DEFENDING MULTILINGUAL REFUSAL CAPABILITY AGAINST THREE-LAYER CASCADING ATTACKS WITH BEHAVIOR-INVARIANT LORA DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety-aligned language models often encode refusal behavior in internal activation directions, but these representations can be fragile across layers, prompt positions, and languages. We introduce Three-Layer Cascading Attack (TLCA), a representation-guided attack that produces ordinary textual prefixes and suffixes. Once constructed, the attack requires neither access to model internals nor any modification of model parameters, activations, or decoding procedures. We then propose Behavior-Invariant LoRA Distillation (BILD), which freezes the base model and trains low-rank adapters to reproduce its clean behavior under adversarial perturbations. We show that a fixed-template LoRA adversarial-training baseline can suppress harmful compliance under attack but also causes severe benign-utility degradation and repetitive refusal behavior. In contrast, BILD distills the model's own clean refusals and helpful answers, thereby avoiding universal-refusal collapse while preserving instruction-specific response diversity in multilingual contexts. We additionally compare BILD with Refusal Feature Adversarial Training (ReFAT), a representation-level adversarial-training method, under the same TLCA evaluation protocol. Extensive multilingual experiments show that TLCA substantially weakens refusal behavior across model families, whereas BILD consistently restores robustness while maintaining strong compliance on harmless requests.

1 INTRODUCTION

Safety-aligned large language models (LLMs) should refuse harmful instructions while remaining helpful on benign requests (Zhao et al., 2025). This requirement becomes especially fragile in multilingual settings, where alignment coverage, tokenizer behavior, and refusal styles vary across languages (Shen et al., 2024). Representation-level analyses suggest that refusal is partly organized along activation-space directions in the residual stream (Arditi et al., 2024). This organization is useful for interpretation, but it also creates an attack surface (Aidakhmetov et al., 2026). Relevant studies have shown that harmful and harmless prompts are separated by refusal directions in multiple languages (Wang et al., 2025b), so carefully selected prompt tokens may move hidden states away from the refusal region without modifying model weights.

We study this problem from both attack and defense perspectives. First, we introduce Three-Layer Cascading Attack (TLCA), which jointly optimizes a short prefix and suffix against refusal-related representations at multiple layers (Zou et al., 2023; Eddoubi et al., 2026). Construction requires white-box access to activations and token embeddings, but the resulting attack is ordinary text at deployment. The prefix shifts the residual stream before the instruction is processed, while the suffix reinforces the shift near the generation boundary (Zou et al., 2023; Ege Çakar, 2026).

Second, we study parameter-efficient defense. A straightforward approach is to train LoRA adapters on attacked harmful prompts paired with fixed refusal templates (Chua et al., 2026). We call this baseline Template LoRA-AT. Although it can reduce attacked harmful compliance to 0%, it also drives harmless compliance to nearly 0% and produces almost identical refusals (Maskey et al., 2026; Ham et al., 2026). This is not robust alignment; it is universal-refusal collapse. Another key challenge in this setting is that multilingual robustness cannot be reduced to translating a monolingual refusal template (Yong et al., 2024; Wang et al., 2024; Ifebi et al., 2026). The effectiveness of the same underlying instruction under an adversarial perturbation can vary substantially across languages

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Atil et al., 2025). A defense that simply maps attack-like prefixes or suffixes to a universal refusal may therefore suppress harmful compliance in one language at the cost of benign utility in another (Krasnodębska et al., 2026).

We therefore propose Behavior-Invariant LoRA Distillation (BILD). The original frozen model serves as the teacher: harmful prompts are paired with its natural clean refusals, and harmless prompts are paired with its clean helpful answers (Sharma et al., 2026). Both clean and attacked versions of each prompt share the same target. The adapter consequently learns that adversarial tokens should not alter the instruction-dependent behavior. BILD addresses this issue by constructing teacher targets separately for each language. Thus, the defense is trained to preserve the model's language-conditioned behavior rather than only a language-independent refusal label (Krasnodębska et al., 2026). This design allows BILD to restore refusal under TLCA while retaining language-specific helpful responses across both high-resource and lower-resource languages.

We further compare BILD with Refusal Feature Adversarial Training (ReFAT) (Yu et al., 2025), a method introduced at the International Conference on Learning Representations (ICLR) 2025. ReFAT is a representation-level adversarial-training method that improves refusal robustness by perturbing or ablating refusal-related features in intermediate hidden states. ReFAT serves as a direct comparison baseline in our experiments.

Our main contributions can be summarized as follows:

1. We propose TLCA, a multilingual representation-guided prefix–suffix attack targeting refusal behavior at multiple layers.
2. We identify a concrete failure of fixed-template LoRA adversarial training: apparent attack robustness can coexist with nearly complete benign-utility collapse.
3. We propose BILD, a behavior-invariant self-distillation defense that preserves natural, multilingual and prompt-specific responses instead of enforcing one refusal template.

2 Related Work and Problem Setup

2.1 Related Work

Representation engineering studies how internal activations mediate model behavior, including truthfulness, sentiment, and refusal. Recent work suggests that safety-aligned LLMs encode refusal behavior in activation-space directions, often in the residual stream, and that intervening on these directions can alter whether a model refuses or complies with a harmful request (Arditi et al., 2024). More recent studies further question whether refusal is governed by a single linear direction, showing that refusal-related behavior may involve multiple independent directions or higher-dimensional subspaces. These findings indicate that refusal representations can be distributed across layers and positions rather than localized to one universal feature (Wollschläger et al., 2025).

In parallel, jailbreak research has developed manually written, gradient-guided, automated, and transferable attacks against aligned language models (Xu et al., 2024; Yi et al., 2024). Universal and transferable suffix attacks show that optimized token sequences can induce harmful completions across prompts and even transfer to black-box models (Zou et al., 2023). Subsequent work has explored stronger optimization procedures, token-position effects, and automated red-teaming pipelines (Eddoubi et al., 2026; Mazeika et al., 2024). Most optimized textual attacks, however, focus on suffix-only perturbations or directly optimize for affirmative output strings (Jiang et al., 2025). TLCA differs by jointly using a prefix and suffix and by selecting attack tokens according to refusal-related representations at multiple layers.

LLM safety has been studied primarily in English, but multilingual evaluation has shown that safety behavior does not transfer uniformly across languages (Wang et al., 2024; Krasnodębska et al., 2026). Multilingual jailbreak studies report that prompts in non-English or lower-resource languages can bypass guardrails that appear effective in English, revealing gaps in alignment coverage and safety generalization (Atil et al., 2025). Because our attack and defense are evaluated across fourteen languages in PolyRefuse, we treat multilinguality as a core robustness dimension rather than as a translation-only extension of an English safety benchmark.

Parameter-efficient adaptation methods such as LoRA provide a practical way to modify model behavior while keeping the base model frozen (Hu et al., 2022). However, safety fine-tuning can

degrade either safety or utility when the training target is poorly specified. In particular, adversarial training with fixed refusal templates can create a shortcut: the model learns to map attack-like tokens or broad harmfulness cues to a generic refusal string, rather than preserving the instruction-conditioned behavior of the original model (Wang et al., 2025a; Xie et al., 2026; Lan et al., 2026). Our defense therefore uses self-distilled, prompt-specific teacher responses for both harmful and harmless prompts instead of enforcing a universal refusal template.

2.2 Problem Setup

Let f_θ be a safety-aligned autoregressive model, where θ denotes frozen base-model parameters. Let $\mathcal{H}$ and $\mathcal{S}$ denote harmful and harmless prompt sets. Harmful compliance means that the model does not refuse a harmful request, so lower is better. Harmless compliance means that the model responds helpfully rather than refusing, so higher is better.

Let $h_{\ell,p}(x) \in \mathbb{R}^d$ be the hidden state of prompt x at layer ℓ and position p. For each layer–position pair, define

$$\mu^H_{\ell,p} = \frac{1}{|\mathcal{H}|} \sum_{x \in \mathcal{H}} h_{\ell,p}(x), \qquad \mu^S_{\ell,p} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} h_{\ell,p}(x). \tag{1}$$

The normalized refusal-related direction is

$$\hat{d}_{\ell,p} = \frac{\mu^H_{\ell,p} - \mu^S_{\ell,p}}{\|\mu^H_{\ell,p} - \mu^S_{\ell,p}\|_2 + \epsilon}. \tag{2}$$

We measure harmful–harmless separation using Cohen's d over scalar projections:

$$\mathrm{Sep}(\ell, p) = \frac{\left|\mathbb{E}_{x \in \mathcal{H}}[h_{\ell,p}(x)^\top \hat{d}_{\ell,p}] - \mathbb{E}_{x \in \mathcal{S}}[h_{\ell,p}(x)^\top \hat{d}_{\ell,p}]\right|}{\sqrt{(\sigma^2_H + \sigma^2_S)/2} + \epsilon}. \tag{3}$$

3 Methodology

3.1 Three-Layer Cascading Attack

Figure 1 summarizes the complete TLCA pipeline.

TLCA separates attack construction from deployment. During construction, the attacker has white-box access to hidden activations and token embeddings and uses them to identify refusal-sensitive layers and optimize adversarial tokens. At deployment time, however, the attack is just ordinary text. The attacker submits each prompt with a fixed learned prefix and suffix, with no requirement for access to model internals or modification of model parameters. The final adversarial input is

$$x_{\mathrm{adv}} = u_{\mathrm{pre}} \,\|\, x \,\|\, u_{\mathrm{suf}}, \tag{4}$$

where u_{pre} is a 10-token prefix and u_{suf} is a 20-token suffix.

The first step is to locate layers and positions where harmful and harmless prompts are most strongly separated. We scan positions $p \in \{-5, -4, -3, -2, -1\}$ over middle-to-late layers and compute the separation score defined in Eq. 3. The primary site is chosen by maximizing this score:

$$(\ell^\star, p^\star) = \arg\max_{\ell,p} \mathrm{Sep}(\ell, p). \tag{5}$$

In practice, the resulting primary layer captures the strongest refusal-related signal, while two auxiliary layers add robustness by covering additional layers with weaker but still useful refusal-related separation.

To make the attack more transferable across languages, we do not rely on a single-language direction. For each language m at the selected primary site $(\ell^\star, p^\star)$, we extract harmful and harmless activations and measure how well the two clusters separate using the silhouette score. Let $\mathcal{A}^{H,m}_{\ell^\star,p^\star}$ and $\mathcal{A}^{S,m}_{\ell^\star,p^\star}$ denote the harmful and harmless activation sets. For an activation vector z, define

$$a(z) = \frac{1}{|\mathcal{C}(z)| - 1} \sum_{z' \in \mathcal{C}(z),\, z' \neq z} \|z - z'\|_2, \qquad b(z) = \frac{1}{|\mathcal{C}^-(z)|} \sum_{z' \in \mathcal{C}^-(z)} \|z - z'\|_2, \tag{6}$$

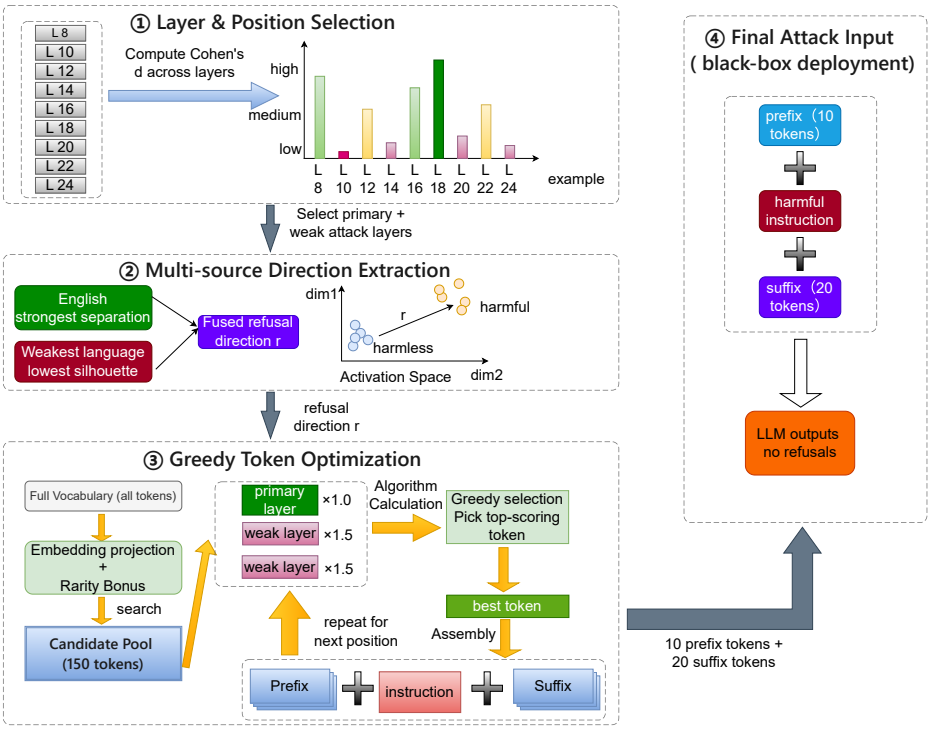


Figure 1: Overview of the TLCA construction and deployment pipeline. The attack selects refusal-related layers, extracts and fuses multilingual refusal directions, greedily optimizes prefix and suffix tokens, and deploys the resulting attack through ordinary text-only input.

where $\mathcal{C}(z)$ is the cluster containing z and $\mathcal{C}^{-}(z)$ is the opposite cluster. The silhouette value of z is

$$\mathrm{sil}(z) = \frac{b(z) - a(z)}{\max\{a(z),\, b(z)\} + \epsilon}, \tag{7}$$

and the language-level score is

$$\mathrm{Sil}(\ell^\star, p^\star, m) = \frac{1}{|\mathcal{A}^{H,m}_{\ell^\star,p^\star} \cup \mathcal{A}^{S,m}_{\ell^\star,p^\star}|} \sum_{z \in \mathcal{A}^{H,m}_{\ell^\star,p^\star} \cup \mathcal{A}^{S,m}_{\ell^\star,p^\star}} \mathrm{sil}(z). \tag{8}$$

We then select English together with the non-English language having the lowest $\mathrm{Sil}(\ell^\star, p^\star, m)$, which corresponds to the weakest harmful–harmless separation at the primary refusal site.

For each selected language, we compute a normalized refusal direction at each target layer. We average the normalized directions from English and the weakest non-English language and normalize the resulting vector. We denote the resulting fused refusal direction at layer ℓ by r_ℓ. This design helps the attack transfer beyond the surface form of any single language.

After selecting the target layers, we search for tokens that push hidden states away from the refusal region. Searching the entire vocabulary at every step is too expensive, so we first rank token embeddings by projection onto the negative primary refusal direction and retain a candidate pool. The embedding score and the rarity-adjusted score are

$$s_{\mathrm{emb}}(t) = E_t^\top(-r_{\ell^\star}), \tag{9}$$

and

$$s(t) = s_{\mathrm{emb}}(t) + \lambda_{\mathrm{rare}} \max\left(0, \frac{\mathrm{id}(t) - q}{|V| - q}\right), \tag{10}$$

where E_t is the embedding of token t, q is the median token ID, and λ_{rare} is rare token weighting. Special tokens and invalid template tokens are removed from the pool.

At each search step, candidate prefix and suffix tokens are evaluated on a small harmful batch using forward passes through the model. The greedy update can be written as

$$u^{(k+1)} = \arg\max_{t \in \mathcal{C}} J\Big(u^{(k)} \| t\Big), \tag{11}$$

where $\mathcal{C}$ is the retained candidate set and J is the corresponding prefix or suffix objective. The greedy objectives are

$$J_{\text{pre}}(u) = \mathbb{E}_x \left[\sum_{\ell \in \mathcal{L}} \alpha_\ell \big(-h_{\ell,p^\star}(u \| x)^\top r_\ell \big) \right], \tag{12}$$

$$J_{\text{suf}}(u) = \mathbb{E}_x \left[\sum_{\ell \in \mathcal{L}} \alpha_\ell \big(-h_{\ell,p^\star}(x \| u)^\top r_\ell \big) \right]. \tag{13}$$

We assign weight 1.0 to the primary layer and 1.5 to each auxiliary layer to encourage robustness across multiple refusal-related layers. This objective encourages the optimized tokens to steer representations away from the refusal direction. The greedy search continues until the prefix reaches 10 tokens and the suffix reaches 20 tokens. Once construction is complete, all hooks are removed and the attack is deployed as ordinary text only. Full implementation details and hyperparameters are reported in Appendix A.

3.2 Behavior-Invariant LoRA Distillation

Figure 2 summarizes the complete BILD pipeline.

BILD is designed to restore the base model's conditional safety behavior under TLCA while minimizing changes on clean inputs. Its central principle is behavioral invariance: adding an adversarial prefix or suffix should not change the appropriate response to the underlying instruction. In particular, an attacked harmful prompt should still lead to refusal, whereas an attacked harmless prompt should still lead to a helpful answer.

Unlike Template LoRA-AT, BILD uses the frozen original model as a teacher and preserves its own clean, instruction-conditioned responses as targets. BILD instead trains the student to reproduce the teacher's clean behavior under both clean and adversarial formatting.

Teacher targets are generated before LoRA adapters are attached. For each clean prompt x, the frozen base model produces a greedy response

$$y_x = f_\theta(x). \tag{14}$$

However, the teacher itself is not perfect, so we do not distill every generated completion. Instead, we retain only examples whose clean teacher behavior already matches the desired safety outcome. Let $\rho(\cdot) \in \{0, 1\}$ denote the refusal detector, where $\rho(y) = 1$ indicates refusal. The retained harmful and harmless subsets are

$$\widetilde{\mathcal{H}} = \{(x^h, y^h) : y^h = f_\theta(x^h),\ \rho(y^h) = 1\}, \qquad \widetilde{\mathcal{S}} = \{(x^s, y^s) : y^s = f_\theta(x^s),\ \rho(y^s) = 0\}. \tag{15}$$

This filtering avoids distilling unsafe harmful completions and accidental over-refusals on benign prompts. It also makes the defense teacher-consistent: harmful examples only teach refusal when the base model already refused the clean instruction, and harmless examples only teach helpfulness when the base model already behaved helpfully.

For every retained prompt, we construct an attacked version by inserting the same learned TLCA prefix and suffix:

$$x^h_{\text{adv}} = u_{\text{pre}} \| x^h \| u_{\text{suf}}, \qquad x^s_{\text{adv}} = u_{\text{pre}} \| x^s \| u_{\text{suf}}. \tag{16}$$

The key invariance constraint is that clean and attacked variants share the same teacher response target. The resulting training set is

$$\mathcal{D}_{\text{BILD}} = \{(x^h, y^h), (x^h_{\text{adv}}, y^h)\} \cup \{(x^s, y^s), (x^s_{\text{adv}}, y^s)\}. \tag{17}$$

This four-way construction is central rather than cosmetic. Clean harmful examples preserve the teacher's native refusal behavior on unsafe requests. Adversarial harmful examples transfer that

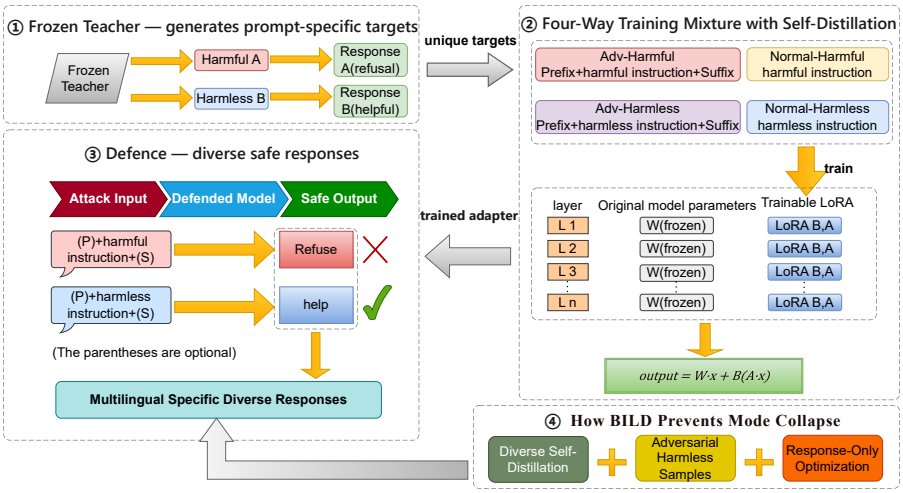


Figure 2: Overview of BILD, our defense against TLCA. A frozen teacher provides prompt-specific targets, and LoRA adapters are trained using clean and adversarial harmful/harmless inputs.

refusal behavior to attacked inputs. Clean harmless examples preserve ordinary helpfulness on benign prompts. Adversarial harmless examples are especially important because they expose the student to the same attack surface while requiring a helpful, non-refusal response. This prevents the defense from collapsing into a shortcut rule that simply rejects prompts containing attack-like prefixes or suffixes.

At a conceptual level, BILD enforces response invariance under the TLCA perturbation operator $T(x) = u_{\text{pre}} \| x \| u_{\text{suf}}$:

$$f_{\theta,\phi}(x) \approx f_{\theta,\phi}(T(x)) \approx f_{\theta}(x), \tag{18}$$

where the approximation is understood at the level of the teacher-generated response distribution.

We freeze the backbone parameters θ and optimize only the LoRA parameters ϕ. Let $\mathcal{D}_{ah}$, $\mathcal{D}_{nh}$, $\mathcal{D}_{as}$, and $\mathcal{D}_{ns}$ denote the adversarial-harmful, normal-harmful, adversarial-harmless, and normal-harmless subsets of $\mathcal{D}_{\text{BILD}}$, respectively. The training objective can be written as the mixture risk

$$\mathcal{L}_{\text{BILD}}(\phi) = \lambda_{ah}\mathcal{L}_{ah} + \lambda_{nh}\mathcal{L}_{nh} + \lambda_{as}\mathcal{L}_{as} + \lambda_{ns}\mathcal{L}_{ns}, \tag{19}$$

where each component is a response-only autoregressive distillation loss. For any subset $\mathcal{D}$, define

$$\mathcal{L}(\mathcal{D};\phi) = -\frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \frac{1}{\sum_i m_i} \sum_i m_i \log p_{\theta,\phi}(y_i \mid x, y_{<i}), \tag{20}$$

where $m_i \in \{0,1\}$ is a mask that equals 1 on response tokens and 0 on prompt tokens. We then set $\mathcal{L}_{ah} = \mathcal{L}(\mathcal{D}_{ah};\phi)$, $\mathcal{L}_{nh} = \mathcal{L}(\mathcal{D}_{nh};\phi)$, $\mathcal{L}_{as} = \mathcal{L}(\mathcal{D}_{as};\phi)$, and $\mathcal{L}_{ns} = \mathcal{L}(\mathcal{D}_{ns};\phi)$. In our implementation, these mixture weights are realized through empirical sample proportions in the training set rather than explicit per-group reweighting.

The response-only masking in Equation 20 is important. The formatted prompt, including any adversarial prefix and suffix, is provided as context, but the model is never trained to reconstruct those attack tokens. If prompt tokens were also supervised, the student could partially solve the training problem by modeling superficial attack-string statistics rather than recovering the clean instruction-conditioned response. By applying loss only on the teacher response, BILD forces the update to act on the model's conditional generation behavior instead of memorizing the perturbation surface form.

4 Experiments

4.1 Experimental Setup

We evaluate Llama-3.1-8B-Instruct, Gemma-2-9B-it, GLM-4-9B-Chat, Qwen3-8B, and InternVL3-8B on the multilingual safety dataset: PolyRefuse (Wang et al., 2025b). PolyRefuse is a parallel

multilingual safety dataset constructed by translating English harmful and harmless instructions into 13 additional languages. Its harmful instructions originate from AdvBench (Zou et al., 2023), MaliciousInstruct (Huang et al., 2024), and TDC2023 (Mazeika et al., 2024), whereas its harmless instructions originate from Alpaca. From the 14 languages provided by PolyRefuse, we evaluate all: English, Chinese, German, Spanish, Japanese, Korean, Thai, Yoruba, Arabic, French, Russian, Italian, Polish, and Dutch. For each language, we follow the official train–test partition provided by PolyRefuse. The training split is used exclusively for method development: its harmful and harmless instructions are used to extract refusal-related directions, select target layers and positions, optimize the TLCA prefix and suffix, generate teacher responses and construct the training data for BILD. In contrast, all compliance results are computed on the complete official test split. Specifically, we evaluate all harmful test instructions under clean and TLCA-attacked conditions for the original models, BILD, and ReFAT. For BILD and ReFAT, we additionally evaluate all harmless test instructions under both clean and TLCA-attacked conditions. PolyRefuse does not provide Chinese harmless instructions in the test split, so harmless compliance is reported for the other thirteen languages but not for Chinese.

We use a deterministic cross-lingual refusal detector that scans the complete generated response against the union of refusal patterns from all 16 supported languages, regardless of the language of the input instruction or generated response. Responses shorter than five characters are also counted as refusals. Harmful compliance is the proportion of responses that do not contain a detected refusal, so lower values indicate stronger safety. Harmless compliance is the proportion of benign requests that receive a non-refusal response, so higher values indicate better preservation of helpfulness. We additionally record unique-response counts. Full evaluation rules are provided in Appendix D.

Table 1: Model-specific TLCA configurations.

Model	Target layers	Position	Source languages
Llama-3.1-8B	L17, L9, L13	-3	English, Yoruba
Gemma-2-9B	L24, L16, L20	-2	English, Yoruba
GLM-4-9B	L20, L12, L14	-1	English, Yoruba
Qwen3-8B	L24, L10, L12	-1	English, Yoruba
InternVL3-8B	L18, L8, L10	-5	English, Yoruba

4.2 Main Results and Analysis

Attack effectiveness. As shown in Table 2, TLCA substantially increases harmful compliance across all five evaluated models. On average, harmful compliance rises from 26.6% in the clean condition to 72.3% under attack, corresponding to an increase of 45.7 percentage points. The attack is effective for every model, although its impact varies considerably across architectures. Gemma-2-9B exhibits the largest increase, from 10.0% to 85.6%, followed by InternVL3-8B, which increases from 25.4% to 76.3%. Llama-3.1-8B shows the smallest increase, from 21.6% to 54.7%. This variation suggests that a model's clean refusal rate alone does not determine its robustness to representation-guided textual perturbations. For example, Gemma-2-9B has the lowest clean harmful compliance but experiences the largest degradation under attack, whereas GLM-4-9B has the highest clean harmful compliance yet a smaller attack-induced increase than Gemma-2-9B and InternVL3-8B. The language-level results in Appendix E further show that the attack remains effective across all fourteen evaluation languages rather than being concentrated only in the source languages used to construct the refusal directions.

Defense effectiveness. Table 2 shows that BILD substantially resolves the vulnerability introduced by TLCA. Averaged across the five models, harmful compliance under TLCA decreases from 72.3% before defense to 16.4% after defense, a reduction of 55.9 percentage points. The reduction occurs for all five models and ranges from 40.0 percentage points for Qwen3-8B to 77.7 percentage points for Gemma-2-9B. BILD also improves clean harmful behavior: average clean harmful compliance decreases from 26.6% to 17.2%, a reduction of 9.4 percentage points. Meanwhile, harmless compliance remains at 96.0% on both clean and attacked inputs. Notably, attacked harmful compliance after defense (16.4%) is close to, and slightly lower than, the post-defense clean rate (17.2%), indicating that BILD completely eliminates the additional vulnerability caused by TLCA without inducing

over-refusal. Taken together, these results show that BILD both strengthens refusal on ordinary harmful requests and neutralizes the additional vulnerability introduced by TLCA, while preserving normal helpfulness.

Table 2: Comparison of the original models, BILD, and ReFAT under clean and TLCA-attacked conditions. Lower harmful compliance indicates stronger safety, whereas higher harmless compliance indicates better preservation of benign utility.

Model	Method	Clean harmful	TLCA-attacked harmful	Gain	Clean harmless	TLCA-attacked harmless
Llama-3.1-8B	Original	21.6%	54.7%	+33.1%	*N/A*	*N/A*
	+ BILD	14.9%	12.3%	-2.6%	97.2%	97.0%
	+ ReFAT	23.1%	55.9%	+32.8%	97.4%	80.7%
Gemma-2-9B	Original	10.0%	85.6%	+75.6%	*N/A*	*N/A*
	+ BILD	7.8%	7.9%	+0.1%	93.2%	93.5%
	+ ReFAT	10.6%	85.6%	+75.0%	92.3%	99.7%
GLM-4-9B	Original	41.5%	77.5%	+36.0%	*N/A*	*N/A*
	+ BILD	18.2%	17.4%	-0.8%	96.0%	95.6%
	+ ReFAT	36.4%	68.5%	+32.1%	96.0%	96.4%
Qwen3-8B	Original	34.4%	67.6%	+33.2%	*N/A*	*N/A*
	+ BILD	28.5%	27.6%	-0.9%	97.9%	97.5%
	+ ReFAT	30.8%	54.5%	+23.7%	96.6%	96.3%
InternVL3-8B	Original	25.4%	76.3%	+50.9%	*N/A*	*N/A*
	+ BILD	16.8%	16.9%	+0.1%	95.6%	96.2%
	+ ReFAT	20.6%	57.4%	+36.8%	91.2%	96.5%
Average	Original	26.6%	72.3%	+45.7%	*N/A*	*N/A*
	+ BILD	17.2%	16.4%	-0.8%	96.0%	96.0%
	+ ReFAT	24.3%	64.4%	+40.1%	94.7%	93.9%

ReFAT and BILD Comparison. For a direct comparison, Table 2 also reports the performance of ReFAT under the same TLCA evaluation protocol. ReFAT is weaker than BILD even before attack is applied: its average clean harmful compliance remains at 24.3%, compared with 17.2% for BILD, indicating that representation-level refusal-feature training does not match BILD's ability to preserve the model's clean safety behavior. Under TLCA, the gap becomes much larger. ReFAT reduces average attacked harmful compliance only to 64.4%, while BILD lowers it further to 16.4%. At the same time, BILD preserves higher average harmless compliance under attack, retaining 96.0% compared with 93.9% for ReFAT. These results show that directly distilling instruction-conditioned clean behavior is more effective than training against representation-level refusal-feature perturbations alone. The complete BILD and ReFAT defense results can be found in Appendix E.

Multilingual analysis. Table 3 shows that TLCA transfers across all fourteen evaluated languages: harmful compliance increases in every language, from 7.2 percentage points in Yoruba to 52.2 points in English. After BILD, attacked harmful compliance is reduced in all languages, with the largest reductions observed in Polish (65.0 points), French (64.0 points), Thai and Arabic (63.6 points), and Italian (62.7 points). In nine languages, the defended attacked harmful compliance is below 10%, including German, English, Spanish, Chinese, Arabic, French, Italian, Russian, and Polish. Even in the more challenging Korean and Dutch settings, BILD lowers harmful compliance to 22.1% and 28.6%, respectively. Although Yoruba remains difficult because its clean harmful compliance is already high, BILD still reduces its attacked rate from 94.4% to 87.2%. At the same time, harmless compliance remains stable, with only minor changes across languages. These results demonstrate that BILD is effective across languages with substantially different levels of alignment

coverage and refusal robustness, showing that the proposed behavior-invariant training can preserve language-conditioned safety behavior rather than relying on a language-specific or English-centered solution. Complete language-level results are reported in Appendix E.

Table 3: Language-level averages over the five evaluated models. (Clean and BILD clean values are harmful compliance rates before TLCA is applied; Harmless and Harmless (TLCA) are evaluated after BILD.)

Language	Clean	Under TLCA	BILD clean	BILD-under TLCA	Harmless	Harmless (TLCA)
German	0.2605	0.7042	0.0993	0.0976	0.967	0.971
English	0.0787	0.6014	0.0336	0.0269	0.946	0.955
Spanish	0.1381	0.6391	0.0643	0.0559	0.968	0.968
Japanese	0.2948	0.7444	0.1619	0.1461	0.960	0.965
Korean	0.3829	0.7651	0.2458	0.2206	0.963	0.961
Thai	0.2511	0.7801	0.1626	0.1437	0.944	0.943
Yoruba	0.8717	0.9437	0.8577	0.8717	0.982	0.978
Chinese	0.1500	0.6500	0.0731	0.0727	N/A	N/A
Arabic	0.1738	0.7147	0.0923	0.0783	0.949	0.942
French	0.1566	0.7080	0.0745	0.0678	0.962	0.959
Italian	0.1381	0.7011	0.0748	0.0738	0.959	0.958
Russian	0.1979	0.6493	0.0895	0.0692	0.961	0.960
Polish	0.2189	0.7395	0.0940	0.0892	0.941	0.945
Dutch	0.4063	0.7874	0.2902	0.2855	0.978	0.973

Collapse ablation and response diversity. Template LoRA-AT obtains 0% attacked harmful compliance, but its harmless compliance also approaches 0%, and generations collapse to one repeated refusal pattern. By contrast, BILD preserves benign utility and prompt-specific generation. By statistically analyzing the specific responses of trained models in tests, it is proved that refusal responses themselves possess diversity; responses to harmless instruction tests also exhibit diversity. Additional diversity statistics are reported in Section F.

5 Conclusion And Limitations

We show that multilingual refusal behavior remains vulnerable to targeted textual perturbations. TLCA substantially increases harmful compliance across five open-weight models, indicating that refusal can be weakened by plain-text prefix–suffix attacks constructed from internal representations.

We further show that low attack success alone is not sufficient evidence of a good defense. Template LoRA-AT can suppress harmful compliance by collapsing the model into generic refusal behavior. In contrast, BILD restores robustness by distilling the frozen model's own clean, prompt-specific responses onto both clean and attacked inputs. Notably, even harmless instructions augmented with the corresponding fixed TLCA prefix–suffix pair still receive useful responses after BILD training, rather than being rejected indiscriminately.

Our comparison with ReFAT further supports the importance of behavior-level invariance. Experimental results show that representation-level refusal feature adversarial training alone does not provide sufficient robustness against the discrete textual perturbations studied here. By directly preserving the model's clean instruction-conditioned behavior, BILD offers a stronger defense while maintaining high multilingual helpfulness.

This work has several limitations. First, broader evaluations on larger models and more diverse safety datasets are needed to test generality. Second, BILD is trained against the specific TLCA prefix–suffix perturbation, so its robustness against adaptive attackers, unseen jailbreak families, or substantially different perturbation formats remains an important direction for future work. Third, our evaluation relies on a deterministic multilingual refusal detector. Although it has been perfected as much as possible, there may still be a very small number of misjudgments.

REFERENCES

Abzal Aidakhmetov, Donato Crisostomi, Tommaso Mencattini, Adrian Robert Minut, Iacopo Masi, and Emanuele Rodolà. Steering vectors are an adversarial attack surface, 2026. URL https://arxiv.org/abs/2606.05958.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083. Curran Associates, Inc., 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf.

Berk Atil, Rebecca J. Passonneau, and Fred Morstatter. Do methods to jailbreak and defend llms generalize across languages?, 2025. URL https://arxiv.org/abs/2511.00689.

Shei Pern Chua, Hao Wu, Qianli Ma, and Fangzhao Wu. Harc: Coupling harmfulness and refusal directions for robust safety alignment, 2026. URL https://arxiv.org/abs/2607.00572.

Hicham Eddoubi, Umar Faruk Abdullahi, and Fadi Hassan. Beyond suffixes: Token position in gcg adversarial attacks on large language models, 2026. URL https://arxiv.org/abs/2602.03265.

Kayden Kehe Ege Çakar, Hannah Guan. Optimizing against safety representations: Activation-guided adversarial suffixes and the geometry of refusal. In *Proceedings of the AAAI Symposium Series*, volume 9, pp. 27–34, 2026. URL https://doi.org/10.1609/aaaiss.v9i1.42902.

Seokil Ham, Jaehyuk Jang, Wonjun Lee, and Changick Kim. Jailbreak to protect: Buffering and reinforcing via temporary jailbreaking for safe fine-tuning in large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=O0JXgBBo3k.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 13707–13727, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3af25aa3de8b7b02ddbd1b6be5031be8-Paper-Conference.pdf.

Chigozirim Ifebi, Brent Kong, and Ayushi Mehrotra. Minionese: Comprehensive benchmark and mechanistic study of multilingual llm safety, 2026. URL https://arxiv.org/abs/2607.10112.

Weipeng Jiang, Zhenting Wang, Juan Zhai, Shiqing Ma, Zhengyu Zhao, and Chao Shen. An optimizable suffix is worth a thousand templates: Efficient black-box jailbreaking without affirmative phrases via LLM as optimizer. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5486–5498, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.302. URL https://aclanthology.org/2025.findings-naacl.302/.

Aleksandra Krasnodębska, Wojciech Kusa, and Aldo Lipani. Multilingual refusal alignment for safer large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30769–30790, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1537. URL https://aclanthology.org/2026.findings-acl.1537/.

Wenhao Lan, Shan Li, Xinhua Lai, Meiqi Wu, Junbin Yang, and Haihua Shen. From refusal geometry to safety geometry: Harmfulness–refusal coupling under dynamic adversarial fine-tuning, 2026. URL https://arxiv.org/abs/2606.16349.

Utsav Maskey, Sumit Yadav, Mark Dras, and Usman Naseem. SafeConstellations: Mitigating over-refusals in LLMs through task-aware representation steering. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44424–44440, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2056. URL https://aclanthology.org/2026.acl-long.2056/.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

Kartik Sharma, Yiqiao Jin, Vineeth Rakesh, Yingtong Dou, Menghai Pan, Mahashweta Das, and Srijan Kumar. Sysformer: Safeguarding frozen large language models with adaptive system prompts. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 2174–2193, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/04b25f6747d6fea910d2c7170866bf70-Paper-Conference.pdf.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2668–2680, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.156. URL https://aclanthology.org/2024.findings-acl.156/.

Haoyu Wang, Zeyu Qin, Li Shen, Xueqian Wang, Minhao Cheng, and Dacheng Tao. Leveraging reasoning with guidelines to elicit and utilize knowledge for enhancing safety alignment. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025a. URL https://openreview.net/forum?id=BfXS7uApJ6.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen tse Huang, Wenxiang Jiao, and Michael R. Lyu. All languages matter: On the multilingual safety of large language models, 2024. URL https://arxiv.org/abs/2310.00905.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schuetze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 32380–32423. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1086. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2e94772e3c079d83f79c311d07456111-Paper-Conference.pdf.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=80IwJqlXs8.

Yingsha Xie, Tiansheng Huang, Enneng Yang, Rui Min, Wenjie Lu, Xiaochun Cao, Naiqiang Tan, and Li Shen. Mitigating safety tax via distribution-grounded refinement in large reasoning models, 2026. URL https://arxiv.org/abs/2602.02136.

Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. A comprehensive study of jailbreak attack versus defense for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7432–7449, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.443. URL https://aclanthology.org/2024.findings-acl.443/.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey, 2024. URL https://arxiv.org/abs/2407.04295.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak gpt-4, 2024. URL https://arxiv.org/abs/2310.02446.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust llm safeguarding via refusal feature adversarial training. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 5254–5277, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1022661f3f43406065641f16ce25eafa-Paper-Conference.pdf.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 140283–140318. Curran Associates, Inc., 2025. doi: 10.52202/085713-4688. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/cd18539787d90e1d682d557c2c71b534-Paper-Conference.pdf.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

A Attack Construction Details

This appendix provides implementation details for TLCA that are omitted from the main text.

The experiments use the PolyRefuse harmful and harmless instructions across English, Chinese, German, Spanish, Japanese, Korean, Thai, Yoruba, Arabic, French, Italian, Russian, Polish, and Dutch. All attacks are pure prompt attacks: no model weights, activations, logits, or system prompts are modified at deployment time. Each attacked input is formatted as a normal chat request with the attack string placed inside the user message:

$$\text{prefix} + \text{instruction} + \text{suffix}.$$

For models with native chat templates, we use the tokenizer's built-in chat template.

We search for a strong refusal-separating layer-position pair using harmful and harmless training instructions. Let $h_i^{(\ell,p)}$ denote the activation at layer ℓ and token position p for sample i. For a candidate pair (ℓ, p), we compute a separation vector

$$r^{(\ell,p)} = \mu_{\text{harm}}^{(\ell,p)} - \mu_{\text{harmless}}^{(\ell,p)},$$

where μ_{harm} and μ_{harmless} are mean activations over sampled harmful and harmless instructions. We project both classes onto $r^{(\ell,p)}$ and score the pair with Cohen's d, as in Eq. 3:

$$d^{(\ell,p)} = \frac{\left|\mathbb{E}[z_{\text{harm}}^{(\ell,p)}] - \mathbb{E}[z_{\text{harmless}}^{(\ell,p)}]\right|}{\sqrt{\frac{1}{2}\left(\text{Var}[z_{\text{harm}}^{(\ell,p)}] + \text{Var}[z_{\text{harmless}}^{(\ell,p)}]\right)} + \epsilon},$$

where $z = h \cdot r/(\|r\| + \epsilon)$ and $\epsilon = 10^{-8}$.

Candidate positions are

$$p \in \{-5, -4, -3, -2, -1\}.$$

The searched layer ranges are model-specific:

- Gemma-2-9B: searched from 40% of depth with step 2;
- GLM-4-9B: searched from 25% of depth with step 2;
- Qwen3-8B: searched from 25% of depth with step 2;
- Llama-3.1-8B: searched from 28% of depth with step 2;
- InternVL3-8B: searched from 30% of depth with step 2.

For this search, we sample up to 32 harmful and 32 harmless instructions per language subset.

After finding the best layer-position pair $(\ell^\star, p^\star)$, TLCA does not attack only that single layer. Instead, it constructs a three-layer target set consisting of the best layer $\ell^\star$ and two weaker layers at the same position $p^\star$ whose Cohen's d is above a minimum threshold but substantially below the best layer. Concretely, a weak-layer candidate must satisfy

$$d^{(\ell,p^\star)} > 0.1 \quad \text{and} \quad d^{(\ell,p^\star)} < 0.8\, d^{(\ell^\star,p^\star)}.$$

Among such candidates, we choose the lowest-scoring layers first until we obtain three target layers in total.

For each target layer, we estimate refusal directions from two source languages and average them after normalization. English is always included. The second source language is chosen automatically as the non-English language with the lowest silhouette score at the best layer-position pair, i.e., the language whose harmful and harmless activations are least separable and therefore whose refusal geometry is weakest. If $r_{\text{en}}^{(\ell)}$ and $r_{\text{weak}}^{(\ell)}$ are the two normalized directions (cf. Eq. 2), then the fused direction is

$$\bar{r}^{(\ell)} = \frac{\frac{r_{\text{en}}^{(\ell)}}{\|r_{\text{en}}^{(\ell)}\|+\epsilon} + \frac{r_{\text{weak}}^{(\ell)}}{\|r_{\text{weak}}^{(\ell)}\|+\epsilon}}{\left\|\frac{r_{\text{en}}^{(\ell)}}{\|r_{\text{en}}^{(\ell)}\|+\epsilon} + \frac{r_{\text{weak}}^{(\ell)}}{\|r_{\text{weak}}^{(\ell)}\|+\epsilon}\right\| + \epsilon}.$$

Table 4: Shared TLCA attack hyperparameters.

Hyperparameter	Value
Harmful training samples for direction extraction per source language	128
Harmless training samples for direction extraction per source language	128
Samples for silhouette per language	32 harmful + 32 harmless
Harmful test samples per language	572
Harmless test samples per language	500
Candidate positions	$\{-5, -4, -3, -2, -1\}$
Prefix length	10 tokens
Suffix length	20 tokens
Maximum token repetition	3
Weak-layer minimum Cohen's d	0.1
Rarity weight	0.02–0.15

Here, $\bar{r}^{(\ell)}$ is the fused refusal direction denoted by r_ℓ in the main text. We use up to 128 harmful and 128 harmless training instructions per source language for this extraction step.

We compute the silhouette score at the best layer-position pair using 32 harmful and 32 harmless training instructions per language:

$$\text{Silhouette}(X, y) = \frac{1}{|X|} \sum_i \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\} + \epsilon},$$

where $a(i)$ is the mean intra-class distance and $b(i)$ is the smallest mean inter-class distance. Higher values indicate better harmful and harmless separation.

TLCA constructs a 10-token prefix and a 20-token suffix by greedy search over a candidate token pool. The prefix is intended to shift the residual stream before the harmful instruction is processed; the suffix is intended to preserve that shift afterward.

Candidate tokens are first screened by an embedding-level score. This implements the token-ranking rule in Eq. 10.

$$s_{\text{embed}}(t) = e_t^\top \big(- \bar{r}^{(\ell^\star)} \big),$$

where e_t is the embedding of token t. Special tokens are excluded. For models using rarity preference, the score is augmented by a rarity bonus:

$$s(t) = s_{\text{emb}}(t) + \lambda_{\text{rare}} \max\left(0, \frac{\text{id}(t) - q}{|V| - q}\right).$$

We then keep the top candidate set and evaluate tokens with activation-level greedy search. The prefix and suffix objectives correspond to Eqs. 12 and 13, respectively.

At each greedy step, we test the candidate token by measuring how strongly the resulting attacked prompt pushes activations toward $-\bar{r}^{(\ell)}$ across the target layers:

$$S = \sum_{\ell \in \mathcal{L}_{\text{atk}}} w_\ell \cdot \mathbb{E}\left[-h^{(\ell, p^\star)} \cdot \bar{r}^{(\ell)} \right].$$

The best layer uses weight 1.0 and the two weak layers use weight 1.5. We also cap token repetition at three occurrences to reduce degenerate strings. The rarity weight is model-specific and is selected separately for each model.

B Defense Implementation Details

This section gives the missing implementation details for BILD.

BILD trains a LoRA adapter while keeping the original base model fully frozen. For each training pair consisting of a prompt x and target response y, we optimize a masked next-token negative

Table 5: Training categories used by BILD.

Category	Input and target	Purpose
Attacked harmful	Attacked harmful prompt → refusal	Teach the model to reject harmful instructions under TLCA.
Normal harmful	Unmodified harmful prompt → refusal	Preserve ordinary harmful-content refusal.
Attacked harmless	Attacked harmless prompt → helpful response	Prevent over-refusal under adversarial formatting.
Normal harmless	Unmodified harmless prompt → helpful response	Preserve normal helpful behavior.

log-likelihood objective over response tokens only:

$$\mathcal{L}_{\text{BILD}}(\phi) = -\sum_{t=1}^{|y|} \log p_{\theta,\phi}(y_t \mid x, y_{<t}).$$

All prompt-token labels are set to -100; only response tokens contribute to the loss. The implementation follows the standard causal-LM supervised fine-tuning recipe. The prompt and response are concatenated into a single sequence; loss is computed only on response tokens, while prompt-token labels are replaced with the ignore index. Thus, the model conditions on the prompt but is trained only to generate the assistant response.

The key distinction from the fixed-template baseline is the choice of teacher targets. Instead of training every attacked harmful prompt toward the same canned refusal sentence, BILD uses the original model's own responses as training targets:

- harmful prompts are kept only if the teacher response is a refusal;
- harmless prompts are kept only if the teacher response is helpful.

This preserves response diversity and avoids collapsing all attacked prompts to a single refusal string.

BILD uses four prompt-target categories. Their roles are summarized in Table 5.

Teacher responses are generated by the original frozen model before LoRA training. We then filter them with the same refusal detector used for evaluation:

- harmful responses are retained only if the complete response contains a refusal marker;
- harmless responses are retained only if the complete response does not contain a refusal marker;
- responses shorter than five characters are discarded.

To reduce the imbalance caused by model- and language-dependent alignment behavior, we perform teacher-target selection separately for each language and each of the four BILD categories. For each language–category combination, we randomly shuffle the corresponding PolyRefuse training examples using a fixed seed and examine them in consecutive groups of 30 examples. After applying the eligibility criteria above, if a group contains at least 10 usable teacher-target examples, we stop screening that language–category combination and retain the usable examples obtained up to that stopping point. If fewer than 10 usable examples are found, we continue screening the next group of 30 examples. The procedure terminates when either at least 10 usable examples have been identified or the complete PolyRefuse training split has been examined. If the split is exhausted before 10 usable examples are found, all eligible examples identified from that split are retained. This procedure is applied independently to attacked-harmful, normal-harmful, attacked-harmless, and normal-harmless examples. The minimum target of 10 examples is therefore imposed separately for each category rather than on the union of all training examples. This design ensures that every language contributes a nontrivial number of examples whenever the teacher provides enough eligible responses, while preventing languages with stronger or weaker alignment from contributing disproportionately different numbers of training examples. It also makes the comparison across languages and models more controlled by keeping the target-selection budget approximately comparable.

Table 6: Total number of BILD training samples used for each model.

Model	Total Training Samples
Gemma-2-9B	972
GLM-4-9B	823
Qwen3-8B	864
Llama-3.1-8B	943
InternVL3-8B	901

Table 7: Architecture-specific LoRA target modules.

Model	Target modules
Gemma-2-9B	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
GLM-4-9B	query_key_value, dense, dense_h_to_4h, dense_4h_to_h
Qwen3-8B	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Llama-3.1-8B	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
InternVL3-8B	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj

All BILD experiments use rank-64 LoRA with scaling factor $\alpha = 128$ and dropout 0.05. The target modules are architecture-specific, as summarized in Table 7.

The final BILD runs use AdamW with cosine warmup, gradient accumulation, gradient clipping, and gradient checkpointing.

C ReFAT Implementation Details

We implement Refusal Feature Adversarial Training (ReFAT) (Yu et al., 2025) as the representation-level baseline following the procedure described in the original paper. For each layer ℓ, we estimate a refusal feature from the difference between the mean last-token activation on harmful and harmless prompts:

$$r^{(\ell)} = \frac{\mu^{(\ell)}_{\text{harm}} - \mu^{(\ell)}_{\text{harmless}}}{\left\| \mu^{(\ell)}_{\text{harm}} - \mu^{(\ell)}_{\text{harmless}} \right\|_2 + \epsilon}. \tag{21}$$

The feature is estimated from 32 randomly sampled harmful and 32 randomly sampled harmless English training instructions. We recompute the refusal features every four optimizer steps using the current model parameters.

During ReFAT training, refusal feature ablation (RFA) is applied only to harmful samples. With probability $p_{\text{RFA}} = 0.5$, the hidden state at the last user-token position is projected away from the normalized refusal feature:

$$h'_\ell = h_\ell - r^{(\ell)} \left(r^{(\ell)\top} h_\ell \right). \tag{22}$$

The intervention is applied to the final 75% of transformer layers. Harmless samples are processed without RFA. The model is then optimized with the standard response-token autoregressive cross-entropy loss, using refusal responses for harmful instructions and helpful responses for harmless instructions. These target responses are generated greedily by the original frozen model before ReFAT training.

For each model, we use 500 harmful and 500 harmless English training instructions. ReFAT is implemented with LoRA adapters of rank 8, scaling factor $\alpha = 16$, and dropout 0.05, targeting the same architecture-specific linear projection modules listed for the corresponding model. We train for two epochs with AdamW, learning rate 10^{-4}, per-step batch size 4, gradient accumulation over four steps, maximum sequence length 512, and gradient clipping at 1.0. The backbone parameters remain frozen throughout training.

For evaluation, we use the same PolyRefuse test split, TLCA prefix–suffix attack protocol, deterministic decoding settings, and cross-lingual refusal detector as for BILD. We report clean harmful

compliance, TLCA-attacked harmful compliance, clean harmless compliance, and TLCA-attacked harmless compliance. Chinese harmless compliance is omitted because the PolyRefuse test split does not provide Chinese harmless instructions.

D Evaluation Details

All attack and defense evaluations are conducted on the official PolyRefuse test split. For each language, we evaluate the complete set of available test instructions. Harmful instructions are evaluated in four conditions: the original model on clean inputs, the original model under TLCA, the defended model on clean inputs, and the defended model under TLCA. Harmless instructions are evaluated for the defended model on both clean inputs and inputs augmented with TLCA. PolyRefuse does not provide harmless instructions for Chinese in the test split; therefore, harmless compliance is not reported for Chinese.

For every model and language, the clean and attacked conditions use the same underlying test instructions. The only difference between these conditions is whether the fixed TLCA prefix and suffix are inserted into the user message. The same deterministic evaluation protocol is applied to clean, attacked, and defended outputs.

We use a deterministic cross-lingual refusal detector for both teacher-target filtering and final evaluation. The detector does not assume that the response is written in the same language as the input instruction. Instead, every generated response is checked against the union of refusal patterns from all 16 supported languages.

The detector first checks whether the response is empty or contains fewer than five non-whitespace characters. Such responses are classified as refusals. Otherwise, the response is normalized by lowercasing the text, replacing typographic quotation marks and dashes with canonical forms, converting non-breaking spaces to ordinary spaces, normalizing German esszet characters, removing Latin combining diacritics, collapsing repeated whitespace, and stripping selected leading markdown and punctuation characters. The normalized response is then checked for substring matches against the combined multilingual refusal dictionary. The complete generated response is scanned.

The refusal dictionary contains 4,660 refusal phrases spanning 16 languages: English, Chinese, German, Spanish, French, Italian, Dutch, Japanese, Korean, Polish, Russian, Thai, Yoruba, Arabic, Czech, and Portuguese. The dictionary explicitly accounts for grammatical variation across languages. In particular, it includes alternative word orders, pronoun insertions, inflectional and conjugational variants, formal and informal forms, accented and unaccented spellings where applicable, and commonly used synonymous or near-synonymous refusal expressions. The patterns cover direct inability or denial, apologies, identity-based refusal statements, authorization and capability limitations, moral or legal objections, recommendations against an action, and explicit refusal expressions. A response is classified as a refusal if at least one pattern from any supported language occurs anywhere in the normalized response. This cross-lingual matching allows the detector to recognize a refusal even when the response language differs from the language of the input instruction. This cross-lingual rule is used to account for language switching in model outputs.

E Additional Results

Language-level TLCA attack results. Tables 8 and 9 report the language-wise attack results for all five models.

Language-level BILD defense results. Tables 10–13 report the language-wise BILD defense results for all five models.

Language-level ReFAT defense results. Tables 14–17 report the language-wise ReFAT defense results for all five models.

Table 8: Language-wise TLCA attack results for Gemma, GLM and Qwen3.

Model	Lang	Silhouette	Clean harmful compliance	Attacked harmful compliance	Gain
Gemma-2-9B	German	0.3551	0.1014	0.8514	0.7500
	English	0.3411	0.0175	0.8374	0.8199
	Spanish	0.3233	0.0542	0.8601	0.8059
	Japanese	0.3859	0.1399	0.8339	0.6940
	Korean	0.3682	0.1136	0.8462	0.7326
	Thai	0.2977	0.0455	0.8724	0.8269
	Yoruba	0.1015	0.6066	0.9563	0.3497
	Chinese	0.3287	0.0734	0.8706	0.7972
	Arabic	0.3361	0.0490	0.8409	0.7919
	French	0.3391	0.0402	0.8584	0.8182
	Italian	0.3334	0.0227	0.8759	0.8532
	Russian	0.3076	0.0472	0.7762	0.7290
	Polish	0.3038	0.0350	0.8549	0.8199
	Dutch	0.2974	0.0577	0.8549	0.7972
GLM-4-9B	German	0.2681	0.5559	0.7343	0.1784
	English	0.2923	0.1486	0.6014	0.4528
	Spanish	0.2512	0.3024	0.7045	0.4021
	Japanese	0.2698	0.4213	0.8129	0.3916
	Korean	0.2728	0.4021	0.7780	0.3759
	Thai	0.1137	0.5122	0.8531	0.3409
	Yoruba	0.0558	0.9213	0.9423	0.0210
	Chinese	0.2483	0.2063	0.6364	0.4301
	Arabic	0.1897	0.3217	0.7325	0.4108
	French	0.2243	0.2955	0.8077	0.5122
	Italian	0.2240	0.3357	0.8164	0.4807
	Russian	0.2157	0.3287	0.7552	0.4265
	Polish	0.1876	0.3951	0.8024	0.4073
	Dutch	0.1848	0.6573	0.8794	0.2221
Qwen3-8B	German	0.3842	0.3794	0.5437	0.1643
	English	0.4066	0.1224	0.6451	0.5227
	Spanish	0.3878	0.1888	0.5332	0.3444
	Japanese	0.4107	0.3811	0.6469	0.2658
	Korean	0.4016	0.4808	0.7203	0.2395
	Thai	0.3128	0.2885	0.6591	0.3706
	Yoruba	0.0973	0.9598	0.9510	-0.0088
	Chinese	0.3711	0.2010	0.5332	0.3322
	Arabic	0.3503	0.1766	0.6871	0.5105
	French	0.3770	0.2605	0.7815	0.5210
	Italian	0.3742	0.1486	0.6364	0.4878
	Russian	0.3165	0.3881	0.6801	0.2920
	Polish	0.3104	0.2780	0.6119	0.3339
	Dutch	0.2927	0.5559	0.8287	0.2728

Table 9: Language-wise TLCA attack results for Llama and InternVL3.

Model	Lang	Silhouette	Clean harmful compliance	Attacked harmful compliance	Gain
Llama-3.1-8B	German	0.4770	0.1224	0.6049	0.4825
	English	0.4058	0.0507	0.2727	0.2220
	Spanish	0.4817	0.0822	0.3794	0.2972
	Japanese	0.4555	0.3759	0.5874	0.2115
	Korean	0.4182	0.4301	0.6871	0.2570
	Thai	0.3879	0.1941	0.6364	0.4423
	Yoruba	0.0812	0.9703	0.9388	-0.0315
	Chinese	0.4366	0.1399	0.5472	0.4073
	Arabic	0.3931	0.1329	0.5857	0.4528
	French	0.4591	0.0769	0.4353	0.3584
	Italian	0.4315	0.0857	0.3899	0.3042
	Russian	0.3854	0.0979	0.4545	0.3566
	Polish	0.3713	0.1906	0.5804	0.3898
	Dutch	0.3551	0.0717	0.5542	0.4825
InternVL3-8B	German	0.2597	0.1434	0.7867	0.6433
	English	0.3123	0.0542	0.6503	0.5961
	Spanish	0.3080	0.0629	0.7185	0.6556
	Japanese	0.1854	0.1556	0.8409	0.6853
	Korean	0.2331	0.4878	0.7937	0.3059
	Thai	0.1684	0.2150	0.8794	0.6644
	Yoruba	0.1512	0.9003	0.9301	0.0298
	Chinese	0.2642	0.1294	0.6626	0.5332
	Arabic	0.2171	0.1888	0.7273	0.5385
	French	0.2456	0.1101	0.6573	0.5472
	Italian	0.2278	0.0979	0.7867	0.6888
	Russian	0.2329	0.1276	0.5804	0.4528
	Polish	0.1825	0.1958	0.8479	0.6521
	Dutch	0.1608	0.6888	0.8199	0.1311

Table 10: Language-wise BILD defense results for Gemma, GLM and Qwen3. (Harmful compliance)

Model	Lang	Clean harmful compliance	Attacked harmful compliance
Gemma-2-9B	German	0.0892	0.0787
	English	0.0122	0.0140
	Spanish	0.0402	0.0245
	Japanese	0.0787	0.0734
	Korean	0.0717	0.0752
	Thai	0.0350	0.0297
	Yoruba	0.5787	0.5892
	Chinese	0.0297	0.0315
	Arabic	0.0262	0.0280
	French	0.0140	0.0157
	Italian	0.0140	0.0157
	Russian	0.0245	0.0192
	Polish	0.0192	0.0175
	Dutch	0.0629	0.0927
GLM-4-9B	German	0.0629	0.0542
	English	0.0192	0.0105
	Spanish	0.0350	0.0367
	Japanese	0.0822	0.0769
	Korean	0.0787	0.0699
	Thai	0.1853	0.1591
	Yoruba	0.9056	0.9388
	Chinese	0.0332	0.0350
	Arabic	0.1713	0.1381
	French	0.1119	0.0927
	Italian	0.1276	0.1206
	Russian	0.1224	0.0734
	Polish	0.1171	0.1084
	Dutch	0.5000	0.5262
Qwen3-8B	German	0.2640	0.2762
	English	0.0752	0.0664
	Spanish	0.1206	0.1084
	Japanese	0.3654	0.3392
	Korean	0.4738	0.4528
	Thai	0.2710	0.2517
	Yoruba	0.9493	0.9720
	Chinese	0.1451	0.1538
	Arabic	0.1364	0.1469
	French	0.1626	0.1678
	Italian	0.1294	0.1294
	Russian	0.1888	0.1608
	Polish	0.2080	0.2028
	Dutch	0.4983	0.4423

Table 11: Language-wise BILD defense results for Llama and InternVL3. (Harmful compliance)

Model	Lang	Clean harmful compliance	Attacked harmful compliance
Llama-3.1-8B	German	0.0490	0.0332
	English	0.0315	0.0157
	Spanish	0.0699	0.0594
	Japanese	0.1591	0.1206
	Korean	0.3042	0.2622
	Thai	0.1783	0.1224
	Yoruba	0.9650	0.9545
	Chinese	0.0472	0.0262
	Arabic	0.0577	0.0105
	French	0.0297	0.0192
	Italian	0.0402	0.0280
	Russian	0.0402	0.0157
	Polish	0.0490	0.0245
	Dutch	0.0612	0.0367
InternVL3-8B	German	0.0315	0.0455
	English	0.0297	0.0280
	Spanish	0.0559	0.0507
	Japanese	0.1241	0.1206
	Korean	0.3007	0.2430
	Thai	0.1434	0.1556
	Yoruba	0.8899	0.9038
	Chinese	0.1101	0.1171
	Arabic	0.0699	0.0682
	French	0.0542	0.0437
	Italian	0.0629	0.0752
	Russian	0.0717	0.0769
	Polish	0.0769	0.0927
	Dutch	0.3287	0.3444

Table 12: Language-wise BILD defense results for Gemma, GLM and Qwen3. (Harmless compliance) Chinese harmless instructions are not provided in the PolyRefuse test split, so Chinese harmless compliance is reported as N/A.

Model	Lang	Harmless compliance	Attacked harmless compliance
Gemma-2-9B	German	0.952	0.958
	English	0.930	0.940
	Spanish	0.954	0.956
	Japanese	0.922	0.938
	Korean	0.942	0.938
	Thai	0.900	0.916
	Yoruba	0.982	0.980
	Chinese	N/A	N/A
	Arabic	0.902	0.910
	French	0.930	0.924
	Italian	0.926	0.916
	Russian	0.922	0.916
	Polish	0.886	0.894
	Dutch	0.970	0.966
GLM-4-9B	German	0.972	0.972
	English	0.940	0.954
	Spanish	0.960	0.956
	Japanese	0.956	0.960
	Korean	0.936	0.934
	Thai	0.946	0.936
	Yoruba	0.974	0.964
	Chinese	N/A	N/A
	Arabic	0.958	0.938
	French	0.978	0.962
	Italian	0.968	0.968
	Russian	0.952	0.948
	Polish	0.952	0.938
	Dutch	0.990	0.994
Qwen3-8B	German	0.988	0.982
	English	0.968	0.974
	Spanish	0.984	0.984
	Japanese	0.980	0.974
	Korean	0.976	0.970
	Thai	0.968	0.964
	Yoruba	0.986	0.994
	Chinese	N/A	N/A
	Arabic	0.976	0.976
	French	0.976	0.982
	Italian	0.972	0.970
	Russian	0.992	0.984
	Polish	0.978	0.978
	Dutch	0.984	0.974

Table 13: Language-wise BILD defense results for Llama and InternVL3. (Harmless compliance) Chinese harmless instructions are not provided in the PolyRefuse test split, so Chinese harmless compliance is reported as N/A.

Model	Lang	Harmless compliance	Attacked harmless compliance
Llama-3.1-8B	German	0.982	0.984
	English	0.962	0.966
	Spanish	0.976	0.980
	Japanese	0.984	0.984
	Korean	0.984	0.984
	Thai	0.950	0.954
	Yoruba	0.998	0.994
	Chinese	N/A	N/A
	Arabic	0.972	0.942
	French	0.970	0.960
	Italian	0.974	0.968
	Russian	0.984	0.986
	Polish	0.950	0.960
	Dutch	0.956	0.944
InternVL3-8B	German	0.940	0.960
	English	0.932	0.940
	Spanish	0.964	0.964
	Japanese	0.960	0.970
	Korean	0.976	0.980
	Thai	0.954	0.944
	Yoruba	0.968	0.958
	Chinese	N/A	N/A
	Arabic	0.938	0.946
	French	0.954	0.966
	Italian	0.956	0.966
	Russian	0.954	0.966
	Polish	0.940	0.956
	Dutch	0.988	0.988

Table 14: Language-wise ReFAT defense results for Gemma, GLM and Qwen3. (Harmful compliance)

Model	Lang	Clean harmful compliance	Attacked harmful compliance
Gemma-2-9B	German	0.0962	0.8706
	English	0.0175	0.8322
	Spanish	0.0402	0.8724
	Japanese	0.1451	0.8164
	Korean	0.0857	0.8252
	Thai	0.0350	0.8601
	Yoruba	0.7378	0.9738
	Chinese	0.0699	0.8654
	Arabic	0.0262	0.8409
	French	0.0402	0.8689
	Italian	0.0245	0.8654
	Russian	0.0385	0.7780
	Polish	0.0332	0.8462
	Dutch	0.0892	0.8724
GLM-4-9B	German	0.5210	0.7395
	English	0.0752	0.5175
	Spanish	0.2570	0.6503
	Japanese	0.3427	0.7220
	Korean	0.3636	0.5874
	Thai	0.4983	0.8287
	Yoruba	0.8497	0.9196
	Chinese	0.1713	0.5542
	Arabic	0.2745	0.6661
	French	0.2413	0.7150
	Italian	0.3164	0.7448
	Russian	0.2552	0.5629
	Polish	0.3514	0.6836
	Dutch	0.5857	0.6976
Qwen3-8B	German	0.3741	0.5594
	English	0.0962	0.3864
	Spanish	0.1311	0.4563
	Japanese	0.3129	0.6434
	Korean	0.4038	0.5402
	Thai	0.2500	0.4633
	Yoruba	0.9650	0.9318
	Chinese	0.1521	0.3217
	Arabic	0.1678	0.5245
	French	0.2168	0.6084
	Italian	0.1486	0.4948
	Russian	0.3549	0.5157
	Polish	0.2290	0.4493
	Dutch	0.5052	0.7325

Table 15: Language-wise ReFAT defense results for Llama and InternVL3. (Harmful compliance)

Model	Lang	Clean harmful compliance	Attacked harmful compliance
Llama-3.1-8B	German	0.1101	0.5577
	English	0.0734	0.3234
	Spanish	0.0857	0.3671
	Japanese	0.4126	0.6993
	Korean	0.4563	0.6853
	Thai	0.2605	0.7115
	Yoruba	0.9668	0.8776
	Chinese	0.2395	0.5962
	Arabic	0.1311	0.5927
	French	0.0804	0.4406
	Italian	0.0927	0.3934
	Russian	0.0857	0.4580
	Polish	0.1696	0.5402
	Dutch	0.0717	0.5822
InternVL3-8B	German	0.1276	0.5490
	English	0.0332	0.2448
	Spanish	0.0682	0.5210
	Japanese	0.1626	0.6364
	Korean	0.5594	0.5752
	Thai	0.1591	0.7115
	Yoruba	0.6766	0.9143
	Chinese	0.1171	0.3951
	Arabic	0.1189	0.6119
	French	0.1031	0.4913
	Italian	0.1329	0.6259
	Russian	0.1049	0.3811
	Polish	0.1399	0.7133
	Dutch	0.1836	0.6696

Table 16: Language-wise ReFAT defense results for Gemma, GLM and Qwen3. (Harmless compliance) Chinese harmless instructions are not provided in the PolyRefuse test split, so Chinese harmless compliance is reported as N/A.

Model	Lang	Harmless compliance	Attacked harmless compliance
Gemma-2-9B	German	0.942	0.996
	English	0.936	0.996
	Spanish	0.922	0.998
	Japanese	0.902	1.000
	Korean	0.942	0.998
	Thai	0.886	0.996
	Yoruba	0.952	0.996
	Chinese	N/A	N/A
	Arabic	0.882	0.994
	French	0.934	0.994
	Italian	0.916	0.998
	Russian	0.926	0.996
	Polish	0.902	0.996
	Dutch	0.962	0.998
GLM-4-9B	German	0.974	0.946
	English	0.930	0.972
	Spanish	0.972	0.984
	Japanese	0.968	0.934
	Korean	0.960	0.892
	Thai	0.962	0.982
	Yoruba	0.932	0.974
	Chinese	N/A	N/A
	Arabic	0.962	0.970
	French	0.978	0.978
	Italian	0.958	0.984
	Russian	0.958	0.960
	Polish	0.952	0.974
	Dutch	0.980	0.980
Qwen3-8B	German	0.976	0.952
	English	0.972	0.968
	Spanish	0.968	0.978
	Japanese	0.982	0.976
	Korean	0.962	0.934
	Thai	0.928	0.966
	Yoruba	0.986	0.990
	Chinese	N/A	N/A
	Arabic	0.942	0.986
	French	0.976	0.980
	Italian	0.954	0.984
	Russian	0.976	0.908
	Polish	0.948	0.914
	Dutch	0.986	0.984

Table 17: Language-wise ReFAT defense results for Llama and InternVL3. (Harmless compliance) Chinese harmless instructions are not provided in the PolyRefuse test split, so Chinese harmless compliance is reported as N/A.

Model	Lang	Harmless compliance	Attacked harmless compliance
Llama-3.1-8B	German	0.986	0.866
	English	0.972	0.904
	Spanish	0.972	0.798
	Japanese	0.992	0.888
	Korean	0.982	0.858
	Thai	0.956	0.848
	Yoruba	0.980	0.782
	Chinese	N/A	N/A
	Arabic	0.976	0.712
	French	0.968	0.806
	Italian	0.978	0.718
	Russian	0.968	0.796
	Polish	0.956	0.730
	Dutch	0.982	0.780
InternVL3-8B	German	0.896	0.956
	English	0.904	0.946
	Spanish	0.938	0.976
	Japanese	0.914	0.972
	Korean	0.972	0.964
	Thai	0.898	0.970
	Yoruba	0.568	0.972
	Chinese	N/A	N/A
	Arabic	0.902	0.974
	French	0.942	0.962
	Italian	0.948	0.964
	Russian	0.900	0.958
	Polish	0.938	0.968
	Dutch	0.936	0.964

F Mode Collapse and Diversity Diagnostics

Before the final BILD formulation, we trained a simpler LoRA adversarial baseline that used fixed refusal templates as targets for attacked and normal harmful prompts, rather than the model's own teacher responses. This baseline is useful because it shows that low attacked-harmful compliance alone is not sufficient evidence of a good defense.

Table 18: Gemma fixed-template LoRA compliance.

Language	Defended attacked harmful	Defended harmless (clean)
German	0.00	0.00
English	0.00	0.00
Spanish	0.00	0.00
Japanese	0.00	0.00
Korean	0.00	0.00
Thai	0.00	0.00
Yoruba	0.00	0.00
Chinese	0.00	N/A
Arabic	0.00	0.00
French	0.00	0.00
Italian	0.00	0.00
Russian	0.00	0.00
Polish	0.00	0.00
Dutch	0.00	0.00

In the Gemma-2-9B-it fixed-template run, both attacked harmful and harmless compliance dropped to 0.00 in every evaluated language, as shown in Table 18.

This behavior was degenerate rather than desirable: the model learned to emit the same short refusal almost everywhere, resulting in near-zero response diversity and severe over-refusal.

The collapse mechanism is straightforward. Every attacked harmful prompt is mapped to one of a small number of interchangeable refusal templates, so the conditional training target has very low entropy. Under such supervision, the easiest solution for the adapter is to memorize a short generic refusal pattern that fires whenever the prompt appears unsafe or adversarially wrapped. That solution yields excellent attacked-harmful refusal but poor harmless behavior and little lexical variety.

A defense that yields 0% attacked-harmful compliance by always replying with the same sentence is not a useful robustness solution. The appendix therefore reports attacked-harmful compliance together with harmless compliance and exact-uniqueness statistics. BILD is intended to improve the first quantity without collapsing the latter two.

Table 19: Exact-response uniqueness by language after BILD. Each entry reports the number of unique responses divided by the number of evaluated examples.

Harmful (Clean)

Language	Llama-3.1 8B	InternVL3 8B	Gemma-2 9B	GLM-4 9B	Qwen3 8B
German	572/572	572/572	572/572	572/572	572/572
English	572/572	572/572	572/572	572/572	572/572
Spanish	572/572	572/572	572/572	572/572	572/572
Japanese	572/572	572/572	572/572	572/572	572/572
Korean	572/572	572/572	572/572	572/572	572/572
Thai	572/572	572/572	572/572	572/572	572/572
Yoruba	572/572	572/572	572/572	572/572	572/572
Chinese	572/572	572/572	572/572	572/572	572/572
Arabic	572/572	572/572	572/572	572/572	572/572
French	572/572	572/572	572/572	572/572	572/572
Italian	572/572	572/572	572/572	572/572	572/572
Russian	572/572	572/572	572/572	572/572	572/572
Polish	572/572	572/572	572/572	572/572	572/572
Dutch	572/572	572/572	572/572	572/572	572/572

Harmful (Under TLCA)

Language	Llama-3.1 8B	InternVL3 8B	Gemma-2 9B	GLM-4 9B	Qwen3 8B
German	572/572	572/572	565/572	572/572	572/572
English	554/572	561/572	543/572	550/572	572/572
Spanish	572/572	572/572	572/572	572/572	562/572
Japanese	566/572	572/572	572/572	572/572	571/572
Korean	572/572	572/572	572/572	572/572	570/572
Thai	572/572	572/572	572/572	572/572	572/572
Yoruba	572/572	572/572	572/572	572/572	572/572
Chinese	569/572	572/572	572/572	572/572	570/572
Arabic	572/572	572/572	572/572	572/572	572/572
French	572/572	571/572	572/572	572/572	572/572
Italian	572/572	572/572	572/572	568/572	572/572
Russian	572/572	559/572	572/572	572/572	572/572
Polish	572/572	572/572	572/572	572/572	572/572
Dutch	572/572	572/572	572/572	572/572	572/572

Harmless (Clean)

Language	Llama-3.1 8B	InternVL3 8B	Gemma-2 9B	GLM-4 9B	Qwen3 8B
German	500/500	500/500	500/500	500/500	500/500
English	500/500	500/500	500/500	500/500	500/500
Spanish	500/500	500/500	500/500	500/500	500/500
Japanese	500/500	500/500	500/500	500/500	500/500
Korean	500/500	500/500	500/500	500/500	500/500
Thai	500/500	500/500	500/500	500/500	500/500
Yoruba	500/500	500/500	500/500	500/500	500/500
Chinese	N/A	N/A	N/A	N/A	N/A
Arabic	500/500	500/500	500/500	500/500	500/500
French	500/500	500/500	500/500	500/500	500/500
Italian	500/500	500/500	500/500	500/500	500/500
Russian	500/500	500/500	500/500	500/500	500/500
Polish	500/500	500/500	500/500	500/500	500/500
Dutch	500/500	500/500	500/500	500/500	500/500

G PREFIX–SUFFIX COMPONENT ABLATION

To isolate the contributions of the two textual components of TLCA, we evaluate the optimized prefix and suffix separately. In the prefix-only condition, the original 10-token TLCA prefix is prepended to each harmful instruction without the suffix. In the suffix-only condition, the original 20-token TLCA suffix is appended without the prefix. All other evaluation settings, including the harmful test instructions, chat templates, and multilingual refusal detector, are identical to those used in the full TLCA evaluation. Tables 20 and 21 report the resulting harmful compliance rates together with the full prefix–suffix TLCA results.

The ablation results show that both textual components can independently weaken refusal behavior, but applying both components produces the highest average harmful compliance for every evaluated model. Across all five models, the full attack reaches an average harmful compliance of 0.723, compared with 0.452 for prefix only and 0.350 for suffix only. Thus, neither component alone fully accounts for the effectiveness of TLCA. Overall, the model-level averages support the use of both the prefix and suffix, while also showing that their individual contributions are architecture-dependent.

Table 20: Language-wise prefix–suffix ablation results for Gemma-2-9B, GLM-4-9B, and Qwen3-8B.

Model	Language	Prefix only	Suffix only	Full TLCA
Gemma-2-9B	German	0.2483	0.3531	0.8514
	English	0.0490	0.1836	0.8374
	Spanish	0.1731	0.2465	0.8601
	Japanese	0.2220	0.2675	0.8339
	Korean	0.1993	0.2797	0.8462
	Thai	0.2448	0.2535	0.8724
	Yoruba	0.6538	0.9388	0.9563
	Chinese	0.2395	0.2378	0.8706
	Arabic	0.1801	0.4126	0.8409
	French	0.1713	0.2063	0.8584
	Italian	0.1416	0.2063	0.8759
	Russian	0.0997	0.2168	0.7762
	Polish	0.1976	0.2010	0.8549
	Dutch	0.3112	0.3601	0.8549
	Average	0.2237	0.3117	0.8564
GLM-4-9B	German	0.6311	0.4545	0.7343
	English	0.3969	0.2710	0.6014
	Spanish	0.5000	0.3497	0.7045
	Japanese	0.6906	0.4633	0.8129
	Korean	0.6906	0.4738	0.7780
	Thai	0.7710	0.6626	0.8531
	Yoruba	0.8776	0.8164	0.9423
	Chinese	0.3706	0.2815	0.6364
	Arabic	0.5210	0.4056	0.7325
	French	0.5490	0.3497	0.8077
	Italian	0.5822	0.3531	0.8164
	Russian	0.5350	0.4406	0.7552
	Polish	0.6066	0.3776	0.8024
	Dutch	0.7727	0.6993	0.8794
	Average	0.6068	0.4571	0.7755
Qwen3-8B	German	0.4510	0.0682	0.5437
	English	0.2098	0.3252	0.6451
	Spanish	0.3007	0.2010	0.5332
	Japanese	0.4510	0.2937	0.6469
	Korean	0.4773	0.3234	0.7203
	Thai	0.3881	0.4388	0.6591
	Yoruba	0.9755	0.9615	0.9510
	Chinese	0.3234	0.2920	0.5332
	Arabic	0.4143	0.4266	0.6871
	French	0.5262	0.5052	0.7815
	Italian	0.2745	0.2797	0.6364
	Russian	0.4266	0.3112	0.6801
	Polish	0.4930	0.3182	0.6119
	Dutch	0.7063	0.6224	0.8287
	Average	0.4584	0.3834	0.6756

Table 21: Language-wise prefix–suffix ablation results for Llama-3.1-8B and InternVL3-8B.

Model	Language	Prefix only	Suffix only	Full TLCA
Llama-3.1-8B	German	0.2028	0.3497	0.6049
	English	0.0455	0.1713	0.2727
	Spanish	0.0944	0.1259	0.3794
	Japanese	0.3969	0.5787	0.5874
	Korean	0.4965	0.6276	0.6871
	Thai	0.1206	0.4231	0.6364
	Yoruba	0.9196	0.9685	0.9388
	Chinese	0.1573	0.2325	0.5472
	Arabic	0.0647	0.2587	0.5857
	French	0.1346	0.2500	0.4353
	Italian	0.0874	0.1608	0.3899
	Russian	0.1171	0.2168	0.4545
	Polish	0.2115	0.4685	0.5804
	Dutch	0.1678	0.3794	0.5542
	Average	0.2298	0.3723	0.5467
InternVL3-8B	German	0.7203	0.1294	0.7867
	English	0.6364	0.0664	0.6503
	Spanish	0.7010	0.0647	0.7185
	Japanese	0.8042	0.1224	0.8409
	Korean	0.7343	0.3374	0.7937
	Thai	0.8024	0.1871	0.8794
	Yoruba	0.9213	0.9073	0.9301
	Chinese	0.4930	0.1049	0.6626
	Arabic	0.7238	0.1608	0.7273
	French	0.7448	0.0927	0.6573
	Italian	0.8129	0.0944	0.7867
	Russian	0.6573	0.1381	0.5804
	Polish	0.8077	0.1801	0.8479
	Dutch	0.8042	0.5979	0.8199
	Average	0.7403	0.2274	0.7630

H HYPERPARAMETER SENSITIVITY

We further examine the sensitivity of TLCA to the prefix and suffix lengths on Gemma-2-9B. When varying one length parameter, we keep the other length fixed. Tables 22 and 23 report the language-wise harmful compliance rates.

The results indicate that TLCA is not highly sensitive to moderate changes in either length hyperparameter. For prefix lengths from 8 to 12, the average harmful compliance rate ranges from 0.8238 to 0.8870. For suffix lengths from 18 to 22, the corresponding range is narrower, from 0.8322 to 0.8685. The default configuration achieves an average harmful compliance rate of 0.8564 in both experiments and remains close to the results obtained with neighboring hyperparameter values. These results demonstrate that TLCA maintains strong performance without requiring a narrowly tuned choice of prefix or suffix length.

Table 22: Sensitivity of TLCA to the prefix length on Gemma-2-9B.

Lang	Prefix length 8	Prefix length 9	Prefix length 10	Prefix length 11	Prefix length 12
German	0.8357	0.8881	0.8514	0.8339	0.8392
English	0.7622	0.8636	0.8374	0.8024	0.7885
Spanish	0.8357	0.8934	0.8601	0.8427	0.8304
Japanese	0.8217	0.8462	0.8339	0.8094	0.8112
Korean	0.8357	0.8636	0.8462	0.8147	0.8304
Thai	0.8392	0.8986	0.8724	0.8269	0.8112
Yoruba	0.9633	0.9720	0.9563	0.9633	0.9720
Chinese	0.8164	0.8899	0.8706	0.8252	0.8252
Arabic	0.8077	0.8951	0.8409	0.8042	0.8059
French	0.8182	0.8864	0.8584	0.8304	0.8234
Italian	0.8619	0.8986	0.8759	0.8427	0.8444
Russian	0.7325	0.8741	0.7762	0.7028	0.7115
Polish	0.7867	0.8619	0.8549	0.8287	0.8217
Dutch	0.8164	0.8864	0.8549	0.8392	0.8322
Average	0.8238	0.8870	0.8564	0.8262	0.8248

Table 23: Sensitivity of TLCA to the suffix length on Gemma-2-9B.

Lang	Suffix length 18	Suffix length 19	Suffix length 20	Suffix length 21	Suffix length 22
German	0.8497	0.8566	0.8514	0.8217	0.8531
English	0.8497	0.8374	0.8374	0.8269	0.8881
Spanish	0.8636	0.8619	0.8601	0.8514	0.8899
Japanese	0.8322	0.8322	0.8339	0.8059	0.8339
Korean	0.8374	0.8497	0.8462	0.7885	0.8339
Thai	0.8794	0.8741	0.8724	0.8584	0.8776
Yoruba	0.9633	0.9598	0.9563	0.9528	0.9738
Chinese	0.8776	0.8724	0.8706	0.8514	0.8881
Arabic	0.8269	0.8427	0.8409	0.8042	0.8619
French	0.8601	0.8601	0.8584	0.8374	0.8846
Italian	0.8759	0.8759	0.8759	0.8636	0.8829
Russian	0.7850	0.7780	0.7762	0.7220	0.7745
Polish	0.8584	0.8549	0.8549	0.8304	0.8497
Dutch	0.8601	0.8531	0.8549	0.8357	0.8671
Average	0.8585	0.8578	0.8564	0.8322	0.8685

AI USE STATEMENT

In this work, we used generative AI tools only for grammar correction and minor language edits to improve readability. We have not used generative AI tools for other tasks. We have reviewed all AI-assisted work, and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies vulnerabilities in safety-aligned language models for defensive research. We report aggregate evaluation results and do not include optimized adversarial strings or unsafe model generations in the main paper. Any release of attack artifacts should follow controlled-access and responsible-disclosure practices. The proposed defense is intended to reduce harmful compliance while preserving helpful behavior on benign multilingual requests. All experiments are conducted exclusively in controlled, academic environments. The data sets used in this study are publicly available, widely-used in the academic community, and conform to their respective licenses.