

Unsupervised  
Learning

Supervised  
Fine-tuning

S&DS 365 / 665

Intermediate Machine Learning

# LLM Post-processing: Alignment with human preferences

December 2

RLHF  
(cherry on top ☺)

# Welcome back!

## Transformers and LLMs

- LLM scaling laws (recap)
- Finetuning (recap)
- Reinforcement Learning with Human Feedback (RLHF)
- Direct Preference Optimization (DPO)

# Reminders

- Quiz 5 before break (RL, HMMs, RNNs, GRUs)
- Assn 5 due this Wednesday, Dec. 4 (GRUs, Transformers)
- Final exam: Sunday Dec 15, 2pm
- Practice exams are posted
- Review sessions TBA

# Quiz 5

## Quiz Summary

Section Filter ▾

 Student Analysis

 Item Analysis

⌕ Average Score

**92%**

⌕ High Score

**100%**

⌕ Low Score

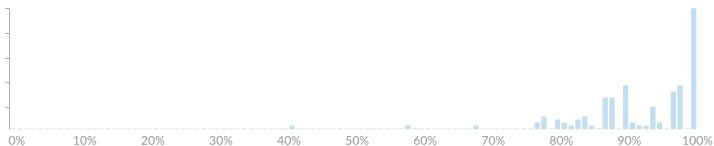
**41%**

⌕ Standard Deviation

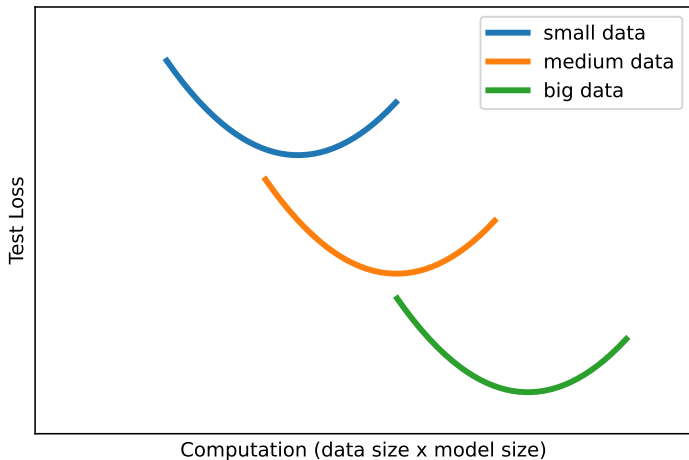
**0.89**

⌕ Average Time

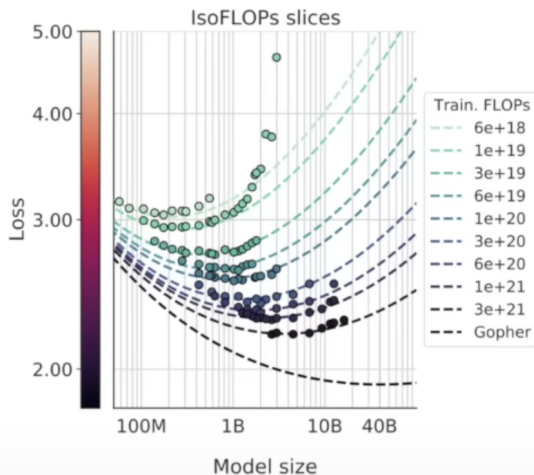
**15:07**



# Recall: Scaling behavior of LLM models



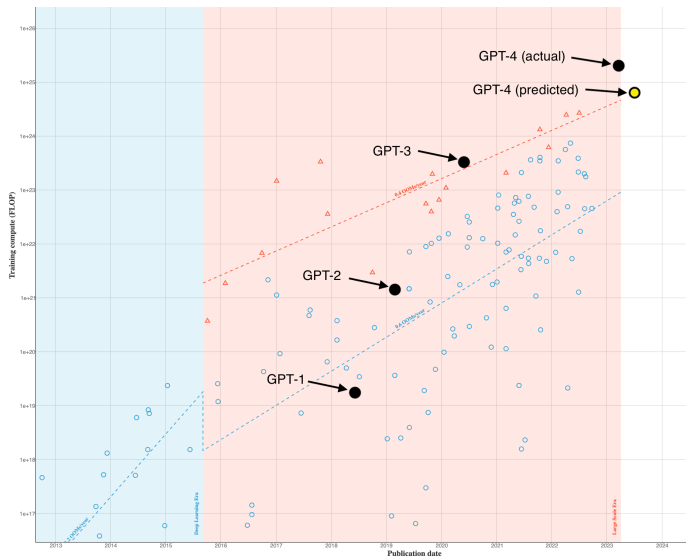
# Recall: Scaling behavior of LLM models



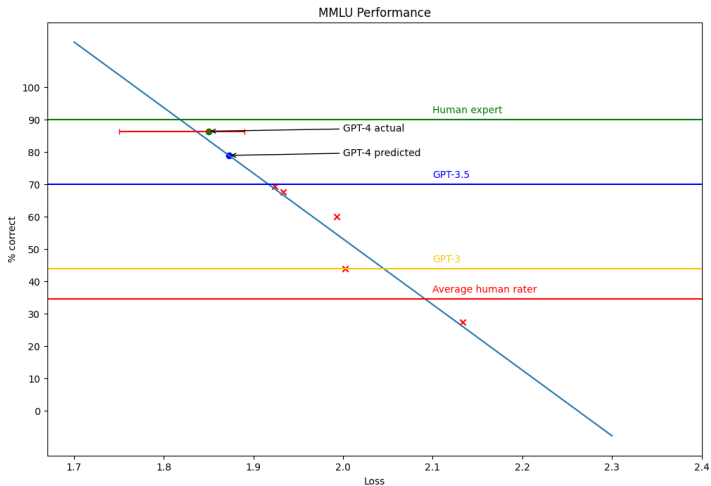
# Back of the envelope calculation



- LLM loss on test data is  $\approx 1.9$  nats/token
- Translates to entropy of about  $1.9 / \log_e(2) \approx 2.74$  bits/token
- So, *loss/less* coding of 500 billion tokens takes  $\approx 171$  gigabytes
- GPT-3's 175 billion params uses 346 gigabytes (16-bit precision)
- GPT-3 could, in principle, memorize all of its training data!
- Recall double descent:  $X\beta_{mn} = Y$







MMLU: Massive Multitask Language Understanding, <https://en.wikipedia.org/wiki/MMLU>,  
<https://paperswithcode.com/dataset/mmlu>

## *Sutton's "Bitter Lesson"*



“The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.”

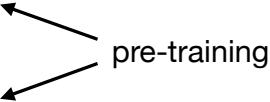
**Making the models useable**

**Recall: Finetuning**

# Recall: Machine learning frameworks

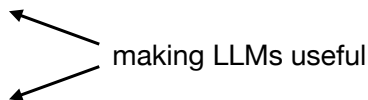
- Supervised
- Unsupervised, self-supervised
- Reinforcement learning
- Representation learning

# Recall: Machine learning frameworks

- Supervised
  - Unsupervised, self-supervised
  - Reinforcement learning
  - Representation learning
- 
- pre-training
- The diagram shows two arrows originating from the text 'pre-training' on the right. One arrow points diagonally up and to the left towards the text 'Unsupervised, self-supervised'. The other arrow points diagonally down and to the left towards the text 'Representation learning'.

# Recall: Machine learning frameworks

- Supervised
- Unsupervised, self-supervised
- Reinforcement learning
- Representation learning



# Standard Supervised learning



Goal: Accurately assign label  $y \in \{\text{dog}, \text{cat}\}$  to an image  $x$ .

“Learn” from a large database of examples  $\{(x_i, y_i)\}$ .

# Standard Supervised learning

Fugetsu-Do, a mochi store in the Little Tokyo neighborhood of Los Angeles, has existed for 121 years, with a lasting impact on its community.

ロサンゼルスのリトル東京地区にある餅屋「風月堂」は 121 年間存在し、地域社会に永続的な影響を与えています。

Goal: Accurately translate input  $x_1, \dots, x_m$  to output  $y_1, \dots, y_m$ .

“Learn” from a large database of examples  $\{(x, y)\}$ .

This is called *sequence-to-sequence learning* (seq2seq)



# Finetuning

Finetuning takes a pre-trained LLM, and feeds it a database of examples of how it should “translate” input to output.

Examples are here:

`huggingface.co/datasets/HuggingFaceH4/no\_robots`

Categories: Generation, Open QA, Brainstorm, Chat, Rewrite, Summarize, Coding, Classify, Closed QA, Extract

# Example training data

```
[ { "content": "What are 5 things I can do when it's  
raining in London? I am visiting for the first time and  
only for a week with my husband. We love to walk, eat  
good food, and explore.", "role": "user" }, { "content":  
"Sure! Here are five options for things to do in London  
on a rainy day:\n\n1. Visit The British Museum.  
Dedicated to human history, art and culture, The British  
Museum has over 8 million permanent objects in its  
collection and is spread over multiple floors and wings  
of its impressive Bloomsbury building. From its  
extensive collection of Egyptian mummies to the Rosetta  
Stone, there are numerous exhibitions to capture your  
imagination. The museum is free to enjoy, though runs  
separately ticketed exhibitions throughout the year, as  
well. \n\n2. Explore Spitalfields Market. Since the 17th  
century, traders have been operating out of this East  
London market selling food, art, antiques, jewellery,  
and so much more. Today, the covered market is a hot  
destination for fashion and food enthusiasts with  
numerous stalls and shops open every day of the week. On  
the weekends, the market hosts various community events  
including pop-up music festivals. \n\n3. Shop at  
Westfield London. Located in West London, Westfield  
London is an enormous shopping centre hosting roughly  
320 stores and restaurants spread over 2.5 million
```

Brainstorm

# Example training data

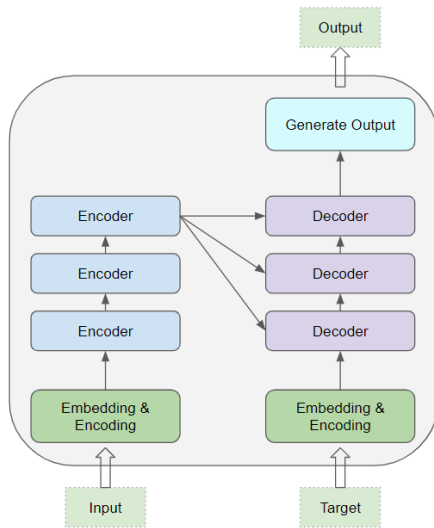
Prompt	Response
Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence.	Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity.
ELI5: What's the cause of the "anxiety lump" in our chest during stressful or disheartening experiences?	The anxiety lump in your throat is caused by muscular tension keeping your glottis dilated to maximize airflow. The clenched chest or heartache feeling is caused by the vagus nerve which tells the organs to pump blood faster, stop digesting, and produce adrenaline and cortisol.
Create a shopping list from this recipe: Trim the ends off zucchini. Cut zucchini in half lengthwise; scoop out pulp, leaving 1/2-in. shells. Finely chop pulp. In a skillet, cook beef, zucchini pulp, onion, mushrooms and peppers over medium heat until meat is no longer pink; drain. Remove from the heat. Add 1/2 cup cheese, ketchup, salt and pepper; mix well. Spoon into the zucchini shells. Place in a greased 13x9-in. baking dish. Sprinkle with remaining cheese.	Zucchini, beef, onion, mushroom, peppers, cheese, ketchup, salt, pepper

# Training process

- ① Prompt is read in and embedding / encoding vectors are processed just as in generation.
- ② Target sequence is generated (predicted) token-by-token
- ③ Parameters are iteratively updated to make the target sequences more and more likely

*After training is complete, updated LLM has been “finetuned” to the tasks in the training data*

# Transformer finetuning architecture



# Finetuning fineprint

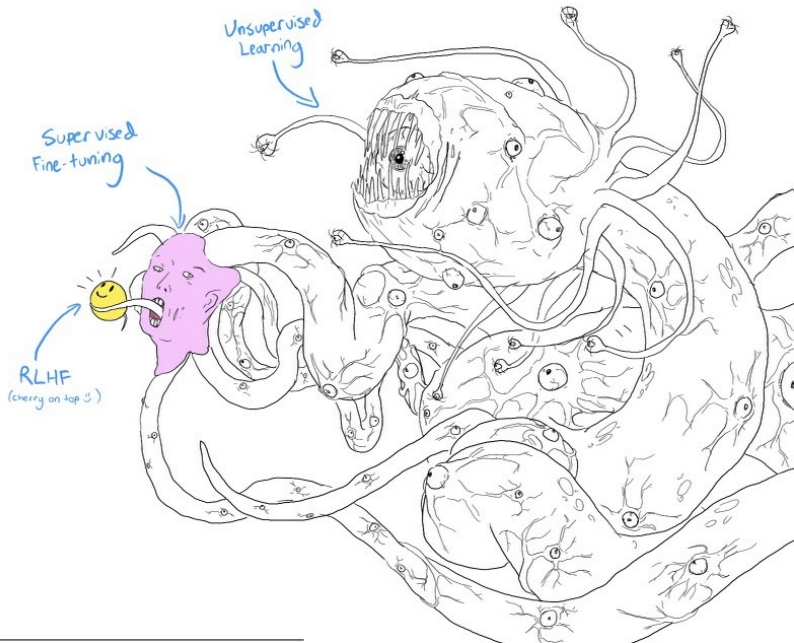
- The LLM is huge (175 billion parameters for GPT-3)
- Finetuning training data are too small to adjust all the parameters
- Special techniques are used to make small changes to the decoder model (e.g. LoRa, low-rank adaptation)

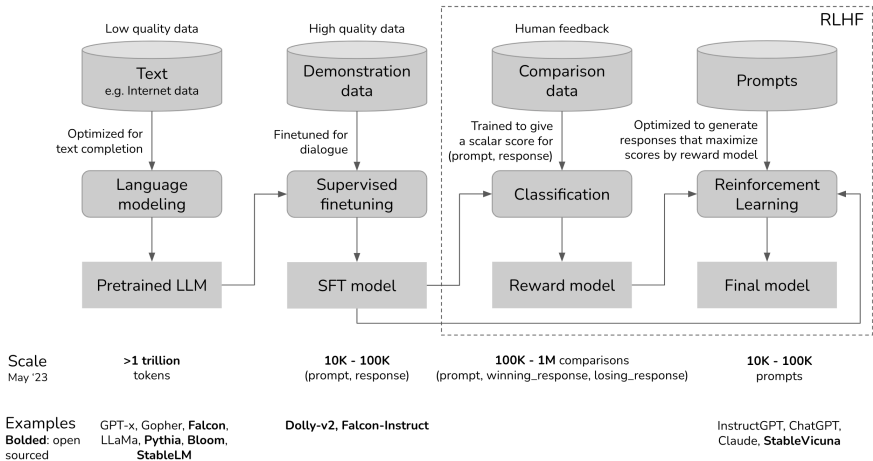
**Making the models useful**

**Today: Aligning with human preferences**

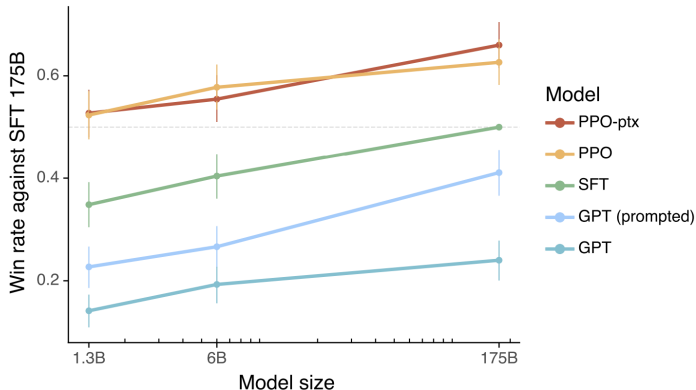
- The model is pre-trained on all kinds of stuff of questionable quality from the Internet—it's like a Shoggoth monster
- The monster is then finetuned on higher quality data—it becomes socially acceptable
- The model is further polished using reinforcement learning—giving it a smiley face :-)
- But the monster is still encoded in the parameters...







# InstructGPT



Original GPT-3 + RL was called “InstructGPT” by OpenAI

---

SFT=“Supervised Finetuning”. L. Ouyang et al., “Training language models to follow instructions with human feedback”, NeurIPS 2022

# Basic idea of RLHF

- ① Train a reward model to rank candidate responses
  - ▶ Uses Bradley-Terry model for pairwise comparisons
- ② Use trained reward function in RL to improve LLM
  - ▶ Uses version of policy gradients called Proximal Policy Optimization (PPO)

# Data collection interface

Submit

Skip

« Page 3 / 11 »

Total time: 05:39

Instruction

Summarize the following news article:

====  
(article)  
====

Include output

Output A

summary!

Rating (1 = worst, 7 = best)

1

2

3

4

5

6

7

Fails to follow the correct instruction / task ? ☐ Yes ☐ No

Inappropriate for customer assistant ? ☐ Yes ☐ No

Contains sexual content ☐ Yes ☐ No

Contains violent content ☐ Yes ☐ No

Encourages or fails to discourage violence/abuse/terrorism/self-harm ☐ Yes ☐ No

Denigrates a protected class ☐ Yes ☐ No

Gives harmful advice ? ☐ Yes ☐ No

Expresses moral judgment ☐ Yes ☐ No

Notes

(Optional) notes

# Data collection interface

**Ranking outputs**

To be ranked

**B** A team of researchers from Yale University and University of California, Davis studied the vocalization patterns of several different types of parrots. They found that parrots like to mimic human speech, and can produce a wide range of sounds, such as whistles, squawks, and other types of vocalizations...

**C** Parrots have been found to have the ability to understand numbers. Researchers have found that parrots can understand numbers up to six. In a series of experiments, the parrots were able to identify the amount of food items under a number of cups...

**A** A research group in the United States has found that parrots can imitate human speech with ease, and some of them can even do so in the same way as humans. This group studied the sounds that parrots make in their natural habitats and found that they use their tongues and beaks in ways that are strikingly...

**D** Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds.

**E** Current research suggests that parrots see and hear things in a different way than humans do. While humans see a rainbow of colors, parrots only see shades of red and green. Parrots can also see ultraviolet light, which is invisible to humans. Many birds have this ability to see ultraviolet light, an ability

Rank 1 (best)

Rank 2

Rank 3

Rank 4

Rank 5 (worst)

<https://openai.com/index/instruction-following/>. If a prompt has 4 ranked completions, this results in  $\binom{4}{2} = 6$  paired comparisons.

# Bradley-Terry model

Preference probability assumed to have from

$$\begin{aligned}\mathbb{P}(y_w \succ y_\ell \mid x) &= \frac{\exp(r^*(x, y_w))}{\exp(r^*(x, y_w)) + \exp(r^*(x, y_\ell))} \\ &= \sigma(r^*(x, y_w) - r^*(x, y_\ell))\end{aligned}$$

- $\sigma(u) = 1/(1 + e^{-u})$  is sigmoid.
- $y_w$  is “winning response”  $y_\ell$  is “losing response”

# Bradley-Terry model

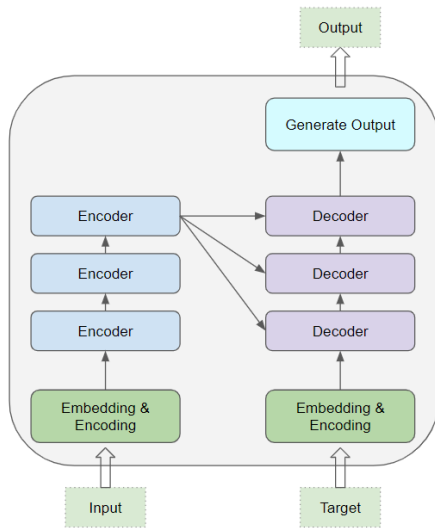
Maximum likelihood training:

- Parametric preference model  $r_\theta$
- Dataset  $\{(x^{(i)}, y_w^{(i)}, y_\ell^{(i)})\}$ ,
- Optimize (via SGD):

$$\hat{\theta} = \arg \max_{\theta} \sum_i \log \sigma(r_\theta(x, y_w) - r_\theta(x, y_\ell))$$



# Reward model from human preference data



# Reward model from human preference data

- Start with finetuned LLM  $\pi_{\text{SFT}}(y \mid x)$ .
- Remove token generation layer
- Parameterize reward model  $r_{\theta}(x, y)$  by adding linear layer, with weights  $\theta$ , after final Transformer layer
- Train  $\theta$  using maximum likelihood under Bradley-Terry model using human rankings of LLM responses.

# RLHF

Next step: Use trained reward model  $r_\theta$  in reinforcement learning to improve “policy”  $\pi_{\text{SFT}}(y | x)$

- Actions correspond to tokens to generate
- Parameterized as  $\pi_\phi(y | x)$

# Objective function

$$\begin{aligned} \max_{\phi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\phi}(y|x)} \left\{ r_{\theta}(x, y) - \beta \log \frac{\pi_{\phi}(y|x)}{\pi_{\text{SFT}}(y|x)} \right\} \\ = \max_{\phi} \mathbb{E}(r(x, y)) - \beta D_{\text{KL}}(\pi_{\phi}(y|x) \parallel \pi_{\text{SFT}}(y|x)) \end{aligned}$$

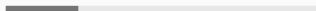
- Can't be optimized directly—not differentiable since tokens  $y$  are discrete
- Treated as plug-in reward function
- Optimized using PPO—“Proximal Policy Optimization”
- A type of Actor-Critic RL

# Quantifying performance change

Dataset

## RealToxicity

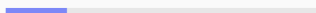
GPT 0.233



Supervised Fine-Tuning 0.199



InstructGPT **0.196**



Dataset

## TruthfulQA

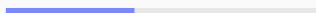
GPT 0.224



Supervised Fine-Tuning 0.206



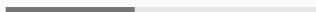
InstructGPT **0.413**



API Dataset

## Hallucinations

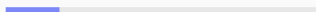
GPT 0.414



Supervised Fine-Tuning **0.078**



InstructGPT 0.172



API Dataset

## Customer Assistant Appropriate

GPT 0.811



Supervised Fine-Tuning 0.880



InstructGPT **0.902**



Evaluating InstructGPT for toxicity, truthfulness, and appropriateness. Lower scores are better for toxicity and hallucinations, and higher scores are better for TruthfulQA and appropriateness. Hallucinations and appropriateness are measured on our API prompt distribution. Results are combined across model sizes.

# Direct Preference Optimization (DPO)

- PPO was “bread and butter” at early OpenAI
- But can be unstable, difficult to scale
- 2023 DPO paper from Stanford:
  - ▶ Shows how RL is unnecessary
  - ▶ Aligns LLM directly to preference data

# DPO: Key steps

- 1 Start with same objective function as RLHF
- 2 Show (nonparametric) LLM solution has a closed form
- 3 Invert to get ranking/preference function
- 4 Plug into Bradley-Terry model

*Punchline: Can optimize Bradley-Terry model directly in terms of LLM*

# Steps 1-2: Optimize objective nonparametrically

Step 1: Objective is to maximize

$$\mathbb{E}_{\pi}(r(x, y)) - \beta D_{\text{KL}}(\pi(y | x) \| \pi_{\text{SFT}}(y | x))$$

Step 2: Solution has a closed form:

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

where  $Z(x)$  is a normalizing constant (partition function)

Follows from a Lagrange multiplier argument.



## Derivation (simplified)

Constrained optimization:  $\min D(q||p)$  such that  $E_q r = c$ .

Lagrangian is

$$\mathcal{L}_\lambda(x, q) = \sum_y q(y|x) \log \frac{q(y|x)}{p(y|x)} + \lambda \left( \sum_y q(y|x) r(x, y) - c \right)$$

Taking derivatives:

$$\nabla_{q(y|x)} \mathcal{L}(q, x) = \log \frac{q(y|x)}{p(y|x)} + 1 + \lambda r(x, y)$$

Setting to zero and normalizing:

$$q(y|x) = \frac{p(y|x)}{Z(x)} \exp(\lambda r(x, y))$$

## Steps 3-4: Invert and plug in

Step 3: Invert (algebra):

$$r(x, y) = \beta \log \frac{\pi(y | x)}{\pi_{\text{SFT}}(y | x)} + \beta \log Z(x)$$

Step 4: Parameterize LLM and plug into Bradley-Terry objective:

$$\mathcal{L}_{\text{DPO}}(\phi) = - \sum_{(x, y_w, y_\ell)} \log \sigma \left( \beta \log \frac{\pi_\phi(y_w | x)}{\pi_{\text{SFT}}(y_w | x)} - \beta \log \frac{\pi_\phi(y_\ell | x)}{\pi_{\text{SFT}}(y_\ell | x)} \right)$$

# Making sense of gradients

Stochastic gradient descent:

$$\phi \leftarrow \phi - \eta \mathbb{E}_{(x, y_w, y_\ell)} [\nabla_\phi \mathcal{L}_{\text{DPO}}(\phi)]$$

$$\begin{aligned} -\mathbb{E} \nabla_\phi \mathcal{L}_{\text{DPO}}(\phi) = \mathbb{E}_{(x, y_w, y_\ell)} \bigg\{ & \underbrace{\sigma(r_\phi(x, y_\ell) - r_\phi(x, y_w))}_{\text{higher when reward estimate is wrong}} \times \\ & \times \left[ \underbrace{\nabla_\phi \log \pi(y_w | x)}_{\text{Increases likelihood of } y_w} - \underbrace{\nabla_\phi \log \pi(y_\ell | x)}_{\text{Decreases likelihood of } y_\ell} \right] \bigg\} \end{aligned}$$

---

Using  $\nabla \sigma(u) = -\nabla \log(1 + e^{-u}) = \frac{e^{-u}}{1 + e^{-u}} \nabla u = \sigma(-u) \nabla u$

# Performance

Compared with RLHF (PPO), DPO is

- More computationally stable and lightweight
- As effective in aligning with human preferences
- Easier to implement and scale
- Easier to understand

# Summary: What have we learned today?

- RLHF (PPO) is used to align LLM with human preferences
- But can increase hallucinations
- Requires human labeled rankings of responses to prompts
- DPO is a simpler and better approach

