KTO: Model Alignment as Prospect Theoretic Optimization

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, Douwe Kiela

RL with Human Feedback

Start with a dataset of preferences: (x, y_w, y_l)

Probability that y_w is preferred over y_l can be captured with a specific function class (e.g., Bradley-Terry model):

$$p^*(y_w \succ y_l|x) = \sigma(r^*(x, y_w) - r^*(x, y_l))$$

Train a reward model:

$$\mathcal{L}_R(r_\phi) = \mathbb{E}_{x,y_w,y_l\sim D}[-\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))]$$

RL with Human Feedback

Train a reward model:

$$\mathcal{L}_R(r_\phi) = \mathbb{E}_{x,y_w,y_l\sim D}[-\log\sigma(r_\phi(x,y_w) - r_\phi(x,y_l))]$$

Train a language model through reward maximization and add a KL divergence w.r.t. to the base model:

$$\mathbb{E}_{x\in D,y\in\pi_\theta}[r_\phi(x,y)] \; - \beta D_{\mathrm{KL}}(\pi_\theta(y|x)\|\pi_{\mathrm{ref}}(y|x))$$

RL with Human Feedback (DPO)

Design a closed-form loss that maximizes the margin between the preferred and dispreferred generations.

Direct Preference Optimization:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x, y_w, y_l \sim D}$$
$$\left[-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]$$

Prospect Theory

Prospect theory explains why, when faced with an uncertain event, humans make decisions that do not maximize their expected value.

Gamble: $100 with 80% and $0 with 20%.

A person might prefer $60 for 100% even though the expected return if they gambled was $80 as they might be **loss-averse**.

Prospect Theory (Tversky and Kahneman, 1992)

A **value function** maps an outcome z, relative to reference point z_0, to its perceived (subjective) value.

Tversky and Kahneman proposed the following functional form for human value:

$$v(z;\lambda,\alpha,z_0)=\begin{cases}(z-z_0)^{\alpha} & \text{if } z\geq z_0\\ -\lambda(z_0-z)^{\alpha} & \text{if } z<z_0\end{cases}$$

α controls the curvature of the function (risk aversion)
λ controls the steepness of the function (loss aversion)

HALOs

HALO: human-aware losses

Implied reward: $r_\theta(x, y) = l(y) \log[\pi_\theta(y|x)/\pi_{\text{ref}}(y|x)]$

Reference point distribution: $Q(Y'|x)$

Human value of (x, y): $v(r_\theta(x, y) - \mathbb{E}_Q[r_\theta(x, y')])$

expected reward from human's perspective

HALOs

A function f is a HALO for v if $\exists\ a_{x,y} \in \{-1, +1\}$ such that:

$$f(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x,y\sim\mathcal{D}}[a_{x,y} v(r_\theta(x,y) - \mathbb{E}_Q[r_\theta(x,y')])] + C_\mathcal{D}$$

where D is the feedback data and $C_D \in \mathbb{R}$ is a data-specific constant.

HALOs Interpretation

$$\mathbb{E}_{x\in D,y\in\pi_\theta}[r_\phi(x,y)] \; - \beta D_{\text{KL}}(\pi_\theta(y|x)\|\pi_{\text{ref}}(y|x)) \longrightarrow \pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r^*(x,y)\right)$$

$$f(\pi_\theta,\pi_{\text{ref}}) = \mathbb{E}_{x,y\sim\mathcal{D}}[a_{x,y}v(r_\theta(x,y) - \mathbb{E}_Q[r_\theta(x,y')])] + C_\mathcal{D}$$

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

$$l(\cdot) = \beta$$

$$\longrightarrow r_{\theta^*}(x,y) = r^*(x,y) - \beta\log Z(x)$$

under θ^*, the HALO-defined reward is the optimal reward shifted by an input-specific term
=> r_{θ^*} is in the same equivalence class as r^*
=> would induce optimal policy π^*

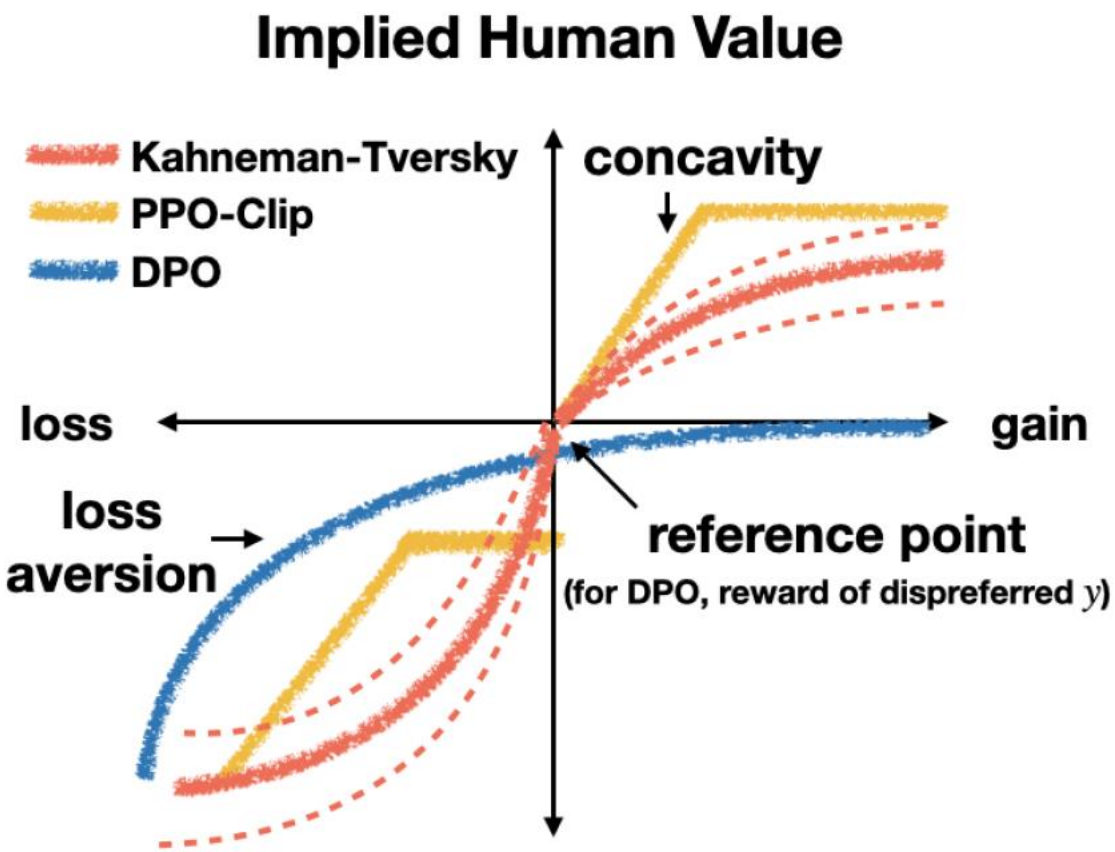
HALO vs non-HALO

Conditional SFT: non-HALO

Sequence Likelihood Calibration(SLiC): non-HALO

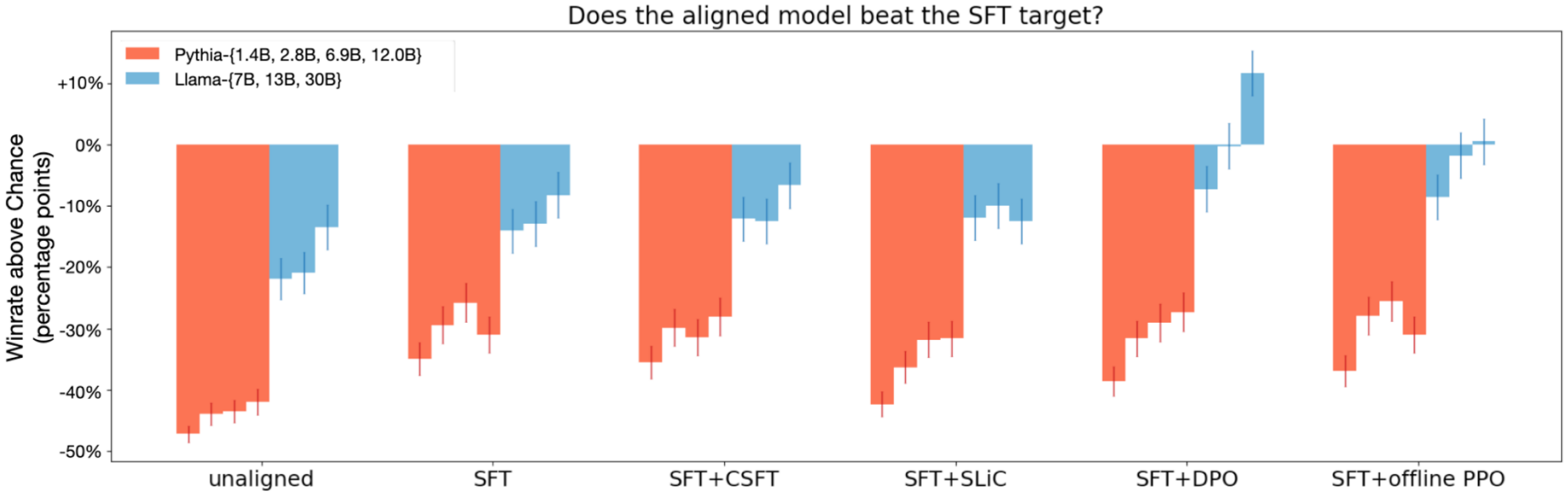
DPO: HALO

PPO (offline): HALO



HALO vs non-HALO

LLM-as-a-judge (GPT-4) to compare the aligned model's response with the SFT target (subset of y_w).

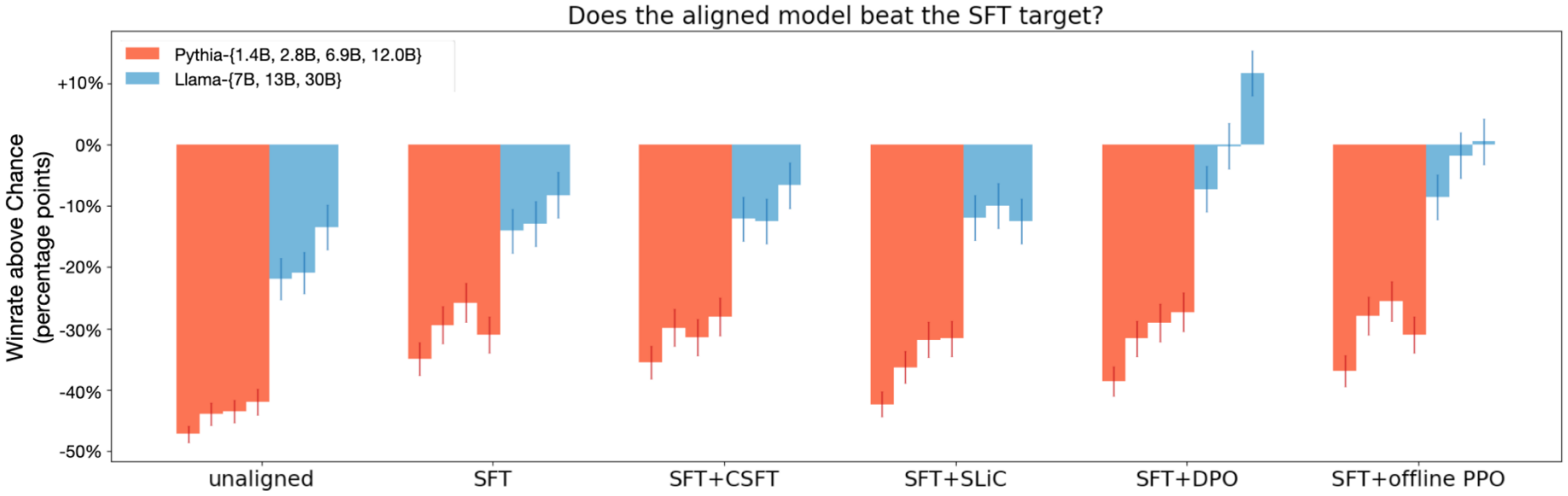


Up to a scale of 7B, alignment provides no gains over SFT alone.

Why?

HALO vs non-HALO

LLM-as-a-judge (GPT-4) to compare the aligned model's response with the SFT target (subset of y_w).



HALOs either match or outperform (13B+) non-HALOs.

Why?

Kahneman-Tversky Optimization

Start with the canonical Kahneman-Tversky value function:

$$v(z;\lambda,\alpha,z_0)=\begin{cases}(z-z_0)^{\alpha} & \text{if } z\geq z_0\\ -\lambda(z_0-z)^{\alpha} & \text{if } z<z_0\end{cases}$$

- Replace exponent α with the logistic function for stability.
- Control the degree of risk aversion, using hyperparameter β (the greater β, the more quickly the value saturates) - similar to effect as β in RLHF and DPO.
- Replace loss aversion coefficient λ with $\{\lambda_D,\lambda_U\}$ for desirable and undesirable outputs, respectively.
- For the reference point z_0, assume humans judge the quality of $y|x$ in relation to *all* possible outcomes.

$$L_{\text{KTO}}(\pi_\theta,\pi_{\text{ref}})=\mathbb{E}_{x,y\sim D}[\lambda_y-v(x,y)]$$

where

$$r_\theta(x,y)=\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

$$z_0=\text{KL}(\pi_\theta(y'|x)\|\pi_{\text{ref}}(y'|x))$$

$$v(x,y)=\begin{cases}\lambda_D\sigma(\beta(r_\theta(x,y)-z_0)) & \text{if } y\sim y_{\text{desirable}}|x\\ \lambda_U\sigma(\beta(z_0-r_\theta(x,y))) & \text{if } y\sim y_{\text{undesirable}}|x\end{cases}$$

Kahneman-Tversky Optimization

KTO loss:

$$L_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x,y\sim D}[\lambda_y - v(x,y)]$$

where

$$r_\theta(x,y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

$$z_0 = \mathbf{KL}(\pi_\theta(y'|x) \| \pi_{\text{ref}}(y'|x))$$

$$v(x,y) = \begin{cases} \lambda_D \sigma(\beta(r_\theta(x,y) - z_0)) \text{ if } y \sim y_{\text{desirable}}|x \\ \lambda_U \sigma(\beta(z_0 - r_\theta(x,y))) \text{ if } y \sim y_{\text{undesirable}}|x \end{cases}$$

Intuition:

If the model increases the reward of a desirable example in a blunt manner, then the KL penalty also rises, and no progress is made.

This forces the model to learn exactly what makes an output desirable, so that the reward can be increased while keeping the KL term flat.

Kahneman-Tversky Optimization

KTO loss:

$$L_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x,y\sim D}[\lambda_y - v(x,y)]$$

where

$$r_\theta(x,y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

$$z_0 = \text{KL}(\pi_\theta(y'|x) \| \pi_{\text{ref}}(y'|x))$$

$$v(x,y) = \begin{cases} \lambda_D \sigma(\beta(r_\theta(x,y) - z_0)) \text{ if } y \sim y_{\text{desirable}}|x \\ \lambda_U \sigma(\beta(z_0 - r_\theta(x,y))) \text{ if } y \sim y_{\text{undesirable}}|x \end{cases}$$

What's wrong with this loss?

Problem: estimating z_0 is impractical because sampling from π_θ is slow and humans do not perceive the full distribution induced by π_θ when making judgements.

Simulate human-perceived reference point: create m pairs (x_i, y_j) where y_j is in the same m-sized batch of offline data as x_i.

$$\hat{z}_0 = \max\left(0, \frac{1}{m}\sum_{i\neq j} \log \frac{\pi_\theta(y_j|x_i)}{\pi_{\text{ref}}(y_j|x_i)}\right)$$

Kahneman-Tversky Optimization

Data

Convert preference data $y_w \succ y_l$ by assuming that y_w is drawn from the desirable and y_l from the undesirable distribution.

What are some problems with this approach? How can we mitigate these problems?

Hyperparameters

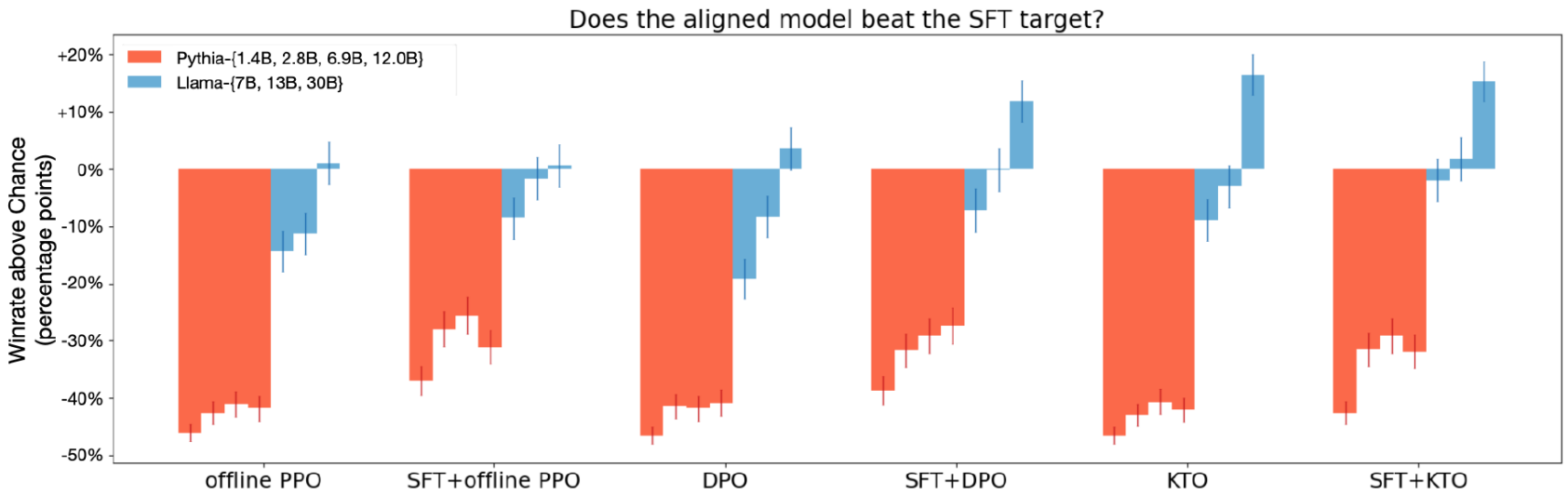
Control the degree of loss aversion with λ_D and λ_U.

$$\frac{\lambda_D n_D}{\lambda_U n_U} \in \left[1, \frac{4}{3}\right]$$

Tune to mitigate class imbalance.

If minimizing the downside more important (e.g., toxicity prevention), set $\lambda_D n_D < \lambda_U n_U$

Evaluation

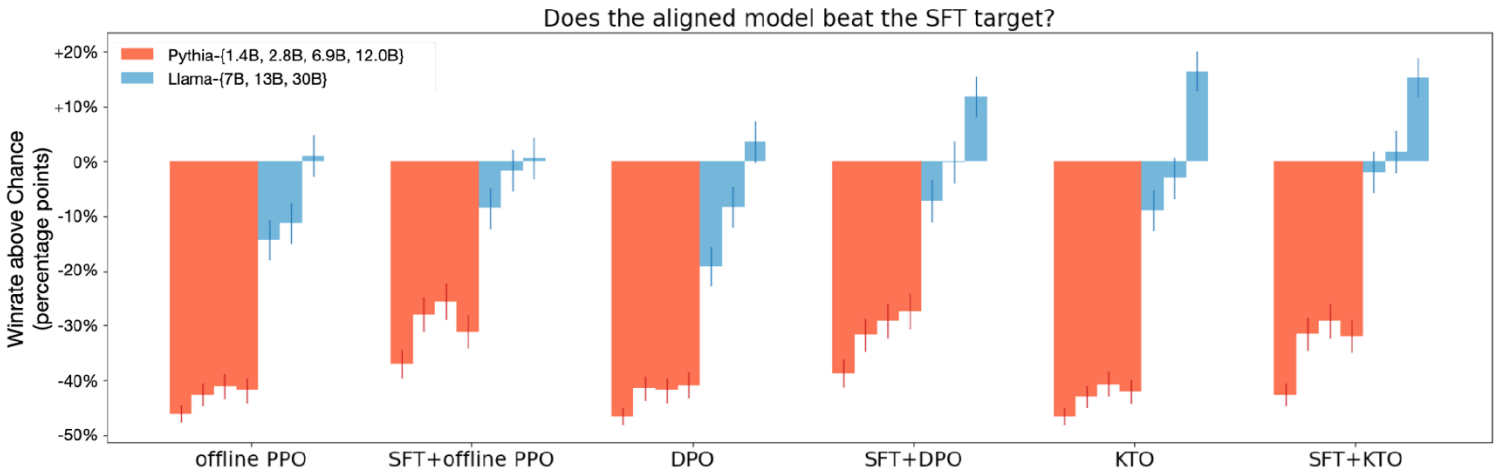


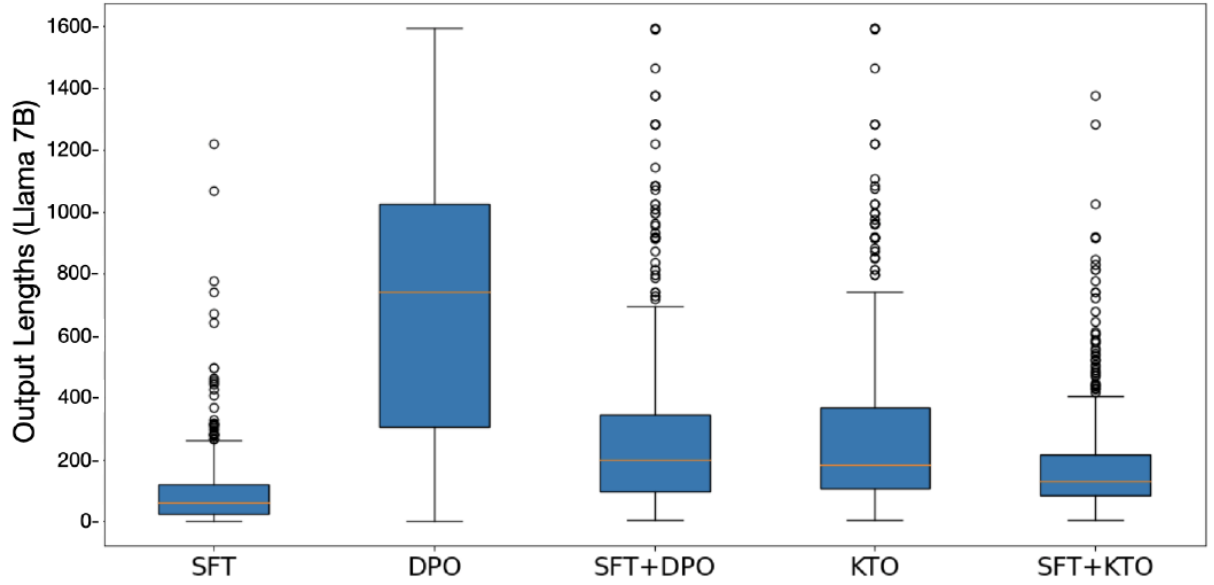
Dataset (→) Metric (→)	MMLU EM	GSM8k EM	HumanEval pass@1	BBH EM
SFT	57.2	39.0	30.1	46.3
DPO	58.2	40.0	30.1	44.1
ORPO ($\lambda = 0.1$)	57.1	36.5	29.5	47.5
KTO ($\beta = 0.1, \lambda_D = 1$)	**58.6**	**53.5**	**30.9**	**52.6**

KTO > DPO

- SFT+KTO comparable to SFT+DPO (1B-30B).
- KTO alone is better than DPO alone for Llama-{7B, 13B, 30B}. No significant difference for Pythia models (why?)

Evaluation



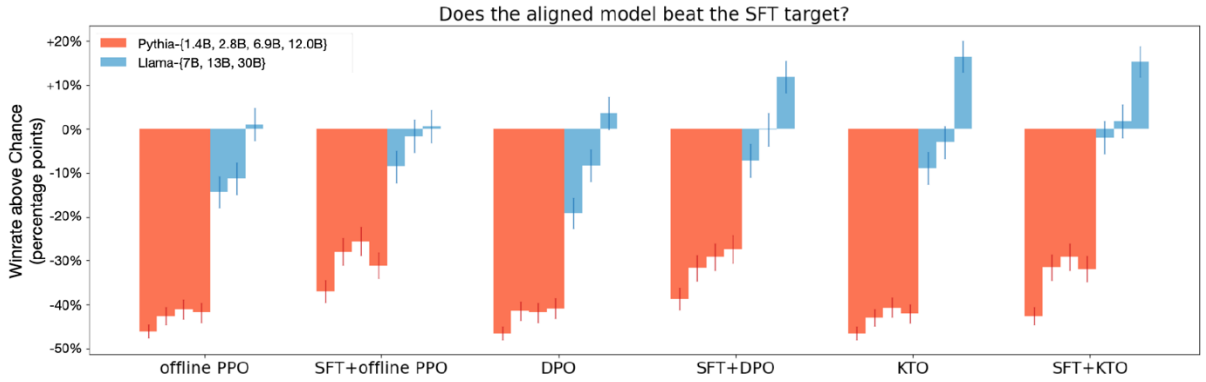


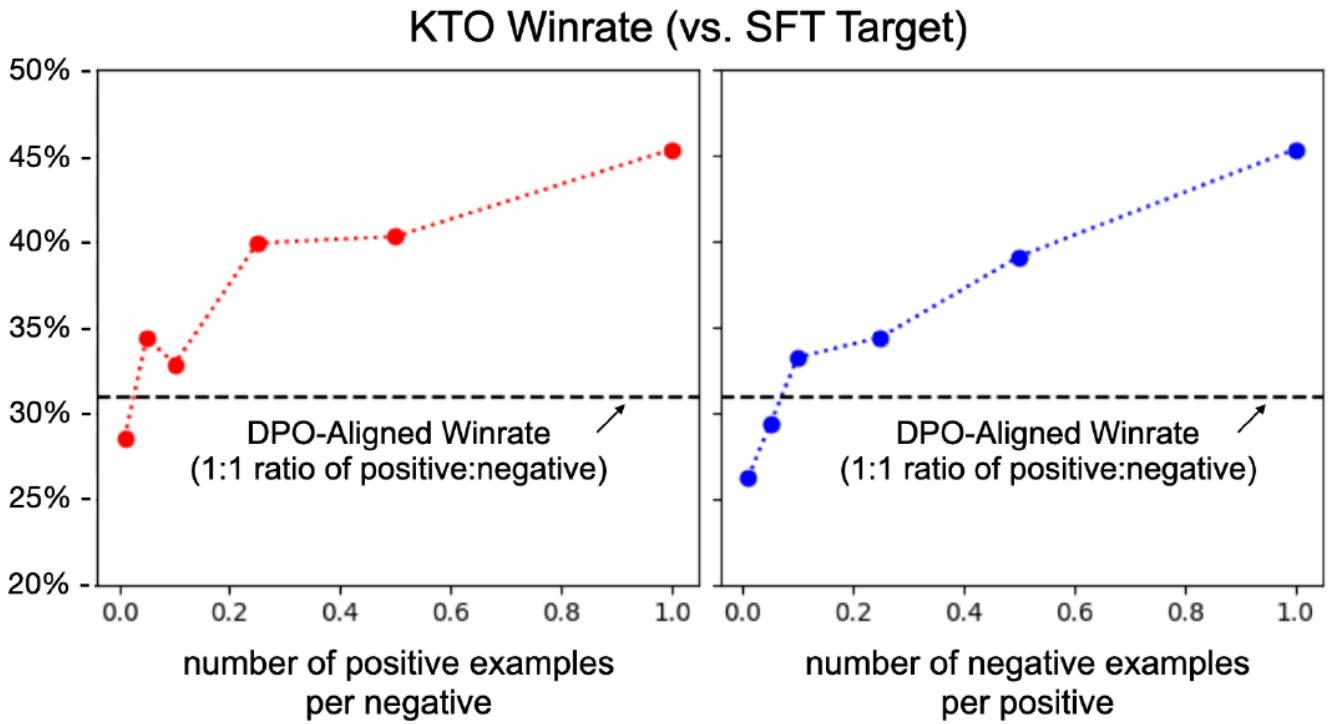
KTO ≈ SFT+KTO

At sufficient scale (Llama-{13B, 30B}) KTO is competitive with SFT+KTO.

This is not case for other methods (e.g., DPO). Why?

Evaluation





Is the difference due to $2n > n$?
For Llama-7B, up to 90% of the desirable data can be discarded while still outperforming DPO.
(tune λ_D and λ_U accordingly)

Method	Winrate vs. SFT Target
Mistral-7B (unaligned)	0.525 ± 0.037
Mistral-7B + DPO	0.600 ± 0.037
Mistral-7B + KTO (all y per x)	$\mathbf{0.652 \pm 0.036}$
Mistral-7B + KTO (one y per x)	0.631 ± 0.036
Mistral-7B-Instruct	0.621 ± 0.031

Evaluation

Remove reference model?

Assume π_{ref} returns a uniform distribution over outputs.

Better than DPO on GSM8k, BBH and worse on MMLU, HumanEval.

Dataset (→) Metric (→)	MMLU EM	GSM8k EM	HumanEval pass@1	BBH EM
SFT	57.2	39.0	30.1	46.3
DPO	58.2	40.0	30.1	44.1
ORPO ($\lambda = 0.1$)	57.1	36.5	29.5	47.5
KTO ($\beta = 0.1, \lambda_D = 1$)	**58.6**	**53.5**	**30.9**	**52.6**
KTO (one-y-per-x)	58.0	50.0	30.7	49.9
KTO (no z_0)	58.5	49.5	30.7	49.0
KTO (concave, $v = \log \sigma$)	58.3	42.5	30.6	43.2
KTO (risk-neutral, $v(\cdot) = \cdot$)	57.3	42.0	28.8	6.1
KTO (no $\pi_{\text{ref}}, \lambda_D = 1.75$)	57.5	47.5	29.5	51.6

Theoretical analysis

Proposition 4.1. *As the reward implied by the current policy tends to $\pm\infty$, the KTO update of π_θ tends to zero.*

$$L_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x,y\sim D}[\lambda_y - v(x,y)]$$

where

$$r_\theta(x,y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

$$z_0 = \text{KL}(\pi_\theta(y'|x) \| \pi_{\text{ref}}(y'|x))$$

$$v(x,y) = \begin{cases} \lambda_D \sigma(\beta(r_\theta(x,y) - z_0)) \text{ if } y \sim y_{\text{desirable}}|x \\ \lambda_U \sigma(\beta(z_0 - r_\theta(x,y))) \text{ if } y \sim y_{\text{undesirable}}|x \end{cases}$$

If a data point (x, y) is implied by the current policy to be too difficult or too easy to learn from, then it is ignored.

Pros: ignore noisy data (mislabeled examples)
Cons: ignore hard-to-learn but necessary data.

tune β (make it smaller)?

Theoretical analysis

Theorem 4.2. *Assuming the value function is logistic, for a reward function r_a^* that maximizes (2), there exists a reward function in its equivalence class (i.e., $r_b^*(x, y) = r_a^*(x, y) + h(x)$ for some $h(x)$) that induces the same optimal policy π^* and the same Bradley-Terry preference distribution but a different human value distribution.*

Value distribution (human utility) is affected by input specific changes – h(x).
Maximizing preference likelihood ≠ maximizing human utility

Human evals:
win rate of KTO: 72.9% (65.2% by GPT-4)
win rate of DPO: 62.1% (60.0% by GPT-4)

Why?

$$L_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x,y\sim D}[\lambda_y - v(x, y)]$$

where

$$r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

$$z_0 = \text{KL}(\pi_\theta(y'|x)\|\pi_{\text{ref}}(y'|x))$$

$$v(x, y) = \begin{cases} \lambda_D \sigma(\beta(r_\theta(x, y) - z_0)) \text{ if } y \sim y_{\text{desirable}}|x \\ \lambda_U \sigma(\beta(z_0 - r_\theta(x, y))) \text{ if } y \sim y_{\text{undesirable}}|x \end{cases}$$

Proof. Following the definition in Rafailov et al. (2023), we say r_a^* and r_b^* are in the same equivalence class if there exists some function $h(x)$ such that $r_b^*(x, y) = r_a^*(x, y) + h(x)$. From Lemma 1 in Rafailov et al. (2023), we know that two functions in the same equivalence class induce the same optimal policy:

$$\begin{aligned}
\pi^*_{r_a}(y|x) &= \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r_a^*(x, y)\right) \\
&= \frac{1}{\sum_y \pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r_a^*(x, y)\right)\exp\left(\frac{1}{\beta}h(x)\right)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r_a^*(x, y)\right)\exp\left(\frac{1}{\beta}h(x)\right) \\
&= \frac{1}{\sum_y \pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}(r_a^*(x, y) + h(x))\right)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}(r_a^*(x, y) + h(x))\right) \\
&= \pi^*_{r_b}(y|x)
\end{aligned}$$

For a Bradley-Terry model of preferences, it is trivial to show that $p(y_w \succ y_l|x)$ is unaffected by $h(x)$ since it is added to the reward of both y_w and y_l. We will now show that the two reward functions do not necessarily induce the same distribution of human values.

A Taylor series expansion of the human value of $r_a^*(x, y)$ would be:

$$\sigma(0) + \sigma'(0)(r_a^*(x, y) - z_0) + \frac{\sigma''(0)}{2}(r_a^*(x, y) - z_0)^2 + ...$$

A Taylor series expansion of the value of $r_a^*(x, y) + h(x)$ around $h(x)$ would be:

$$\sigma(h(x)) + \sigma'(h(x))(r_a^*(x, y) - z_0) + \frac{\sigma''(h(x))}{2}(r_a^*(x, y) - z_0)^2 + ...$$

Since σ is strictly monotonic, for these series to be equal, we must have $h(x) = 0$. If this is not the case, then the values of $r_a^*(x, y)$ and $r_b^*(x, y)$ will be different. Thus two arbitrary reward functions in the same equivalence class do not induce the same distribution of human values. □

Theoretical analysis

Theorem 4.3. *For input x with outputs $\{y_a, y_b\}$, let dataset D comprise contradictory preferences $y_a \succ y_b$ and $y_b \succ y_a$ in proportion $p \in (0.5, 1)$ and $(1-p) \in (0, 0.5)$ respectively. If $p^{1/\beta}\pi_{ref}(y_a|x) < (1-p)^{1/\beta}\pi_{ref}(y_b|x)$, then the optimal DPO policy is more likely to produce the minority-preferred y_b; the optimal KTO policy will strictly produce the majority-preferred y_a for a loss-neutral value function ($\lambda_D = \lambda_U$).*

KTO has better worst-case outcomes when handling feedback intransitivity.

$$L_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x,y\sim D}[\lambda_y - v(x,y)]$$

where

$$r_\theta(x,y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

$$z_0 = \text{KL}(\pi_\theta(y'|x)\|\pi_{\text{ref}}(y'|x))$$

$$v(x,y) = \begin{cases} \lambda_D \sigma(\beta(r_\theta(x,y) - z_0)) \text{ if } y \sim y_{\text{desirable}}|x \\ \lambda_U \sigma(\beta(z_0 - r_\theta(x,y))) \text{ if } y \sim y_{\text{undesirable}}|x \end{cases}$$

Proof. Where $u = \beta(r_\theta(x,y_a) - r_\theta(x,y_b))$, we can write the total DPO loss for x as

$$\mathcal{L}_{\text{DPO}}(x) = p(-\log\sigma(u)) + (1-p)(-\log\sigma(-u))$$

Taking the derivative with respect to u and setting to zero, we get

$$0 = -p\frac{\sigma(u)\sigma(-u)}{\sigma(u)} + (1-p)\frac{\sigma(-u)\sigma(u)}{\sigma(-u)} = -p(1-\sigma(u)) + (1-p)\sigma(u) = -p + \sigma(u)$$

$$\implies u = \sigma^{-1}(p)$$

$$\beta r^*_\theta(x,y_a) = \sigma^{-1}(p) + \beta r^*_\theta(x,y_b)$$

$$\beta\log\frac{\pi^*_\theta(y_a|x)}{\pi_{\text{ref}}(y_a|x)} = \log\frac{p}{1-p} + \beta\log\frac{\pi^*_\theta(y_b|x)}{\pi_{\text{ref}}(y_b|x)}$$

$$\pi^*_\theta(y_a|x) = \left(\frac{p}{1-p}\right)^{1/\beta} \cdot \frac{\pi_{\text{ref}}(y_a|x)}{\pi_{\text{ref}}(y_b|x)} \cdot \pi^*_\theta(y_b|x)$$

Thus when $p^{1/\beta}\pi_{\text{ref}}(y_a|x) < (1-p)^{1/\beta}\pi_{\text{ref}}(y_b|x)$, we have $\pi^*_\theta(y_a|x) < \pi^*_\theta(y_b|x)$, meaning the optimal DPO policy is more likely to produce the minority-preferred y_b.

Where $u_a = \beta(r_\theta(x,y_a) - \mathbb{E}_Q[r_\theta(x,y')])$ and $u_b = \beta(r_\theta(x,y_b) - \mathbb{E}_Q[r_\theta(x,y')])$, noting that $1-\sigma(-u) = \sigma(u)$, we can write the total KTO loss for x as

$$\begin{aligned}\mathcal{L}_{\text{KTO}}(x) &= p\lambda_D(1-\sigma(u_a)) + (1-p)\lambda_U\sigma(u_a) + p\lambda_U\sigma(u_b) + (1-p)\lambda_D(1-\sigma(u_b)) \\ &= p\lambda_D + ((1-p)\lambda_U - p\lambda_D)\sigma(u_a) + (1-p)\lambda_D + (p\lambda_U - (1-p)\lambda_D)\sigma(u_b) \\ &= \lambda_D + ((1-p)\lambda_U - p\lambda_D)\sigma(u_a) + (p\lambda_U - (1-p)\lambda_D)\sigma(u_b) \\ &= \lambda_D + \lambda_D((1-2p)\sigma(u_a) + (2p-1)\sigma(u_b)) \quad \text{(under loss neutrality)}\end{aligned}$$

Given that $p > 0.5$ by assumption and $\lambda_D > 0$ by definition, the KTO loss is decreasing in u_a and increasing in u_b—and thus decreasing in $r_\theta(x,y_a)$ and increasing in $r_\theta(x,y_b)$ respectively. The optimal KTO policy is thus $\pi^*_\theta(y|x) = \mathbb{1}[y = y_a]$. □

Which one?

KTO

- Binary-formatted imbalanced human feedback.
- Preference data: noisy feedback with intransitivity (theorem 4.3).

DPO

- Little noise and little intransitivity (KTO might underfit – proposition 4.1).

Discussion

KTO has a default loss. The parameters (e.g., loss aversion) differ across individuals in behavioral studies.

How can we efficiently adapt the current KTO formulation to account for this? Would this make a big difference?

KTO (value function) was inspired from behavioral studies on monetary gambles, which differ from how humans perceive and interpret text.

What analogous settings in human behavior could provide closer inspiration for AI alignment? And how can we refine the KTO formulation accordingly?