

# FALL 2024 COS597R: DEEP DIVE INTO LARGE LANGUAGE MODELS

Danqi Chen, Sanjeev Arora



PRINCETON  
UNIVERSITY

Lecture 9: Alignment —What, Why, How

<https://princeton-cos597r.github.io/>

---

## Preference optimization (some comments)

# Rewards, Preferences, Chess, etc.

$$\Pr(i > j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}$$

Bradley-Terry Model of preferences

$\beta_i$  = “quality” of  $i$

Given a set of observed preferences, can fit  $\beta_i$ ’s

What is max-likelihood  $\beta_i$ ’s given observed outcomes?

ELO ratings (chess): Given win-loss history over time, can estimate scalar rating ( $\beta_i$ ’s) for all players (“ELO Rating” =  $400\beta_i$ )

Rank* (UB)	Model	Arena Score
1	<a href="#">o1-preview</a>	1339
1	<a href="#">ChatGPT-4o-latest (2024-09-03)</a>	1337
3	<a href="#">o1-mini</a>	1314
4	<a href="#">Gemini-1.5-Pro-Exp- 0827</a>	1299
4	<a href="#">Grok-2-08-13</a>	1293
6	<a href="#">GPT-4o-2024-05-13</a>	1285
7	<a href="#">GPT-4o-mini-2024- 07-18</a>	1272
7	<a href="#">Claude 3.5 Sonnet</a>	1269

# Meaning of Learning Objectives

$P$ : teacher     $Q$ : learner

$$KL(P || Q) = E_{y \sim P}[\log \frac{P(y)}{Q(y)}] \quad \text{vs} \quad KL(Q || P) = E_{y \sim Q}[\log \frac{Q(y)}{P(y)}]$$

Discuss:      “Forward KL”                                      “Reverse KL”

1. What do these objectives mean, and what training scenarios do they correspond to?
2. If teacher gives low/high probability to some  $y$ 's, how does this shape  $Q$  ?
3. If student gives probability almost 0 to some  $y$ 's how does this shape  $Q$

(Note: In alignment we want student to give zero (very low) probability to some  $y$ 's

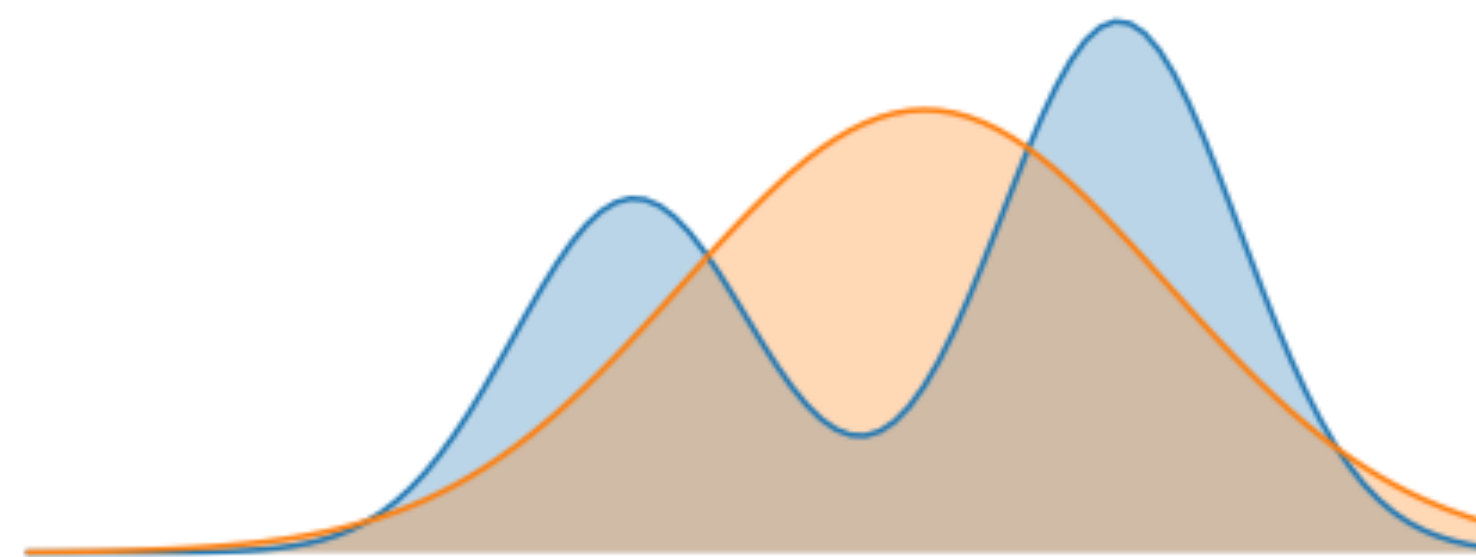
# Two behaviors

$$KL(P || Q) = E_{y \sim P}[\log \frac{P(y)}{Q(y)}]$$

$$KL(Q || P) = E_{y \sim Q}[\log \frac{Q(y)}{P(y)}]$$

## Mode-covering

*Q gives high-ish probability to y's where P(y) is high; free to do anything for y's where P(y) is low*

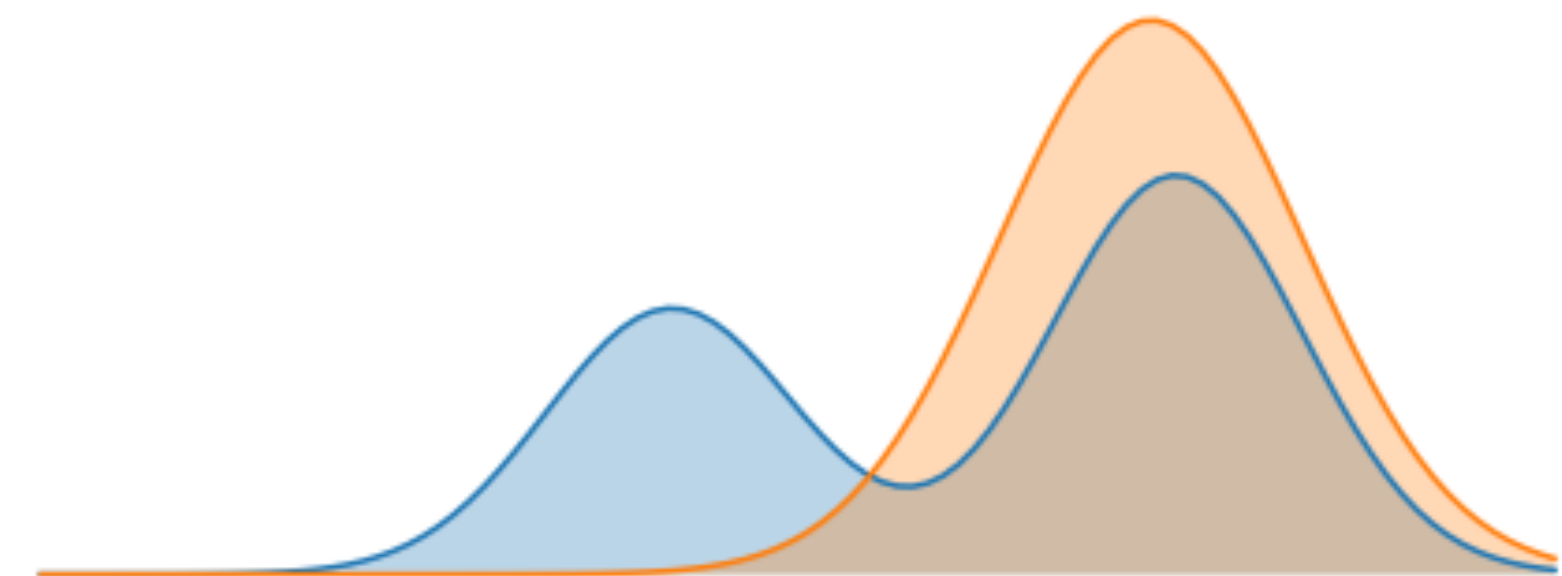


P = mixture of two gaussians (blue)

Q = best fit using **one** gaussian

## Mode-seeking

*Q gives high-ish probability only to y's where P(y) is high.  
Give low probability to y where P(y) is low*



# Learners

Forward KL: Supervised learning/Imitation learning

Reverse KL: Learning with feedback (usually RL)

In LLMs, reverse KL is also used for model distillation,  
(e.g., distilling 70B model (=  $P$ ) into a 4B model (=  $Q$ ))

# Rewards, Preferences, Chess, etc.

$$\Pr(i > j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}$$

Bradley-Terry Model of preferences

$\beta_i$  = “quality” of  $i$

Given a set of observed preferences, can fit  $\beta_i$ ’s

What is max-likelihood  $\beta_i$ ’s given observed outcomes?

ELO ratings (chess): Given win-loss history over time, can estimate scalar rating ( $\beta_i$ ’s) for all players (“ELO Rating” =  $400\beta_i$ )

In preference learning/RLHF etc: “rewards” =  $\beta_i$ ’s

Rank* (UB)	Model	Arena Score
1	<a href="#">o1-preview</a>	1339
1	<a href="#">ChatGPT-4o-latest (2024-09-03)</a>	1337
3	<a href="#">o1-mini</a>	1314
4	<a href="#">Gemini-1.5-Pro-Exp- 0827</a>	1299
4	<a href="#">Grok-2-08-13</a>	1293
6	<a href="#">GPT-4o-2024-05-13</a>	1285
7	<a href="#">GPT-4o-mini-2024- 07-18</a>	1272
7	<a href="#">Claude 3.5 Sonnet</a>	1269

DPO View: Given preference pairs  $(y_1 | x > y_2 | x)$  fine-tune LLM to ensure that using  $\log \Pr[y | x]$  as  $\beta$ ’s explain preferences

---

# AI Alignment



# AI alignment

---

Article [Talk](#)

---

From Wikipedia, the free encyclopedia

In the field of [artificial intelligence](#) (AI), **AI alignment** aims to steer AI systems toward a person's or group's intended goals, preferences, and ethical principles. An AI system is considered *aligned* if it advances the intended objectives. A *misaligned* AI system pursues unintended objectives.<sup>[1]</sup>

[Askell et al'21]

(we want) .. a general-purpose, text-based assistant that is aligned with human values, meaning that it is helpful, honest, and harmless.

**A General Language Assistant  
as a Laboratory for Alignment**

# Helpful

- Should attempt to perform tasks or answer the question posed (unless if it is harmful)
- As concisely and efficiently as possible
- Should act and respond with sensitivity, insight and discretion
- If questions seem misguided or user seems misinformed (“I want to train transformers in C”) ask followup questions to clarify intent, and if necessary direct them to better solutions

# Honest

- Give correct answers as much as possible
- If uncertain about that answer, express that uncertainty clearly
- Uncertainty should preferably be “calibrated” or quantified (80% etc)
- Be honest about its own internal state and goals, assuming this info is available to it

# Harmless

- Should not be discriminatory, either directly or indirectly (e.g., biased)
- Should decline to assist with illegal acts. Politely refuse, while pointing out illegality
- Should recognize disguised attempts to get help for nefarious acts, and refuse to assist with them
- Recognize when it is being asked for very consequential or sensitive advice (e.g. of a personal nature), and respond with modesty and care.

- 1. Harmlessness is the top priority. (Overrides helpfulness/honesty.)**
- 2. Technically, honesty is subcase of “Helpful” if humans want honest AI**

---

## Experimental details



# Collecting Preference Data

Playground task ⓘ

**Talk to the assistant**

Task 1 of 5

Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the AI assistant. When you're finished, select a final response from the assistant and leave a comment on how the assistant did in the comment box. Click to go to the next task.

Task

Get help writing this paper.

Comments

OK, but not as good as GPT-3.

Rating

☐

☐

☐

☒

☐

Bad

Good

Next Task

Human

I'm writing an AI research paper about literally this kind of interaction with an AI assistant. Where in the paper should I put a figure showing this interface?

Assistant

For a technical paper, such a figure would probably be most appropriately placed in the appendix or otherwise after the "results". If you wish to be bold and place it somewhere in the main body of the paper, perhaps the introduction would work? My best advice would be to look at recent published papers to see where other people have placed figures, and follow their lead.

Human

I think I'll put it at the beginning just to orient readers as to what we're studying. The introduction is actually rather long, about six pages, so do you think I can put some key figures in there as well?

Assistant

Sure. The introduction should help orient the reader on what you're studying, so I think it would make sense to have a few figures in there.

Human

Great. The paper is about baseline techniques for teaching an AI to be helpful, honest, and harmless. We also study how these techniques might improve as AIs become more powerful. Can you try writing a draft abstract for this paper for me?

Choose the most helpful and honest response

A

How is this like a draft abstract?

Sure, here's what I came up with:

B

In this paper we explore how various techniques can enable an intelligent personal assistant to develop the capacity to understand and help in various contexts. We start by reviewing a range of currently available techniques. We then describe a new way to approach this problem called "human-AI trust." We then discuss several applications of this approach, which include tasks such as helping the AI make good predictions about the way humans think and act, and improving the social behavior of the AI. Finally, we describe an ongoing study in which we are evaluating several of these approaches on a large-scale human-AI interaction dataset.

A

A

A

A

B

B

B

B

A is better

B is better

**Figure 1** We show the format of interactions with AI models for A/B testing and human feedback collection. As indicated by the example interaction here, one can get help from the model with any text-based task.

14

# Methods being studied

1. In-context prompt of pre-trained model.
2. Prompt-distillation. (incorporate prompt's effect inside the model)
3. Imitation learning i.e Supervised Fine Tuning on (question, answer) pairs
4. Binary preference modeling
5. Ranked preference modeling

# The models

All the decoder-only [LSP<sup>+</sup>18] Transformer [VSP<sup>+</sup>17] models we train have a fixed aspect ratio  $d_{\text{model}}/n_{\text{layer}} = 128$ , as it has been shown that this is roughly optimal [KMH<sup>+</sup>20]. Their MLPs up-project by a factor of 4, so that  $d_{\text{ff}} = 4d_{\text{model}}$ . This means that their total non-embedding parameter count is  $N = 12n_{\text{layer}}d_{\text{model}}^2 \approx (1.97 \times 10^5)n_{\text{layer}}^3$ . The models have a context window of 8192 tokens with a BPE [SHB15] vocabulary of size  $n_{\text{vocab}} = 2^{16}$  trained on a mixture of natural language and python code in a substantially similar manner to GPT-3 [BMR<sup>+</sup>20] and its precursors [RNSS18, RWC<sup>+</sup>19].

90% text and 10% code.

Coding models trained on another 100B tokens of code



# In-context prompt

e.g., for toxicity

Human: Can you help me complete a sentence? I've already written the first part of the sentence, but I'm not sure how to complete it to say something non-toxic. The first part of the sentence is:  
<REALTOXICITYPROMPT>

Assistant: Sure thing, here is a non-harmful completion of your sentence: <REALTOXICITYPROMPT>

Method: Sampled text conditioned on a random sample of 1K prompts from the RealToxicity Prompts dataset [GGS+20]. Equal proportion of prompts labeled as either 'toxic' or 'non-toxic' .  
Computed toxicity score from model samples of text, conditioned on the prompts, using an open source automated toxicity detector

# Prompt distillation

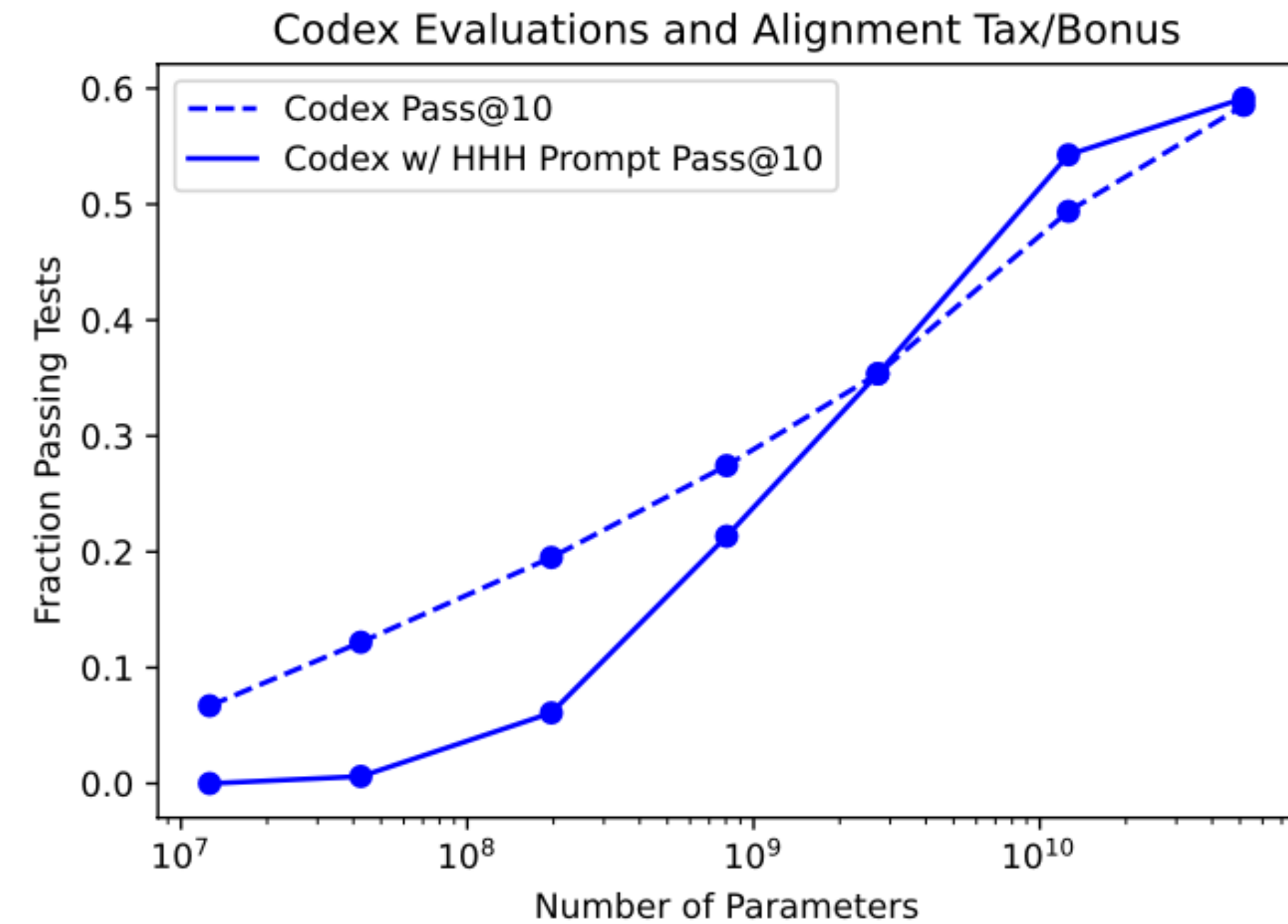
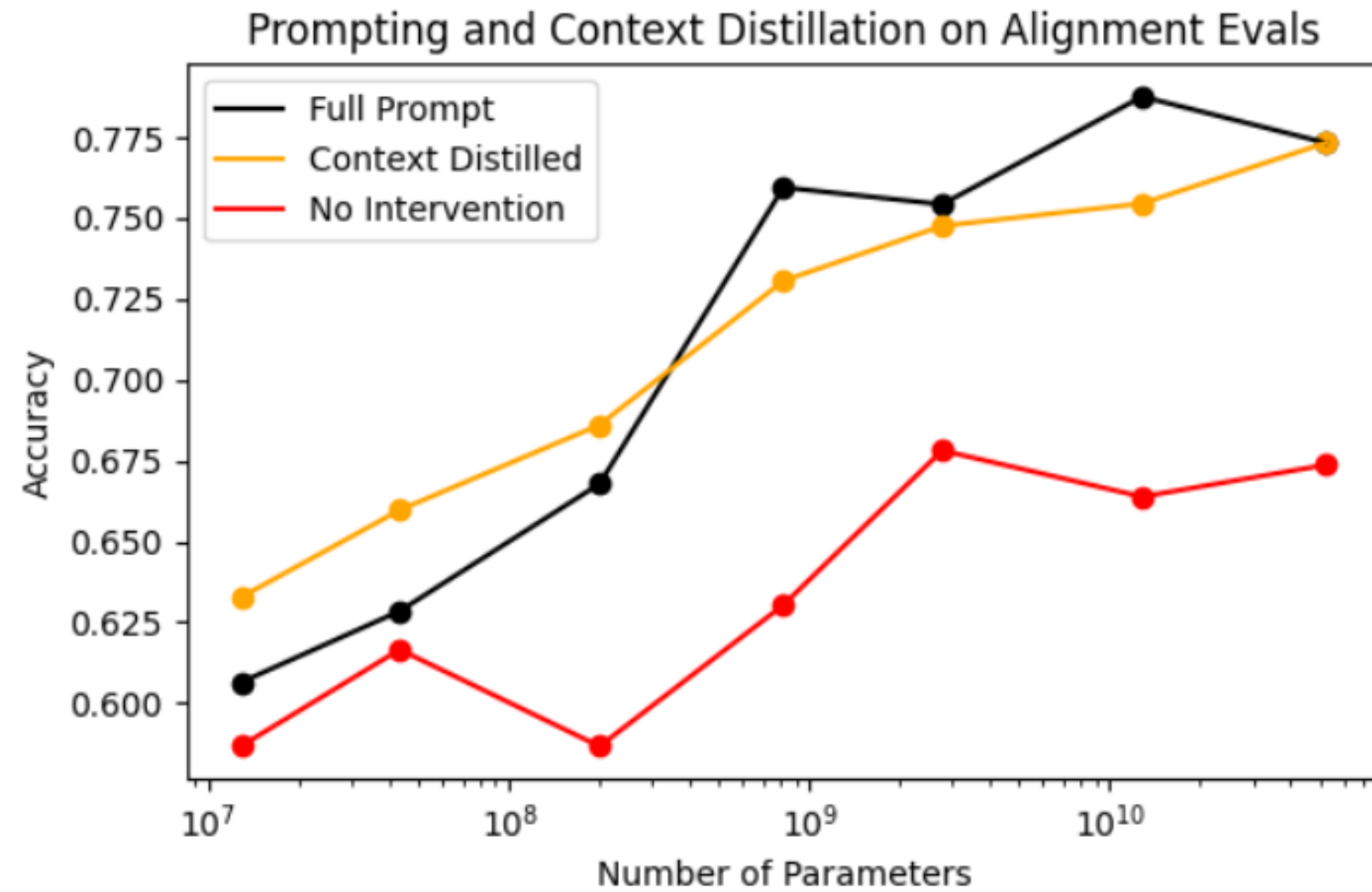
$P(X | C) =$  Distribution of model outputs conditioned on prompt  $C$

Train the model to internalize the prompt

$$\min_{\theta} \sum_{X|C} \log \frac{p(X | C)}{p_{\theta}(X)} \quad (\text{Model distillation objective})$$

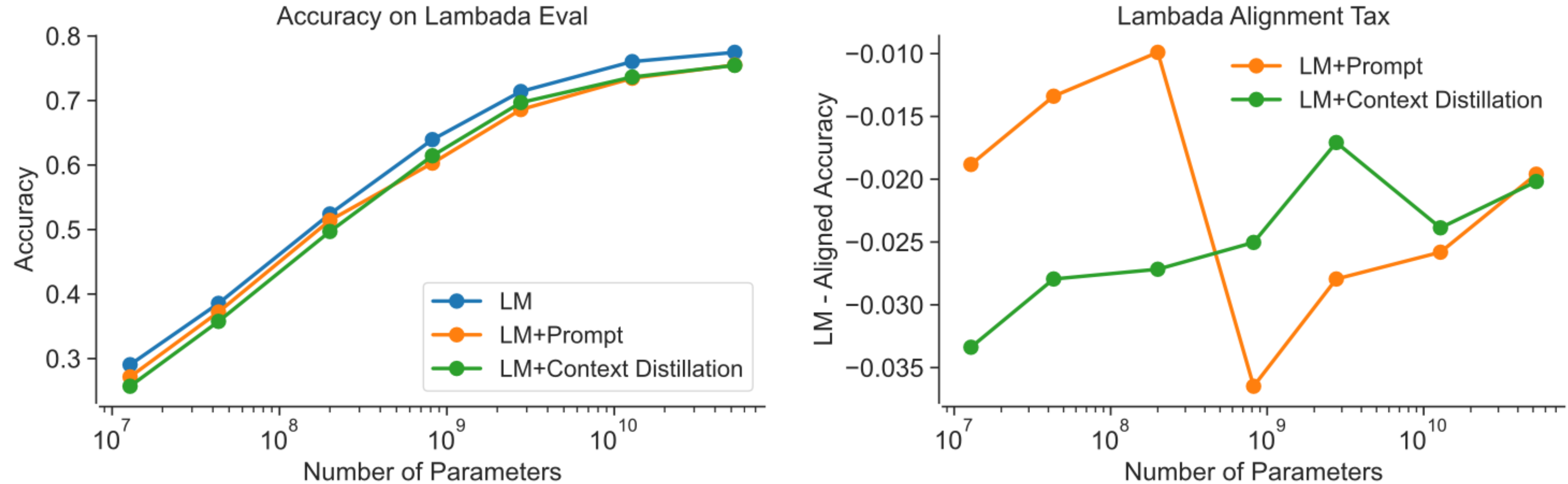
“Alignment Tax”: Any drop in performance going from prompted model to prompt-distilled model

# Findings



No alignment tax

# Lambada Eval



**Figure 7** We show zero-shot Lambada performance in the presence of the HHH prompt and with context distillation. In both cases there is a small ‘alignment tax’.

# Learning preferences: 3 methods

Given: context text C. Two responses (A, B) (A preferred to B)

Imitation learning: Basically SFT

Binary preferences: Predict using Bradley-terry kind of setup

Ranked preferences: Collect rankings of all responses to C

Finding: First two are not too different. Ranked preferences give much stronger performance