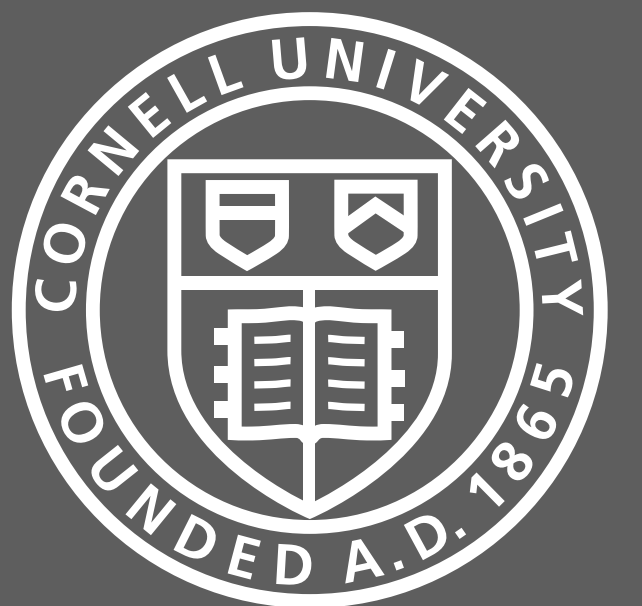


Ranking Models Beyond Voting

Paul Gözl

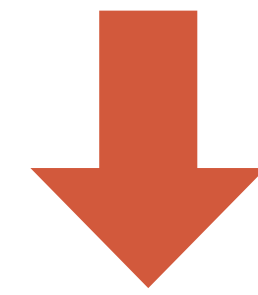
Lecture 8, ORIE 4360/5360, Fall '26



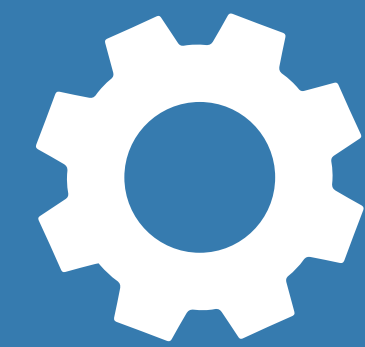
Recap: Generative Models

2

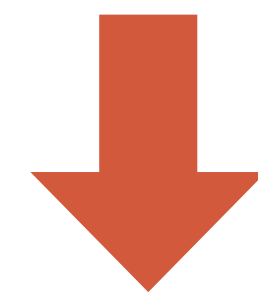
Latent parameter: true ranking π .



Generative
model



Defines, for each
observation y ,
likelihood $\mathbb{P}[y \mid \pi]$.



Observation y from voter, e.g., reported ranking σ_i .

Recap: MLE

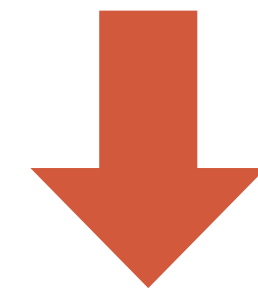
Given observations $y_1, \dots, y_n$, we estimate π through the **maximum likelihood estimator**:

$$\hat{\pi} = \operatorname{argmax}_{\pi} \prod_{i=1}^n \mathbb{P}[y_i \mid \pi] = \operatorname{argmax}_{\pi} \sum_{i=1}^n \log \mathbb{P}[y_i \mid \pi].$$

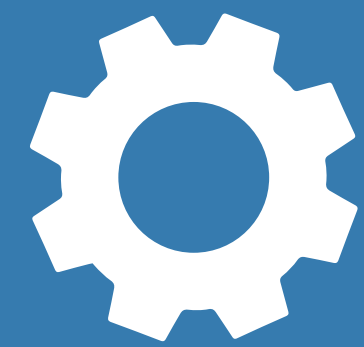
Today: New Generative Agent Model

4

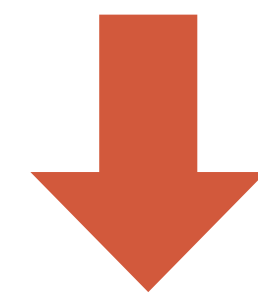
Unobservable parameter: alternative scores $\mathbf{s} \in \mathbb{R}_{>0}^m$.



Generative
model



Defines, for each
observation y ,
likelihood $\mathbb{P}[y \mid \mathbf{s}]$.



Observation y from agent: pairwise comparisons.

The Bradley-Terry Model

Let $s_x > 0$ be the **score** of alternative x .

Agent (independently) compares alternatives x, y ; prefers x with probability $\frac{s_x}{s_x + s_y}$.

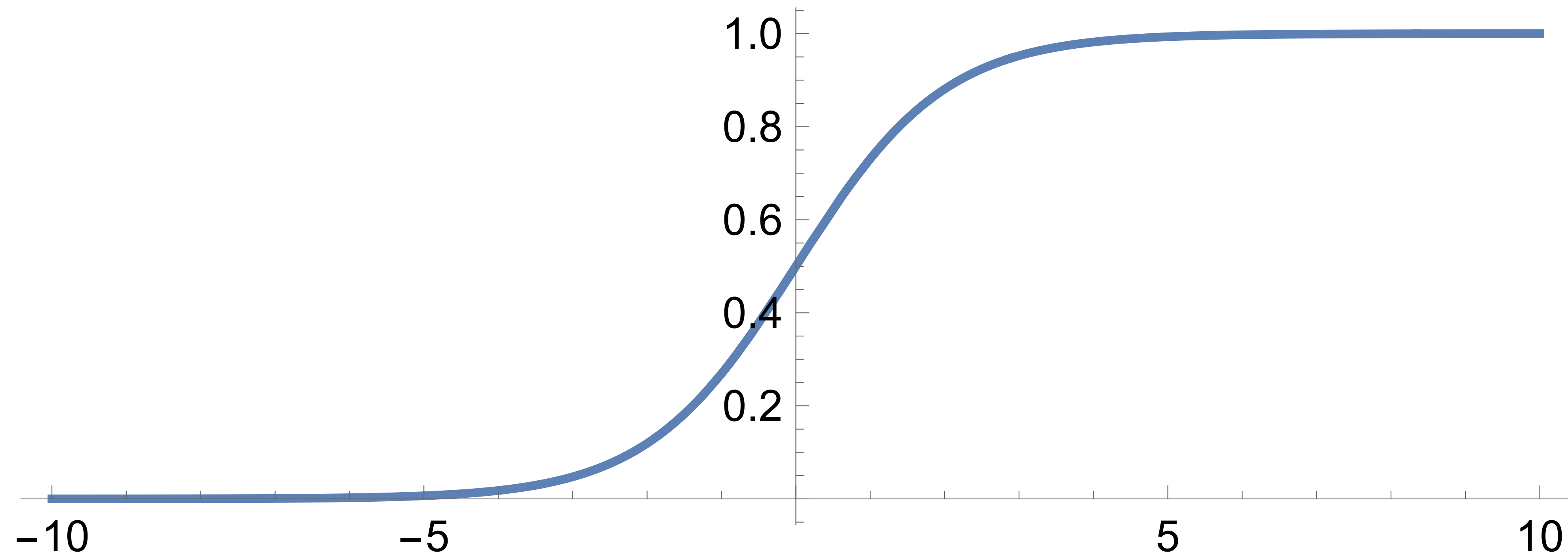
We rewrite this model by setting $s_x = e^{\beta_x}$. Then,

$$\mathbb{P}[x \succ y] = \frac{\exp(\beta_x)}{\exp(\beta_x) + \exp(\beta_y)} = \frac{1}{1 + \exp(-(\beta_x - \beta_y))} = \sigma(\beta_x - \beta_y),$$

where $\sigma : \mathbb{R} \rightarrow [0, 1], z \mapsto 1/(1 + e^{-z})$ is the **logistic sigmoid**.

The Logistic Sigmoid Function

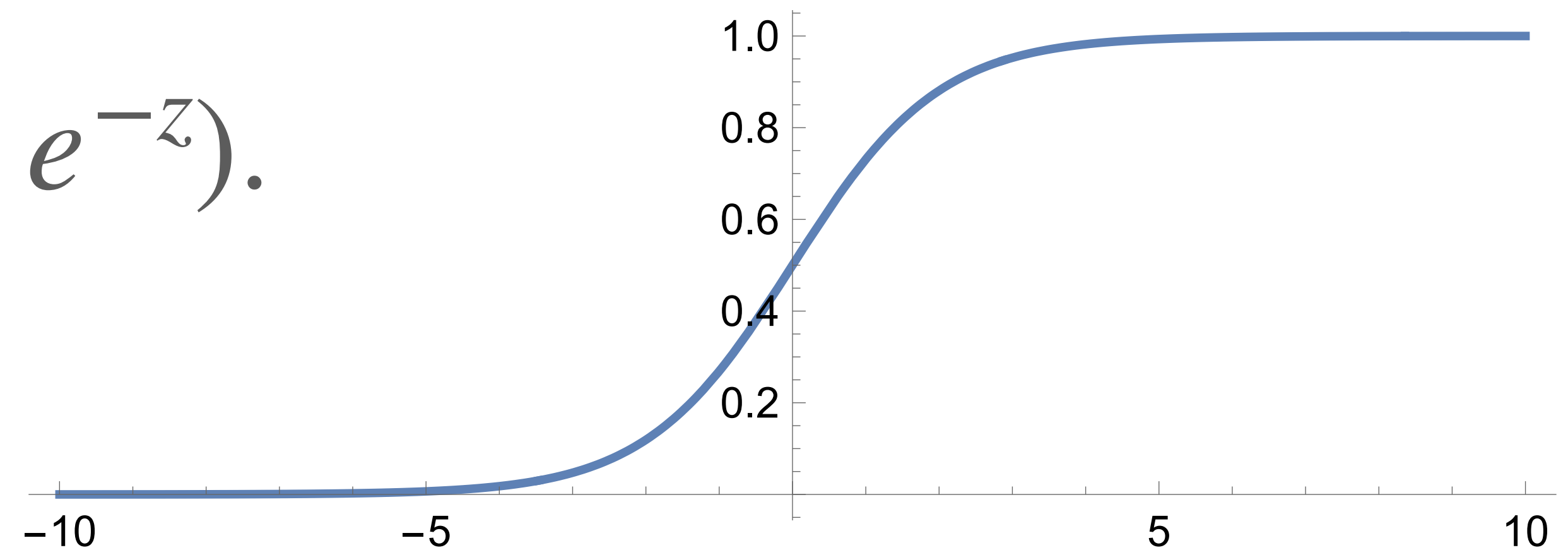
6



$$\sigma : \mathbb{R} \rightarrow [0,1], z \mapsto 1/(1 + e^{-z}).$$

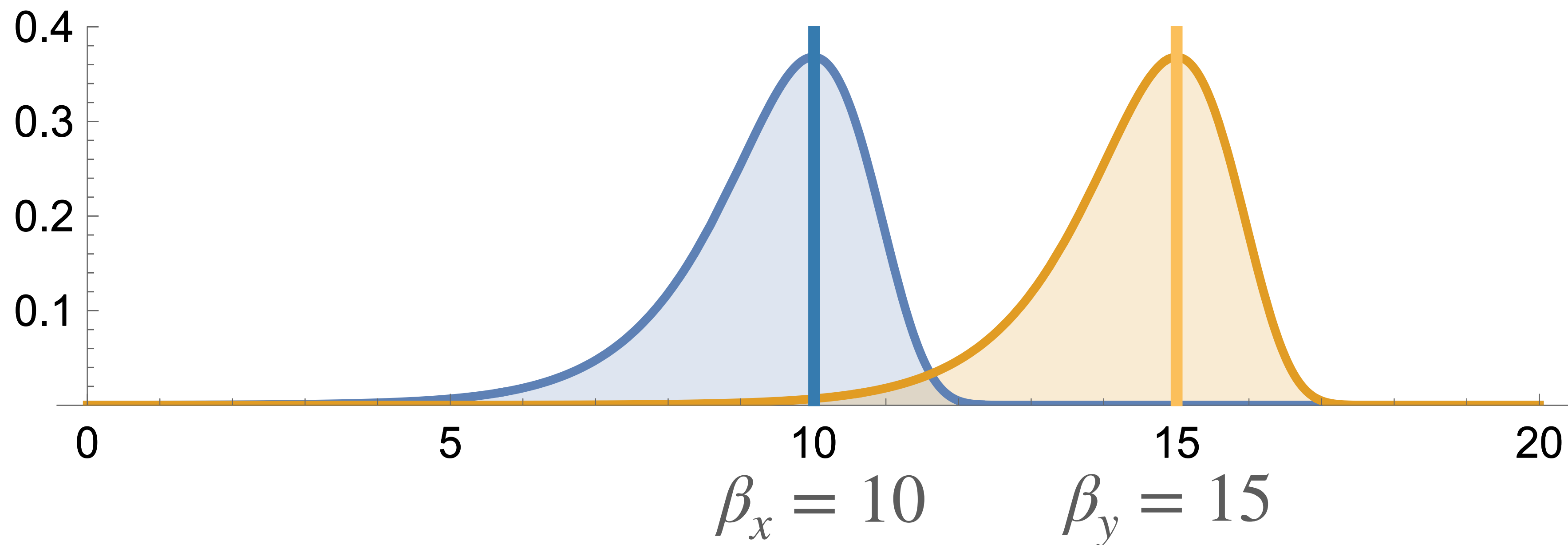
Useful Properties of the Sigmoid

$$\sigma : \mathbb{R} \rightarrow [0,1], z \mapsto 1/(1 + e^{-z}).$$



- Q: ▶ Can you show that $\sigma(z) = 1 - \sigma(-z)$?
- ▶ Can you show that $\sigma(z)/\sigma(-z) = e^z$?
- ▶ Can you derive $\frac{d}{dz} \log \sigma(z)$?

One Justification for Bradley-Terry



True quality β_x but agent perceives with noise
 $\hat{\beta}_x \sim \text{Gumbel}(\beta_x, 1)$.

Then, $\mathbb{P}[\hat{\beta}_x > \hat{\beta}_y] = \sigma(\beta_x - \beta_y)$.

Applications of Bradley-Terry Model

9



Which product will a consumer buy? Generally, drawn towards better product/lower price, but might pick a worse one with some probability.

Applications of Bradley-Terry Model

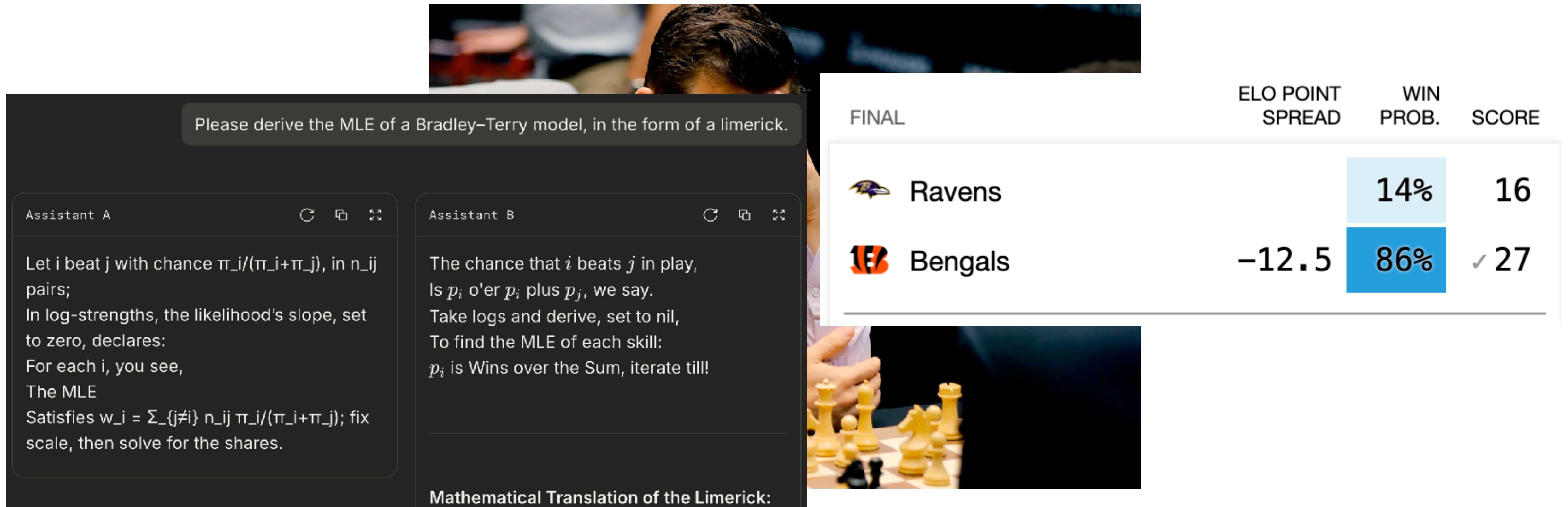
10



Which chess player will win? Generally, the better one, but sometimes the underdog wins.

Applications of Bradley-Terry Model

11



Please derive the MLE of a Bradley-Terry model, in the form of a limerick.

Assistant A

Let i beat j with chance $\pi_i/(\pi_i+\pi_j)$, in n_{ij} pairs;
In log-strengths, the likelihood's slope, set to zero, declares:
For each i , you see,
The MLE
Satisfies $w_i = \sum_{\{j \neq i\}} n_{ij} \pi_j / (\pi_i + \pi_j)$; fix scale, then solve for the shares.

Assistant B

The chance that i beats j in play,
Is p_i o'er p_i plus p_j , we say.
Take logs and derive, set to nil,
To find the MLE of each skill:
 p_i is Wins over the Sum, iterate till!

Mathematical Translation of the Limerick:

FINAL	ELO POINT SPREAD	WIN PROB.	SCORE
 Ravens		14%	16
 Bengals	-12.5	86%	✓ 27

Same for competitions between LLMs (arena.ai), NFL teams (FiveThirtyEight), ...

Likelihood for Bradley-Terry Model

In our observation, let $n_{x,y}$ denote how often x was preferred to y . Then, the likelihood for $\mathbf{s}$ is

$\prod_{x,y} \sigma(\beta_x - \beta_y)^{n_{x,y}}$ and the log-likelihood is $\sum_{x,y} n_{x,y} \log \sigma(\beta_x - \beta_y)$.

MLE estimates the strengths β_x of alternatives.

Elo Scores

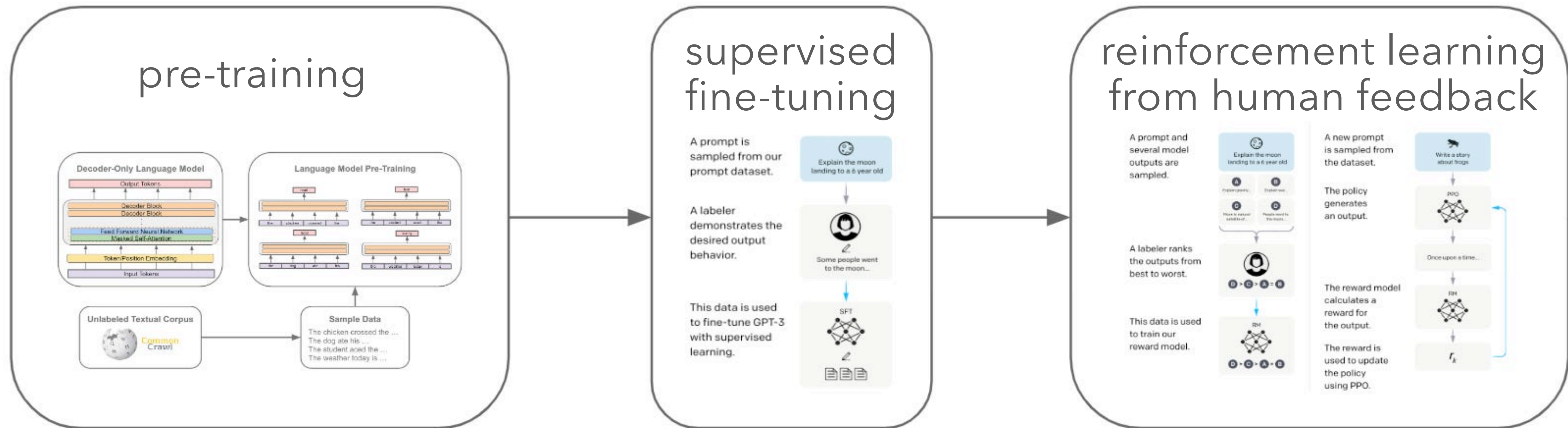
For MLE, we would want to observe the outcomes of all games, then maximize the log-likelihood $\sum_{x,y} n_{x,y} \log \sigma(\beta_x - \beta_y)$.

To estimate strengths in sports/chess, this is inconvenient, would like to slowly evolve scores over time.

Idea: optimize via stochastic gradient descent.
[Blackboard]

RLHF in LLM Development

14

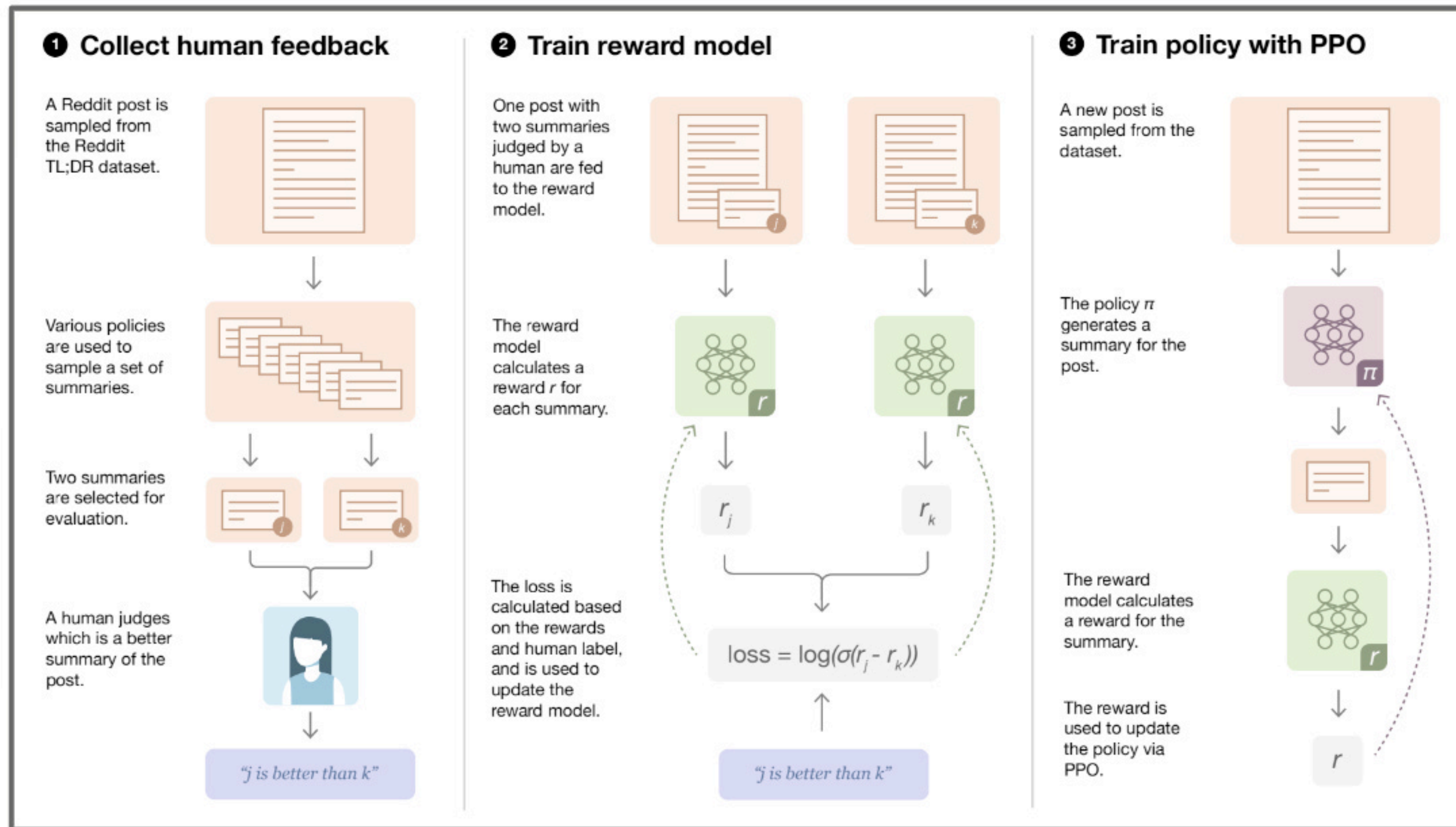


nowadays, part of a larger
post-training stage

RLHF in LLM Development

15

“Reinforcement Learning from Human Feedback”.



Training the Reward Model

16

Annotators compare outputs for many prompts (e.g., summaries for many texts). For theoretical tractability, pretend that there is only one prompt.

Familiar setting: Responses are the alternatives.

Assume that alternative x has a true reward β_x , and agents order based on Bradley-Terry.

Rewards are estimated with MLE:

$$\hat{\beta} = \operatorname{argmax}_{\beta} \sum_{x,y} n_{x,y} \log \sigma(\beta_x - \beta_y). \leftarrow \text{concave!}$$

Heterogenous Preferences?

17

What will MLE give if Bradley-Terry assumption is wrong? I.e., if agents have heterogeneous preferences, not just noise around “true” reward.

Theorem [Anderson et al. 2009]. If each agent ranks all pairs (not following any generative model), the $\hat{\beta}_x$ calculated as Bradley-Terry MLE are ordered in the same way as the alternatives' Borda scores.

Theorem [Anderson et al. 2009]. If each agent ranks all pairs (not following any generative model), the $\hat{\beta}_x$ calculated as Bradley-Terry MLE are ordered in the same way as the alternatives' Borda scores.

$$\triangleright \sigma(z) = 1 - \sigma(-z).$$

Useful properties: $\triangleright \sigma(z)/\sigma(-z) = e^z.$

$$\triangleright \frac{d}{dz} \log \sigma(z) = \sigma(-z).$$

Implications?

19

Theorem [Anderson et al. 2009]. If each agent ranks all pairs (not following any generative model), the $\hat{\beta}_x$ calculated as Bradley-Terry MLE are ordered in the same way as the alternatives' Borda scores.

Q: Based on what we know about the Borda score, what concerns does this result raise?

Bibliography

Anderson et al. (2009): *Relationships between Borda Voting and Zermelo Ranking*. Social Choice and Welfare, 2009.

Recently rediscovered in RLHF setting by:

Siththaranjan, Cassidy & Hadfield-Menell:
Distributional Preference Learning. ICLR 2024
([link](#)).