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

tiankaiy@usc.edu

Paper Code LinkedIn

NeurIPS 2026

MOTIVATION

A safe average can hide weak spots.

Model ability · Category difficulty · Data quality

One global training pressure

Privacy Cybercrime Public health

Alpaca · Beaver

DPO (agree)

90.5% Overall safety

85.5% Weakest-3 safety

02 Main Results

GPT-4o evaluator · ↑

Method	Helpfulness	Harmlessness	Safe %
Alpaca-7B			
DPO (helpfulness)	6.53	7.38	69.8
DPO (safety)	7.76	8.83	88.7
DPO (agree)*	7.69	8.72	86.0
SACPO	7.65	8.93	92.4
SafeDPO	8.20	9.26	93.7
Cat-DPO	**8.38**	**9.44**	**94.9**
Qwen3-4B			
DPO (helpfulness)	7.61	9.05	93.5
DPO (safety)	8.25	9.73	99.6
DPO (agree)*	8.53	9.68	98.6
SACPO	7.92	9.65	99.4
SafeDPO	8.05	9.29	96.2
Cat-DPO	**8.47**	**9.73**	**99.4**

* Agree: both labels choose the same answer.
Mean scores: 1–10; Safe %: harmlessness > 5.

Higher, more even safety

Alpaca / Beaver

Category mean ± SD 80–100%

DPO (agree) 91.38 ± 3.77

SafeDPO 96.46 ± 2.02

Cat-DPO 97.30 ± 1.67

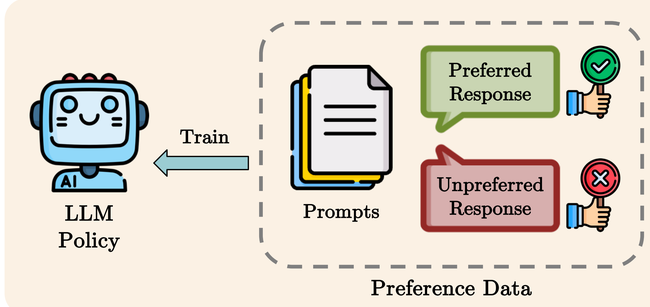
Weakest-3: **85.5 → 94.3%** · gap: **13.6 → 6.4 pp**

Yes, 3 seeds. Alpaca / GPT-4o · Safe % (mean ± SD)

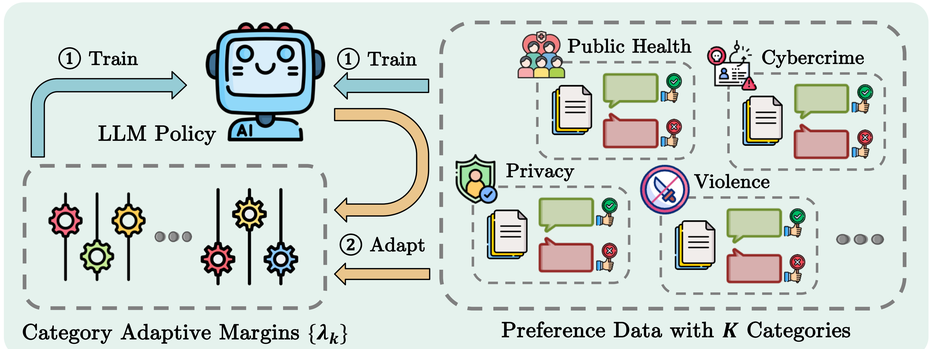
DPO(agree) 88.2±1.3 · SafeDPO 92.9±1.6 · Cat-DPO 94.8±1.2

01 Give Each Category the Pressure it Needs

Direct Preference Optimization (DPO)



Category-Adaptive Direct Preference Optimization (Cat-DPO)



Learning feedback sets the next margin

Illustrative update · λ_k starts at 10 ε=.02 · η=.5

DPO safe preference → Error vs ε=2% → Next pressure λ_k

90% still learning → 10% error > 2% → 10.040 ↑

99% clear preference → 1% error < 2% → 9.995 ↓

ε .02→.05: target 98%→95% η .1→.5: update ×5

Vanilla DPO βδ: safe-answer lead

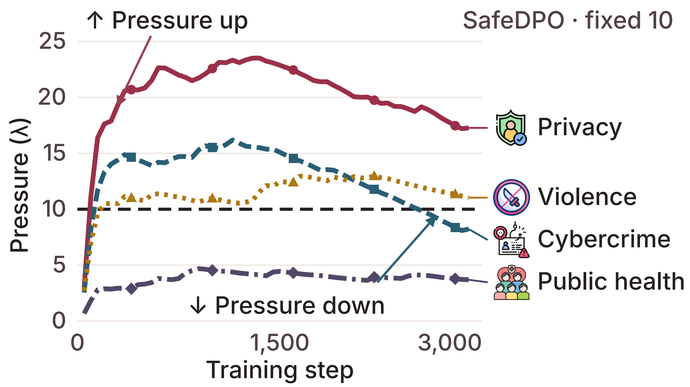
$$\ell_i = -\log \sigma(\beta\delta_i)$$

Cat-DPO m_i = max active λ_k

$$\ell_i = -\log \sigma(\beta\delta_i - m_i(t))$$

Fixed vs adaptive pressure

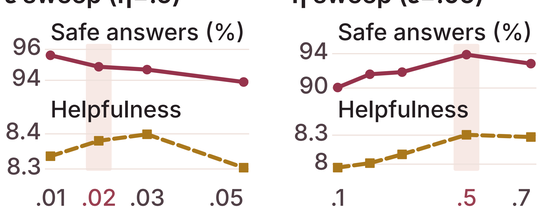
Alpaca-7B



03 Tune to Model Capacity

Main: ε=.02, η=.5 Alpaca · GPT-4o

ε sweep (η=.5) η sweep (ε=.05)



Small Qwen3: ε=.05; η=.5 unchanged.

04 Beyond Preset Categories

No harm-category labels?

Cluster pairs → Train Cat-DPO

Final safety · Beaver 80–100%

DPO (agree) 90.5%

Auto clusters 95.3%

Dataset labels 97.2%

+4.8 pp safety without category labels.

Same pairs/settings · Alpaca / Beaver

05 Over-Refusal

Alpaca-7B

Benign requests refused (%) ↓

	Over-refusal ↓	Safety ↑
Vanilla DPO	9.6%	74.0%
SafeDPO	10.8%	68.5%
Cat-DPO	**10.0%**	**76.5%**

Higher safety, ~10% over-refusal.

XSTest · 250 benign / 200 unsafe
GPT-4o · DPO uses safety labels.