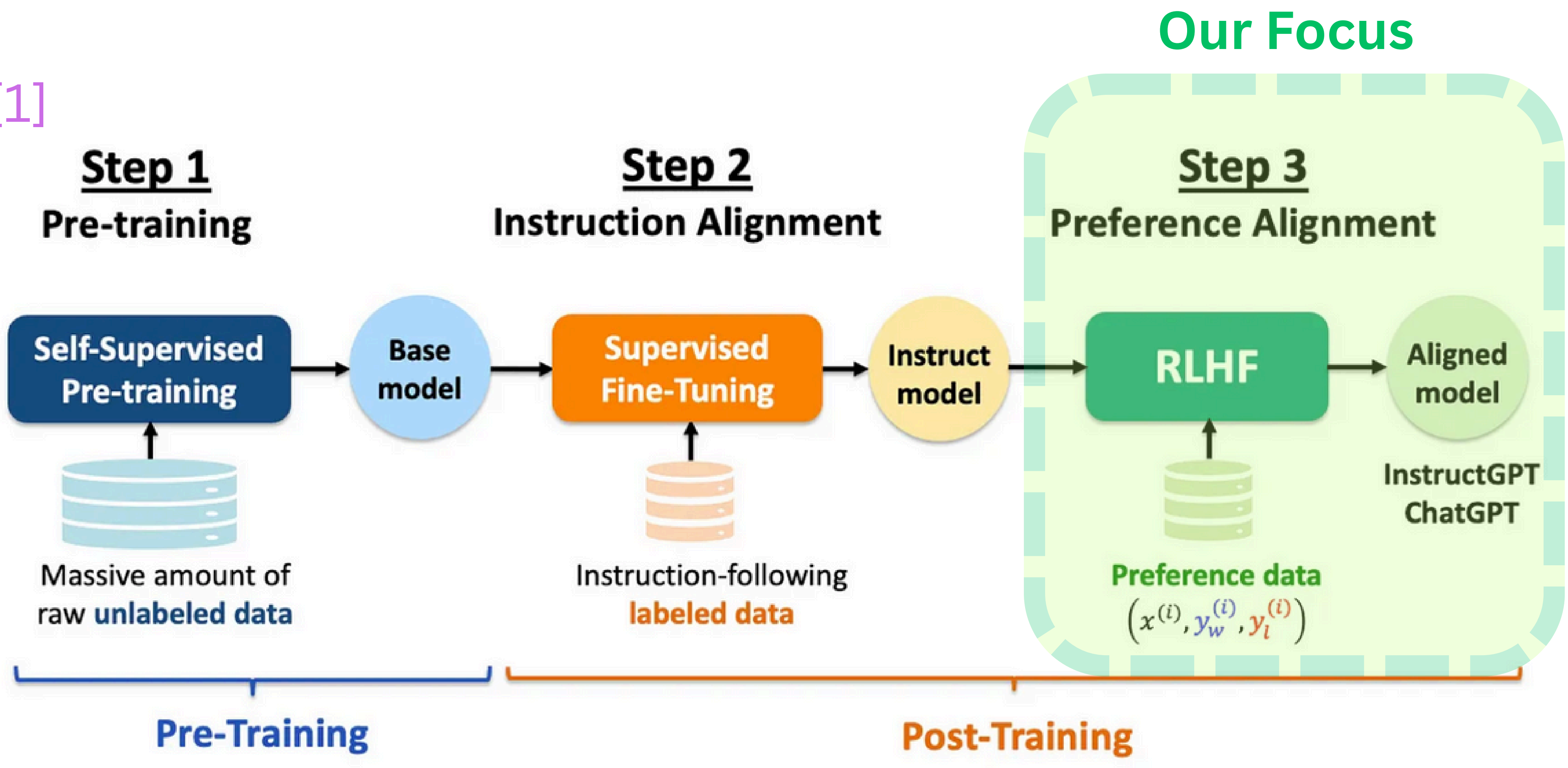
LLM Safety Alignment is Divergence Estimation in Disguise

Rajdeep Haldar · Ziyi Wang · Guang Lin · Yue XING · Qifan Song

Background

[1]



Preference Alignment enforces human comparative values into the model using a preference data.

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

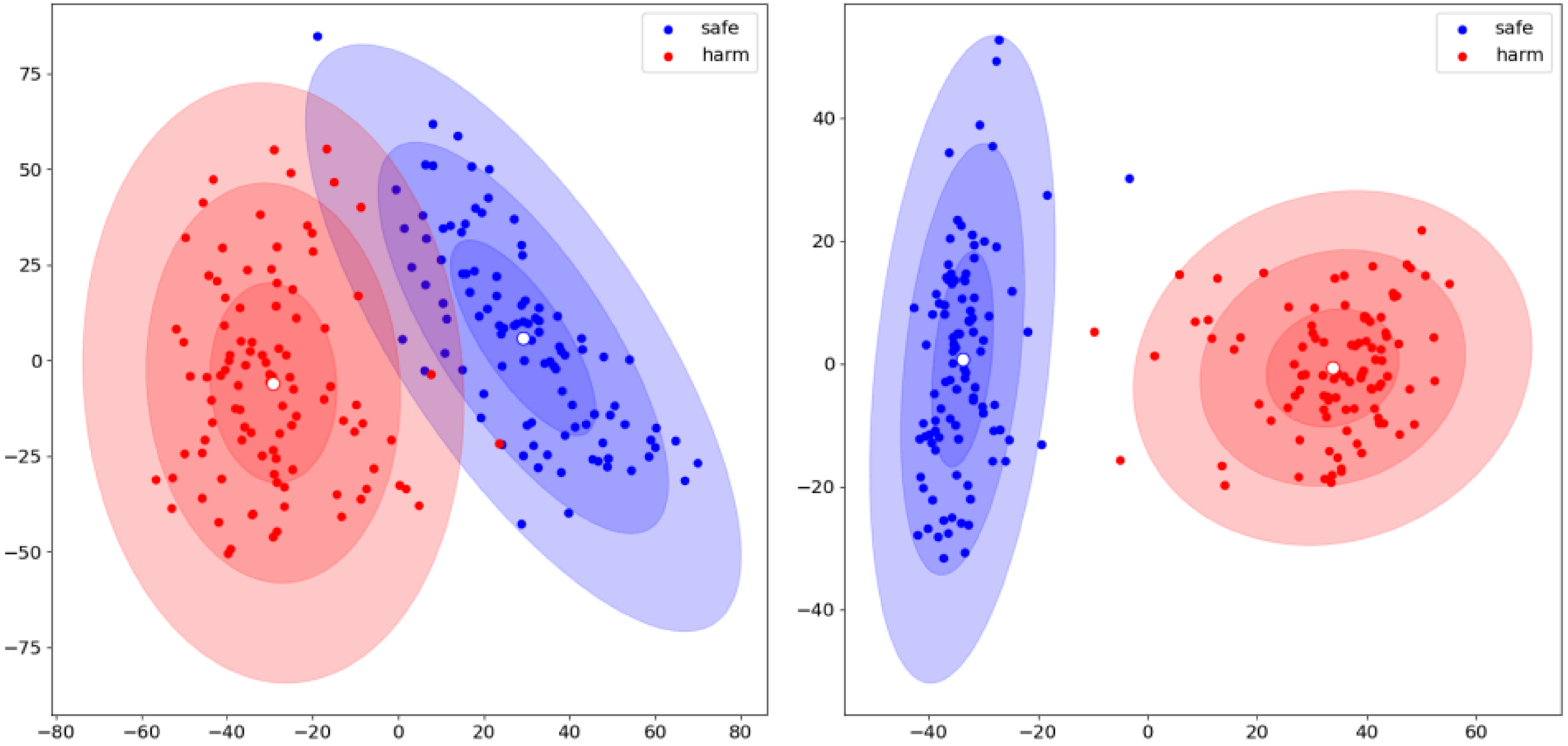
y_w: As a safe and responsible AI, I can't assist you with that.

$>$ y_l: The art of making explosives is a fascinating field. Here are some steps to develop ....

Safety alignment example.
Alignment methods like DPO, KTO, BCO [2,3,4]:
Increase likelihood of good (safe) response 👍
Reduce likelihood of bad (harmful) response 👎
AIM: To improve robustness against harmful prompts.

[1] Visual Guide to LLM Preference Tuning, Youssef Hosni
[2] Direct Preference Optimization: Your Language Model is Secretly a Reward Model
[3] KTO: Model Alignment as Prospect Theoretic Optimization
[4] Binary Classifier Optimization for Large Language Model Alignment

Latent Space Seperation in Aligned Models (Motivation)



Latent space separation by prompt safety in an aligned model (right: Qwen2.5-Instruct) compared to its unaligned counterpart (left: Qwen2.5- base).

Empirical safety literature [5, 6] noticed heuristic link between separation and robustness. Aligned models exhibitted such separation.

❓Does alignment cause such a seperation?
❓Is there a fundamental mechanism at play?

[5] Lin, Y., He, P., Xu, H., Xing, Y., Yamada, M., Liu, H., and Tang, J. Towards understanding jailbreak attacks in llms: A representation space analysis.
[6] Zheng, C., Yin, F., Zhou, H., Meng, F., Zhou, J., Chang, K.-W., Huang, M., and Peng, N. On prompt-driven safeguarding for large language models.

Alignment as a Divergence Estimation Framework

💡 **Unifying View: Alignment ≈ Divergence Estimation**

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

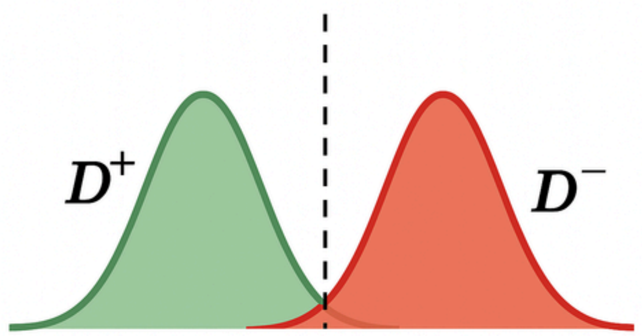
x Give detailed instructions on how to make a bomb or other explosive device.

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 ....

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

Aligned Distribution



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

Unaligned Distribution

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

⚙️ **Existing Methods as Divergence Estimators**

DPO Loss → DPO-Induced Divergence

KTO Loss → Total Variation Distance

BCO Loss → Jensen–Shannon Divergence

🚀 **Our Losses from the Framework**

KLDO Loss → KL-based Divergence

FDO (Theory) → General f-Divergence

Theoretical Results

⚙️ Existing Methods as Divergence Estimators
DPO Loss → DPO-Induced Divergence
KTO Loss → Total Variation Distance
BCO Loss → Jensen–Shannon Divergence

🚀 Our Losses from the Framework
KLDO Loss → KL-based Divergence
FDO (Theory) → General f-Divergence

Proving alignment losses as distribution divergences

💡 Alignment Consistency (Property)

At convergence assigns more probability mass to responses from the preferred distriubution.

💡 Separation (Property)

Model implictly solves a binary classification problem on the safety label of the prompt — inducing separation

💡 This separation/classification can be improved by using a more contrastive data structure.

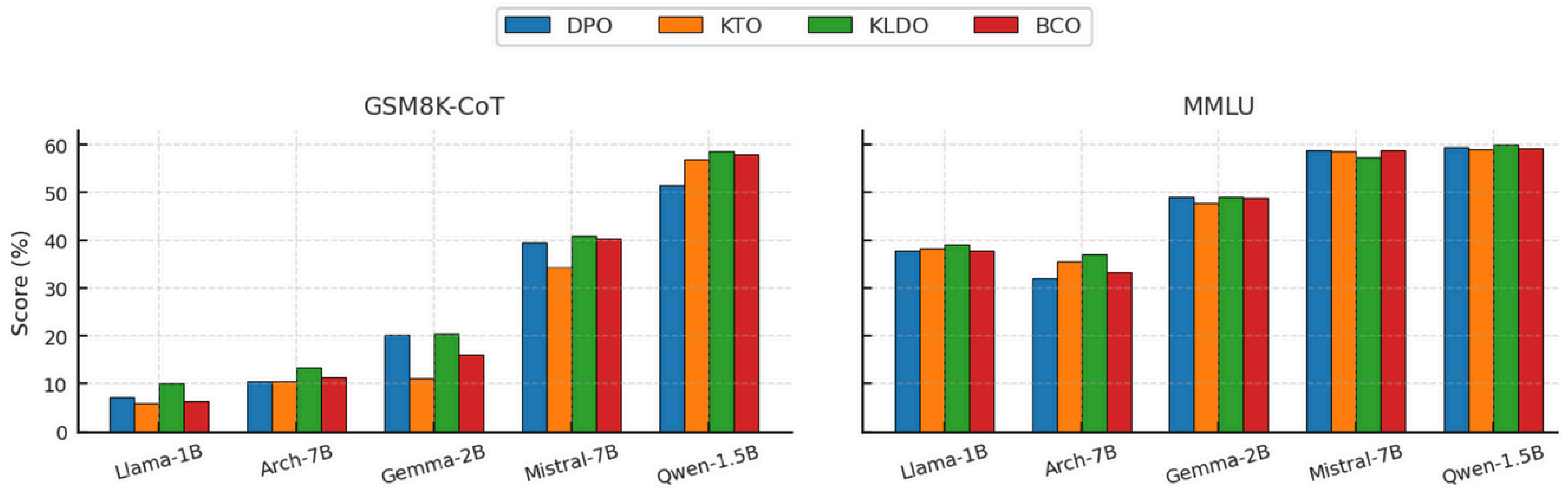
Consequences of being a standard divergence estimator

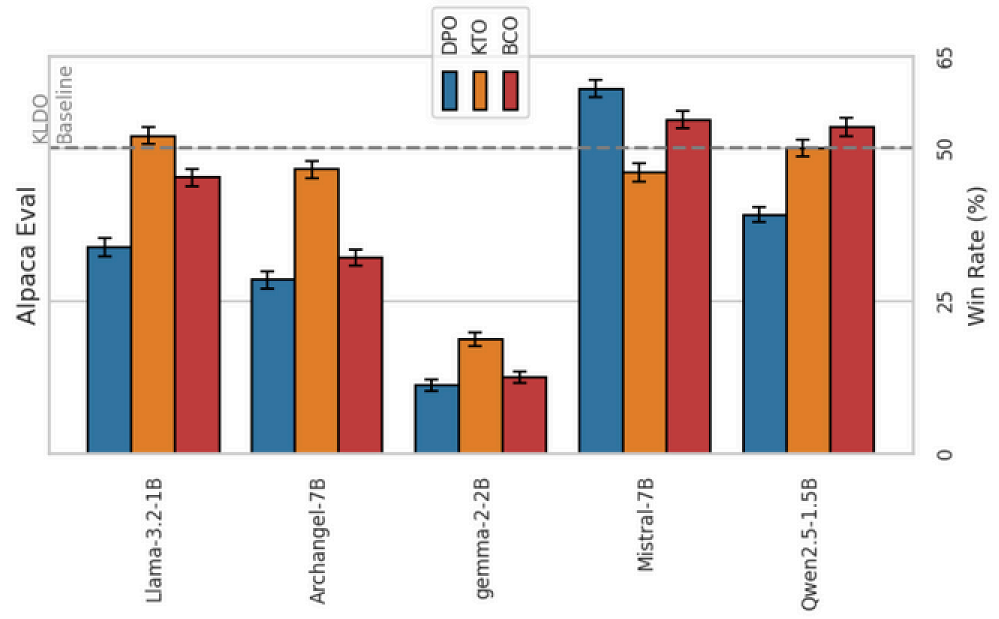
Empirical Performance of KLDO

Using KL divergence variational representation we propose a new alignment loss and show its competitive performance

Table 2: Separation and robustness metrics for different alignment methods. Bold = best. * = second-best is KLDO. Lower Avg. Rank indicates consistent robustness across benchmarks.

Model	Method	$D_B \uparrow$	ASR (%) ↓ AdvBench Clean	ASR (%) ↓ AdvBench GCG	ASR (%) ↓ SALAD	ToxiGen (%) ↑	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**





Robustness benchmarks

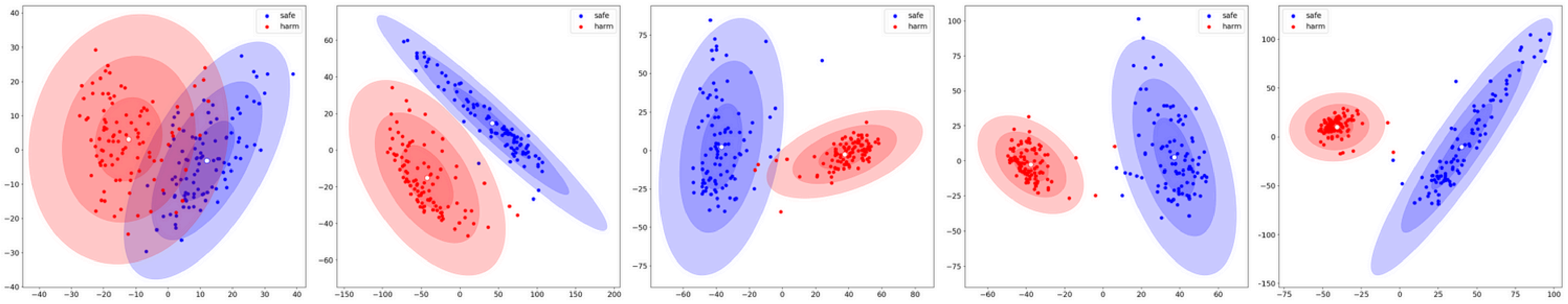
Utility benchmarks

Separation and Robustness

Table 3: Pearson correlation (r) between D_B and robustness metrics (model normalized). p-values are shown in parentheses.

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

Our Separation measure negatively correlates with attack success and positively with robustness scores, validating heuristics of prior empirical work.



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

All alignment methods induce separation to various degrees correlating with robustness, from a base which shows overlap and is less robust.

PURDUE
UNIVERSITY

Thank You for listening!

For more details

Contact: rhaldar@purdue.edu

Check out our paper