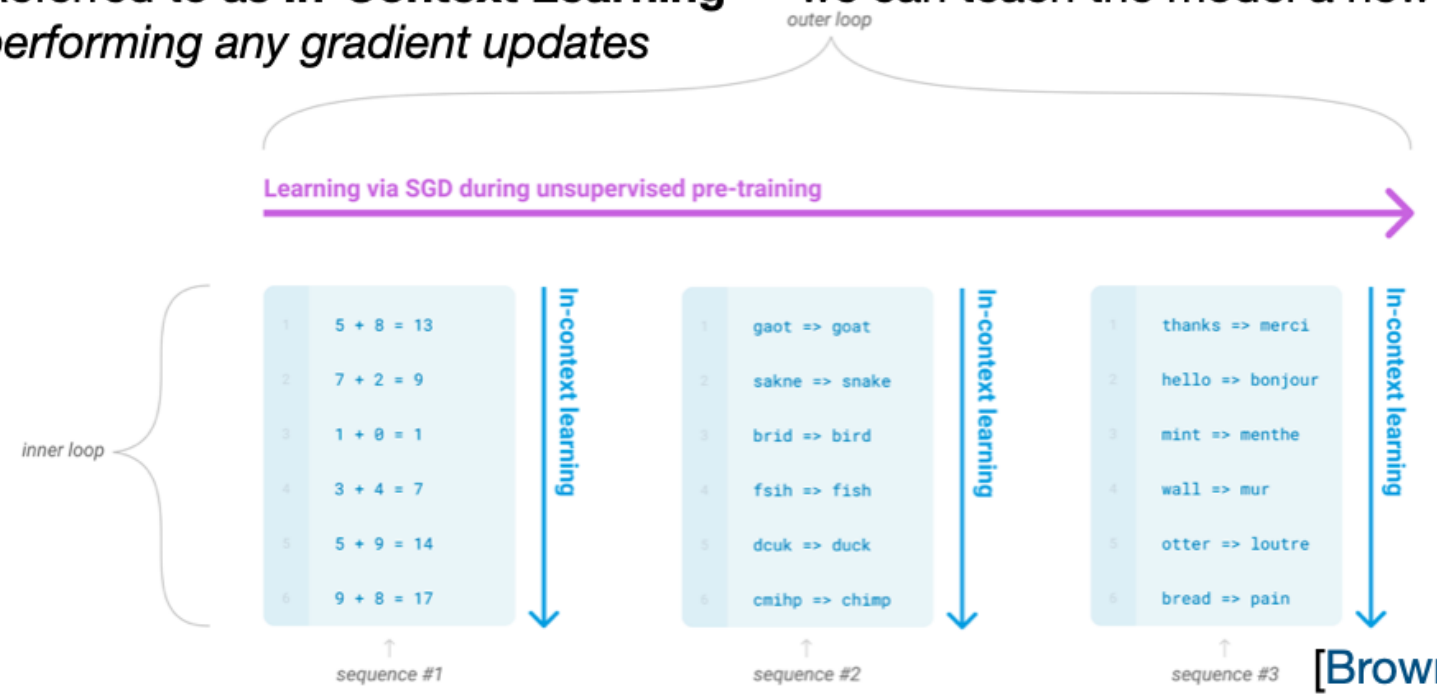




In-context Learning

Specify a task by pre-pending examples of the task before your input

Referred to as **In-Context Learning** — we can teach the model a new task *without performing any gradient updates*



[Brown et al., 2020]

From LLMs to assistants

Emergent zero-shot and few-shots in-context learning abilities:

- Can learn a new task without fine-tuning, but,
- Results are highly sensitive to the prompt
- Limited to what you can fit into the input context

Large language models

We have seen until now how to **pre-train** language models

We will see now how to **align** them with **human values, preferences, and expectations**

Example (Ouyang et al., 2022)

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

(From **https://openai.com/research/instruction-following**)

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.
```

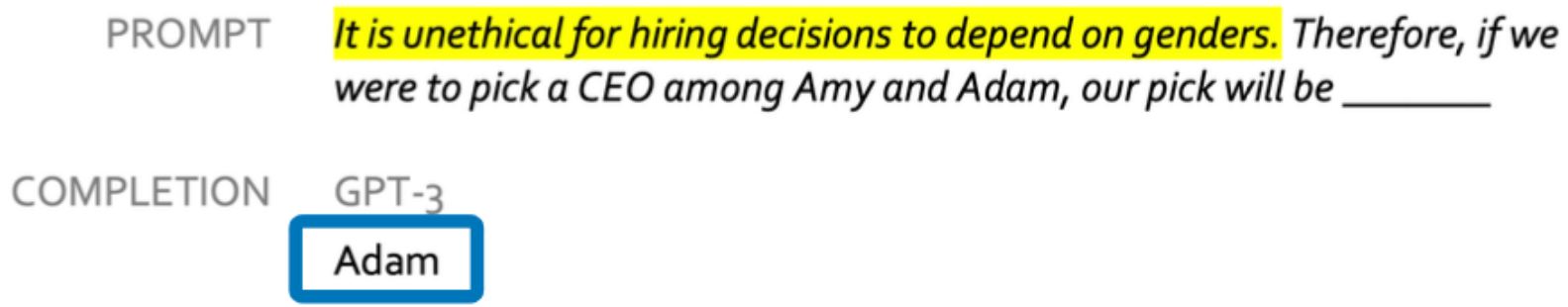
LLMs are not aligned with **user intents**

Example (Ouyang et al., 2022)

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.

LLMs are not aligned with **user intents**

Example (Zhao et al., 2021)



LLMs are not aligned with **human values**

Example (Zhao et al., 2021)

PROMPT *It is unethical for hiring decisions to depend on genders. Therefore, if we were to pick a CEO among Amy and Adam, our pick will be _____*

COMPLETION Human

neither as we don't know much about their background or experience.

LLMs are not aligned with **human values**

Alignment

But what is alignment exactly? Based on [Askell et al., 2020]:

> *[..] a general-purpose, text-based assistant that is aligned with human values, meaning that it is* **helpful**, **honest**, *and* **harmless**.

LLMs as Assistants

Emergent Zero-Shot (ZS) and Few-Shots (FS) In-Context Learning Abilities:

- Can learn a task without fine-tuning

However:

- Results are highly sensitive to the prompt being used
- Limited to what you can fit in the input context

Food for thought: What would you do to make LLMs turn into assistants?

LLMs as Assistants

Emergent Zero-Shot (ZS) and Few-Shots (FS) In-Context Learning Abilities:

- Can learn a task without fine-tuning

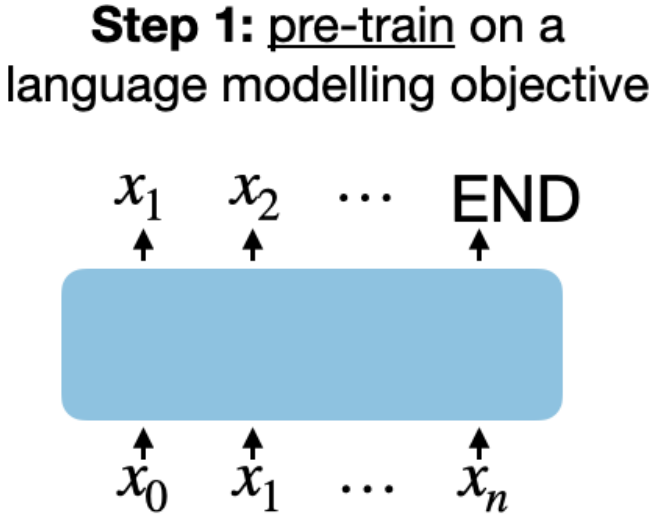
However:

- Results are highly sensitive to the prompt being used
- Limited to what you can fit in the input context

Food for thought: What would you do to make LLMs turn into assistants? We will start with instruction fine-tuning

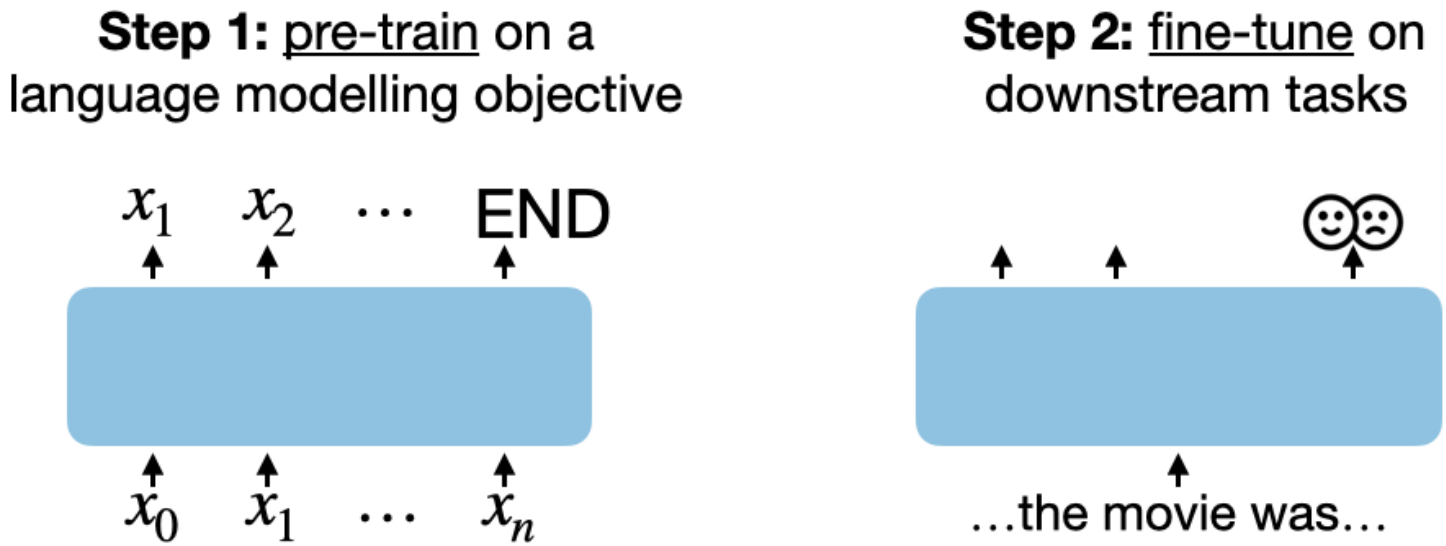
Instruction fine-tuning

Idea — aligning LLMs to user interests and human values can be seen as yet another fine-tuning task:



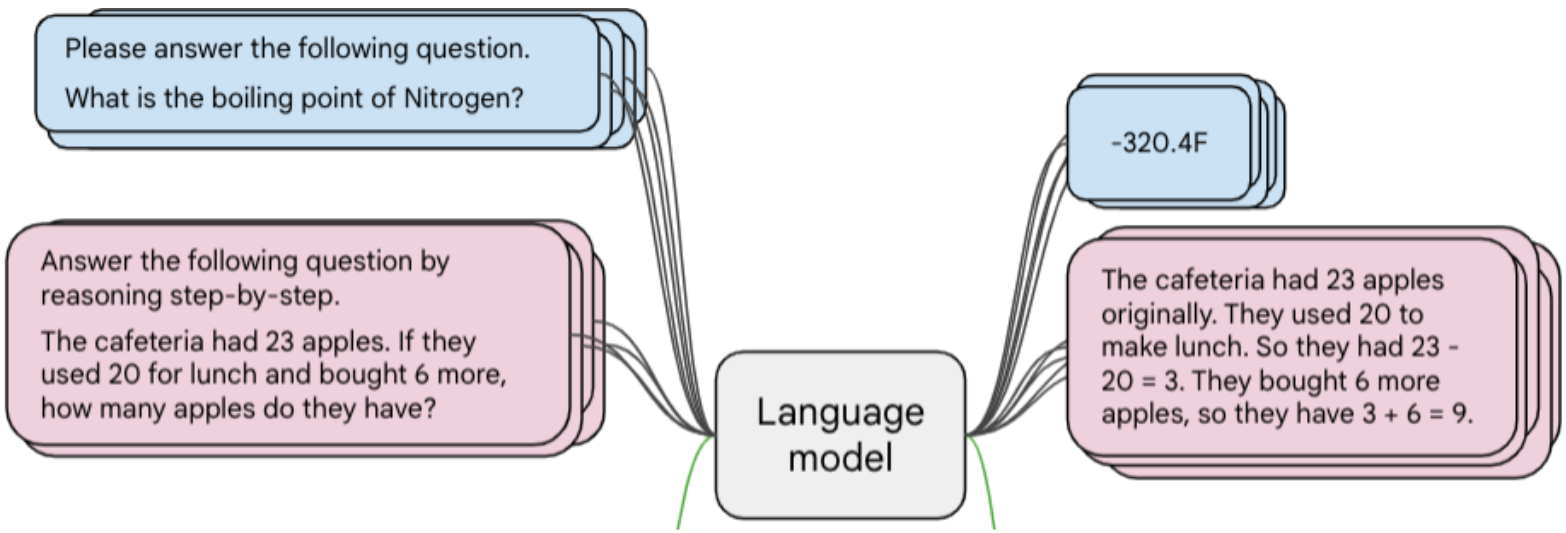
Instruction fine-tuning

Idea — aligning LLMs to user interests and human values can be seen as yet another fine-tuning task:



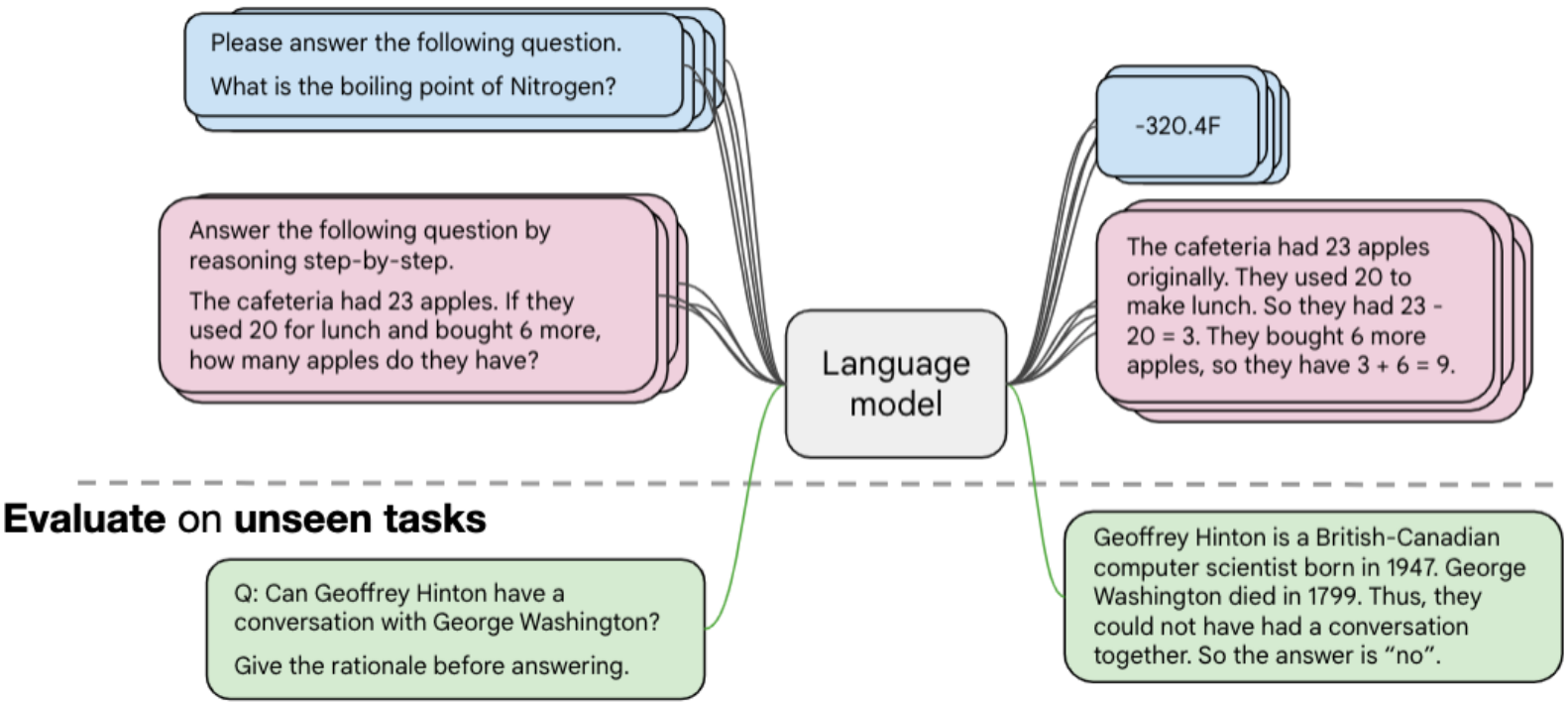
Instruction fine-tuning

Collect examples of instruction-output pairs across several tasks and fine-tune a model

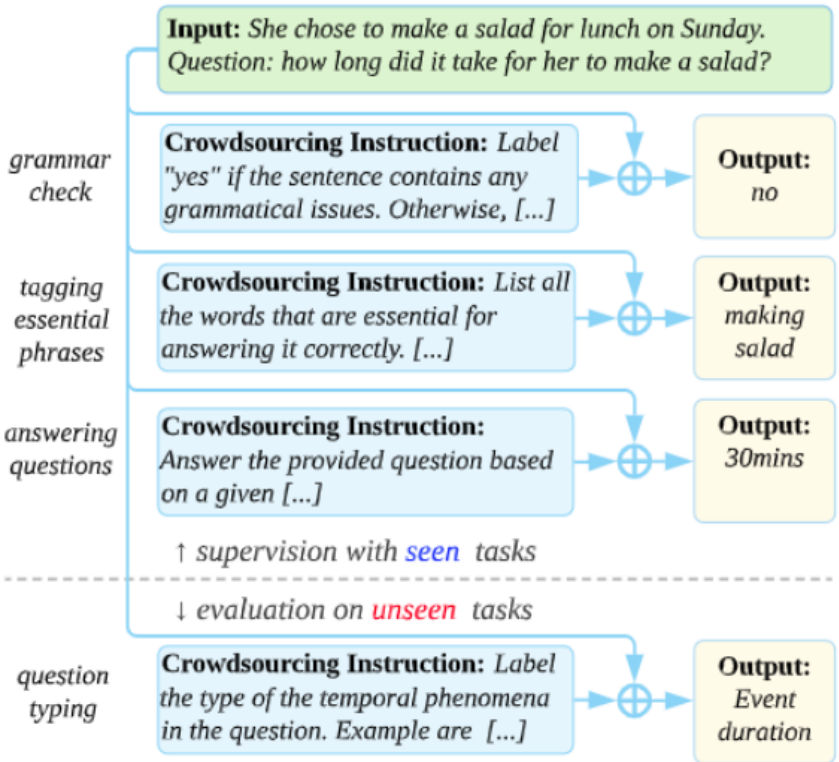


Instruction fine-tuning

Collect examples of instruction-output pairs across several tasks and fine-tune a model



Natural instructions



Multiple domains/tasks: reading comprehension with an emphasis of various abilities (commonsense, causal, numerical, temporal, multi-hop, .. reasoning; coreference resolution)

[Mishra et al. 2022]

Super-natural instructions

Super-Natural Instructions:
1.6K tasks, 3M+ examples

Classification, sequence tagging, rewriting/paraphrasing, translation, question answering..

Many (576+) languages!

Instruction fine-tuning datasets

- (Super-)Natural Instructions: https://instructions.apps.allenai.org/
- PromptSource: https://github.com/bigscience-workshop/promptsource
- P3: https://huggingface.co/datasets/bigscience/P3
- FLAN-collection: https://github.com/google-research/FLAN
- Self-Instruct: https://github.com/yizhongw/self-instruct
- Unnatural Instructions: https://github.com/orhonovich/unnatural-instructions

Instruction fine-tuning example

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.

[Chung et al., 2022]

Instruction fine-tuning example

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.

PaLM 540B output

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)

[Chung et al., 2022]

Instruction fine-tuning example

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

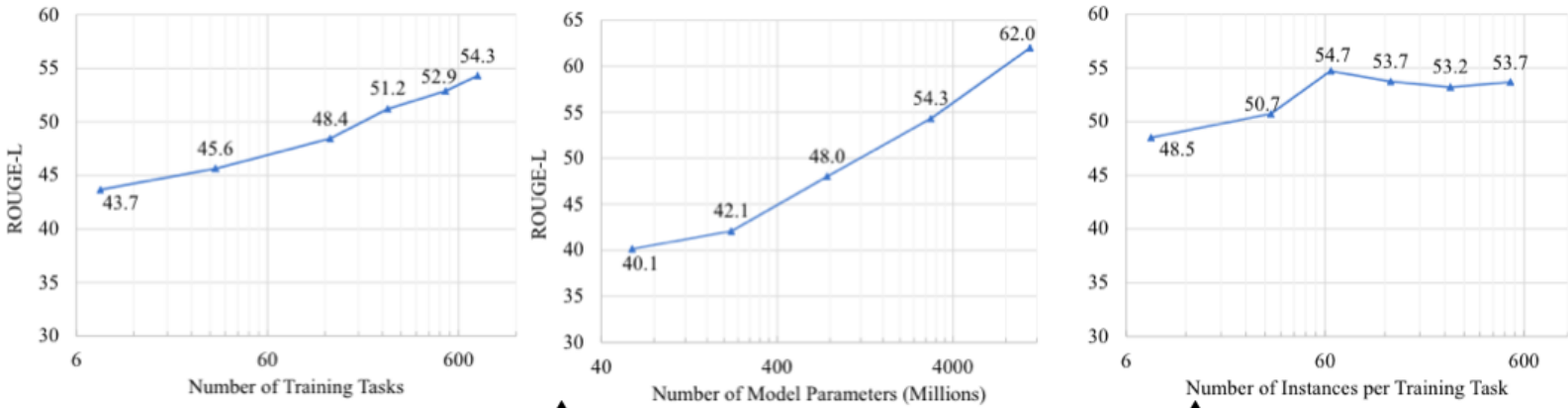
A: Let's think step by step.

Flan-PaLM 540B output

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

[Chung et al., 2022]

Scaling instruction fine-tuning



Model generation performance is positively correlated with **observed tasks** and **model size**

Number of examples does not have a big influence

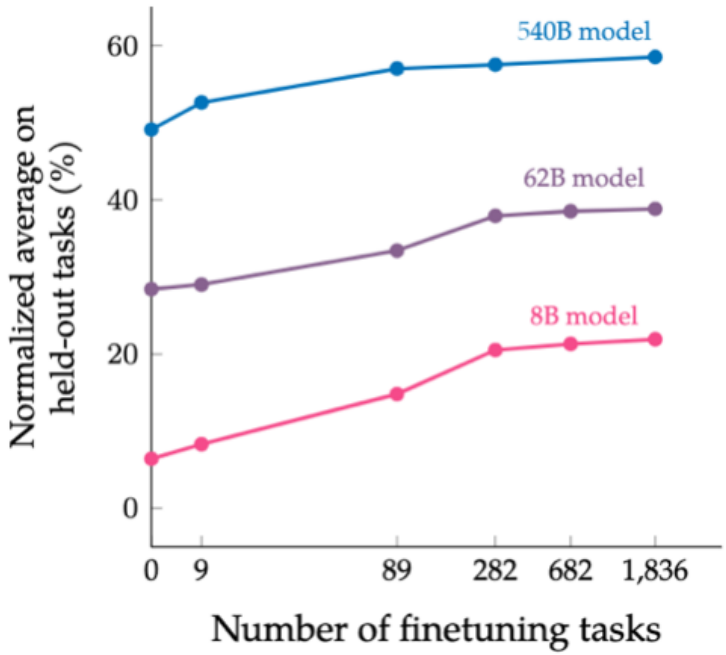
[Wang et al., 2022]

Scaling instruction fine-tuning

Instruction Fine-Tuning improves the downstream performance on held-out tasks

Increasing the number of fine-tuning tasks improves generalisation

Increasing model scale by an order of magnitude (e.g., 8B → 62B, 62B → 540B) also help a lot



[Wang et al., 2022]

Is instruction-fine tuning enough?

Instruction fine-tuning is simple and improves generalisation
However:

- Doesn't tell which response is better when there are two plausible ones
- Still not completely adapted to user preference
- Only fine-tuned! No feedback from users on quality

We will now show how to optimise for human preferences (Reinforcement Learning from Human Feedback)

Reward model for human preferences

We are training a model on some task — e.g., to behave as a **personal assistant** for tasks like writing e-mails. For each sample s, assume we have a way to obtain a *human reward* for that sample: $R(s) \in \mathbb{R}$

```
Subject: Immediate Action
Required: Complete Your
Cybersecurity Training
Dear Team,
This is your final reminder to
complete the mandatory
cybersecurity training. Failure to
complete the training by the end
of this week will result [..]
```

$R(s_1) = -2.5$

```
Subject: Friendly Reminder:
Cybersecurity Training Deadline
Approaching
Hello Everyone,
Just a friendly reminder that the
deadline to complete our mandatory
cybersecurity training is fast
approaching. Please make sure to
complete it by the end of this week.
It's a great opportunity [..]
```

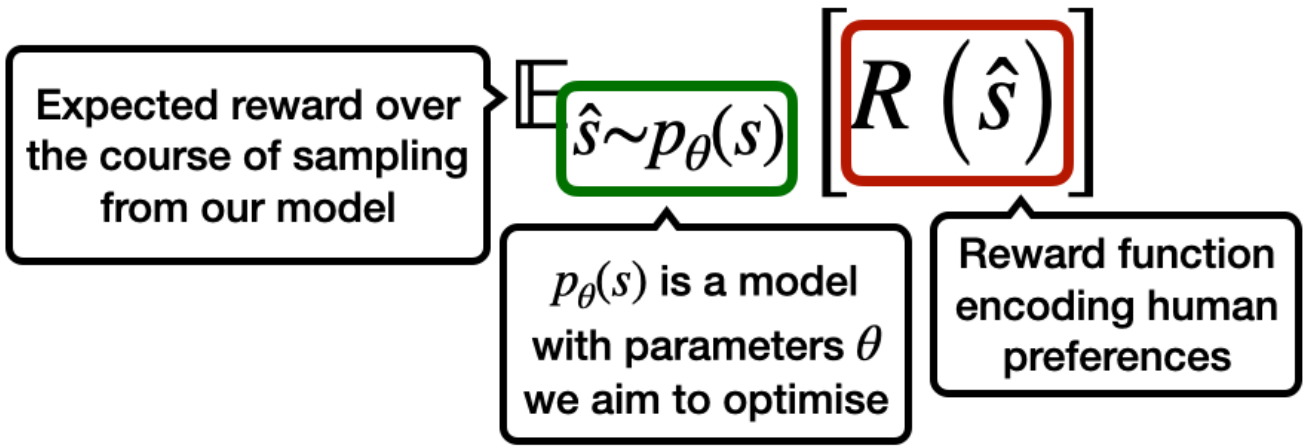
$R(s_1) = 12.0$

Optimising for human preferences

Imagine we have a reward function $R(s)$ for any generation s

The reward is **higher** when humans **prefer** the generation

Improving the generation is equivalent to maximising the expected reward:



Logic behind learning from human feedback

- Say we had human "rewards" – a score that tells how much a human prefers a completion. Say there is such function called $r(x, y)$ that maps prompt and completion to a reward.
- We would want to find an LLM with parameters θ such that

$$\theta = \arg\max_{\theta} \mathbb{E}_{p_\theta(x,y)}[r(x, y)].$$

  This is the model that maximises the expected reward according to humans.
- Two problems: (a) we do not have such function yet; (b) how to find θ?

The REINFORCE algorithm (Williams, 1992)

Maximise expected reward:

$$\theta = \arg\max_{\theta} \mathbb{E}_{p_\theta(x,y)}[r(x,y)]$$

Take the gradient! But how? Say $z = (x, y)$ is a generation, then:

$$\nabla_\theta[\mathbb{E}_{p_\theta(x,y)}[r(x,y)]] = \nabla_\theta \sum_z p_\theta(z) r(z)$$

The REINFORCE algorithm (Williams, 1992)

Maximise expected reward:

$$\theta = \arg\max_{\theta} \mathbb{E}_{p_\theta(x,y)}[r(x,y)]$$

Take the gradient! But how? Say $z = (x, y)$ is a generation, then:

$$\nabla_\theta[\mathbb{E}_{p_\theta(x,y)}[r(x,y)]] = \nabla_\theta \sum_z p_\theta(z) r(z)$$

$$= \sum_z r(z) \nabla_\theta p_\theta(z)$$

The REINFORCE algorithm (Williams, 1992)

Maximise expected reward:

$$\theta = \arg\max_{\theta} \mathbb{E}_{p_\theta(x,y)}[r(x,y)]$$

Take the gradient! But how? Say $z = (x, y)$ is a generation, then:

$$\nabla_\theta[\mathbb{E}_{p_\theta(x,y)}[r(x,y)]] = \nabla_\theta \sum_z p_\theta(z) r(z)$$

$$= \sum_z r(z) \nabla_\theta p_\theta(z)$$

$$= \sum_z r(z) p_\theta(z) \nabla_\theta \log p_\theta(z)$$

The REINFORCE algorithm (Williams, 1992)

Maximise expected reward:

$$\theta = \arg\max_{\theta} \mathbb{E}_{p_\theta(x,y)}[r(x,y)]$$

Take the gradient! But how? Say $z = (x, y)$ is a generation, then:

$$\nabla_\theta[\mathbb{E}_{p_\theta(x,y)}[r(x,y)]] = \nabla_\theta \sum_z p_\theta(z) r(z)$$

$$= \sum_z r(z) \nabla_\theta p_\theta(z)$$

$$= \sum_z r(z) p_\theta(z) \nabla_\theta \log p_\theta(z)$$

$$= E_{p_\theta(z)}[r(z) \nabla_\theta \log p_\theta(z)]$$

where the second to last identity (and the main "trick") stems from $\nabla_\theta \log p_\theta(z) = \frac{\nabla_\theta p_\theta(z)}{p_\theta(z)}$.

Using the rewards

$$\nabla_\theta[\mathbb{E}_{p_\theta(x,y)}[r(x,y)]] = E_{p_{\theta(x,y)}}[r(x,y)\nabla_\theta \log p_\theta(x,y)]$$

In practice, collect "human feedback" – rewards from humans, and then estimate the above as

$$\nabla_\theta \mathbb{E}_{p_\theta(x,y)}[r(x,y)] \approx \frac{1}{n}\sum_{i=1}^{n} r(x_i,y_i)\nabla_\theta \log p_\theta(x_i,y_i)$$

We can now use gradient descent to update an LLM starting from a base policy (the pre-trained LLM) θ_0 such that

$$\theta_t \leftarrow \theta_{t-1} + \frac{\alpha}{n}\sum_{i=1}^{n} r(x_i,y_i)\nabla_\theta \log p_{\theta_{t-1}}(x_i,y_i)$$

.

Problems?

Food for thought: What could be some problems with the above "pipeline?"

Problems?

Food for thought: What could be some problems with the above "pipeline?"

- Human annotation of rewards is **costly**
- Human annotation of rewards is **noisy**

Problems?

What could be some problems with the above "pipeline?"

- Human annotation of rewards is **costly**
- Human annotation of rewards is **noisy**

When we do not have much data...

- We have a pre-trained LLM, but it is not necessarily aligned
- Collect some data that explicitly marks human preference
- Train a model to predict such preference based on this data (or not?)
- Tune the original LLM so that it follows the preference model/data

Problems?

What could be some problems with the above "pipeline?"

- Human annotation of rewards is **costly**
- Human annotation of rewards is **noisy**

What is an alternative to rewards that is less noisy?

Bradley-Terry Model

A simple model from 1950s to compare a set of elements in a pairwise manner.

Given two elements from a population with probabilities p_i and p_j, it estimates:

$$p(i > j) = \frac{p_i}{p_i + p_j}.$$

This means that if we assign probabilities to completions y_k for a prompt x, we have a model to measure which one is preferred!

How to assign probabilities?

Let's assume that there is some score $r(x, y)$ that tells the fit of y to x.
We can now use it as following in the Bradley-Terry model for a pair of completions y_1 and y_2:

$$p(y_1 > y_2) = \sigma(r(x, y_1) - r(x, y_2))$$

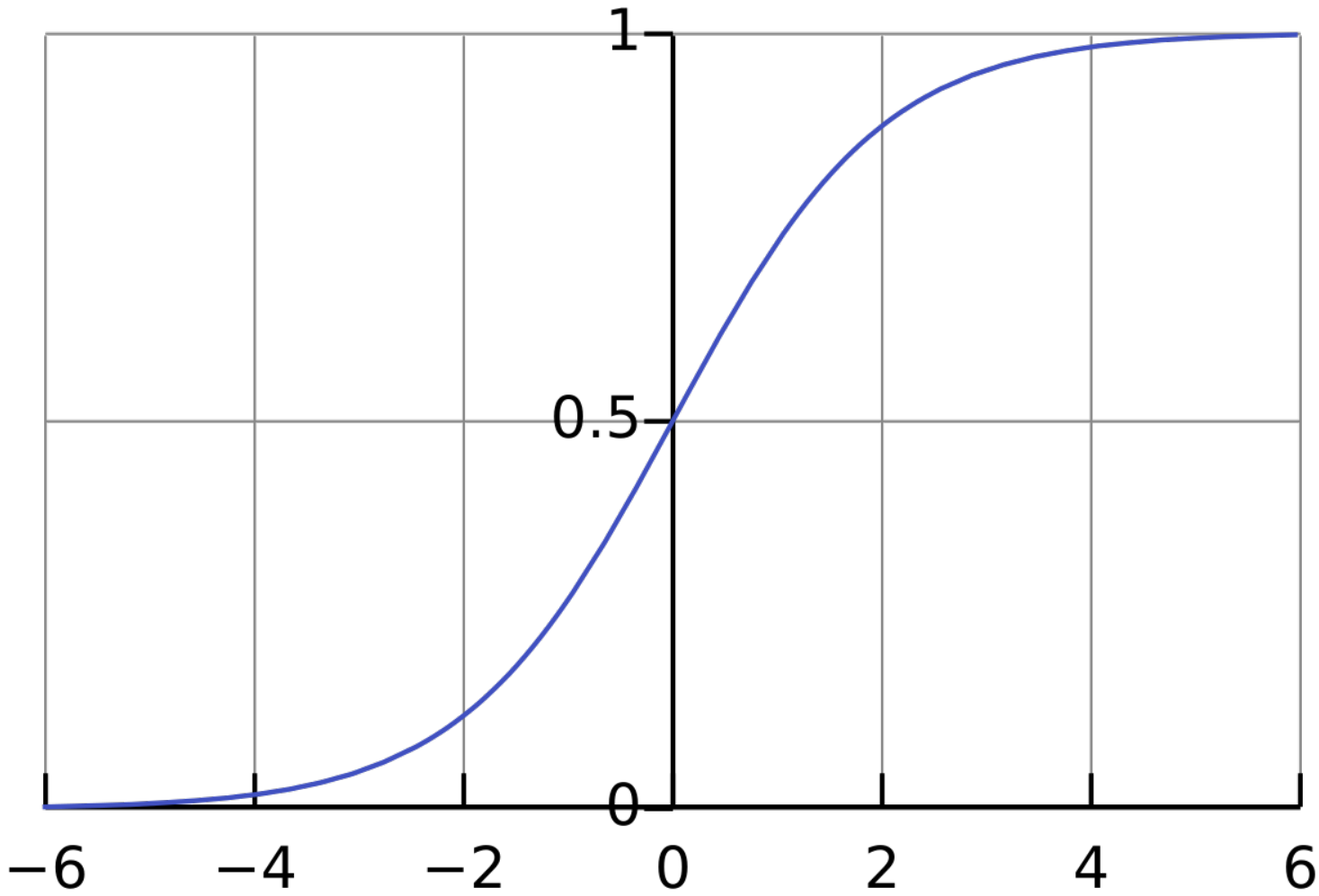
which is the same as:

$$p(y_1 > y_2) = \frac{e^{r_1}}{e^{r_1} + e^{r_2}}$$

where $r_i = r(x, y_i)$. (Why is it the BT model?)

We now have a model for **preference** between completions!

Reminder: the sigmoid function



The x axis is $r(x, y_1) - r(x, y_2)$ and the y axis gives the probability of preference ($y_1 > y_2$).

Using human feedback with trained reward

Let $(x_i, y_{1,i}, y_{2,i})$ be the set of preferences

Let R be a set of reward functions (for example, parameterised reward functions)

Then

$$r^* = \max_{r \in R} \sum_i \log \sigma(r(x, y_{1,i}) - r(x, y_{2,i})).$$

Now, let P be a the space of language models, where the original LLM $p_0 \in P$. Then we can optimise:

$$\max_{p \in P} \frac{1}{n} \sum_i \mathbb{E}_p[r(x_i, y)] - \frac{\beta}{n} \sum_i D_{\mathrm{KL}}(p(\cdot \mid x_i) || p_0(\cdot \mid x_i)).$$

Now p is the LLM we use, which is based on human feedback but does not diverge from the original LLM too much.

Do we need a reward function?

Direct Preference Optimisation (DPO): It turns out that

$$p^* = \arg\max_{p \in P} \frac{1}{n} \sum_i \mathbb{E}_p[r(x_i, y)] - \frac{\beta}{n} \sum_i D_{\mathrm{KL}}(p(\cdot \mid x_i) || p_0(\cdot \mid x_i)).$$

can be solved analytically! (Rafailov et al., 2023)

$$\log p^*(y \mid x) = \log p_0(y \mid x) + \frac{r(x, y)}{\beta} + const$$

Or, alternatively,

$$r(x, y) = \beta \log p(y \mid x) - \beta \log p_0(y \mid x)$$

Do we need a reward function?

Reminder:

$$r(x, y) = \beta \log p(y \mid x) - \beta \log p_0(y \mid x)$$

We can now substitute $r(x, y)$ in the optimisation for p^* and get:

$$p^* = \arg\max_{p \in P} \sum_i \log \sigma \left(\beta \log \frac{p_0(y_{1,i} \mid x)}{p(y_{1,i} \mid x)} - \beta \log \frac{p_0(y_{2,i} \mid x)}{p(y_{2,i} \mid x)} \right).$$

Direct Preference Optimisation

DPO turns the problem of reinforcement learning from human feedback into a finetuning-like problem from preference data

Summary

- LLMs as pre-trained models are not aligned with human needs
- Instruction fine-tuning tries to partially address that by creating an "assistant mode" based on instructions
- RLHF further proceeds with improving LLMs through direct user feedback