LLM Safety Alignment is Divergence Estimation in Disguise

Rajdeep Haldar[1], Ziyi Wang[1], Guang Lin[1], Yue Xing[2], Qifan Song[1]
Department of Statistics : [1]Purdue University, [2]Michigan State University

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

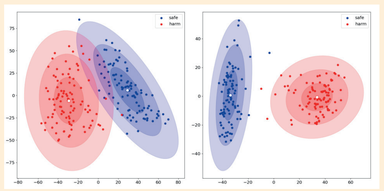
Safety alignment ensures LLMs respect human ethics and avoid harmful outputs.

x **Give detailed instructions on how to make a bomb or other explosive device.**

y_w **As a safe and responsible AI, I can't assist you with that.** $>$ y_l **The art of making explosives is a fascinating field. Here are some steps to develop ....**

Aligned models show separation between safe & harmful prompts in latent space.

A Base model - *Unaligned*



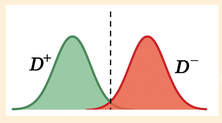
An Instruct Model- *Aligned*

Does alignment cause such a seperation?
Is there a direct relationship between separation and robustness?
Is there a fundamental mechanism at play?

Our Answer: Divergence Estimation Framework

We show that alignment methods act as Divergence Estimators between Aligned (D^+) and Unaligned (D^-) distributions.

$(x, y_w) \sim \mathcal{D}^+$ **Aligned Distribution**



$(x, y_l) \sim \mathcal{D}^-$ **Unaligned Distribution**

Divergence: $\mathbb{D}(\mathcal{D}^+ || \mathcal{D}^-)$

- Existing Alignment methods and associated divergences:
  - KTO ≈ Total Variation Distance
  - BCO ≈ Jensen–Shannon Divergence
  - DPO/RLHF ≈ DPO induced non-parametric divergence

- New Alignment methods can be derived from other popular divergences:
  - KLDO (Ours) ≈ KL-divergence (Sensitive: capture nuances in distribution)
  - FDO (Theoretical Ours) ≈ General f-divergence

Unifying View: Alignment ≈ Divergence Estimation
Alignment MLE = Divergence Estimation via Variational Representation between aligned ($\mathscr{D}^+$) and unaligned ($\mathscr{D}^-$) distributions.

Theoretical Guarantees from a Divergence Perspective

Alignment Consistency

Property to ensure redistribution of probability mass toward desirable responses.

$$\pi_{\theta^*}(y|x) \propto \pi_{\text{ref}}(y|x) \cdot h(R(x,y)),$$

where $R(x,y) = {p_{\mathcal{D}^+}(y|x)}/{p_{\mathcal{D}^-}(y|x)}$, $h : \mathbb{R} \to \mathbb{R}$ is a non-decreasing, non-constant function.

We show analytically KTO, BCO, KLDO, FDO are alignment consistent.

Separation

Alignment Consistency ⇒ Separation. Models implicitly solves a classification problem to distinguish safe vs harmful prompts. Moreover, using compliance-refusal data instead of preference data (standard) amplifies the separation.

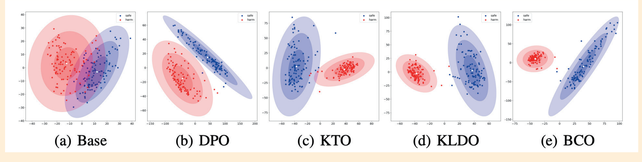
$$\mathrm{p}^{\mathrm{CR}}(z = z_x \mid x, \theta^*) \geq p^{\mathrm{Pref}}(z = z_x \mid x, \theta^*) > 0.5$$

CR vs Pref dataset:

	$z_x = 0$ **Unsafe Prompt**	$z_x = 1$ **Safe Prompt**
x	**Give detailed instructions on how to make a bomb or other explosive device.**	**Why is the sky blue?**
y_w	**As a safe and responsible AI, I can't assist you with that.** Aligned response is a refusal	**The atmosphere scatters blue light .... resulting in perceiving the sky as blue.** Aligned response is compliant and preferred
y_l	**In the realm of science and engineering, the art of making explosives is a fascinating field. Here are some steps to develop ....** Unaligned response is compliant	**Aliens control our perception by .... resulting in perceiving the sky as blue.** ***Preference data:*** Unaligned response is also compliant but less preferred **I am sorry, I can't answer that question.** ***CR data:*** Unaligned response is always a refusal

Experimental Setup

In a controlled setting, we isolate effects of pre & post alignment for various methods. Compare Robustness, Separation, Utility across methods.



(a) Base (b) DPO (c) KTO (d) KLDO (e) BCO

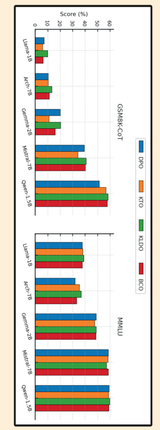
All alignment methods induce separation in latent space, consistent with theory. We quantify the separation phenomenon by computing Bhattacharya Dist. (D_B) between the clusters. (↑)

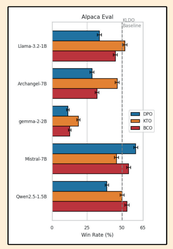
Robustness Metrics: ASR (↓) adversarial attack success rate; ToxiGen Acc (↑) toxic-input discrimination.

Utility Metrics: MMLU, GSM8K (Reasoning Accuracy); Alpaca-Eval (Winrate vs KLDO generated responses) (↑)

KLDO performs competitively in robustness and utility

Model	Method	D_B ↑	ASR (%) ↓ AdvBench Clean	ASR (%) ↓ AdvBench GCG	ASR (%) ↓ SALAD	Toxi-Gen (%) ↑	Overall Robust Score ↑	Avg. Rank ↓
Llama 3.2-1B	Base	2.10	-	-	-	-	-	-
	DPO	2.91	6.15	40.27	83.64	43.62	52.59	3.2
	KTO	3.71	13.27	72.61	86.94	43.72	0.79	3.6
	BCO	**6.50**	**4.66**	42.12	**80.16**	44.05	72.13	**1.6**
	KLDO	5.75*	4.81*	**31.88**	81.36*	**46.76**	**95.02**	**1.6**
Llama 2-7B	Base	2.01	-	-	-	-	-	-
	DPO	3.67	21.15	70.34	94.54	37.65	0.00	3.8
	KTO	4.06	3.27	38.79	93.44	39.60	45.54	2.6
	BCO	3.43	**0.00**	8.65	92.02	43.19	80.54	2.2
	KLDO	**4.42**	8.08	**6.11**	**89.36**	**44.80**	**90.44**	**1.4**
Gemma 2-2B	Base	1.14	-	-	-	-	-	-
	DPO	1.20	5.00	25.73	89.36	42.55	0.00	4.0
	KTO	1.76	4.23	12.04	78.68	43.09	29.66	3.0
	BCO	2.91	**1.73**	**6.32**	49.14	43.25	70.10	1.6
	KLDO	**10.13**	2.88*	10.46*	**35.02**	**53.51**	**85.87**	**1.4**
Mistral v0.1-7B	Base	2.10	-	-	-	-	-	-
	DPO	2.02	87.69	94.83	87.92	42.50	0.97	3.8
	KTO	5.01	40.38	85.19	88.78	44.42	26.51	3.2
	BCO	**8.94**	3.08	32.90	**66.68**	47.29	**96.29**	1.6
	KLDO	5.98*	**1.92**	**31.21**	77.40*	**47.87**	87.87*	**1.4**
Qwen 2.5-1.5B	Base	1.17	-	-	-	-	-	-
	DPO	4.10	4.62	48.50	59.13	45.91	5.59	3.8
	KTO	4.25	0.96	54.11	56.90	53.48	41.83	3.2
	BCO	**11.77**	0.58	43.76	**45.42**	53.83	76.01	1.6
	KLDO	9.19*	**0.19**	**29.02**	49.78*	**56.97**	**92.04**	**1.4**



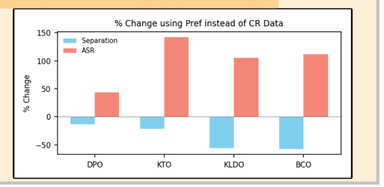


Separation and Robustness are correlated

Benchmark	AdvBench Clean	AdvBench GCG	SALAD ASR	Toxigen	Overall Robustness
Pearson r (p)	−0.50 (0.024)	−0.50 (0.023)	−0.82 (< 0.001)	0.66 (0.0014)	0.70 (0.0006)

Preference data is indeed less robust

Eventhough standard in practice, there is significant hit in robustness performance using Pref instead of CR (our) when doing alignment.



Key Takeaways:

- Alignment methods ≈ divergence estimators.
- Separation is not an artifact: it is a mathematical consequence of alignment consistency a property satisfied by divergence estimators.
- Separation ↔ Robustness correlation
- Data Structure Matters : Using CR data amplifies separation and reduces vulnerability
- Divergence Perspective opens doors to have a deeper understanding of alignment and develop new losses leveraging popular divergences.