Alignment in Large Language Models: Statistical and Game-Theoretic Perspectives

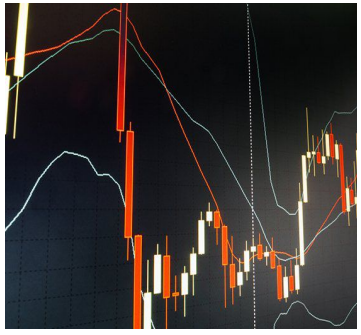
Weijie Su

Wharton, University of Pennsylvania

A brief history of statistics

Experimental design

High-dimensional statistics

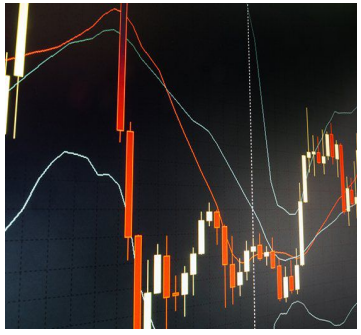


Time series

A brief history of statistics

Experimental design

High-dimensional statistics



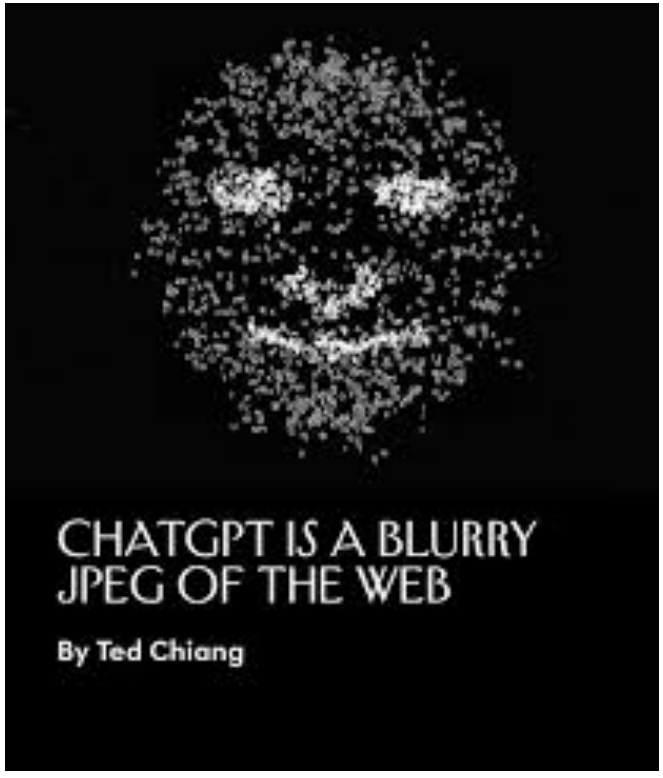
Time series

Statistics repeatedly reinvents itself when a new data modality emerges

A story about an octopus

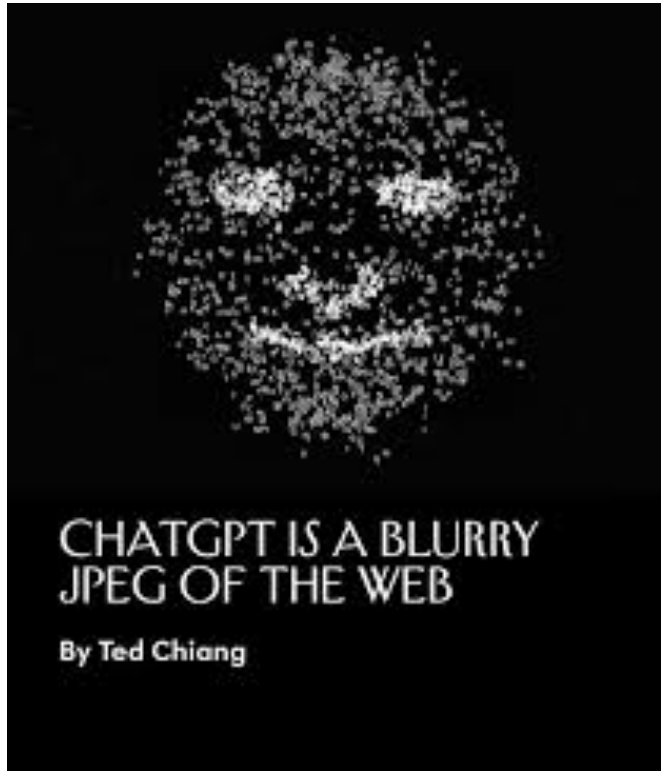
- The modern LLM paradigm starts from *next-token prediction*
- At first, this seemed almost too simple to support rich language understanding
- Yet scale and data turned this simple objective into a remarkably powerful engine

The Bitter Lesson in AI and why it is a blessing for statisticians



- The *bitter lesson*: scalable computation often beats handcrafted structure

The Bitter Lesson in AI and why it is a blessing for statisticians



- The *bitter lesson*: scalable computation often beats handcrafted structure
- So the opportunity is not to fight scale, but to build statistical principles that survive scale

LLMs do need statistical foundations

The Annals of Applied Statistics
2026, Vol. 20, No. 1, 724–743
https://doi.org/10.1214/26-AOAS2151

DO LARGE LANGUAGE MODELS (REALLY) NEED STATISTICAL FOUNDATIONS?

BY WEIJIE SU[a]

Department of Statistics and Data Science, University of Pennsylvania, [a]*suw@wharton.upenn.edu*

Large language models (LLMs) represent a new paradigm for processing unstructured data, with applications across an unprecedented range of domains. In this paper we address, through two arguments, whether the development and application of LLMs would genuinely benefit from foundational contributions from the statistics discipline. First, we argue affirmatively, beginning with the observation that LLMs are inherently statistical models due to their profound data dependency and stochastic generation processes, where statistical insights are naturally essential for handling variability and uncertainty. Second, we argue that the persistent black-box nature of LLMs—stemming from their immense scale, architectural complexity, and development practices often prioritizing empirical performance over theoretical interpretability—renders closed-form or purely mechanistic analyses gener-

LLMs do need statistical foundations

The Annals of Applied Statistics
2026, Vol. 20, No. 1, 724–743
https://doi.org/10.1214/26-AOAS2151

DO LARGE LANGUAGE MODELS (REALLY) NEED STATISTICAL FOUNDATIONS?

BY WEIJIE SU[a]

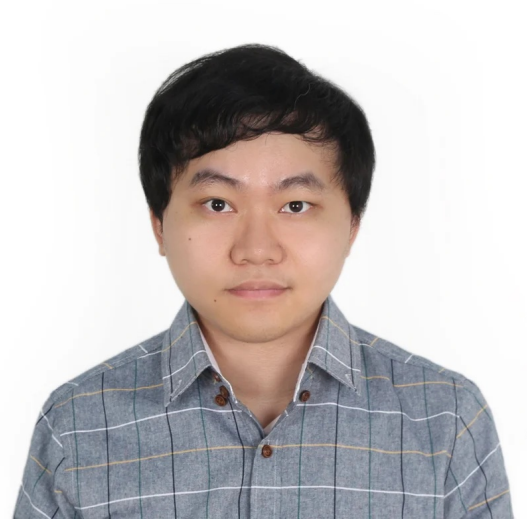
Department of Statistics and Data Science, University of Pennsylvania, [a]*suw@wharton.upenn.edu*

Large language models (LLMs) represent a new paradigm for processing unstructured data, with applications across an unprecedented range of domains. In this paper we address, through two arguments, whether the development and application of LLMs would genuinely benefit from foundational contributions from the statistics discipline. First, we argue affirmatively, beginning with the observation that LLMs are inherently statistical models due to their profound data dependency and stochastic generation processes, where statistical insights are naturally essential for handling variability and uncertainty. Second, we argue that the persistent black-box nature of LLMs—stemming from their immense scale, architectural complexity, and development practices often prioritizing empirical performance over theoretical interpretability—renders closed-form or purely mechanistic analyses gener-

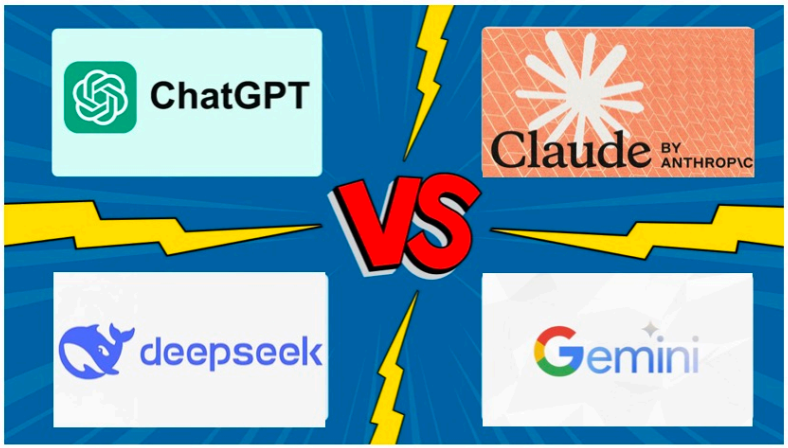
- Alignment, uncertainty quantification, hallucination, watermarking, evaluation...

Collaborators

- **Jiancong Xiao (Penn)**
- Ziniu Li (CUHK-Shenzhen)
- Xingyu Xie (NUS)
- Emily Getzen (Apple)
- Cong Fang (Peking University)
- Qi Long (Penn)
- Kaizhao Liu (MIT)
- Zhekun Shi (Princeton)
- Bohan Zhang (Peking University)



Diverse human preference

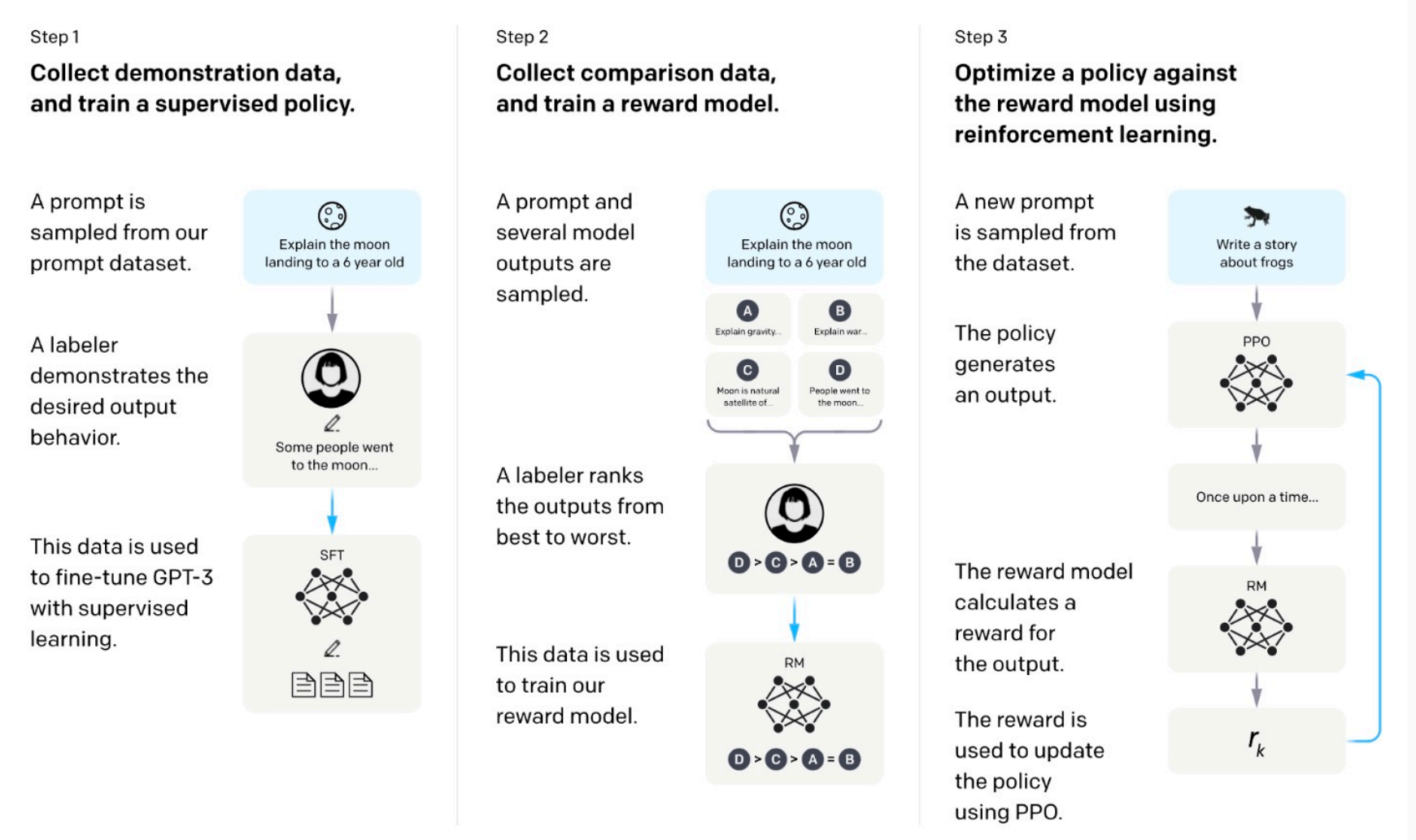


Can large language models (LLMs) faithfully represent or align with diverse human preference?

Diverse human preference

- It's personal choice but important: coffee vs tea, football vs basketball
- Annotator–annotator and annotator–researcher agreement rates ranging from 63% to 77% [Ouyang et al., 2022]

Preliminaries: Reinforcement learning from human feedback (RLHF) [Ouyang et al., 2022]



RLHF

Step 1: Supervised Fine-Tuning (SFT)

- MLE from prompt-answer (x, y) pairs

$$\min_{\phi} -\mathbb{E}_{x,y}[\log \pi_\phi(y|x)]$$

- Gives a reference model π_{ref}

RLHF

Step 1: Supervised Fine-Tuning (SFT)

- MLE from prompt-answer (x, y) pairs

$$\min_{\phi} -\mathbb{E}_{x,y}[\log \pi_\phi(y|x)]$$

- Gives a reference model π_{ref}

Step 2: Reward Modeling

- BT model [Bradley and Terry, 1952]

$$\mathcal{P}(y_1 \succ y_2|x) = \frac{\exp(r(x, y_1))}{\exp(r(x, y_1)) + \exp(r(x, y_2))}$$

- Learn a reward model r by MLE

$$\min_{\theta} -\mathbb{E}_{x,y_w,y_l} \log \sigma \left(r_\theta(x, y_w) - r_\theta(x_i, y_l)\right)$$

RLHF

Step 1: Supervised Fine-Tuning (SFT)

- MLE from prompt-answer (x, y) pairs

$$\min_{\phi} -\mathbb{E}_{x,y}[\log \pi_\phi(y|x)]$$

- Gives a reference model π_{ref}

Step 2: Reward Modeling

- BT model [Bradley and Terry, 1952]

$$\mathcal{P}(y_1 \succ y_2|x) = \frac{\exp(r(x, y_1))}{\exp(r(x, y_1)) + \exp(r(x, y_2))}$$

- Learn a reward model r by MLE

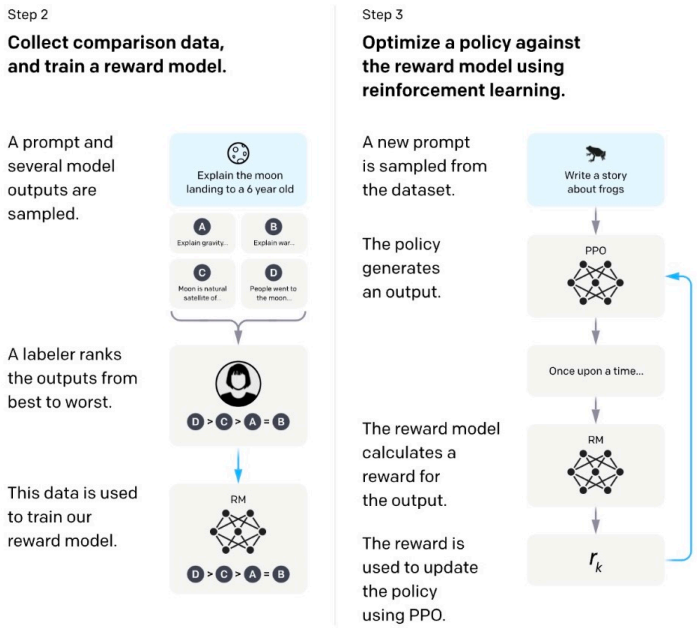
$$\min_{\theta} -\mathbb{E}_{x,y_w,y_l} \log \sigma \left(r_\theta(x, y_w) - r_\theta(x_i, y_l)\right)$$

Step 3: RLHF Fine-tuning

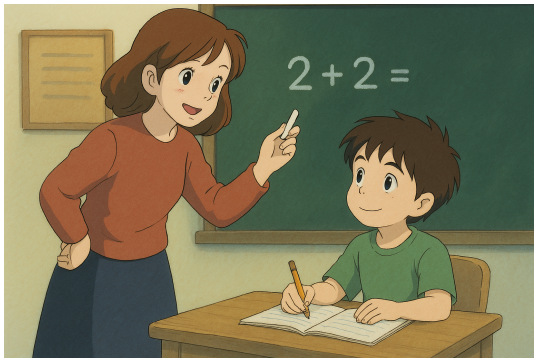
$$\max_{\phi} \mathbb{E}_{y\sim\pi_\phi(y|x)} r(x, y) - \beta \cdot D_{\text{KL}}(\pi_\phi(y|x) \| \pi_{\text{ref}}(y|x))$$

Two steps warrant statistical thinking

Good teacher

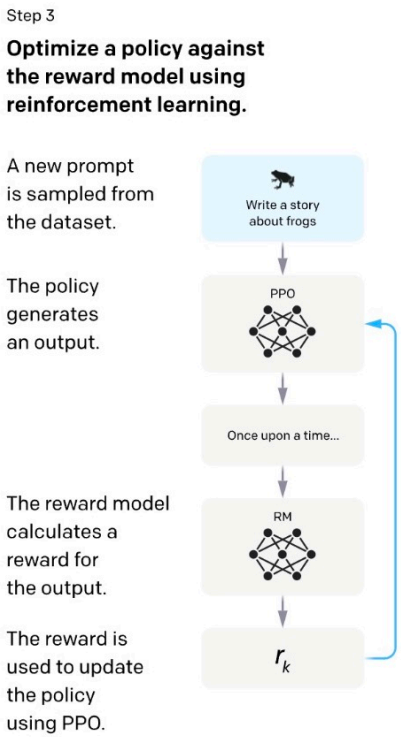


Preference transfer

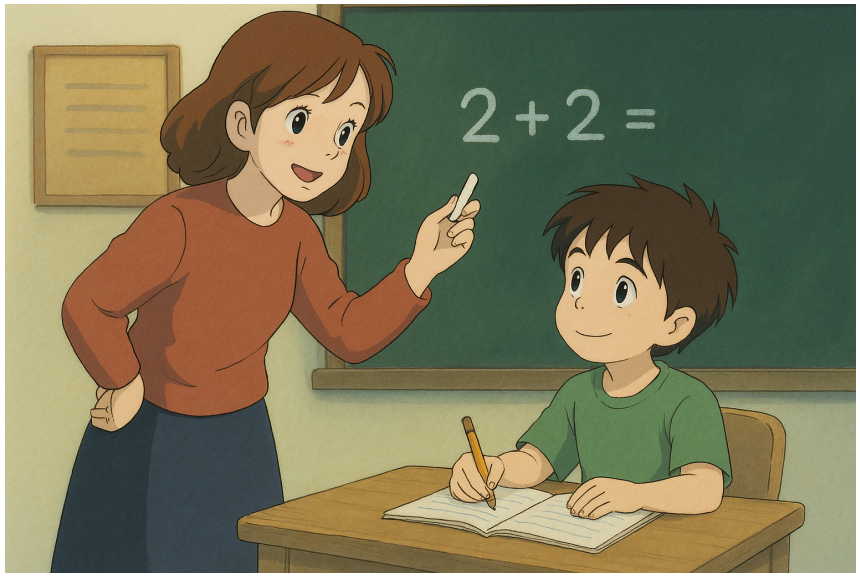


- Step 2 asks whether the reward model is a *good teacher*
- Step 3 asks whether optimization yields faithful *preference transfer*

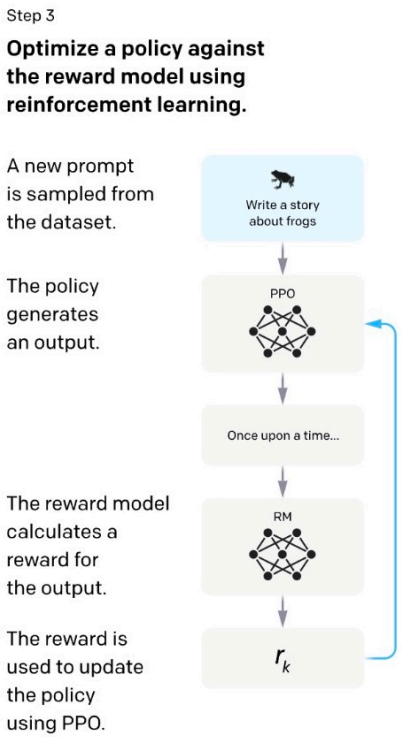
Is OpenAI's Step 3 statistically sound?



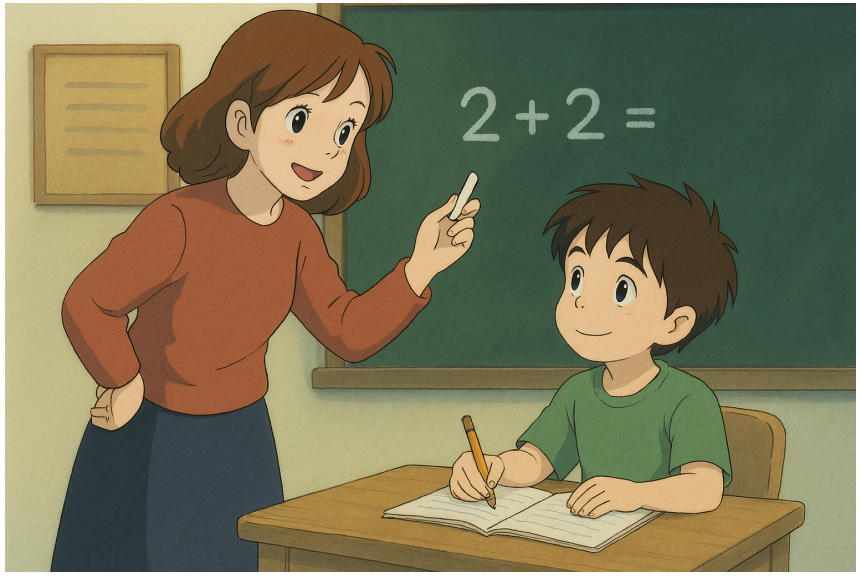
Preference transfer



Is OpenAI's Step 3 statistically sound?



Preference transfer

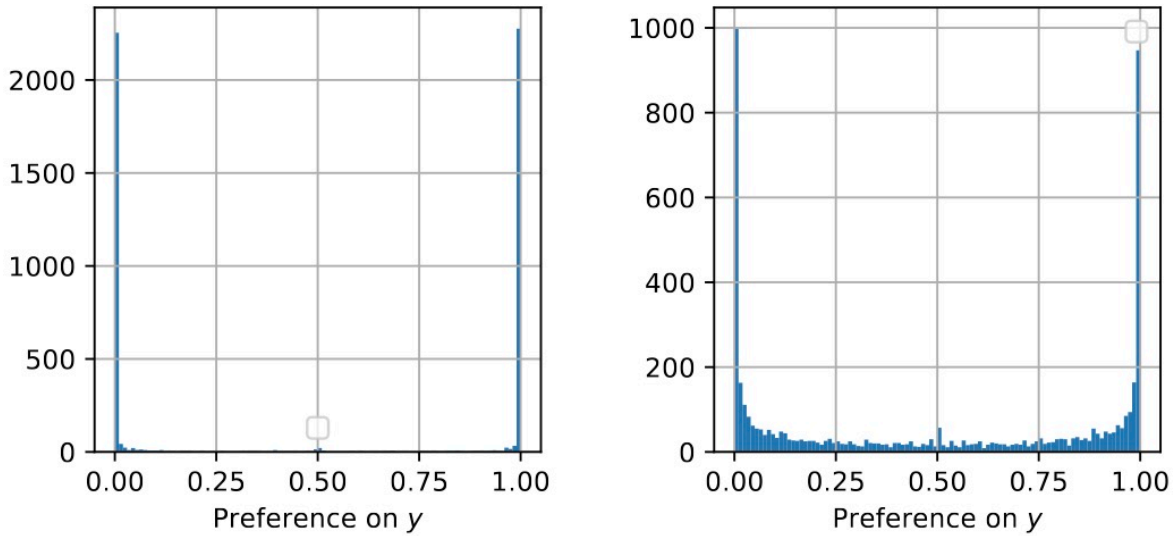


Question

Does reward maximization preserve diverse human preferences?

A simple experiment on Llama-7B

- Given a prompt x, report LLM/reward model's preference on y (vs y')
- Repeat multiple times



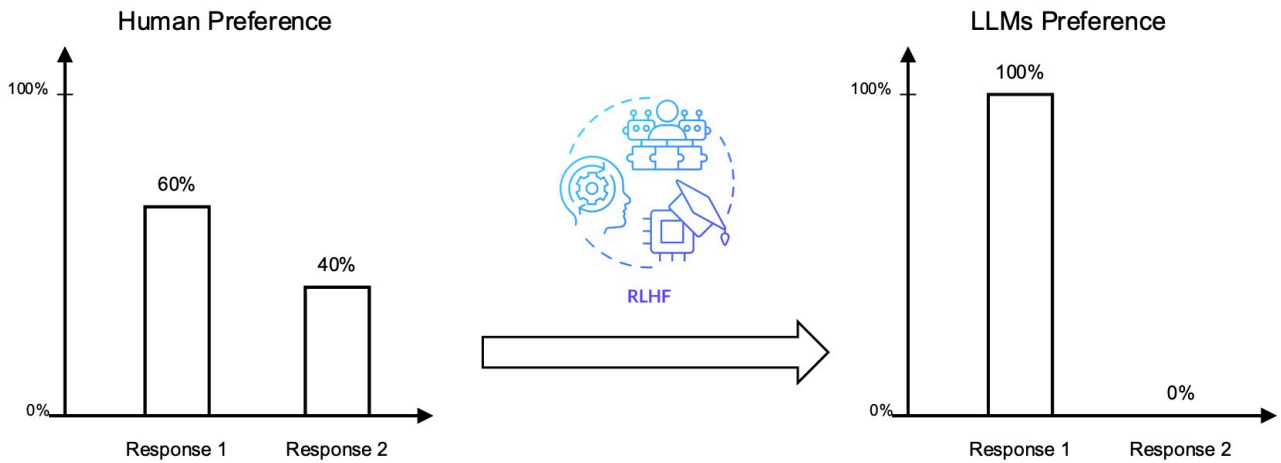
(a) LLMs (b) Reward Models

- Preference of LLM (*student*): 0 or 1
- Preference of reward model (*teacher*): smooth

Algorithmic bias of RLHF

A practical observation:

- $\pi_\phi(y|x)$ will strongly favor y_1, with a probability close to 100%
- Preference of the minority are effectively disregarded



- Call this phenomenon *preference collapse*

Tyranny of the majority

Even with perfect reward models, LLMs may not align with human preferences

Tyranny of the majority

Even with perfect reward models, LLMs may not align with human preferences

Unregularized

$$\max_{\pi_\phi} \mathbb{E}_{x,y}\big[r(x,y)\big], \qquad y \sim \pi_\phi(y \mid x)$$

Tyranny of the majority

Even with perfect reward models, LLMs may not align with human preferences

Unregularized

$$\max_{\pi_\phi} \mathbb{E}_{x,y}\big[r(x,y)\big], \qquad y \sim \pi_\phi(y \mid x)$$

If $y^\star$ has the highest reward, the optimal policy outputs $y^\star$ with probability essentially 1

Tyranny of the majority

Even with perfect reward models, LLMs may not align with human preferences

Unregularized

$$\max_{\pi_\phi} \mathbb{E}_{x,y}\big[r(x,y)\big], \qquad y \sim \pi_\phi(y \mid x)$$

If $y^\star$ has the highest reward, the optimal policy outputs $y^\star$ with probability essentially 1

- Casper et al. (2023): when preferences differ, the majority wins and under-represented groups can be disadvantaged

Tyranny of the majority

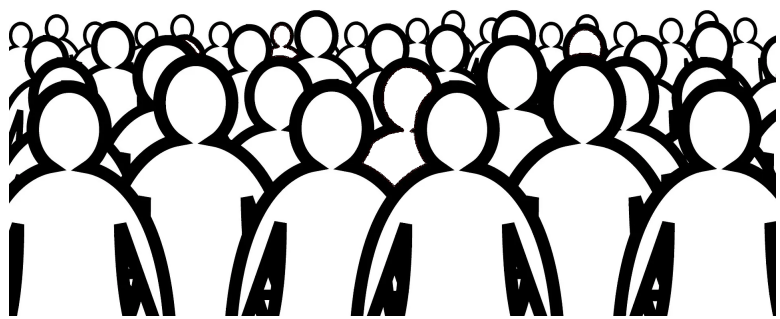
Even with perfect reward models, LLMs may not align with human preferences

Unregularized

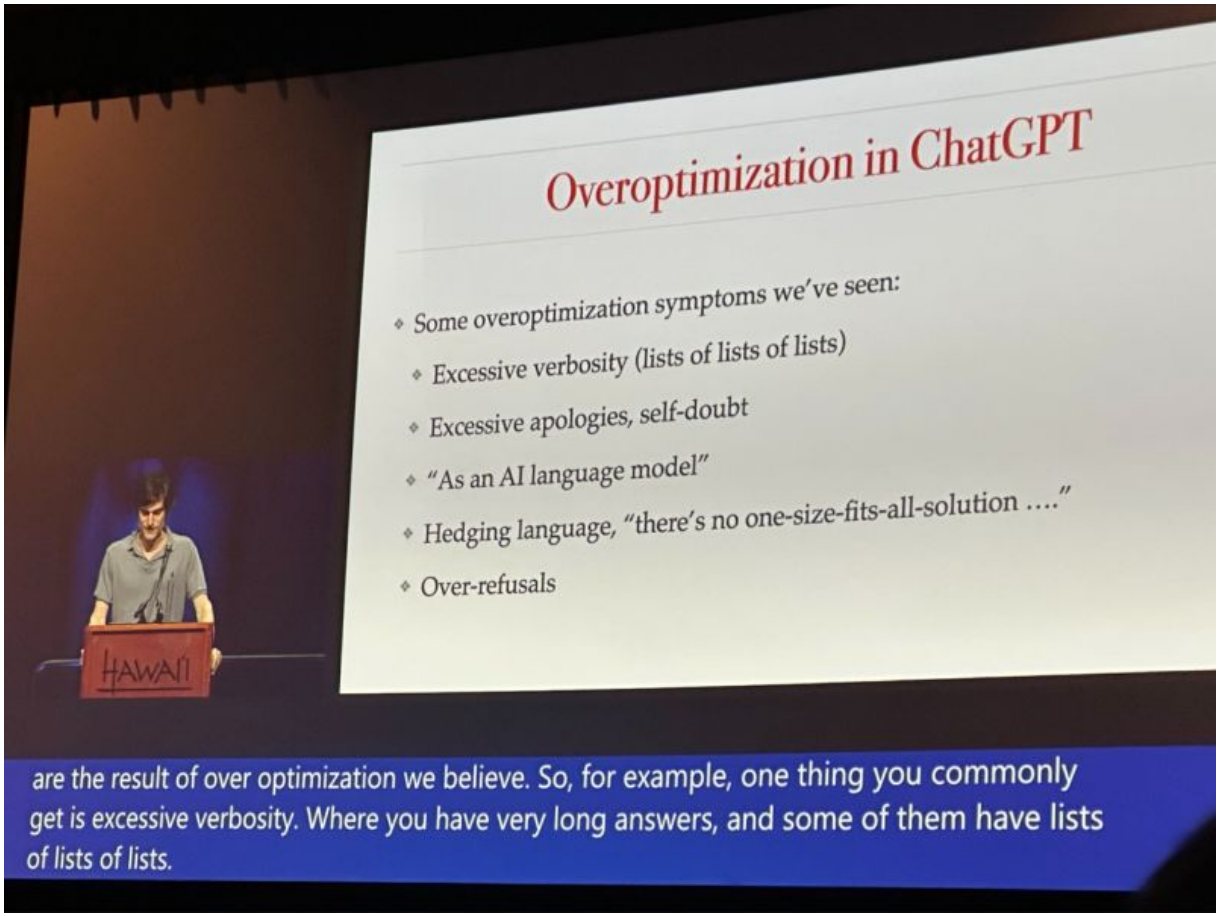
$$\max_{\pi_\phi} \mathbb{E}_{x,y}\big[r(x,y)\big], \qquad y \sim \pi_\phi(y \mid x)$$

If $y^\star$ has the highest reward, the optimal policy outputs $y^\star$ with probability essentially 1

- Casper et al. (2023): when preferences differ, the majority wins and under-represented groups can be disadvantaged
- A 51% vs. 49% preference split can become a near-deterministic policy output

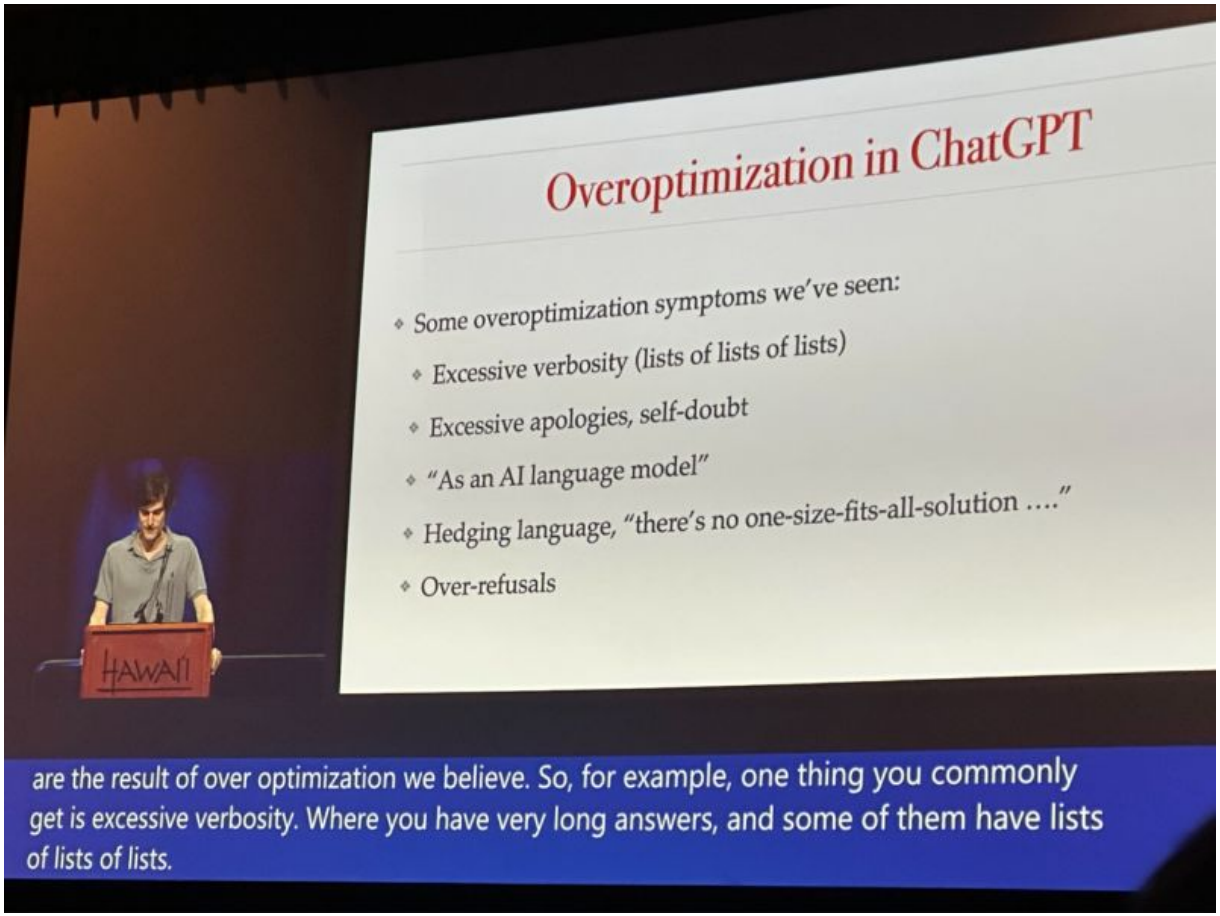


Does OpenAI's solution (early stopping) solve the problem?



John Schulman, ICML 2023 invited talk

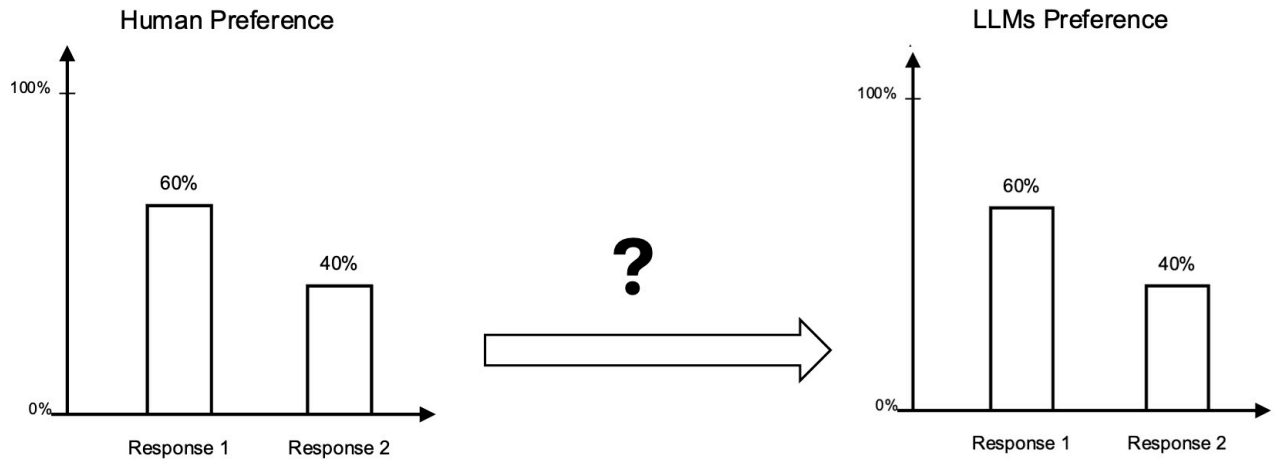
Does OpenAI's solution (early stopping) solve the problem?



Ad hoc, bias remains

John Schulman, ICML 2023 invited talk

Preference matching RLHF



Question 1: When and how can LLMs faithfully represent or align with diverse human preference?

Desideratum for aligning diverse human preference

Definition (Preference Matching)

Given an LLM π, let $p_1, \ldots, p_n$ be the output probabilities of generating $y_1, \ldots, y_n$. π is called preference matching, if for all $i \neq j$,

$$p_i : p_j = \mathcal{P}(y_i \succ y_j) : \mathcal{P}(y_j \succ y_i)$$

- 60% of individuals prefer y_1, 40% prefer y_2
- The LLM π fully preserves human preference if $p_1 : p_2 = 60\% : 40\%$

Finding the preference matching regularizer

Consider adding a regularizer $R(\pi_\phi(y|x))$ to the reward maximization problem:

$$\max_\phi \mathbb{E}_{y\sim\pi_\phi(\cdot|x)}\left[r(x,y) + R(\pi_\phi(y|x))\right]$$

Goal: find $R(\cdot)$

Finding the preference matching regularizer

Consider adding a regularizer $R(\pi_\phi(y|x))$ to the reward maximization problem:

$$\max_\phi \mathbb{E}_{y\sim\pi_\phi(\cdot|x)}\left[r(x,y) + R(\pi_\phi(y|x))\right]$$

Goal: find $R(\cdot)$

Theorem

Under mild assumption, the solution is PM if and only if

$$R(\pi) = -\log\pi + C_{1,x} + \frac{C_{2,x}}{\pi},$$

where $C_{1,x}$ and $C_{2,x}$ are arbitrary constants depending on x

Finding the preference matching regularizer

Consider adding a regularizer $R(\pi_\phi(y|x))$ to the reward maximization problem:

$$\max_\phi \mathbb{E}_{y\sim\pi_\phi(\cdot|x)}\left[r(x,y) + R(\pi_\phi(y|x))\right]$$

Goal: find $R(\cdot)$

Theorem

Under mild assumption, the solution is PM if and only if

$$R(\pi) = -\log\pi + C_{1,x} + \frac{C_{2,x}}{\pi},$$

where $C_{1,x}$ and $C_{2,x}$ are arbitrary constants depending on x

Proposition

The unregularized and KL RLHF does not yield the PM solution

Proof sketch for finding the PM regularizer I

Condition on x, and write $r_i = r(x, y_i)$ and $\pi_i = \pi(y_i \mid x)$

- Preference matching requires the solution to be

$$\pi_i^\star = \frac{e^{r_i}}{\sum_l e^{r_l}}, \qquad i = 1, \ldots, n$$

  for arbitrary $r_1, \ldots, r_k$
- Use the Lagrangian

$$L = \sum_{i=1}^{n} \pi_i \big(r_i + R(\pi_i)\big) + \lambda \Big(\sum_{i=1}^{n} \pi_i - 1 \Big)$$

- First-order optimality at $\pi^\star$ gives

$$r_i + R(\pi_i^\star) + \pi_i^\star R'(\pi_i^\star) + \lambda = 0$$

- Since $\pi_i^\star \propto e^{r_i}$, we can write

$$r_i = \log \pi_i^\star + c_x$$

  with $c_x = \log \sum_j e^{r_j}$ independent of i

Proof sketch for finding the PM regularizer II

- Hence

$$R(\pi) + \pi R'(\pi) = -\log \pi + c'_x$$

Differentiating yields

$$\pi R''(\pi) + 2R'(\pi) + \frac{1}{\pi} = 0$$

whose general solution is

$$R(\pi) = -\log \pi + C_{1,x} + \frac{C_{2,x}}{\pi}$$

PM RLHF

$$\max_{\phi} \mathbb{E}\left[r(x,y) - \log \pi_\phi(y|x) + C_{1,x} + \frac{C_{2,x}}{\pi_\phi(y|x)}\right]$$

- The first term: $-\log \pi$. It is the entropy:

$$\mathcal{H}(\pi) = -\mathbb{E}_{y\sim\pi_\phi(\cdot|x)}[\log(\pi_\phi(y|x))]$$

- The second term $C_{1,x}$: depends only on x but not on y. Possible choice
  - length penalty
  - Normalize the reward with zero means
- The last term $C_{2,x}/\pi$: zero

A dual perspective on PM RLHF

Dual objective of the PM RLHF objective

$$\max_{\phi} \min_{r'(x,y)} \mathbb{E}_{y \sim \pi_\phi(\cdot|x)} r'(x,y)$$
$$\text{s.t. log-sum-exp}(r(x,\cdot) - r'(x,\cdot)) \leq \epsilon$$

for some constant ϵ

- log-sum-exp(): $\log \left[\sum_y \exp(r(x,y) - r'(x,y)) \right]$
- Adversarial perturbation the rewards
- Robust optimization problem

Two practical variants

Conditional PM RLHF

$$\max_{\phi} \mathbb{E}_{y\sim\pi_\phi(\cdot|x)}\left[r(x,y) - \log(\pi_\phi(y|x))\mathbb{1}(\pi_{\text{ref}}(y|x) \geq \alpha) - \log\left(\frac{\pi_\phi(y|x)}{\pi_{\text{ref}}(y|x)}\right)\mathbb{1}(\pi_{\text{ref}}(y|x) < \alpha)\right]$$

- Provably Preference Matching on $\mathcal{Y} = \{\text{ref}(y|x) \geq \alpha\}$

Two practical variants

Conditional PM RLHF

$$\max_{\phi} \mathbb{E}_{y\sim\pi_\phi(\cdot|x)}\left[r(x,y) - \log(\pi_\phi(y|x))\mathbb{1}(\pi_{\text{ref}}(y|x) \geq \alpha) - \log\left(\frac{\pi_\phi(y|x)}{\pi_{\text{ref}}(y|x)}\right)\mathbb{1}(\pi_{\text{ref}}(y|x) < \alpha)\right]$$

- Provably Preference Matching on $\mathcal{Y} = \{\text{ref}(y|x) \geq \alpha\}$

Preference Matching Optimization

$$-\mathbb{E}_{(x,y_w,y_l)} \log\sigma\left((\alpha+\beta)\log\frac{\pi_\phi(y_w|x)}{\pi_{\text{ref}}(y_w|x)^{\frac{\beta}{\alpha+\beta}}} - (\alpha+\beta)\log\frac{\pi_\phi(y_l|x)}{\pi_{\text{ref}}(y_l|x)^{\frac{\beta}{\alpha+\beta}}}\right)$$

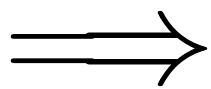
- The DPO-variants of PM RLHF

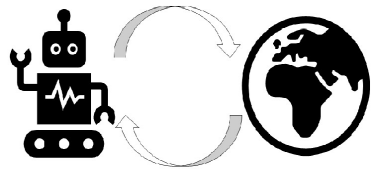
The pipeline of PM fine-tuning

Train on all data

The pipeline of PM fine-tuning

Train on all data

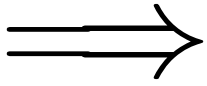


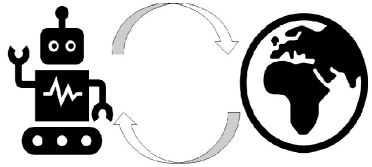


RLHF

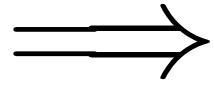
The pipeline of PM fine-tuning

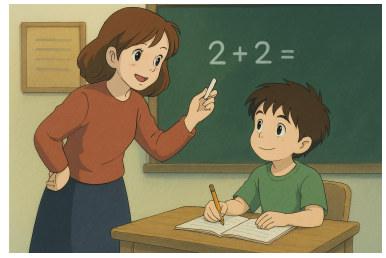
Train on all data





RLHF

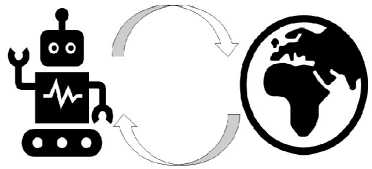




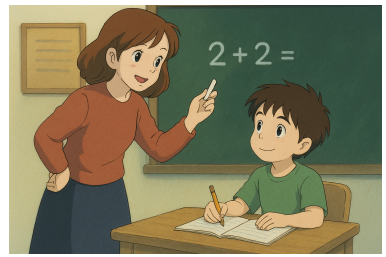
PM RLHF

The pipeline of PM fine-tuning

Train on all data ⟹



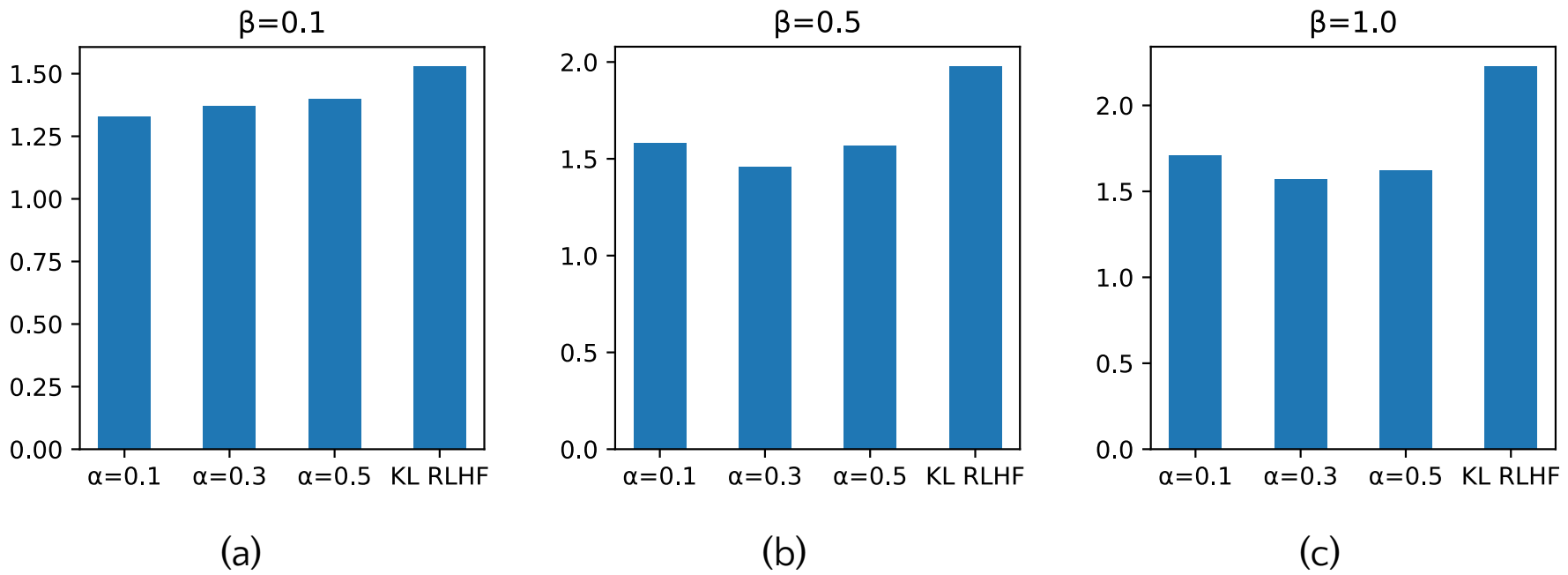
RLHF ⟹



PM RLHF

- Preference matching aims to preserve the full preference distribution rather than collapsing to a single winner

Experiments on Llama-2-7B



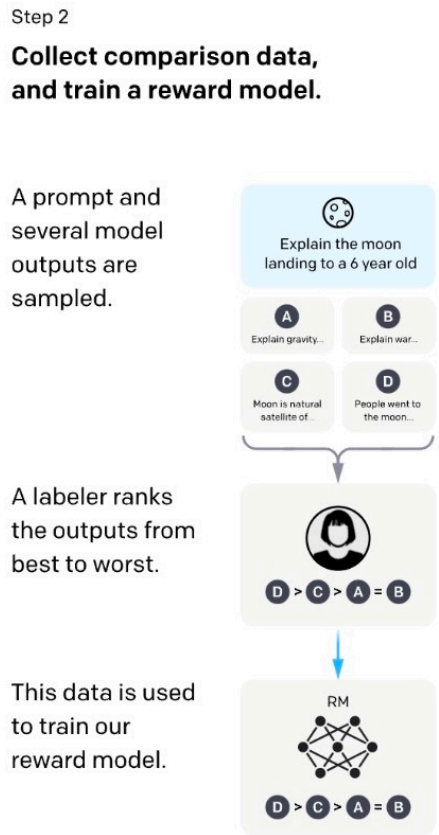
- y-axis: KL divergence between LLM and reward model preference

It doesn't compromise accuracy on five tasks

model	arc_challenge	hellaswag	mmlu	truthfulqa	winogrande	average
Gemma-DPO	0.3524	0.4776	0.2614	0.2938	0.5943	0.3959
Gemma-CPO	0.3498	0.4721	0.2551	0.2925	0.5927	0.3925
Gemma-SimPO	0.3609	0.4786	0.2695	0.3060	0.5880	0.4006
Gemma-PMO	0.3737	0.4568	0.2621	0.2987	0.6014	0.3985
Llama-DPO	0.3208	0.4442	0.4414	0.2546	0.5896	0.4101
Llama-CPO	0.3336	0.4538	0.4346	0.2619	0.5983	0.4164
Llama-SimPO	0.3387	0.4500	0.3945	0.2583	0.5998	0.4083
Llama-PMO	0.3507	0.4500	0.4526	0.2656	0.5912	0.4220
Qwen-DPO	0.4471	0.5015	0.5983	0.2387	0.6448	0.4861
Qwen-CPO	0.4078	0.5115	0.5927	0.2546	0.6417	0.4817
Qwen-SimPO	0.4471	0.5014	0.5978	0.2387	0.6440	0.4858
Qwen-PMO	0.4394	0.5023	0.5984	0.2521	0.6417	0.4868

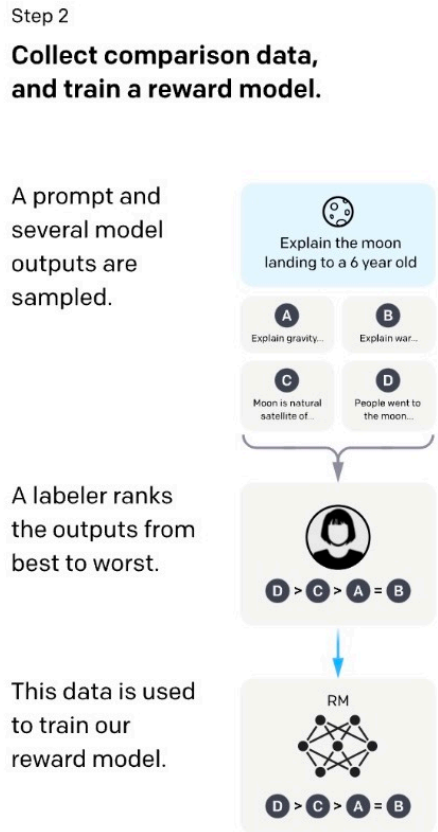
Taking a step back: is OpenAI's reward model good?

Good teacher



Taking a step back: is OpenAI's reward model good?

Good teacher

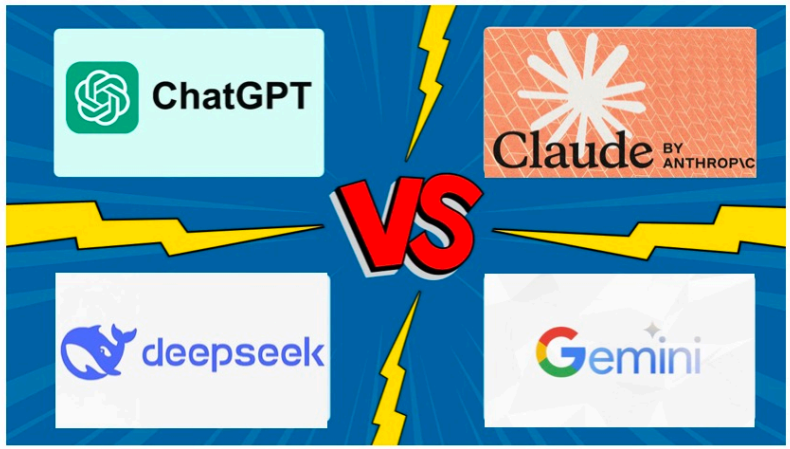


- Before asking whether Step 3 is good optimization, we should first ask whether Step 2 produces a faithful teacher

Beyond the BT assumptions

Question 1: When and how can LLMs faithfully represent or align with diverse human preference?

Under BT assumptions, PM RLHF gives a positive answer



Question 2: Can LLMs partially represent or align with diverse human preference beyond BT assumptions?

Beyond BT reward models

A Simplified Model:

- One prompt x
- n responses $y_1, y_2, \ldots, y_n$
- m labelers
- Pairwise Comparison:

$$\mathcal{P}(y_i \succ y_j) = \frac{\#(y_i \succ y_j)}{m}$$

- $\mathcal{P}(y_i \succ y_j) + \mathcal{P}(y_j \succ y_i) = 1$ (no tie)

Definition (Reward-consistent preference)

A reward model $r : \mathcal{Y} \to \mathbb{R}$ captures a preference $\mathcal{P}(y \succ y')$ if for any two distinct responses y and y', $r(y) > r(y')$ implies $\mathcal{P}(y \succ y') > \frac{1}{2}$

Beyond BT reward models

A Simplified Model:

- One prompt x
- n responses $y_1, y_2, \ldots, y_n$
- m labelers
- Pairwise Comparison:

$$\mathcal{P}(y_i \succ y_j) = \frac{\#(y_i \succ y_j)}{m}$$

- $\mathcal{P}(y_i \succ y_j) + \mathcal{P}(y_j \succ y_i) = 1$ (no tie)

Definition (Reward-consistent preference)

A reward model $r : \mathcal{Y} \to \mathbb{R}$ captures a preference $\mathcal{P}(y \succ y')$ if for any two distinct responses y and y', $r(y) > r(y')$ implies $\mathcal{P}(y \succ y') > \frac{1}{2}$

- Weakest possible constraint on reward-based preferences

Condorcet paradox

Definition (Condorcet paradox)

Three labelers (A, B, C) and three responses $\{y_1, y_2, y_3\}$.
Each gives a complete ranking:

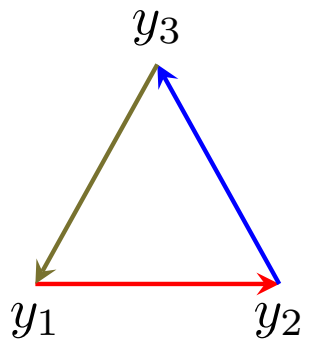
- A: $y_1 \succ y_2 \succ y_3$
- B: $y_2 \succ y_3 \succ y_1$
- C: $y_3 \succ y_1 \succ y_2$

This gives

$$P(y_1 \succ y_2) = \tfrac{2}{3}, P(y_2 \succ y_3) = \tfrac{2}{3}, P(y_3 \succ y_1) = \tfrac{2}{3}$$

Marquis de Condorcet

- Cyclic human preferences cannot be captured by any reward model



Condorcet paradox

Definition (Condorcet paradox)

Three labelers (A, B, C) and three responses $\{y_1, y_2, y_3\}$.
Each gives a complete ranking:

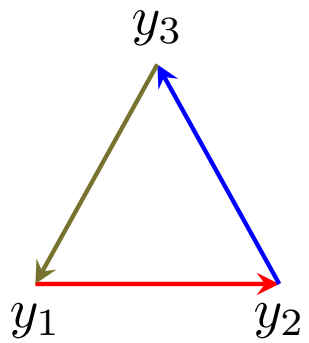
- A: $y_1 \succ y_2 \succ y_3$
- B: $y_2 \succ y_3 \succ y_1$
- C: $y_3 \succ y_1 \succ y_2$

This gives

$$P(y_1 \succ y_2) = \tfrac{2}{3}, P(y_2 \succ y_3) = \tfrac{2}{3}, P(y_3 \succ y_1) = \tfrac{2}{3}$$

Marquis de Condorcet

- Cyclic human preferences cannot be captured by any reward model
- *Question: How likely is it that human preferences contain Condorcet cycles?*



Assumption

Assumption (Preference as ranking)

Given a prompt x and its associated n responses $y_1, y_2, \ldots, y_n$, an individual (labeler) expresses a preference in the form of a ranking of these n responses

- Example: Individual 1 has: $y_n \succ \cdots \succ y_2 \succ y_1$

Assumption

Assumption (Preference as ranking)

Given a prompt x and its associated n responses $y_1, y_2, \ldots, y_n$, an individual (labeler) expresses a preference in the form of a ranking of these n responses

- Example: Individual 1 has: $y_n \succ \cdots \succ y_2 \succ y_1$

Assumption (Impartial culture condition [Guilbaud, 1952])

Each individual independently samples a linear preference ranking with equal probability $1/n!$

- The most widely used assumption in social choice theory
- Mathematically non-trivial
- Can be relaxed to non-uniform settings: with a probability $\geq \epsilon$

Our result

Theorem

Let the number of responses $n \geqslant 3$ and the number of labelers $m \geqslant 3$. Then,

$$\mathbb{P}_{m,n}(\textit{Condorcet cycle}) \geqslant 1 - c_1 e^{-c_2 n},$$

where $c_1, c_2 > 0$ are universal constants

- Holds for all $m \geq 3$

Our result

Theorem

Let the number of responses $n \geqslant 3$ and the number of labelers $m \geqslant 3$. Then,

$$\mathbb{P}_{m,n}(\textit{Condorcet cycle}) \geqslant 1 - c_1 e^{-c_2 n},$$

where $c_1, c_2 > 0$ are universal constants

- Holds for all $m \geq 3$
- Converge to 1 exponentially

Our result

Theorem

Let the number of responses $n \geqslant 3$ and the number of labelers $m \geqslant 3$. Then,

$$\mathbb{P}_{m,n}(\textit{Condorcet cycle}) \geqslant 1 - c_1 e^{-c_2 n},$$

where $c_1, c_2 > 0$ are universal constants

- Holds for all $m \geq 3$
- Converge to 1 exponentially
- With high probability, human preference cannot be represented by a reward model

Limitations of reward models

Reward-model-based approaches face fundamental limitations
Are there better alternatives?

Preference model and Nash learning from human feedback (NLHF)

Preference model

$$\mathcal{P}\left(y \succ y' \mid x\right)$$

- Keeps pairwise human preference information instead of collapsing it to one scalar reward
- Define the preference between two policies π and π' conditioned on a prompt x by

$$\mathcal{P}\left(\pi \succ \pi' \mid x\right) := \mathbb{E}_{y \sim \pi(\cdot|x), y' \sim \pi'(\cdot|x)} \mathcal{P}\left(y \succ y' \mid x\right)$$

NLHF objective [Munos et al., 2024]

$$\max_{\pi} \min_{\pi'} \mathcal{P}(\pi \succ \pi')$$

Condorcet winning response (winner)

Definition (Condorcet Winning Response)

A response $y^\star$ is called a Condorcet winning response if $\mathcal{P}(y^\star \succ y) > \frac{1}{2}$ for all $y \neq y^\star$

- Corresponds to the winner of an election in traditional social choice theory
- Various names, including the outright winner [May, 1971] and the pairwise majority rule winner [Gehrlein, 2006]

Condorcet consistency

Theorem (NLHF and Condorcet winners)

1. *A pure-strategy Nash equilibrium exists if and only if there exists a Condorcet winning response*
2. *When it exists, the equilibrium is unique and equals that response*

Condorcet consistency

Theorem (NLHF and Condorcet winners)

1. *A pure-strategy Nash equilibrium exists if and only if there exists a Condorcet winning response*
2. *When it exists, the equilibrium is unique and equals that response*

- If no Condorcet winning response exists, Nash equilibria are mixed

Condorcet consistency

Theorem (NLHF and Condorcet winners)

1. *A pure-strategy Nash equilibrium exists if and only if there exists a Condorcet winning response*
2. *When it exists, the equilibrium is unique and equals that response*

- If no Condorcet winning response exists, Nash equilibria are mixed
- NLHF can therefore retain diversity instead of collapsing to one response

Probability of no Condorcet winning response

Existing Research:

- In 1968, [Garman and Kamien, 1968] conjectured that the probability converges to 0 as $n \to \infty$ for all $m \geq 3$
- The conjecture was later resolved by [May, 1971]
- The probability decreases at a rate of $O(1/\sqrt{n})$ for all $m \geq 3$
- At a rate of $O(1/n)$ for $m = \infty$
- Not optimal
- As noted in a series of subsequent papers, this question remains open
- In Gehrlein's book *Condorcet Paradox* (Section 7), it has been listed as an open problem (for more than 50 years)

Our results

Theorem (Upper Bound)

For $m \geqslant 3$, $l = \lceil \frac{m}{2} \rceil$, given prompt with its n responses,

$$\mathbb{P}_{m,n}(\textit{Condorcet winning response})$$

$$\leqslant (n+1)^{\frac{m+3}{2}} e^{-(\log n)^2} + (n+1)(m+1)\, n^{-\frac{m}{l}} (\log n + 1)^{\left(\frac{2}{l}+1\right)m} + \frac{m(m-1)}{n+1}$$

Theorem (Lower Bound)

For $m \geqslant 3$, $l = \lceil \frac{m}{2} \rceil$, there exists a universal constant c such that

$$\mathbb{P}_{m,n}(\textit{Condorcet winning response}) \geqslant c \cdot n^{1-\frac{m}{l}} \cdot \left(\frac{n}{n+1}\right)^m \cdot \left(\frac{1}{2\binom{m}{l}}\right)^{\frac{m}{l}}$$

Our results

In the LLM regime, where m is fixed and $n \to \infty$, our main results simplify to the following corollary

Corollary

For $m \geqslant 3$, given a prompt with its n responses,

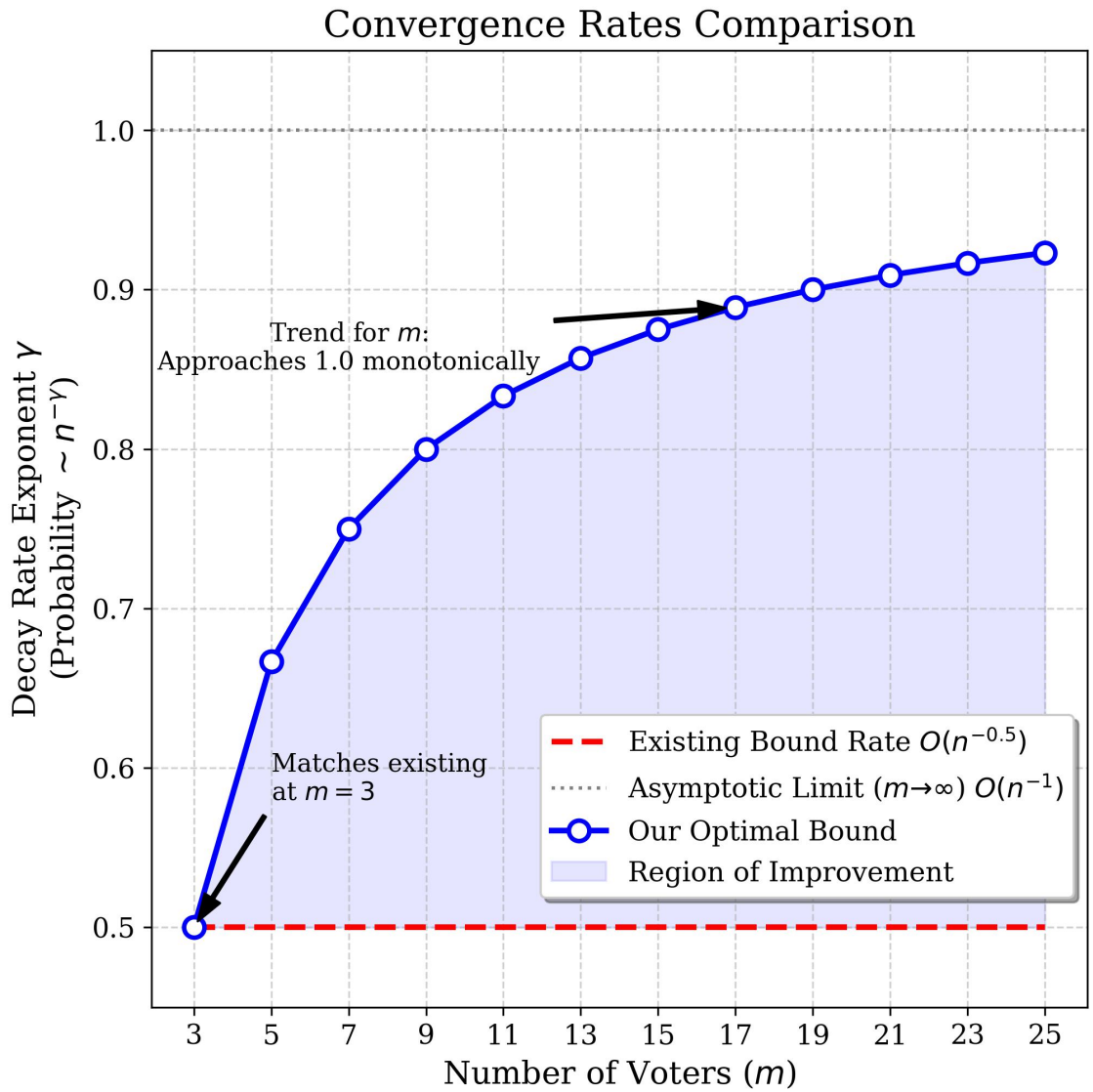
$$\mathbb{P}_{m,n}(\textit{Condorcet winning response}) = \tilde{\Theta}\left(n^{1-\frac{m}{l}}\right) \to 0 \quad \textit{as } n \to \infty$$

Furthermore, the Nash equilibria of NLHF are mixed strategies

This resolves the open problem

- The probability converges to 0 polynomially

How the exponent grows



Extension to non-uniform rankings

Assumption

Each individual independently samples a linear preference ranking with probability $\geq \epsilon$

Corollary

For $m \geqslant 3$*, given a prompt with its* n *responses,*

$$\mathbb{P}_{m,n}(\textit{Condorcet winning response}) = \tilde{\Theta}\left(n^{1-\frac{m}{l}}\right) \to 0 \quad \textit{as } n \to \infty$$

- It still converges to 0 polynomially

A broader game-theoretic view

Instead of using the raw preference:

$$\max_{\pi} \min_{\pi'} \; \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi(\cdot|x),\, y'\sim\pi'(\cdot|x)}\Big[\Psi\big(\mathcal{P}(y \succ y' \mid x)\big)\Big]$$

A broader game-theoretic view

Instead of using the raw preference:

$$\max_{\pi} \min_{\pi'} \; \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi(\cdot|x),\, y'\sim\pi'(\cdot|x)}\Big[\Psi\big(\mathcal{P}(y \succ y' \mid x)\big)\Big]$$

- NLHF is the special case $\Psi(t) = t$
- A game-theoretic analogue of RLHF uses $\Psi(t) = \log \frac{t}{1-t}$

A broader game-theoretic view

Instead of using the raw preference:

$$\max_{\pi} \min_{\pi'} \; \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi(\cdot|x),\, y'\sim\pi'(\cdot|x)}\Big[\Psi\big(\mathcal{P}(y \succ y' \mid x)\big)\Big]$$

- NLHF is the special case $\Psi(t) = t$
- A game-theoretic analogue of RLHF uses $\Psi(t) = \log \frac{t}{1-t}$
- Question: *which properties depend on the choice of Ψ, and which are robust?*

Which properties survive the choice of Ψ?

Three target properties

- Condorcet consistency: output the Condorcet winner when it exists
- Mixed-strategy diversity: avoid collapsing to one response when no winner exists
- Smith consistency: support only on the smallest top set that beats everything outside it

Which properties survive the choice of Ψ?

Three target properties

- Condorcet consistency: output the Condorcet winner when it exists
- Mixed-strategy diversity: avoid collapsing to one response when no winner exists
- Smith consistency: support only on the smallest top set that beats everything outside it

Theorem

- *Condorcet consistency:* $\Psi(t) \geq \Psi(1/2)$ *for* $t \geq 1/2$, *and* $\Psi(t) < \Psi(1/2)$ *for* $t < 1/2$
- *Mixed strategies:* $\Psi(t) + \Psi(1-t) \geq 2\Psi(1/2)$
- *Smith consistency:* $\Psi(t) + \Psi(1-t) = 2\Psi(1/2)$

Which properties survive the choice of Ψ?

Three target properties

- Condorcet consistency: output the Condorcet winner when it exists
- Mixed-strategy diversity: avoid collapsing to one response when no winner exists
- Smith consistency: support only on the smallest top set that beats everything outside it

Theorem

- *Condorcet consistency:* $\Psi(t) \geq \Psi(1/2)$ *for* $t \geq 1/2$, *and* $\Psi(t) < \Psi(1/2)$ *for* $t < 1/2$
- *Mixed strategies:* $\Psi(t) + \Psi(1-t) \geq 2\Psi(1/2)$
- *Smith consistency:* $\Psi(t) + \Psi(1-t) = 2\Psi(1/2)$

- So raw preference is *not* unique: a whole family of payoffs leads to the same qualitative alignment guarantees

Fundamental limitation: no exact preference matching

Theorem (Impossibility)

No smooth payoff built only from pairwise preference data can perfectly match a specific target human policy as a unique Nash equilibrium

Fundamental limitation: no exact preference matching

Theorem (Impossibility)

No smooth payoff built only from pairwise preference data can perfectly match a specific target human policy as a unique Nash equilibrium

- Exact recovery of the full preference distribution is impossible in general
- What remains possible are qualitative guarantees such as mixed strategies and Smith-set support

Fundamental limitation: no exact preference matching

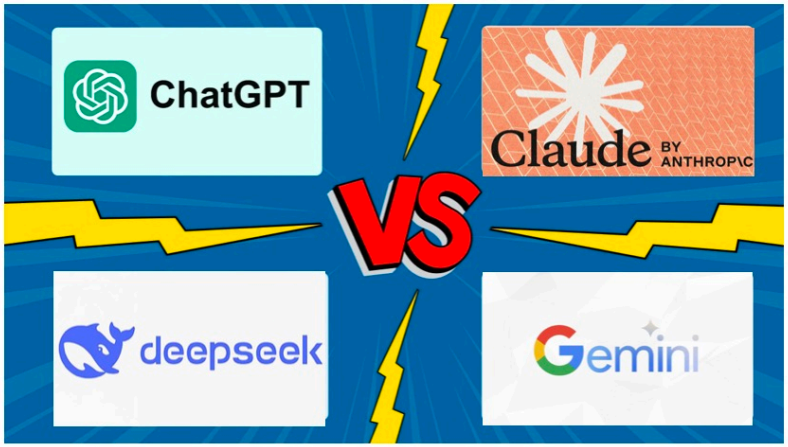
Theorem (Impossibility)

No smooth payoff built only from pairwise preference data can perfectly match a specific target human policy as a unique Nash equilibrium

- Exact recovery of the full preference distribution is impossible in general
- What remains possible are qualitative guarantees such as mixed strategies and Smith-set support

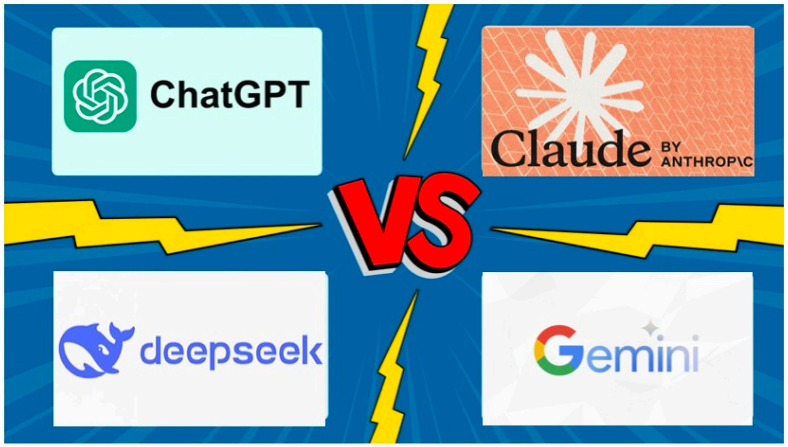
Takeaway: game-theoretic alignment preserves structure, not every detail

Beyond the BT assumptions



Question 2: Can LLMs partially represent or align with diverse human preference beyond BT assumptions?

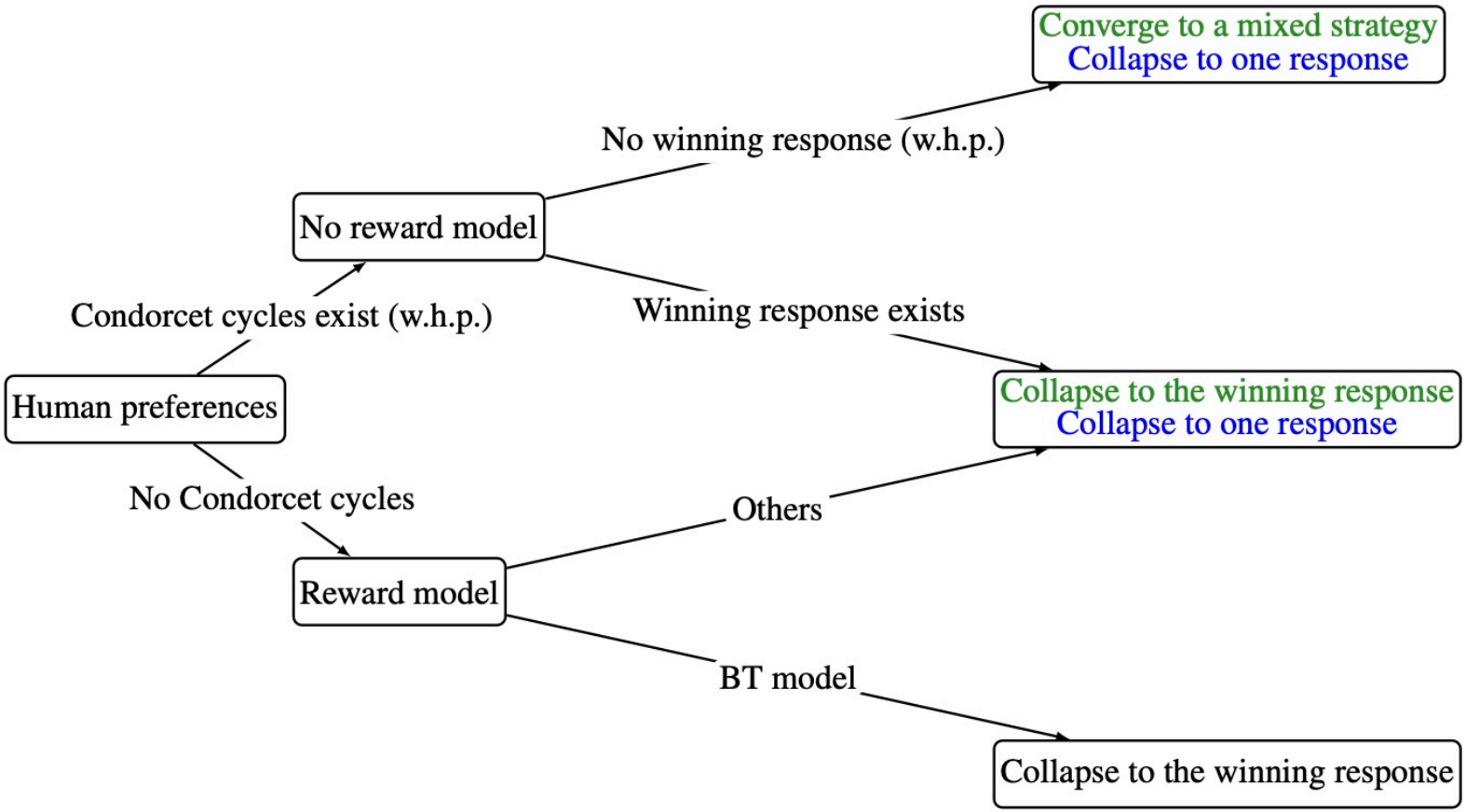
Beyond the BT assumptions



Question 2: Can LLMs partially represent or align with diverse human preference beyond BT assumptions?

NLHF gives a meaningful partial answer

Summary



- Green: NLHF
- Blue: RLHF

From partial alignment to misspecification

- Beyond BT, pairwise preference data need not come from one global score vector
- Cyclic structure is exactly where reward modeling begins to break
- So the next question is quantitative: how much loss is forced when we fit the best BT model anyway?

This turns a qualitative objection into a measurable notion of misspecification

Quantifying Bradley–Terry misspecification under cyclic preferences

- If preferences are globally rankable, Bradley–Terry should fit well
- If preferences contain genuine cycles, Bradley–Terry is statistically misspecified

Quantifying Bradley–Terry misspecification under cyclic preferences

- If preferences are globally rankable, Bradley–Terry should fit well
- If preferences contain genuine cycles, Bradley–Terry is statistically misspecified

Diagnostic

$$\min_{r \in \mathbb{R}^n} L(r), \qquad L(r) = \sum_{i<j} D_{\mathrm{KL}}(p_{ij} \,\|\, \sigma(r_i - r_j))$$

Geometric viewpoint: the cycle component is the obstruction

Set-up

$$\Delta_{ij} = p_{ij} - \frac{1}{2}, \qquad (Br)_{ij} = r_i - r_j$$

$$\mathcal{C} = \mathrm{im}(B), \qquad \mathcal{C}^{\perp} = \text{cycle space}$$

- $\mathcal{C}$ contains the globally rankable component
- $\mathcal{C}^{\perp}$ contains the cyclic inconsistency

Main heuristic

$$\min_r L(r) \approx 2\left\|\Pi_{\mathcal{C}^{\perp}} \Delta\right\|_2^2,$$

up to higher-order nonlinear remainder terms

Uniform random rankings

Model

$$\pi^{(1)}, \ldots, \pi^{(m)} \overset{\text{i.i.d.}}{\sim} \text{Unif}(S_n)$$

$$p_{ij} = \frac{1}{m} \sum_{s=1}^{m} \mathbb{1}\left\{\pi^{(s)}(i) < \pi^{(s)}(j)\right\}$$

$$\Delta_{ij} = p_{ij} - \frac{1}{2}$$

- The population cycle component is zero
- Misspecification comes entirely from finite-sample cycle noise

Uniform random rankings

Model

$$\pi^{(1)}, \dots, \pi^{(m)} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(S_n)$$

$$p_{ij} = \frac{1}{m} \sum_{s=1}^{m} \mathbb{1}\left\{ \pi^{(s)}(i) < \pi^{(s)}(j) \right\}$$

$$\Delta_{ij} = p_{ij} - \frac{1}{2}$$

- The population cycle component is zero
- Misspecification comes entirely from finite-sample cycle noise

Main bounds

$$M_{n,m} := \mathbb{E}\big[\min_{r} L(r) \big]$$

$$M_{n,m} \geq \frac{(n-1)(n-2)}{12m} - C \frac{n^2 \log n}{m^2}$$

$$M_{n,m} \leq \frac{(n-1)(n-2)}{12m} + C \frac{n^2 \log n}{m^2}$$

$$\mathbb{E}\Big[\big\| \Pi_{\mathcal{C}^\perp} \Delta \big\|_2^2 \Big] = \frac{(n-1)(n-2)}{24m}$$

leading term of $M_{n,m} \asymp \dfrac{(n-1)(n-2)}{12m}$

References I

- *On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization*
  Xiao, Xie, Li, Getzen, Fang, Long, and Su
- *Statistical Impossibility and Possibility of Aligning LLMs with Human Preferences: From Condorcet Paradox to Nash Equilibrium*
  Liu, Long, Shi, Su, and Xiao
- *Fundamental Limits of Game-Theoretic LLM Alignment: Smith Consistency and Preference Matching*
  Shi, Liu, Long, Su, and Xiao
- *The Price of Transitivity: Sharp KL Bounds for Bradley–Terry Approximation of Pairwise Preferences*
  Zhang, Xiao, Qi, and Su

References III


Bradley, R. A. and Terry, M. E. (1952).
Rank analysis of incomplete block designs: I. the method of paired comparisons.
Biometrika, 39(3/4):324–345.


Garman, M. B. and Kamien, M. I. (1968).
The paradox of voting: Probability calculations.
Behavioral Science, 13(4):306–316.


Gehrlein, W. V. (2006).
Condorcet's paradox.
Springer.


Guilbaud, G. T. (1952).
Les théories de l'intérêt général et le problème logique de l'agrégation.
Economie appliquée, 5(4):501–584.


May, R. M. (1971).
Some mathematical remarks on the paradox of voting.
Behavioral Science, 16(2):143–151.

References IV

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022).
Training language models to follow instructions with human feedback.
Advances in Neural Information Processing Systems, 35:27730–27744.