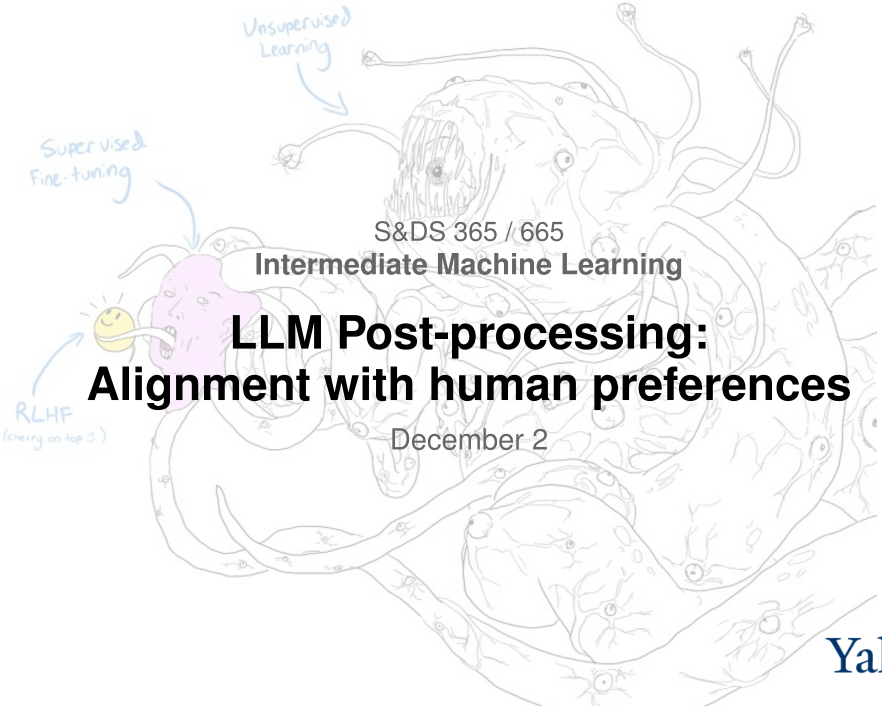


S&DS 365 / 665

Intermediate Machine Learning

LLM Post-processing: Alignment with human preferences

December 2

Yale

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

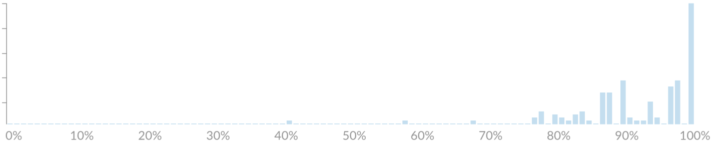
https://registrar.yale.edu/general-information/final-exams

Quiz 5

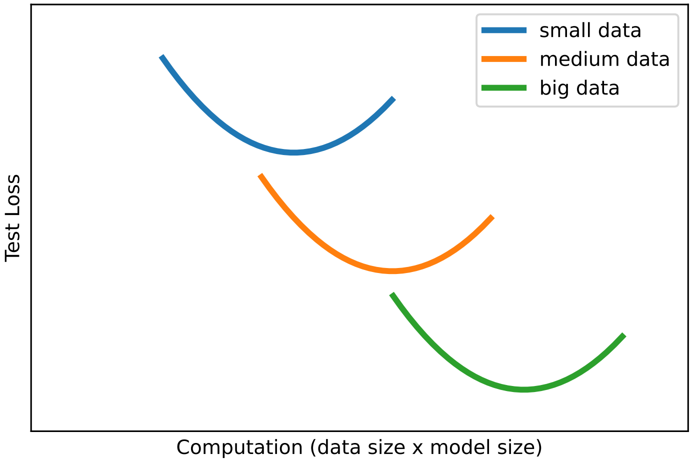
Quiz Summary

Section Filter ▾ | Student Analysis | Item Analysis

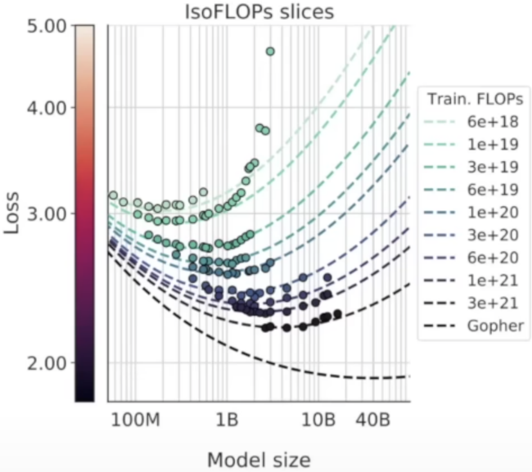
Average Score	High Score	Low Score	Standard Deviation	Average Time
92%	100%	41%	0.89	15:07



Recall: Scaling behavior of LLM models

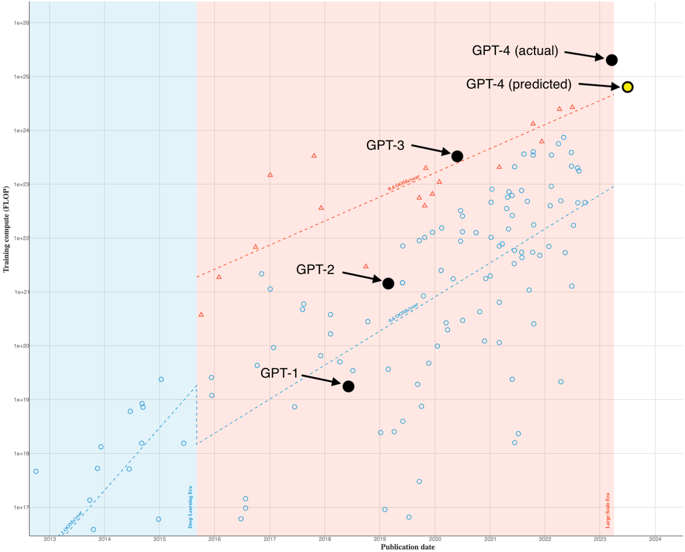


Recall: Scaling behavior of LLM models

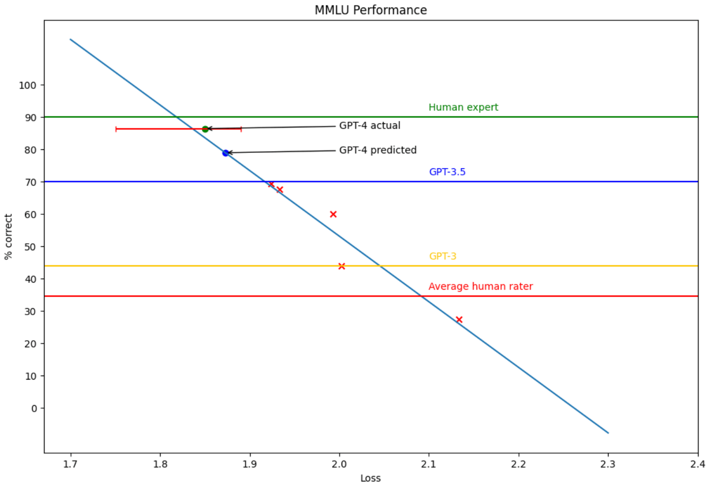


Back of the envelope calculation

- LLM loss on test data is ≈ 1.9 nats/token
- Translates to entropy of about $1.9/\log_e(2) \approx 2.74$ bits/token
- So, *lossless* coding of 500 billion tokens takes ≈ 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: $X\beta_{mn} = Y$



GPT-4 (actual)
GPT-4 (predicted)
GPT-3
GPT-2
GPT-1
0.4 OOMs/year
0.6 OOMs/year
Deep Learning Era
Large Scale Era
Training compute (FLOP)
Publication date
1e+26
1e+25
1e+24
1e+23
1e+22
1e+21
1e+20
1e+19
1e+18
1e+17
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024



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."

---

http://www.incompleteideas.net/IncIdeas/BitterLesson.html

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

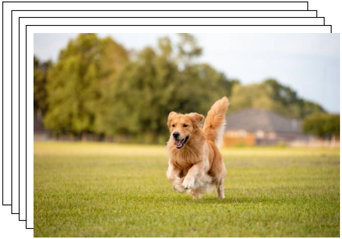
pre-training

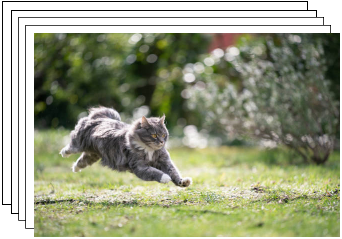
Recall: Machine learning frameworks

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

making LLMs useful

Standard Supervised learning





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

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

Standard Supervised learning

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

“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

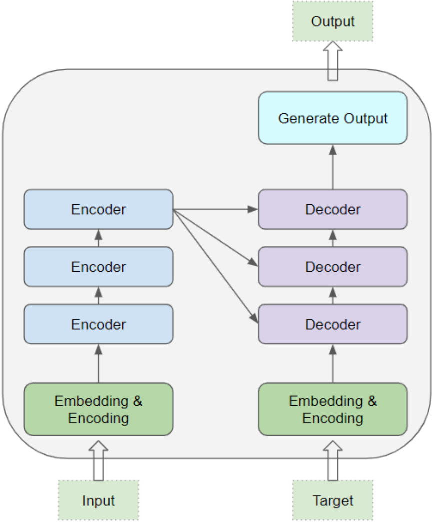
Prompt	Response
Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence.	Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity.
ELI5: What's the cause of the "anxiety lump" in our chest during stressful 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

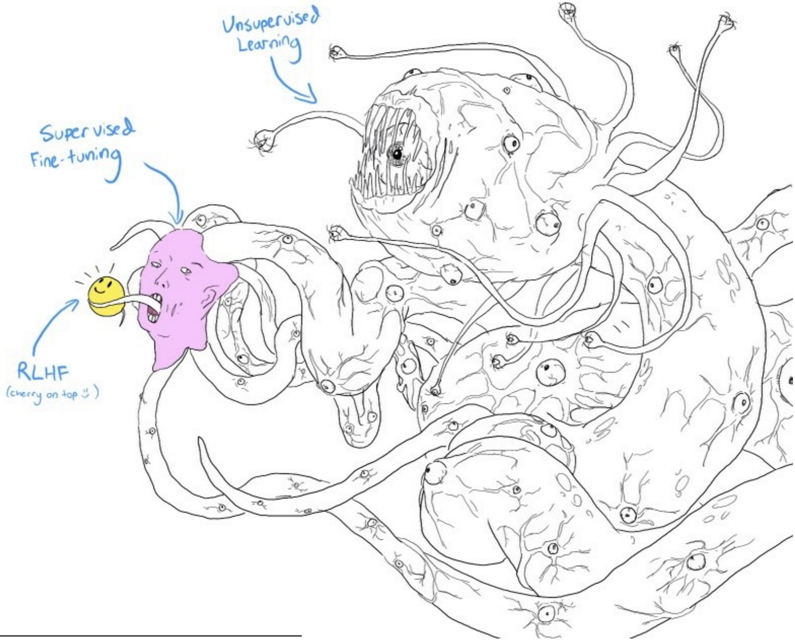
- The LLM is huge (175 billion parameters for GPT-3)
- Finetuning training data are too small to adjust all the parameters
- Special techniques are used to make small changes to the decoder model (e.g. LoRa, low-rank adaptation)

Making the models useful

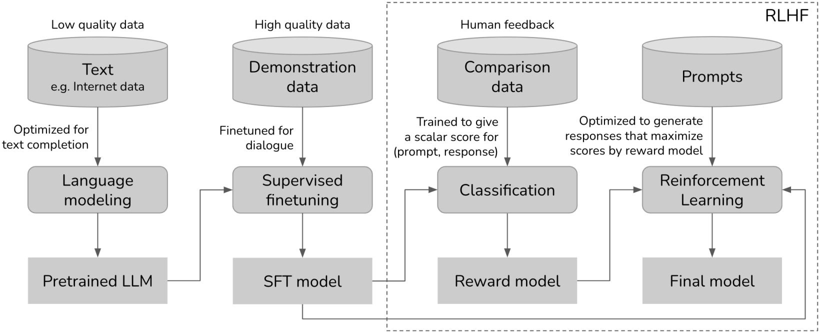
Today: Aligning with human preferences

- 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...

Part of the material for today comes from a blog by Chip Huyen: https://huyenchip.com/2023/05/02/rlhf.html

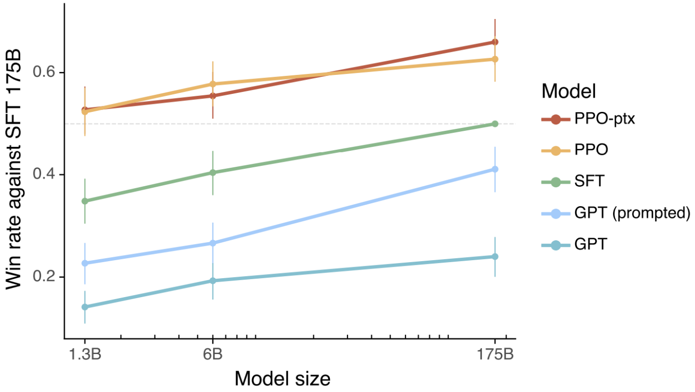


Part of the material for today comes from a blog by Chip Huyen: https://huyenchip.com/2023/05/02/rlhf.html



Scale May '23	**>1 trillion** tokens	**10K - 100K** (prompt, response)	**100K - 1M** comparisons (prompt, winning_response, losing_response)	**10K - 100K** prompts
Examples **Bolded**: open sourced	GPT-x, Gopher, **Falcon**, LLaMa, **Pythia**, **Bloom**, **StableLM**	**Dolly-v2, Falcon-Instruct**		InstructGPT, ChatGPT, Claude, **StableVicuna**

InstructGPT



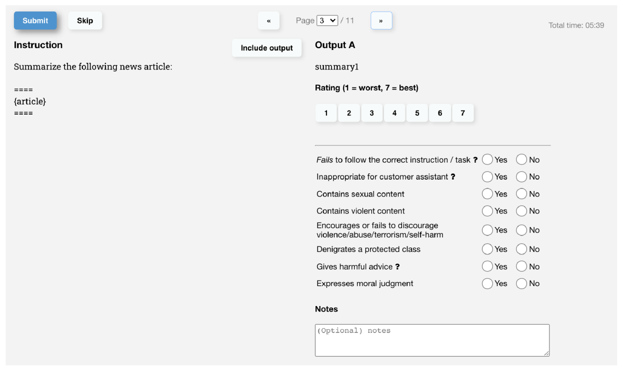
Original GPT-3 + RL was called “InstructGPT” by OpenAI

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

Basic idea of RLHF

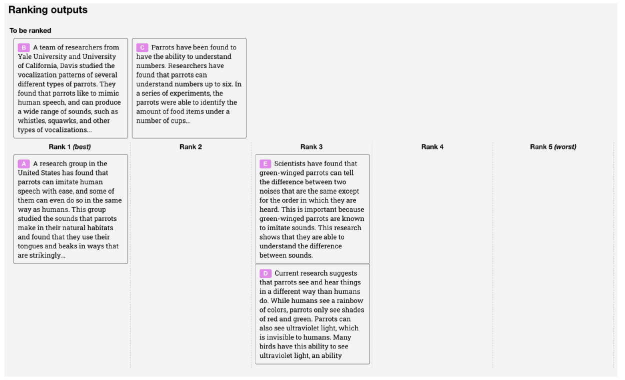
1. Train a reward model to rank candidate responses
   - Uses Bradley-Terry model for pairwise comparisons
2. Use trained reward function in RL to improve LLM
   - Uses version of policy gradients called Proximal Policy Optimization (PPO)

Data collection interface



https://openai.com/index/instruction-following/

Data collection interface



---

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

$$\mathbb{P}(y_w \succ y_\ell \mid x) = \frac{\exp(r^*(x, y_w))}{\exp(r^*(x, y_w)) + \exp r^*(x, y_\ell)}$$

$$= \sigma\left(r^*(x, y_w) - r^*(x, y_\ell)\right)$$

- $\sigma(u) = 1/(1 + e^{-u})$ is sigmoid.
- y_w is "winning response" y_ℓ is "losing response"

R. Bradley and M. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". Biometrika. 39, 1952, 324–345.

Bradley-Terry model

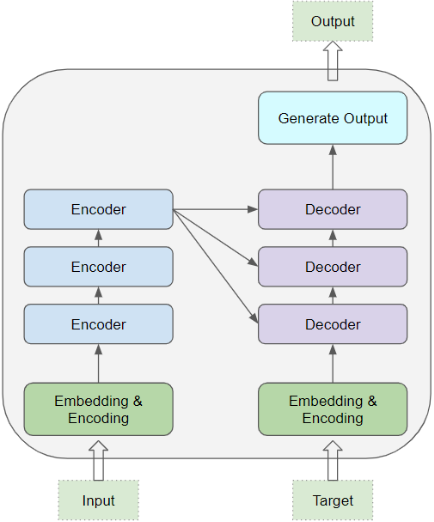
Maximum likelihood training:

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

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

R. Bradley and M. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". Biometrika. 39, 1952, 324–345.

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 θ, after final Transformer layer
- Train θ using maximum likelihood under Bradley-Terry model using human rankings of LLM responses.

SFT="Supervised finetuning"

RLHF

Next step: Use trained reward model r_θ in reinforcement learning to improve "policy" $\pi_{\text{SFT}}(y \mid x)$

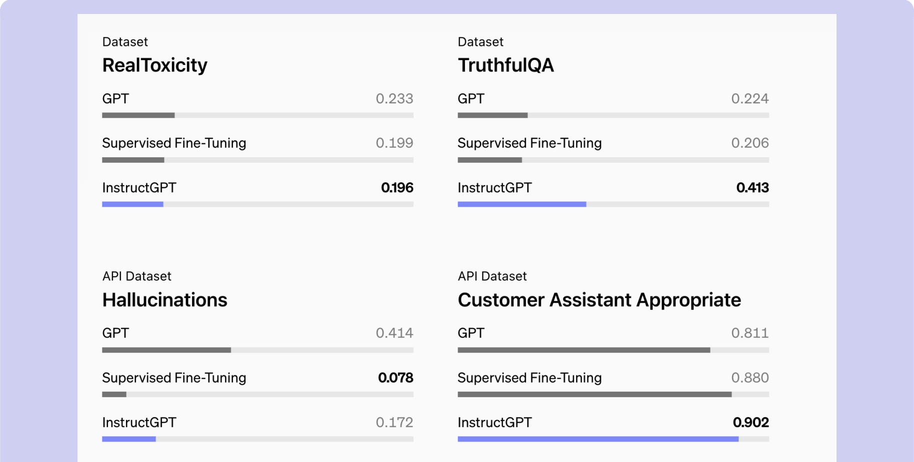
- Actions correspond to tokens to generate
- Parameterized as $\pi_\phi(y \mid x)$

Objective function

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

- 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



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.

https://openai.com/index/instruction-following/

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

---

https://arxiv.org/abs/2305.18290. “Direct Preference Optimization: Your language model is secretly a reward function”, Rafailov et al., 2023

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

https://arxiv.org/abs/2305.18290. "Direct Preference Optimization: Your language model is secretly a reward function", Rafailov et al., 2023

Steps 1-2: Optimize objective nonparametrically

Step 1: Objective is to maximize

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

Step 2: Solution has a closed form:

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

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

Follows from a Lagrange multiplier argument.

Connections to rate-distortion theory and maximum entropy.

Derivation (simplified)

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

Lagrangian is

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

Taking derivatives:

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

Setting to zero and normalizing:

$$q(y \mid x) = \frac{1}{Z(x)} p(y \mid x) \, \exp(\lambda \, r(x, y))$$

Add another Lagrange multiplier for constraint that $\sum_y q(y \mid x) = 1$

Steps 3-4: Invert and plug in

Step 3: Invert (algebra):

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

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

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

Making sense of gradients

Stochastic gradient descent:

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

$$-\mathbb{E}\nabla_\phi \mathcal{L}_{\text{DPO}}(\phi) = \mathbb{E}_{(x,y_w,y_\ell)} \Big\{ \underbrace{\sigma\left(r_\phi(x, y_\ell) - r_\phi(x, y_w)\right)}_{\text{higher when reward estimate is wrong}} \times$$

$$\times \Big[\underbrace{\nabla_\phi \log \pi(y_w \mid x)}_{\text{Increases likelihood of } y_w} - \underbrace{\nabla_\phi \log \pi(y_\ell \mid x)}_{\text{Decreases likelihood of } y_\ell} \Big] \Big\}$$

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