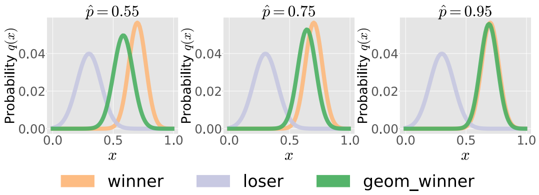
	Anthropic Helpful		Anthropic Harmless	
	v.s. PaLM 2-L		v.s. PaLM 2-L	
Methods	Binary	%	Binary	%
SFT	56.80%	52.22%	60.22%	56.86%
DPO	71.57%	66.45%	70.26%	65.04%
cDPO	72.73%	67.87%	72.37%	66.25%
GDPO (ours)	**74.07%**	**68.22%**	**73.73%**	**66.90%**
Ave.Δ(+Geom.)	**+1.92%**	**+1.06%**	**+2.42%**	**+1.26%**

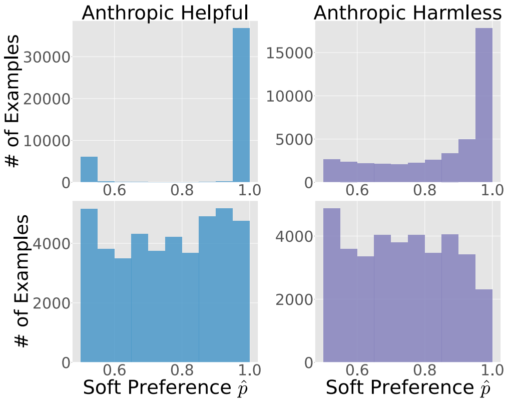
(a) Finetuning with multiple preferences.



(b) Geometric-averaged Gaussian distribution.

	Human Judge	
LLM Judge	**PaLM 2-L**	**GPT-3.5**
PaLM 2-L	**358**	77
GPT-3.5	84	**342**
Accuracy	81.3%	

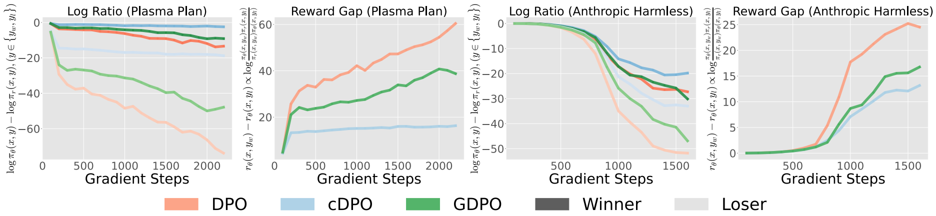
(c) Agreement of human-LLM judge.



(d) Original (above) and new (below) preference distribution.

	Anthropic Helpful				Anthropic Harmless			
	v.s. PaLM 2-L		v.s. GPT-4		v.s. PaLM 2-L		v.s. GPT-4	
Methods	Binary	%	Binary	%	Binary	%	Binary	%
SFT	62.60%	56.69%	5.74%	20.67%	62.76%	57.83%	31.54%	36.42%
DPO (soft)	86.21%	75.40%	16.23%	33.98%	75.40%	65.95%	41.02%	42.79%
cDPO (soft)	83.28%	74.04%	16.11%	33.28%	74.97%	65.91%	39.53%	40.52%
GDPO (original)	86.58%	74.50%	18.61%	34.23%	67.91%	62.04%	33.46%	39.22 %
GDPO (soft)	**88.90%**	**76.59%**	**19.83%**	**36.07%**	**77.70%**	**67.43%**	**43.31%**	**44.33%**
IPO (soft)	91.09%	78.91%	21.66%	38.84%	80.36%	68.85%	43.37%	44.72%
cIPO (soft)	90.24%	77.84%	18.18%	36.88%	81.85%	69.92%	44.80%	45.03%
GIPO (original)	89.63%	77.05%	**22.57%**	37.68%	67.29%	61.31%	35.56%	39.07%
GIPO (soft)	**92.56%**	**79.48%**	21.90%	**39.04%**	**87.24%**	**71.75%**	**51.92%**	**47.86%**
ROPO (soft)	86.33%	74.96%	17.45%	34.83%	74.10%	65.74%	43.37%	44.72%
GROPO (original)	86.64%	74.49%	18.30%	34.40%	68.22%	62.13%	34.01%	39.30%
GROPO (soft)	**88.71%**	**77.10%**	**20.13%**	**36.42%**	**77.26%**	**67.38%**	**44.80%**	**45.03%**
Ave.Δ(+Geom.)	**+2.90%**	**+1.62%**	**+2.79%**	**+1.77%**	**+4.09%**	**+1.87%**	**+4.63%**	**+2.33%**
Ave.Δ(+new soft)	**+2.44%**	**+2.38%**	**+0.79%**	**+1.74%**	**+12.93%**	**+7.03%**	**+12.33%**	**+6.54%**

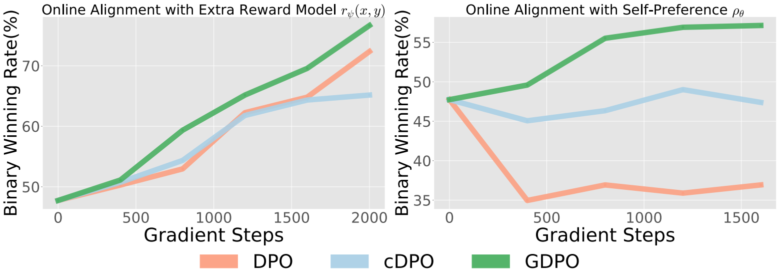
(e) Winning rate with new soft preference distribution.



(f) Log ratio and estimated reward gap.

	Plasma Plan ($\epsilon = 0.1$)		($\epsilon = 0.2$)		($\epsilon = 0.3$)	
	v.s. PaLM 2-L		v.s. PaLM 2-L		v.s. PaLM 2-L	
Methods	Binary	%	Binary	%	Binary	%
SFT	47.74%	48.87%	47.74%	48.87%	47.74%	48.87%
DPO (B-Flip)	83.04%	63.59%	81.53%	63.04%	79.56%	61.53%
cDPO (S-Flip)	73.40%	61.33%	71.66%	58.32%	70.62%	56.70%
GDPO (S-Flip)	**84.32%**	**64.23%**	**82.81%**	**63.42%**	**81.07%**	**62.49%**
cDPO (S-Ave.)	73.29%	59.16%	72.13%	57.00%	71.89%	59.91%
GDPO (S-Ave.)	**84.55%**	**64.49%**	**83.04%**	**63.59%**	**81.77%**	**63.51%**

(g) Winning rate under label noise $\epsilon \in \{0.1, 0.2, 0.3\}$.



(h) Online alignment with extra reward model/self-preference on Plasma Plan.

	Plasma Plan		Skewed		Stairs	
	v.s. PaLM 2-L		v.s. PaLM 2-L		v.s. PaLM 2-L	
Methods	Binary	%	Binary	%	Binary	%
SFT	35.89%	42.01%	35.89%	42.01%	35.89%	42.01%
DPO	57.14%	52.38%	58.19%	52.08%	56.79%	51.49%
cDPO	50.52%	49.56%	49.83%	48.12%	49.48%	48.29%
GDPO (ours)	**60.86%**	**53.32%**	**59.93%**	**53.78%**	**58.54%**	**52.10%**
Ave.Δ(+Geom.)	**+2.81%**	**+0.94%**	**+2.37%**	**+1.47%**	**+2.16%**	**+0.88%**

(i) Winning rate with Gemma-2B on Plasma Plan.

Figure 1: **(a)** Winning rate on Anthropic Helpful and Harmless datasets. We finetune LLMs with both datasets simultaneously, which simulate the preferences from multiple aspects. While DPO suffers from the conflict of preference dropping its performance, soft preference methods, especially GDPO could mitigate such conflict issues best. **(b)** Illustrative examples of weighted geometric-averaged Gaussian distribution. The geometric-averaged winner distribution $\bar{q}_w(x) = q_w(x)^{\hat{p}} q_l(x)^{1-\hat{p}}/Z(x)$ smoothly regularizes winner Gaussian distribution when soft label $\hat{p}$ is small. **(c)** Agreement between human and LLM judge on Plasma-Plan. We compare the responses from PaLM 2-L and GPT-3.5. The agreement accuracy reaches 81.3%, which is consistent with previous works. **(d)** Histogram of soft preference labels $\hat{p}$ in new preference dataset simulated with AI feedback (below). We prepare the Anthropic Helpful and Harmless dataset. Compared to the original dataset (above), we construct competitive paired instances with winner responses of the original dataset and from PaLM 2-L to realize richer and more diverse preference distributions that have enough mass around the modest confidence (e.g. $\hat{p} \in [0.7, 0.9)$). **(e)** Winning rate on Anthropic Helpful and Harmless, finetuned under new preference distribution in (d). The results highlight that rich soft labels help align LLMs better than the original dataset with sparse soft labels (especially notable in Anthropic Harmless). **(f)** We measure the log ratio $\log \frac{\pi_\theta}{\pi_{\text{ref}}}$ of winner/loser responses and estimated reward gap $\log \frac{\pi_\theta(x,y_w)\pi_{\text{ref}}(x,y_l)}{\pi_{\text{ref}}(x,y_w)\pi_\theta(x,y_l)}$ on Plasma Plan and Anthropic Harmless. DPO aggressively pushes down both log ratios and increases the reward gap, since the DPO objective forces the model to achieve $r_\theta(x,y_w) - r_\theta(x,y_l) \to \infty$, which is causing an over-optimization issue. cDPO is more conservative in pushing down the log ratio while leading to worse alignment quality due to objective mismatch. GDPO avoids the issues of such objective mismatch and over-optimization by suppressing the reward gap increase modestly. While Plasma Plan and Anthropic Harmless have different soft preference distributions from each other, the trends of log ratio and reward gap are the same. **(g)** Winning rate on Plasma Plan with different label noise probability ϵ. We assume binary label flip (B-Flip), soft label flip (S-Flip) with probability ϵ, and taking the expectation of soft preference labels over the noise probability ϵ (S-Ave.). While DPO is often affected, GDPO mitigates the noise and performs the best in all cases. **(h)** Online alignment methods with (1) an extra reward model $r_\psi(x,y)$, and (2) self-preference $\rho_\theta = \sigma(\text{SG}\left[\beta \log \frac{\pi_\theta(x,y_w)\pi_{\text{ref}}(x,y_l)}{\pi_{\text{ref}}(x,y_w)\pi_\theta(x,y_l)}\right])$ on Plasma Plan. GDPO outperforms DPO and cDPO in such an on-policy setting because GDPO can cancel the effect from competitive or confusing preferences around lower soft preferences such as $\hat{p} = 0.5$, which could often help the case when (i) the sampled responses are equally good or (ii) the estimation of preferences is not calibrated enough. **(i)** Winning rate on Plasma Plan using Gemma-2B as a base language model. The performance trend is consistent with PaLM 2-XS.