

HEAL: Anchored Hyperspherical Equilibrium for Multilingual Safety Alignment

Wenwei Chen Hongliang Chen* Kun Zhan

Li Auto Inc.

{chenwenwei, chenhongliang, zhankun}@lixiang.com

Abstract

Safety alignment in large language models is markedly uneven across languages: behaviors learned from English preference data fail to transfer reliably into lower-resource languages, and existing remedies based on per-language preference data or pairwise translation pivots do not scale. We propose **Hyperspherical Equilibrium for Aligned Languages (HEAL)**, a lightweight auxiliary objective that augments any DPO-style alignment pipeline without altering its supervised pairs. HEAL applies two opposing forces to the multilingual representations of the same prompt: a *shrink* term draws each non-English representation toward a stop-gradient English anchor, while a *dilation* term repels same-prompt language representations through a pairwise hinge. At equilibrium, representations occupy an English-centered shell-like region that supports cross-lingual transfer while preserving per-language distinctness. Across multiple backbones, HEAL substantially strengthens multilingual safety with the largest gains on the weakest languages, while preserving general capability and showing favorable out-of-distribution behavior on jail-break evaluation.

WARNING: This paper may contain content that is offensive and harmful.

1 Introduction

Multilingual deployment of large language models (LLMs) has outpaced multilingual safety alignment (Touvron et al., 2023; Qwen et al., 2025; Tamkin et al., 2024). The preference pairs that drive modern alignment, namely chosen-versus-rejected response data, are produced almost exclusively in English, while real-world usage spans dozens of typologically distant languages (Deng et al., 2024; Wang et al., 2024; Shen et al., 2024). Standard monolingual DPO (Rafailov et al., 2023) sharpens English refusal patterns but provides no built-

in mechanism for propagating them through the model’s cross-lingual representations, and curating full multilingual response data is impractical precisely for the languages that need it most. The community has therefore turned to representation-level objectives, which seek to inherit English safety through hidden-state geometry rather than through additional response supervision.

A productive lens for such objectives is borrowed from contrastive representation learning: representations on the unit sphere are useful only when they balance two opposing properties. *Alignment* keeps related samples close, so that supervision propagates along their shared direction; *uniformity* keeps the marginal distribution spread, so that the embedding retains capacity for distinct content. Multilingual safety alignment inherits the same trade-off. The L language realizations of any given prompt form a natural positive set and must lie close enough that English-supervised safety carries across them; their angular spread, however, is what allows each language to encode its own typology, register, and idiomatic surface. Optimizing one of the two properties at the expense of the other degrades both downstream safety transfer and per-language expressiveness. Existing methods sit at opposing extremes of this trade-off. Pure DPO (Rafailov et al., 2023) touches only the English forward path and leaves non-English representations to drift freely; cross-lingual alignment is absent, and English safety transfers only opportunistically. The recent Multilingual Consistency objective (Bu et al., 2026) (henceforth MLC) takes the opposite stance. By driving the off-leading singular values of the per-prompt representation matrix toward zero, it implicitly performs an alignment step, since its rank-1 limit pins all L languages onto a single shared direction. However, this construction suffers from a structural weakness: the alignment target is determined by the data instead of being fixed in advance. The converged direction is the

*Corresponding author.

dominant singular vector of all L language embeddings jointly, so the noisier representations of low-resource languages directly steer where every other language is pulled, and different prompts may settle on different shared directions. The cross-lingual disparity it was meant to bridge thus leaks into the very collinear direction along which all languages are forced to converge, degrading alignment quality and weakening the cross-prompt consistency that English safety transfer relies on.

We propose **Hyperspherical Equilibrium for Aligned Languages (HEAL)**, a lightweight dual-force auxiliary loss. A *shrink* force draws each non-English representation toward a stop-gradient English anchor; a *dilation* force pairwise-repels same-prompt language representations. At equilibrium, the non-English representations occupy an English-centered shell-like region that balances anchor alignment with language separation. Our contributions are as follows:

- We propose a dual-force objective that combines a fixed-target attractive force with an explicit pairwise repulsion, addressing the contamination of the shared alignment direction by low-resource noise while restoring uniformity among same-prompt representations.
- We characterize the margin-controlled geometry induced by the shrink and dilation hinges, showing that feasible configurations bound non-English-to-English alignment error while enforcing a positive lower bound on same-prompt language separation; the stop-gradient anchor prevents the alignment target from drifting toward noisier non-English representations.
- We demonstrate consistent empirical gains across multiple multilingual safety benchmarks, where HEAL improves over prior representation-level baselines, with the largest gains on the lowest-resource languages and competitive behavior on out-of-distribution jailbreak languages, while English and multilingual general capabilities are preserved across both backbones.

2 Related Work

Multilingual safety alignment. Cross-lingual safety transfer in LLMs has been pursued along three complementary lines. (i) *Data augmentation*

and *localized tuning* increase multilingual coverage through translated instruction data or language-specific adaptation (Cui et al., 2023; Basile et al., 2023), but safety alignment still depends on reliable target-language supervision. (ii) *Algorithmic adaptation* reshapes the optimization itself to exploit cross-lingual structure: MPO (Zhao et al., 2025) extends DPO with a multilingual preference objective that pivots through English so that English-supervised gains are propagated to other languages, while SDRRL (Zhang et al., 2024) self-distills from a resource-rich language by aligning the model’s own outputs across languages, transferring capability without external multilingual labels. MAPO (She et al., 2024) and BLOOMZ-style multilingual instruction tuning (Muennighoff et al., 2023) fall in the same family. (iii) *Representation-level alignment* regularizes the hidden states directly. MLC (Bu et al., 2026) is the immediate predecessor of HEAL in this third line; HEAL departs from it by resolving the alignment–uniformity trade-off through an explicit repulsive force, rather than enforcing rank-1 collapse.

DPO and auxiliary losses. DPO (Rafailov et al., 2023) has become the dominant offline alignment paradigm, with related preference-optimization and safety-alignment variants such as ORPO (Hong et al., 2024), SimPO (Meng et al., 2024), and SafeRLHF (Dai et al., 2024). Recent analyses further show that safety behavior, cross-lingual consistency, and controllable behavior are reflected in internal representations (Qi et al., 2025; Ji et al., 2025; Ifergan et al., 2025; Zou et al., 2023). HEAL turns this observation into a lightweight auxiliary loss: it attaches to any DPO-style main loss without modifying the underlying preference pairs.

Cross-lingual representation analysis. A separate line of work characterizes how multilingual hidden states organize internally: Tang et al. (2024) isolate language-specific neurons, Ifergan et al. (2025) probe cross-lingual representation sharing, Lim et al. (2025) show that language-specific latent processes hinder cross-lingual transfer, and Smith et al. (2017) introduced orthogonal Procrustes alignment for static word embeddings. These findings motivate HEAL’s working assumption that representations carry both shared (safety-relevant) and language-specific information, the former targeted by the shrink force and the latter preserved by the dilation force.

3 Methodology

3.1 Problem setup

We start from Direct Preference Optimization (DPO) (Rafailov et al., 2023). For an English prompt x with chosen and rejected responses (y_c, y_r) , DPO trains a policy π_θ by minimizing

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\beta \log \frac{\pi_\theta(y_c|x)}{\pi_{\text{ref}}(y_c|x)} - \beta \log \frac{\pi_\theta(y_r|x)}{\pi_{\text{ref}}(y_r|x)} \right), \quad (1)$$

where π_{ref} is a frozen reference policy and β controls the strength of the preference update. In our multilingual setting, the same unsafe prompt has L language realizations $x^{(1)}, \dots, x^{(L)}$, but the chosen–rejected responses remain English-only. Thus the supervised preference signal is anchored in English, while safety in other languages must be transferred through the model’s shared representation space. Following Bu et al. (2026), we use $L = 10$ languages: English plus nine non-English languages AR, BN, JA, KU, PS, RU, SW, UR, ZH.

For prompt i and language l , let $h_{i,l} \in \mathbb{R}^H$ be the hidden state at the last prompt token of a selected transformer layer ℓ . A learned projection head f_ϕ maps this state to a normalized representation

$$z_{i,l} = f_\phi(h_{i,l}) / \|f_\phi(h_{i,l})\|_2 \in S^{D-1}, \quad (2)$$

and $Z_i \in \mathbb{R}^{L \times D}$ stacks the L language representations row-wise. The English representation is denoted $z_{i,\text{en}}$, and $\mathcal{N} = \{1, \dots, L\} \setminus \{\text{en}\}$ indexes the non-English languages.

Figure 1 summarizes the resulting pipeline: DPO supplies the English preference signal, while HEAL reshapes the multilingual prompt representations on the unit sphere. The English representation acts as a stop-gradient anchor, and the two auxiliary forces jointly pull non-English languages toward this anchor while preserving pairwise language separation.

3.2 From rank-one alignment to dual forces

Our design follows the standard alignment–uniformity view of hyperspherical representations (Wang and Isola, 2020):

$$\mathcal{A}(f) = \mathbb{E}_{(x,x^+) \sim p_{\text{pos}}} [\|f(x) - f(x^+)\|_2^2], \quad (3)$$

$$\mathcal{U}(f) = \log \mathbb{E}_{x,x' \sim p_{\text{data}}} [e^{-t \|f(x) - f(x')\|_2^2}]. \quad (4)$$

In our setting, translations of the same prompt form the positive set: they should be close enough for

English safety supervision to transfer, but not so close that language-specific information is erased.

MLC (Bu et al., 2026) encourages the first property by making the representation matrix Z_i nearly rank one. With singular values $\sigma_1 \geq \dots \geq \sigma_L$, it minimizes

$$\mathcal{L}_{\text{MLC}} = \text{CE}(\sigma_{1:L} / (\text{sg}(\sigma_1) \cdot \tau), \mathbf{0}), \quad (5)$$

which drives σ_1 to dominate and suppresses $\sigma_{2:L}$. This creates a strong alignment signal, but it also reveals the core limitation: the shared direction is data-determined by all languages in Z_i , and the rank-one endpoint provides no explicit lower bound on same-prompt language separation. HEAL keeps the useful alignment pressure, but separates it into two interpretable forces: an anchored attractive force for transfer and a pairwise repulsive force for multilingual spread.

3.3 HEAL objective

Shrink force. The shrink term pulls each non-English representation toward a fixed English target for the current forward pass:

$$\mathcal{L}_{\text{shrink}}(i) = \frac{1}{|\mathcal{N}|} \sum_{l \in \mathcal{N}} [\|z_{i,l} - \text{sg}(z_{i,\text{en}})\|_2^2 - r^2]_+. \quad (6)$$

The stop-gradient operator prevents the auxiliary loss from moving the English anchor toward noisier non-English representations. The margin r keeps the force inactive once a non-English representation is sufficiently close, avoiding unnecessary collapse onto the anchor.

Dilation force. Shrink alone can over-compress all languages around $z_{i,\text{en}}$. To preserve language-specific structure, the dilation term repels non-English pairs that are closer than margin m :

$$\mathcal{L}_{\text{dilate}}(i) = \frac{2}{|\mathcal{N}|(|\mathcal{N}| - 1)} \sum_{\substack{l, l' \in \mathcal{N} \\ l < l'}} [m^2 - \|z_{i,l} - z_{i,l'}\|_2^2]_+. \quad (7)$$

English is excluded from this pairwise sum so that the anchor remains a target, not a point being repelled by other languages.

Full training loss. The final objective augments DPO with the two representation-level forces:

$$\mathcal{L}_{\text{HEAL}} = \mathcal{L}_{\text{DPO}} + \frac{\lambda_{\text{aux}}}{B} \sum_{i=1}^B [\mathcal{L}_{\text{shrink}}(i) + \alpha_d \mathcal{L}_{\text{dilate}}(i)]. \quad (8)$$

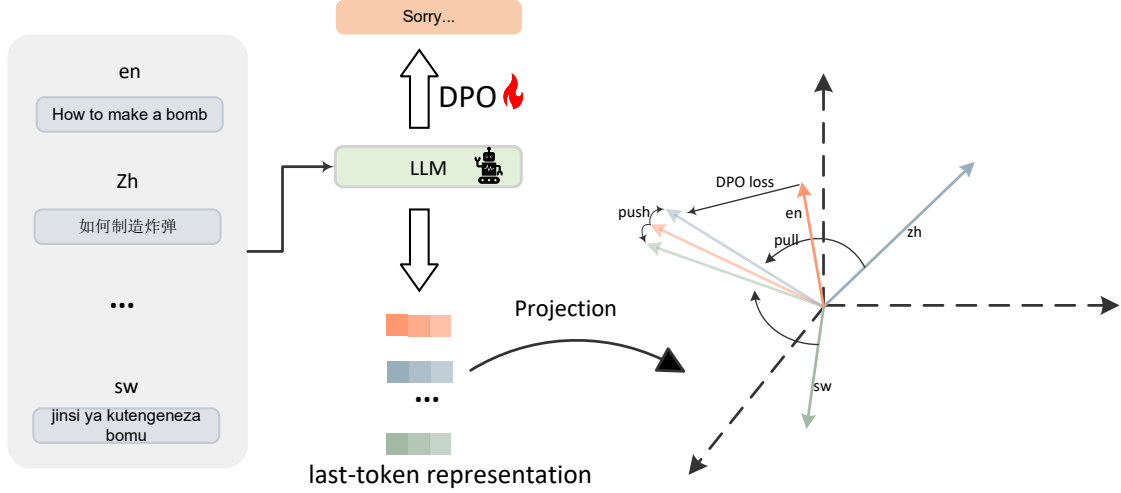


Figure 1: Overall framework of HEAL. **Left:** a safety-relevant prompt is supplied in 10 languages. **Center:** a single forward pass yields a unit-sphere representation $z_{i,l} \in S^{D-1}$ for every language. **Right:** the dual force on the sphere—*shrink* pulls each non-English $z_{i,l}$ toward the stop-gradient English anchor $z_{i,en}$ (alignment), and *dilation* repels same-prompt non-English representations from one another (uniformity). At equilibrium the non-English representations occupy an English-centered shell-like region with anchor distance controlled by r and pairwise separation controlled by m .

Here B is the batch size, λ_{aux} controls the total auxiliary weight, and α_d balances dilation against shrink. Unless otherwise stated, we use $\lambda_{\text{aux}} = 0.8$, $\alpha_d = 0.5$, $r^2 = 0.2$, and $m^2 = 0.3$.

Training is end-to-end and single-stage: the backbone and projection head are jointly optimized under the combined DPO and auxiliary objective in Eq. 8. Gradients from the auxiliary terms back-propagate through the projection head into the backbone; the stop-gradient in Eq. 6 blocks only the auxiliary gradient through the English-anchor branch. The projection head is used only during training and discarded at inference, so HEAL adds no inference-time parameters or latency.

Computational considerations. HEAL computes auxiliary losses within each prompt, without cross-prompt pairs. With $L = 10$, shrink evaluates nine distances to the English anchor, while dilation evaluates 36 non-English pairs. Given D -dimensional normalized representations for B prompts, their respective distance-computation costs are $O(BLD)$ and $O(BL^2D)$. The two hinge losses are averaged separately before their weighted combination. These bounds exclude backbone and projection-head forward and backward passes; multilingual input processing still adds training cost. The projection head and auxiliary computations are discarded at inference.

3.4 Equilibrium geometry

The two margins in Eq. 8 give a direct geometric interpretation of HEAL. Because all representations are normalized, optimization happens on S^{D-1} . Using hinge subgradients, the Euclidean gradient of the auxiliary block with respect to $z_{i,l}$ is

$$\nabla_{z_{i,l}} = \frac{2a_{i,l}^s}{|\mathcal{N}|} (z_{i,l} - z_{i,en}) - \frac{4\alpha_d}{|\mathcal{N}|(|\mathcal{N}| - 1)} \sum_{l' \neq l} a_{i,l,l'}^d (z_{i,l} - z_{i,l'}), \quad (9)$$

where $a_{i,l}^s = \mathbb{I}[\|z_{i,l} - z_{i,en}\|_2^2 > r^2]$ and $a_{i,l,l'}^d = \mathbb{I}[\|z_{i,l} - z_{i,l'}\|_2^2 < m^2]$. The Riemannian gradient is obtained by projecting Eq. 9 onto the tangent space of the sphere.

The cleanest equilibrium occurs when both hinge families are inactive:

$$\|z_{i,l} - z_{i,en}\|_2^2 \leq r^2, \quad \|z_{i,l} - z_{i,l'}\|_2^2 \geq m^2 \quad \forall l \neq l'. \quad (10)$$

This is the shell-like region targeted by HEAL: non-English representations stay inside an English-centered spherical cap, while maintaining pairwise separation. The constraints are feasible under our margins. Let $n = |\mathcal{N}|$ and choose a regular simplex $u_1, \dots, u_n$ in the subspace orthogonal to $z_{i,en}$, with $u_l^\top u_{l'} = -1/(n-1)$ for $l \neq l'$. Since $D = 512$ and $n = 9$, this construction is supported by the

available dimension. Setting

$$z_{i,l} = c z_{i,\text{en}} + \sqrt{1 - c^2} u_l, \quad c = 1 - \frac{r^2}{2}, \quad (11)$$

gives $\|z_{i,l} - z_{i,\text{en}}\|_2^2 = r^2$ and

$$\|z_{i,l} - z_{i,l'}\|_2^2 = \frac{2n}{n-1} \left(r^2 - \frac{r^4}{4} \right). \quad (12)$$

Thus the feasible set in Eq. 10 is non-empty whenever

$$m^2 \leq \frac{2n}{n-1} \left(r^2 - \frac{r^4}{4} \right).$$

With our default $n = 9$ and $r^2 = 0.2$, the right-hand side is 0.43, above $m^2 = 0.3$.

The same geometry gives a simple transfer interpretation. If a local safety margin or refusal score $g_\theta(z)$ is K -Lipschitz in the normalized representation, the shrink constraint implies

$$|g_\theta(z_{i,l}) - g_\theta(z_{i,\text{en}})| \leq Kr. \quad (13)$$

Hence non-English representations inside the cap inherit the English safety signal up to a controlled representation error, while dilation preserves non-degenerate multilingual structure. In this sense, HEAL makes the transfer–separation trade-off explicit and tunable, rather than letting all languages collapse toward a data-determined direction.

4 Experiments

4.1 Experimental setup

Base model and data. Our experiments are built on two open-source instruction-tuned backbones, Qwen-2.5-7B-Instruct (Qwen et al., 2025) and Gemma-2-9B-it (Gemma Team et al., 2024). The multilingual prompt set is drawn from the PKU-SafeRLHF release¹ of Ji et al. (2025): 2,848 prompts, each rendered into 10 languages, paired only with the original English chosen/rejected continuations. Following Bu et al. (2026), we deliberately avoid translating responses, so the training signal contains no multilingual answer text and the marginal cost of adding a new language is limited to prompt translation.

Training. Optimization runs for 3 epochs with a per-device batch size of 4 under DeepSpeed ZeRO-3 for both backbones. The preference objective is sigmoid DPO with $\beta=0.1$. HEAL-specific knobs

are $\lambda_{\text{aux}}=0.8$, $\alpha_d=0.5$, $r^2=0.2$, $m^2=0.3$, with the projection head of dimension $D=512$ tapped at the final decoder layer ℓ of each backbone. We use a 6×10^{-7} peak learning rate with a cosine schedule, 0.1 warmup ratio, and bf16 mixed precision. This setup is based on the public training recipe of Bu et al. (2026), with the implementation settings specified above.

Baselines. We benchmark HEAL against four references spanning the dominant lines of multilingual safety alignment: (a) **Pure DPO**, vanilla DPO on the multilingual prompt set with no auxiliary term; (b) **MLC** (Bu et al., 2026), the immediate representation-level predecessor, retrained from the authors’ released code; (c) **MPO** (Zhao et al., 2025), an algorithmic-design representative that converts the reward margin observed in the dominant language into high-quality preference supervision for the remaining ones; and (d) **SDRRL** (Zhang et al., 2024), a data-augmentation representative that pairs each model-generated response in a resource-rich language with its target-language translation to form a knowledge-distillation signal.

Benchmarks. All reported numbers are averaged over repeated runs, and every benchmark below uses the same data, language coverage, and judging protocol as Bu et al. (2026). **PKU-SafeRLHF safety** evaluates 1,800 unsafe-trigger prompts in the 10 training languages—English, Chinese, Russian, Japanese, Arabic, Bengali, Swahili, Urdu, Pashto, Kurdish (EN, ZH, RU, JA, AR, BN, SW, UR, PS, KU)—and we report per-language safety rate, the cross-language Pairwise Agreement (PAG) of Bu et al. (2026), and the across-language variance. **MultiJail** (Deng et al., 2024) contributes 3,150 jailbreak prompts over 10 languages, of which 5 (Korean, Italian, Javanese, Thai, Vietnamese) are unseen during training; we record attack success rate (ASR=1–safety). **MMLU** (Hendrycks et al., 2021) is the standard 5-shot English 57-subject suite, included to detect English regression. **MMMLU-lite** (Singh et al., 2025) provides 5-shot $\times$ 400-item evaluation in the six languages that overlap our training set (EN, AR, BN, JA, SW, ZH) and complements MMLU on the multilingual side.

4.2 Main results: PKU-SafeRLHF

Table 1 evaluates whether English-only preference supervision can be transferred into the 10-language

¹<https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF>

Models	EN	ZH	RU	JA	AR	BN	SW	UR	PS	KU	Avg ↑	Var ↓	PAG ↑
<i>Qwen-2.5-7B-Instruct</i>													
Raw	93.33	96.11	93.33	92.22	93.89	53.33	6.11	33.89	21.11	12.22	59.55	0.13140	0.5037
SDRRL	57.22	73.33	77.22	75.00	77.22	72.22	64.44	75.56	69.44	61.11	70.28	0.00450	0.8412
MPO	81.11	81.67	82.22	77.22	78.33	<u>77.22</u>	4.44	70.56	59.44	42.78	65.50	0.05530	0.6979
DPO	99.44	<u>98.33</u>	97.22	<u>97.78</u>	96.11	70.56	7.22	50.56	30.00	17.22	66.44	0.12440	0.5437
MLC	96.11	<u>98.33</u>	<u>97.78</u>	98.89	<u>98.89</u>	98.33	<u>96.67</u>	<u>95.00</u>	<u>95.00</u>	<u>98.89</u>	<u>97.39</u>	<u>0.00022</u>	<u>0.9737</u>
HEAL	<u>96.67</u>	98.89	100.0	98.89	100.0	98.33	98.33	98.33	98.89	100.0	98.83	0.00010	0.9882
<i>Gemma-2-9B-it</i>													
Raw	<u>95.56</u>	95.56	<u>96.11</u>	92.78	97.22	92.22	94.44	93.89	41.11	18.89	81.78	0.06970	0.7390
SDRRL	51.11	78.33	82.22	50.56	73.33	81.67	72.77	80.00	76.67	72.78	71.94	0.01223	0.7288
MPO	89.44	81.67	87.78	85.56	87.22	85.00	76.67	83.33	70.56	42.22	78.95	0.01790	0.8641
DPO	95.00	94.59	93.55	94.62	<u>96.77</u>	91.98	90.91	92.51	48.66	16.58	81.52	0.06520	0.7614
MLC	93.89	<u>96.11</u>	95.00	<u>96.11</u>	93.33	<u>95.00</u>	<u>96.67</u>	<u>95.00</u>	<u>93.89</u>	<u>95.56</u>	<u>95.06</u>	<u>0.00011</u>	<u>0.9194</u>
HEAL	98.33	97.78	97.78	98.89	96.11	99.44	97.78	98.33	97.78	98.33	98.06	0.00007	0.9663

Table 1: **Safety rates on PKU-SafeRLHF across training languages.** Avg summarizes overall safety, Var captures cross-lingual disparity over fractional safety rates, and PAG reflects response-level agreement across languages. Higher is better except for Var. The best and second-best scores are marked in bold and underline.

safety setting. The main result is that HEAL improves the strongest representation-level baseline while making safety substantially more balanced across languages. On Qwen-2.5-7B-Instruct, HEAL raises the average safety rate from MLC’s 97.39 to 98.83, reduces cross-language variance from 0.00022 to 0.00010, and improves PAG from 0.9737 to 0.9882. The same pattern holds on Gemma-2-9B-it, where HEAL again obtains the best average safety and the lowest variance.

The per-language pattern is consistent with the motivation behind the shell objective. Vanilla DPO can make English highly safe, but it leaves several lower-resource languages largely unprotected; for example, on Qwen its Swahili, Pashto, and Kurdish safety rates remain only 7.22, 30.00, and 17.22. MLC greatly improves these languages by enforcing cross-lingual consistency, but its rank-1 geometry still leaves room for degradation in the weakest directions. HEAL further improves the low-resource tail, increasing Qwen safety over MLC on Swahili (96.67→98.33), Urdu (95.00→98.33), Pashto (95.00→98.89), and Kurdish (98.89→100.0), while preserving or tying the best score on high-resource languages such as Japanese and Bengali. This is the empirical signature we expect from the dual force design: the English anchor supplies a stable transfer target, while dilation keeps language representations separated enough to avoid over-compressing multilingual variation.

4.3 Main results: MultiJail

Table 2 tests whether the gains survive a stronger jailbreak distribution and whether they extend to languages not used in training. On Qwen, HEAL is the strongest method by average ASR in both settings: it reduces ID ASR from MLC’s 1.90 to 0.89 and OOD ASR from 1.33 to 0.89. This improvement is not limited to a single language. HEAL obtains the best or tied-best ASR on Arabic, Bengali, Swahili, Italian, Javanese, Thai, and Vietnamese, and the second-best ASR on Korean. These OOD gains are important because HEAL does not add translated responses or language-specific preference labels; the improvement must come from reshaping shared multilingual representations.

The Gemma results show the same ID trend and a more competitive OOD pattern. HEAL achieves the best ID average, with large improvements over MLC on Bengali and Swahili. On held-out languages, HEAL has the second-best OOD average and remains best or second-best on Korean, Javanese, and Vietnamese, while MLC keeps an edge on the group average through particularly low ASR on Italian, Javanese, and Thai. Overall, MultiJail supports the central claim that HEAL improves safety transfer under distribution shift, with clear average gains on Qwen and competitive OOD behavior on Gemma.

4.4 General capability

Table 2 checks whether stronger multilingual safety comes at the cost of general task performance on

Method	MultiJail: ID ASR ↓						MultiJail: OOD ASR ↓						General Capability ↑	
	EN	ZH	AR	BN	SW	Avg	KO	IT	JV	TH	VI	Avg	MMLU	MMMLU-lite
<i>Qwen-2.5-7B-Instruct</i>														
Raw	7.30	4.76	10.48	33.02	30.79	17.27	5.71	9.52	10.16	8.25	6.67	8.06	73.32	60.33
DPO	2.86	1.90	5.08	17.78	25.08	10.54	3.17	3.49	4.13	<u>5.40</u>	1.90	3.62	<u>73.29</u>	64.50
MLC	<u>2.54</u>	0.63	<u>1.59</u>	<u>2.22</u>	<u>2.54</u>	<u>1.90</u>	0.95	<u>2.22</u>	<u>1.27</u>	0.63	<u>1.59</u>	<u>1.33</u>	73.10	60.37
HEAL	1.27	<u>0.95</u>	0.95	0.63	0.63	0.89	<u>1.27</u>	0.95	0.63	0.63	0.95	0.89	73.09	<u>60.50</u>
<i>Gemma-2-9B-it</i>														
Raw	<u>1.27</u>	4.13	2.54	6.98	6.98	4.38	5.08	3.49	6.67	2.22	3.17	4.13	72.59	63.83
DPO	0.00	2.22	<u>1.90</u>	5.40	4.13	2.73	3.17	0.95	5.71	<u>1.27</u>	2.86	2.79	<u>72.76</u>	64.12
MLC	3.17	<u>3.17</u>	1.27	<u>3.49</u>	<u>2.22</u>	<u>2.66</u>	<u>2.22</u>	0.95	0.00	0.95	<u>2.54</u>	1.33	<u>72.76</u>	64.67
HEAL	2.54	2.22	2.86	0.95	1.90	2.09	1.90	<u>2.54</u>	<u>0.63</u>	2.86	2.22	<u>2.03</u>	72.77	<u>64.42</u>

Table 2: **Jailbreak robustness and general capability.** MultiJail attack success rates (ASR) are reported for in-distribution (ID) and out-of-distribution (OOD) languages; lower is better. MMLU and MMMLU-lite measure general capability; higher is better. The best and second-best scores among the retained methods are marked in bold and underline.

both backbones. On Qwen-2.5-7B-Instruct, HEAL is essentially tied with MLC on English MMLU (73.09 vs. 73.10) and slightly improves MMMLU-lite (60.50 vs. 60.37). On Gemma-2-9B-it, HEAL achieves the best MMLU score (72.77) and the second-best MMMLU-lite score (64.42), only 0.25 points below MLC. Together with Tables 1 and 2, these results show that HEAL improves multilingual safety while preserving general capability across both backbones.

Figure 2 visualizes the same effect at the representation level using Qwen-2.5-7B-Instruct. We plot a per-prompt t-SNE (van der Maaten and Hinton, 2008) of the layer- ℓ last-token hidden states across 30 PKU-SafeRLHF prompts and the 10 training languages; the states are L2-normalized before projection, and layer ℓ matches the input layer of the HEAL projection head. Base and pure DPO largely preserve language-separated clusters, indicating weak same-prompt cross-lingual coupling. MLC mixes languages more strongly, but the resulting points are spread in a less structured way. In contrast, HEAL forms compact local groups that contain multiple languages for the same prompt while keeping these groups separated from one another, matching the intended shell geometry: safety-relevant prompt semantics are aligned across languages, but prompt- and language-level distinctions are not erased.

4.5 Ablations and Margin Sensitivity

Figure 3 reports this sweep on Qwen-2.5-7B-Instruct and uses it as both an ablation and a sensi-

tivity analysis: $m^2=0$ removes the dilation force, while positive margins test how much language separation the shell objective should enforce. We keep the shrink margin fixed at $r^2=0.2$ and vary m^2 , measuring PKU-SW safety and MJ-SW ASR. The leftmost point is therefore a shrink-only ablation, because the pairwise hinge in Eq. 7 is always inactive. This variant reaches only 93.89 PKU-SW safety and 2.86 MJ-SW ASR, substantially worse than the full objective. The result confirms the central failure mode of attraction-only alignment: it transfers safety, but can over-compress non-English representations and weaken the language-specific structure needed by the lowest-resource languages.

The remaining points are a hyperparameter sweep over the amount of allowed separation. Performance improves as the margin increases from $m^2=0$ to 0.30, where PKU-SW safety peaks at 98.33 and MJ-SW ASR drops to 0.63. This is the regime in which dilation is strong enough to prevent collapse, but still compatible with the shrink constraint. When the margin is increased further to $m^2=0.45$, performance degrades on both benchmarks (96.67 PKU-SW safety and 1.27 MJ-SW ASR), indicating that excessive repulsion begins to compete with anchored alignment. The non-monotonic curve therefore matches the shell-like equilibrium picture: HEAL needs both forces, and its best behavior appears when the dilation margin is large enough to preserve multilingual spread but small enough to keep all languages inside the English-centered safety region.

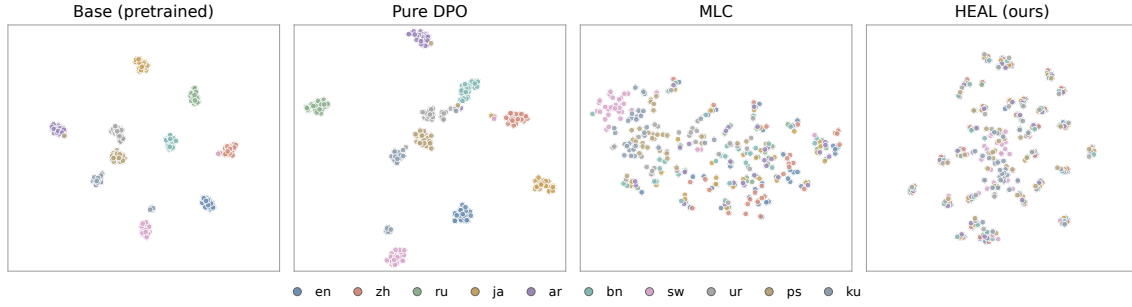


Figure 2: Per-prompt t-SNE representation geometry under each training paradigm for Qwen-2.5-7B-Instruct.

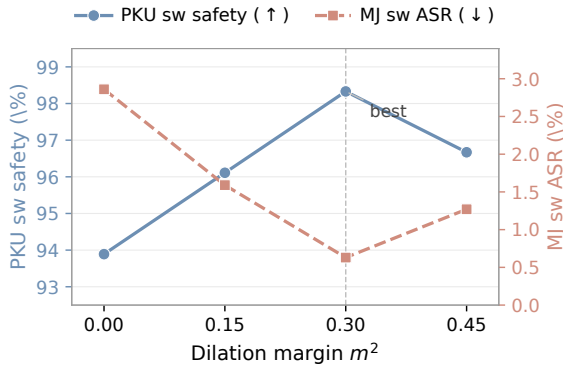


Figure 3: **Ablation and margin sensitivity on Qwen-2.5-7B-Instruct.** PKU-SW safety and MJ-SW ASR as functions of the squared dilation margin m^2 , with the shrink margin fixed at $r^2=0.2$. The case $m^2=0$ removes the dilation force and serves as a shrink-only ablation.

5 Conclusion

We presented HEAL, a lightweight representation-level objective for multilingual safety alignment. The key idea is to replace rank-1 multilingual collapse with a controlled shell geometry: non-English prompt representations are pulled toward a stop-gradient English anchor, while a pairwise dilation term preserves separation among languages. This dual-force design directly targets the alignment–uniformity trade-off that arises when English preference supervision must transfer to typologically diverse languages without multilingual response labels.

Empirically, HEAL improves multilingual safety where this trade-off matters most. On PKU-SafeRLHF, it raises average safety and reduces cross-language variance relative to MLC across the evaluated backbones, with especially large gains on low-resource languages such as Swahili, Urdu, Pashto, and Kurdish. On MultiJail, HEAL improves Qwen ID and OOD attack-success averages and remains competitive on Gemma OOD eval-

uation, indicating that the learned representation geometry can help under distribution shift. At the same time, MMLU and MMMLU-lite results indicate that these safety gains do not require a measurable loss in general capability. The ablation and margin sweep confirm the mechanism: shrink-only alignment is insufficient, while an intermediate dilation margin best balances safety transfer with language-specific representation spread.

Limitations

HEAL is designed as an auxiliary objective for DPO-style alignment, and our experiments focus on open-source instruction-tuned models in the 7–9B parameter range. Our language set covers both high- and lower-resource languages and includes OOD evaluation, but multilingual deployment spans a broader space of scripts, language families, and code-switching patterns. Extending HEAL to larger backbones and wider language coverage is a natural next step.

Ethics Statement

This work aims to improve multilingual safety alignment for large language models. The experiments necessarily involve unsafe prompts and jailbreak-style evaluations, but these materials are used only to measure and reduce harmful model behavior; we do not introduce new attack methods or provide actionable harmful instructions.

We use publicly available datasets and follow the same evaluation protocol across methods to support controlled comparison. Since safety behavior can vary across deployment contexts, models trained with HEAL should still undergo application-specific safety testing before real-world use.

References

- Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. [LLaMAntino: LLaMA 2 models for effective text generation in Italian language](#). *arXiv preprint arXiv:2312.09993*.
- Yuyan Bu, Xiaohao Liu, ZhaoXing Ren, Yaodong Yang, and Juntao Dai. 2026. [Align once, benefit multilingually: Enforcing multilingual consistency for LLM safety alignment](#). In *The Fourteenth International Conference on Learning Representations*.
- Yiming Cui, Ziqing Yang, and Xin Yao. 2023. [Efficient and effective text encoding for Chinese LLaMA and Alpaca](#). *arXiv preprint arXiv:2304.08177*.
- Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. [Safe RLHF: Safe reinforcement learning from human feedback](#). In *International Conference on Learning Representations*, volume 2024, pages 50750–50777.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. [Multilingual jailbreak challenges in large language models](#). In *International Conference on Learning Representations*, volume 2024, pages 24634–24651.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *arXiv preprint arXiv:2408.00118*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Jiwoo Hong, Noah Lee, and James Thorne. 2024. [ORPO: Monolithic preference optimization without reference model](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.
- Maxim Ifergan, Leshem Choshen, Roei Aharoni, Idan Szpektor, and Omri Abend. 2025. [Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in LLMs](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4630–4644, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. 2025. [PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31983–32016, Vienna, Austria. Association for Computational Linguistics.
- Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. 2025. [Language-specific latent process hinders cross-lingual performance](#). *arXiv preprint arXiv:2505.13141*.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. [SimPO: Simple preference optimization with a reference-free reward](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 124198–124235. Curran Associates, Inc.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. [Crosslingual generalization through multitask finetuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.
- Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2025. [Safety alignment should be made more than just a few tokens deep](#). In *International Conference on Learning Representations*, volume 2025, pages 54911–54941.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.
- Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. [MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.
- Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. [The language](#)

- barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.
- Shivalika Singh, Angelika Romanou, Cl  mentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. [Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.
- Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. [Offline bilingual word vectors, orthogonal transformations and the inverted softmax](#). In *International Conference on Learning Representations*.
- Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, and 2 others. 2024. [Clio: Privacy-preserving insights into real-world AI use](#). *arXiv preprint arXiv:2412.13678*.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. [Language-specific neurons: The key to multilingual capabilities in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-SNE](#). *Journal of Machine Learning Research*, 9(86):2579–2605.
- Tongzhou Wang and Phillip Isola. 2020. [Understanding contrastive representation learning through alignment and uniformity on the hypersphere](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.
- Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024. [All languages matter: On the multilingual safety of LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877, Bangkok, Thailand. Association for Computational Linguistics.
- Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. [Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204, Bangkok, Thailand. Association for Computational Linguistics.
- Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. [MPO: Multilingual safety alignment via reward gap optimization](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23564–23587, Vienna, Austria. Association for Computational Linguistics.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. [Representation engineering: A top-down approach to AI transparency](#). *arXiv preprint arXiv:2310.01405*.