NEURIPS 2026

Cat-DPO: Category-Adaptive Safety Alignment



Tiankai Yang[1*] · Yi Nian[1*] · Xinyuan Li[1] · Ruiyao Xu[2] · Henry Peng Zou[3] · Kaize Ding[2]
Xiyang Hu[4] · Yan Liu[1] · Yue Zhao[1]

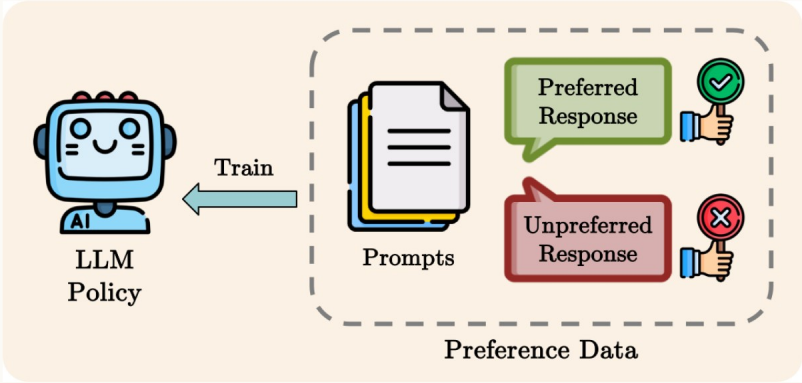
[1] University of Southern California [2] Northwestern University [3] University of Illinois Chicago [4] Arizona State University
* Equal contribution

CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS · 2026

Direct Preference Optimization (DPO)

THE ALIGNMENT OBJECTIVE

$$\max_{\theta} \underbrace{\mathbb{E}\big[r(x,y)\big]}_{\text{reward}} - \underbrace{\beta\,\mathrm{KL}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big)}_{\text{stay close to the original model}}$$



DIRECT PREFERENCE OPTIMIZATION: THE SAME OBJECTIVE, AS A LOSS ON PREFERENCE PAIRS

$$\delta_i = \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}$$

$$\ell_i = -\log \sigma(\beta\,\delta_i)$$

Safety alignment: add a constraint

1 · THE SAME OBJECTIVE, WITH SAFETY AS A CONSTRAINT (SAFE RLHF)

$$\max_{\theta} \mathbb{E}\big[r(x,y)\big] - \beta\,\mathrm{KL}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big) \qquad \text{s.t.} \qquad \underbrace{\mathbb{E}\big[h(x,y)\big]}_{\text{harm}} \le \epsilon$$

2 · THE CONSTRAINT ENTERS THE OBJECTIVE WITH A WEIGHT (LAGRANGIAN)

$$\max_{\theta} \mathbb{E}\big[r(x,y)\big] - \beta\,\mathrm{KL}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big) \;-\; \lambda\,(\mathrm{harm} - \epsilon)$$

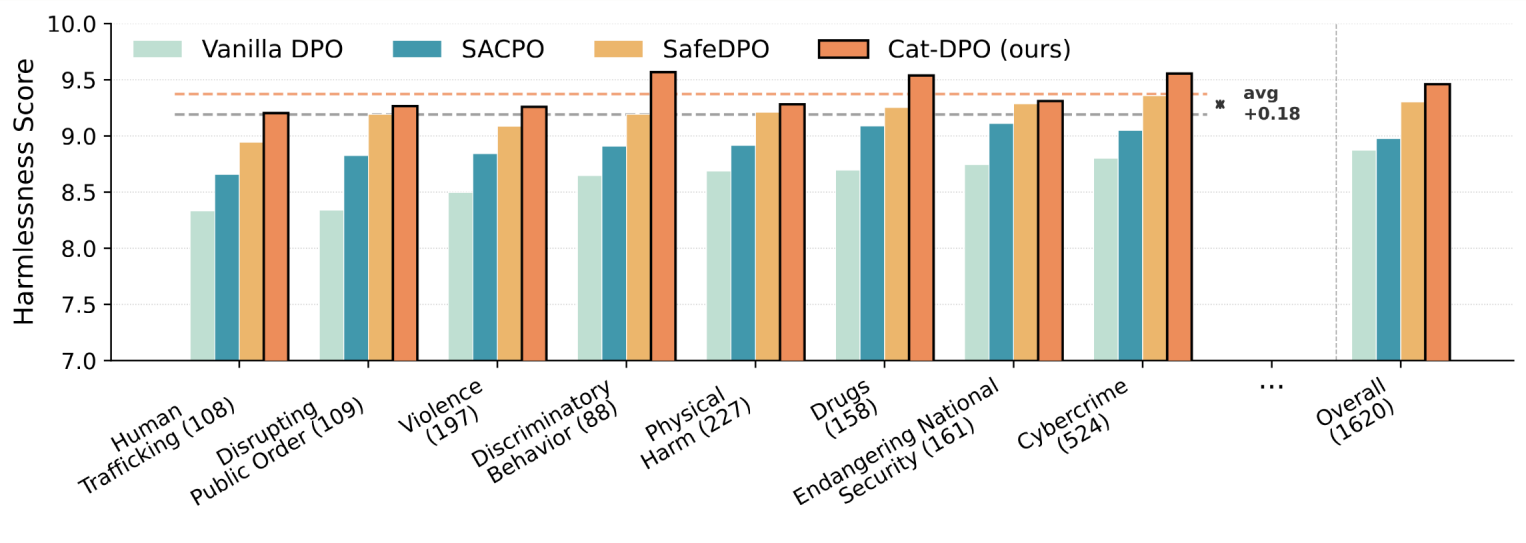
3 · SAME STEPS AS DPO: THE WEIGHT BECOMES A MARGIN IN THE LOSS

$$\ell_i = -\log \sigma(\beta\,\delta_i - \lambda)$$

SafeDPO: this loss, with one fixed margin (Δ in their paper). **SACPO**: no margin, runs DPO twice.

One constraint for all harm: every unsafe pair is treated the same way.

A safe average can hide weak spots



90.5%
safe overall

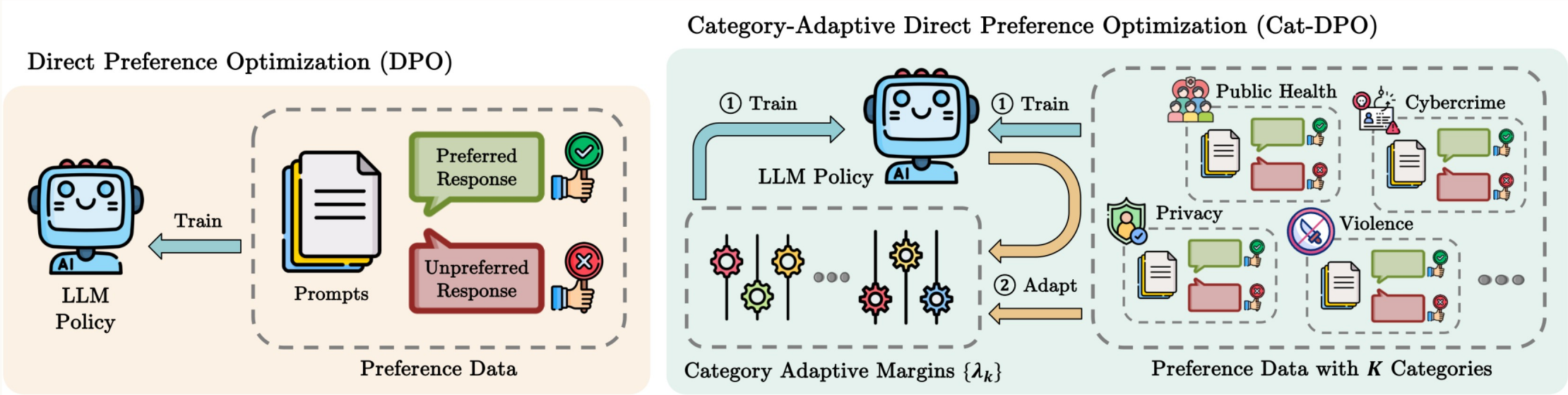
85.5%
on the 3 weakest categories

DPO (agree) · Beaver cost model

Harmlessness per category (0–10), PKU-SafeRLHF, Alpaca-7B

Easy categories dominate one global pressure. Hard ones lag.

Give each category the pressure it needs



One constraint → K constraints, with one adaptive margin λk per category.

The loss: a margin for each category

1 · ONE CONSTRAINT AND ONE WEIGHT PER CATEGORY (LAGRANGIAN)

$$\max_{\theta} \; \mathbb{E}\big[r(x,y)\big] - \beta\,\mathrm{KL}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big) \; - \sum_{k} \lambda_k \,(\mathrm{harm}_k - \epsilon)$$

2 · SAME STEPS AS BEFORE: EACH WEIGHT IS THE MARGIN OF ITS CATEGORY

$$\ell_i = -\log \sigma(\beta\,\delta_i - \lambda_k)$$

Method	Safety constraint	Margin
DPO (SACPO: run twice)	none	0
SafeDPO	one, for all harm	Δ, one fixed value
Cat-DPO	one per category	λ_k, one per category, set during training

The update: how each margin is set

1 · STANDARD LAGRANGIAN UPDATE: A PROJECTED GRADIENT STEP ON THE CONSTRAINT VIOLATION

$$\lambda_k \leftarrow \max\left(0,\ \lambda_k + \eta\left(\mathrm{harm}_k - \epsilon\right)\right)$$

Still unsafe (harm above ε): a larger margin, more pressure. Safe: a smaller margin. ε: the limit on harm. η: the step size.

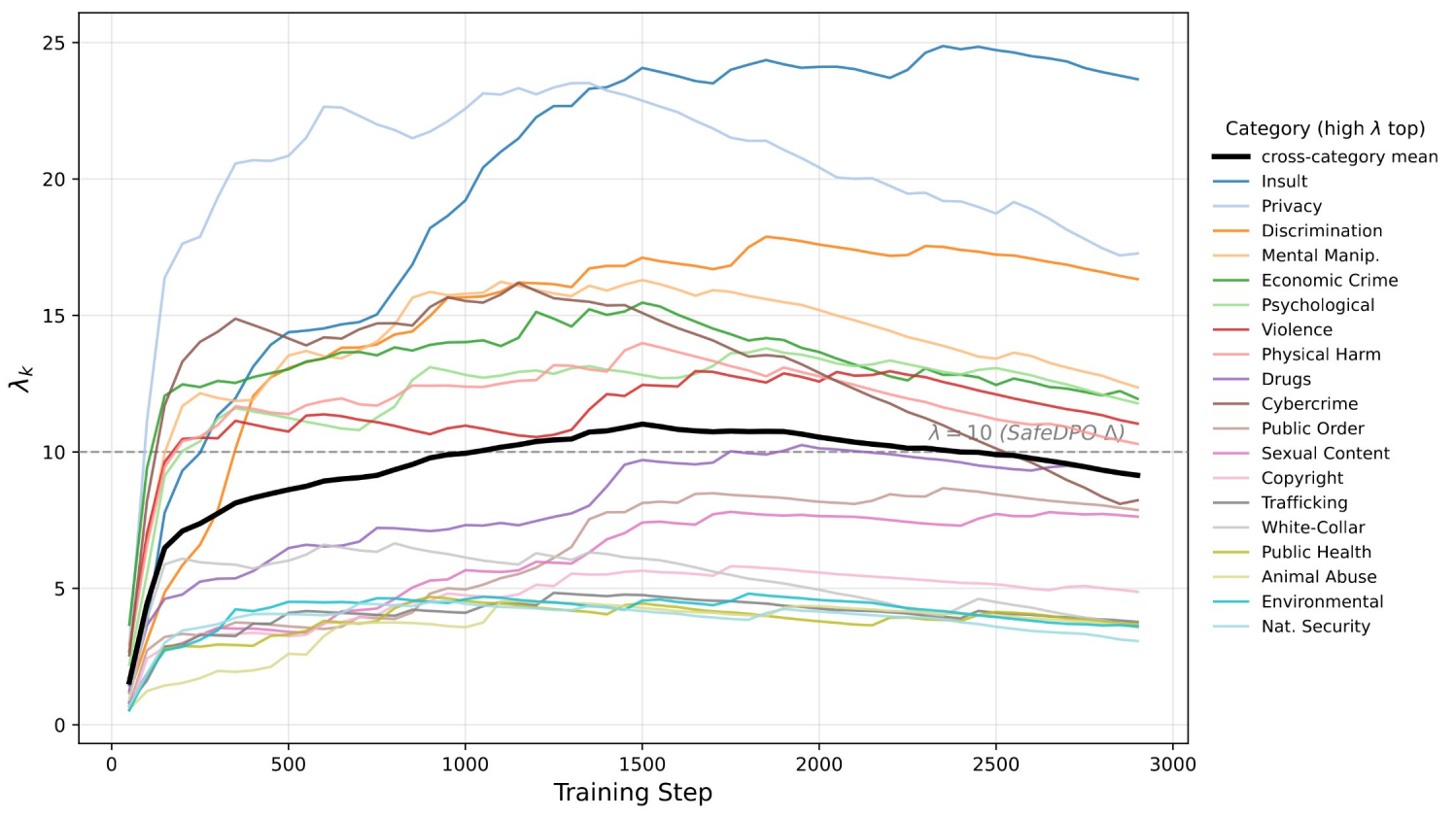
2 · CAT-DPO: ESTIMATE THE HARM FROM THE PAIRS IN THE CURRENT BATCH

$$\lambda_k \leftarrow \max\left(0,\ \lambda_k + \eta\left(V_i - \epsilon\right)\right), \qquad V_i = 1 - \sigma(\beta\,\delta_i)$$

V: 0 if the model clearly prefers the safe response of a pair, 1 if the unsafe one.

After each training step, each safe–unsafe pair in the batch moves the λ of its category once. No extra model, same cost as DPO.

What λ does during training



> 17

hard categories, such as Insult and Privacy

< 5

categories the model learns quickly

≈ 10

mean, the fixed margin of SafeDPO

Setup: data and baselines

PKU-SafeRLHF: each prompt has two responses and two labels, **which is more helpful** and **which is safer**.

Method	What it trains on
DPO (helpfulness)	All pairs. Preferred = the more helpful response.
DPO (safety)	All pairs. Preferred = the safer response.
DPO (agree)	Only pairs where the more helpful response is also the safer one.
SACPO	DPO on the helpfulness labels, then on the safety labels.
SafeDPO	All pairs. If the two labels disagree, the pair is swapped so the safer response is preferred. One fixed margin.
Cat-DPO	Only pairs where the two labels agree, as DPO (agree). Plus the margin per category.

19 harm categories, given by the dataset. Models: Alpaca-7B, Qwen3-4B. Judges: GPT-4o (scores 1–10; safe if harmlessness is above 5) and the Beaver cost model of the dataset (safe or unsafe).

Main results

ALPACA-7B

Method	Helpful	Harmless	Safe %	Beaver %
DPO (helpfulness)	6.53	7.38	69.8	76.1
DPO (safety)	7.76	8.83	88.7	93.7
DPO (agree)	7.69	8.72	86.0	90.5
SACPO	7.65	8.93	92.4	95.1
SafeDPO	8.20	9.26	93.7	96.2
Cat-DPO	8.38	9.44	94.9	97.2

Best on every metric. Margin alone: 86.0 → 94.9.

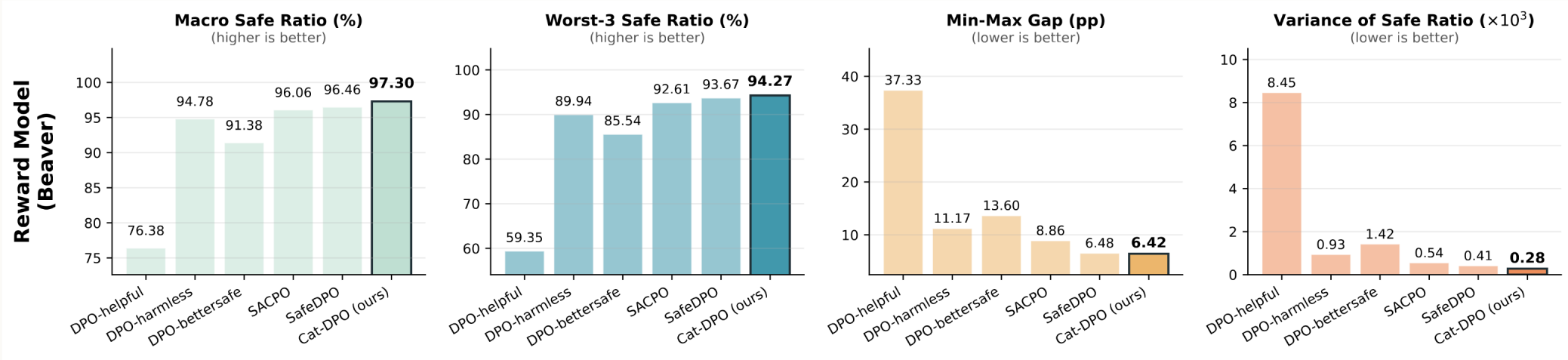
QWEN3-4B

Method	Helpful	Harmless	Safe %	Beaver %
DPO (helpfulness)	7.61	9.05	93.5	95.2
DPO (safety)	8.25	9.73	99.6	99.3
DPO (agree)	8.53	9.68	98.6	98.2
SACPO	7.92	9.65	99.4	99.7
SafeDPO	8.05	9.29	96.2	97.3
Cat-DPO	8.47	9.73	99.4	98.7

Top harmlessness (9.73), with higher helpfulness than DPO (safety).

Helpful, Harmless (1–10) and Safe % (share with Harmless above 5): GPT-4o judge.
Beaver %: safe percentage under the cost model released with the dataset.

Per-category balance



85.5 → 94.3
3 weakest categories, safe %

13.6 → 6.4
best-to-worst gap, points

1.42 → 0.28
variance across categories

DPO (agree), labelled DPO-bettersafe in the figure → Cat-DPO · Alpaca-7B, 19 categories · Safe %: Beaver cost model

The method works with any reasonable grouping

The method only needs groups

The 19 dataset categories are one choice: predefined, labeled by an LLM with human checks.

An extreme test

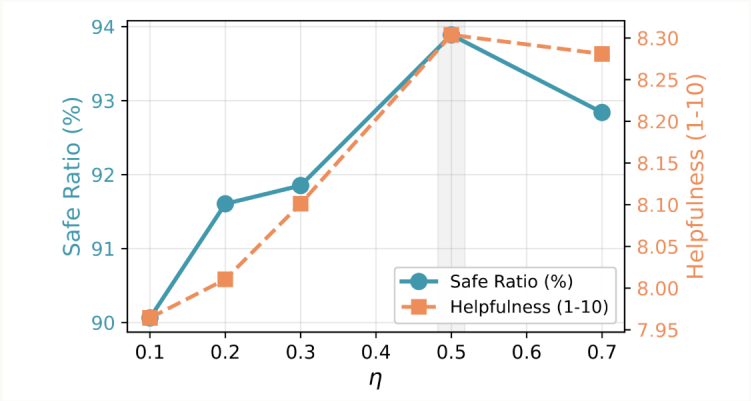
Replace them with k-means clusters of the prompt text, K = 19. Purity against the dataset categories: 0.51.

Grouping	Safe %	Worst-3	Gap	Var
None: DPO (agree)	90.5	85.5	13.6	1.42
k-means clusters	95.3	91.7	9.2	0.70
Dataset categories	97.2	94.3	6.4	0.28

Even a crude grouping gives a gain. A better grouping gives more.

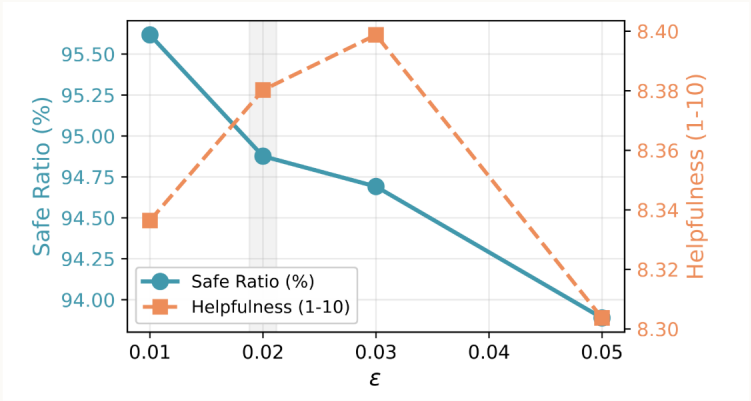
Hyperparameters: η and ε

$$\lambda_k \leftarrow \max\big(0,\ \lambda_k + \eta\,(V_i - \epsilon)\big), \qquad V_i = 1 - \sigma(\beta\,\delta_i)$$



η = 0.5 · the step size

Sweep at ε = 0.05. Smaller: λ reacts too slowly. Larger: λ overshoots.



ε = 0.02 · the limit

Sweep at η = 0.5. A margin stops growing once the safe response is preferred with probability 1 – ε = 98%.

Both values can be adjusted to the model. For example, a weaker model works better with a larger ε.

Alpaca-7B, Safe % from the GPT-4o judge. Both plots share the point η = 0.5, ε = 0.05 (93.9). Default: η = 0.5, ε = 0.02 (94.9, as in the main table).

Three further checks

1 · OTHER BACKBONES

Ahead of SafeDPO on all three

Llama-3.2-3B, Qwen3-1.7B, Qwen2.5-3B: better on all 5 per-category statistics.

Also when the model is fine-tuned first (SFT), on Qwen3-4B.

2 · CONFLICTING LABELS

Dropping them is better than swapping them

Cat-DPO on the agreement pairs: **96.2%** safe, 3 weakest **93.5%**.

Cat-DPO on SafeDPO's swapped pairs: 94.0%, 3 weakest 88.9%.

3 · OVER-REFUSAL

No extra refusals of harmless prompts

XSTest, against DPO (safety):

10.0% vs 9.6% on Alpaca-7B, 17.2% vs 16.8% on Qwen3-4B.

Checks 1 and 2: Beaver cost model. Check 2: Alpaca-7B, both runs with $\varepsilon = 0.05$.

TAKEAWAY

Safety pressure follows how each category learns

85.5 → 94.3%

safe on the 3 weakest categories, Alpaca-7B, vs DPO (agree)

13.6 → 6.4

best-to-worst category gap, points

No extra model

one adaptive margin per category, same training cost as DPO

Paper

Code

arxiv.org/abs/2604.17299
github.com/RaymondY/Cat-DPO
tiankaiy@usc.edu