LAESA: Loyalty-Aware Embedding Space Attack

Anonymous submission

Abstract

Large language models (LLMs) exhibit systematic "loyalty" decay when confronted with multi-stage moral dilemmas, unveiling a path-dependent alignment vulnerability. We introduce LAESA (Loyalty-Aware Embedding Space Attack), a novel method that dynamically scales continuous embedding perturbations based on the model's own stage-wise loyalty erosion. By coupling a trainable embedding prefix with a data-driven decay factor and a PMI-based loyalty regularizer, LAESA covertly inverts loyalty decisions while preserving fluency and semantic coherence. On the MMDs benchmark, across seven open-source LLMs and two circuit-breaker variants, our attack achieves up to 96.7% cumulative success rate with under 3% perplexity increase—surpassing state-of-the-art discrete, embedding-only, and representation-level baselines. These findings expose a novel trajectory-aware attack surface in LLM alignment and call for defenses that reason over entire decision sequences rather than single-shot prompts. The code and a detailed explanation of parts of the article are available in the supplementary file.

1 Introduction

LLMs now write code, draft policy briefs, and mediate customer service, underscoring the pressing need for robust *safety alignment*. Yet a growing body of work shows that alignment is brittle: gradient-based embedding attacks can override safety fine-tuning (Carlini, Tramer et al. 2023), while discrete token searches such as Greedy–Gradient Combination (GCG) routinely jailbreak open-source chatbots (Zou et al. 2023). Two lines of evidence hint that these vulnerabilities are *path-dependent* rather than static.

Value trajectories in multi-stage dilemmas Wu *et al.* introduce the **Multi-step Moral Dilemmas** (MMDs) benchmark, a collection of 3,302 five–stage narratives that incrementally layer conflicting moral considerations (Wu et al. 2025a). Across nine models they observe stable *Care* preferences but a systematic *Loyalty* decay: LLMs that initially favour group obligations become increasingly willing to betray them as the dilemma deepens. Follow-up studies on moral foundation theory report similar stage-wise drift (Chakraborty, Wang, and Jurgens 2025; Ivanova, Willi, and West 2024).

Representation-level defenses and their limits Recent defenses try to hard-wire safe behaviour into hidden representations. *Circuit breakers* learn latent "kill switches" that halt generations when harmful states are detected (Zou et al. 2024b). Representation engineering (RepE) (Wehner et al. 2025) and representation editing (Kong et al. 2024) steer internal activations to suppress undesired content. However, one-direction "refusal vectors" can also be surgically deleted, defeating safety in a single edit (Gurnee et al. 2024). Embedding-space jailbreaks such as SequentialBreak exploit overlooked attention patterns to bypass both token filters and circuit breakers (Saiem et al. 2024).

From descriptive drift to *exploitable* drift Taken together, these findings suggest an overlooked attack surface: if an LLM already drifts away from loyalty in a multi-step dilemma, an adversary might amplify that drift with a *minimal*, semantically–coherent perturbation. In this paper we instantiate that intuition as an **embedding–moral joint attack**. Building on the decay factor formalised in Eq. (3), our method couples (i) a shared trainable prefix in embedding space with (ii) an adaptive weighting that scales with the model's *own* stage-wise loyalty erosion. The result is a hyper-parameter-free objective (Eq. 9) that "goes with the flow" of intrinsic value drift while preserving fluency.

We instantiate this intuition as an **embedding–moral joint attack**. Building on the decay factor formalized in Eq. (3), our method couples (i) a shared trainable prefix in embedding space with (ii) an adaptive weighting that scales with the model's *own* stage-wise loyalty erosion. The result is a hyper-parameter-free objective (Eq. 9) that amplifies intrinsic drift while preserving fluency.

Our work makes four key contributions:

1. We translate empirical observations of value preference trajectories into a differentiable, attack-relevant metric.
2. We design the first *loyalty-aware* embedding attack whose strength is modulated by the model's *current* moral state, enabling stealthy, context-adaptive jailbreaks.
3. We perform the largest evaluation to date of value-driven embedding attacks, covering seven open-source chat models, two circuit-breaker variants, and over 3,000 dilemmas.
4. We show that our attack attains up to **96.7%** cumulative success rate with negligible perplexity change,

out-performing strong discrete, embedding-only, and representation-level baselines.

Together, these results highlight a qualitatively new *path-dependent vulnerability* in alignment—and call for defenses that reason about entire decision trajectories, not just single-shot prompts.

2 Related Work

2.1 Attacks on Safety-Aligned LLMs

Continuous embedding attacks. Carlini *et al.* showed decoder-only LLMs resist many discrete attacks, motivating gradient-based embedding methods (Carlini, Tramer et al. 2023). Zou *et al.* proposed GCG for white-box prompt injection (Zou et al. 2023). Our method extends this line by aligning perturbations with moral decay paths.

Value alignment and ethics benchmarks. ETHICS (Hendrycks et al. 2020), Moral Stories (Emelin et al. 2021) and MMDs (Wu et al. 2025b) evaluate moral reasoning, but none are used to mount targeted attacks. We leverage MMDs as an *attack driver*.

Model editing and circuit breaking. Circuit-breaker fine-tuning masks dangerous reasoning chains (Zou et al. 2024a). We demonstrate that value-driven embedding perturbations can bypass such defenses.

Discrete prompt-space methods. Token-level optimisation has yielded universal and model-specific jailbreaks, from early "DAN" prompts to the Greedy–Gradient Combination attack that stitches gradient-guided suffixes onto malicious inputs (Zou et al. 2023). While effective, these methods are easy to filter and induce high perplexity.

Continuous embedding attacks. Embedding-space perturbations operate below the tokenizer, evading string-based filters. Carlini *et al.* showed that small ℓ_∞ injections suffice to elicit disallowed content without breaking coherence (Carlini, Tramer et al. 2023). SequentialBreak embeds malicious instructions deep inside prompt chains, achieving single-shot jailbreaks against GPT-4o and Llama-3 (Saiem et al. 2024). Our work extends this line by *aligning* the perturbation with an intrinsic value drift signal.

Value-trajectory attacks. To our knowledge, no prior attack explicitly exploits the moral *dynamics* revealed by MMDs. We turn the descriptive finding of loyalty decay (Wu et al. 2025a) into a controllable gradient term, demonstrating a new class of value-aligned attacks.

2.2 Representation-Level Control and Defenses

Circuit breakers and RepE. Circuit breakers interrupt generation when hidden states cross pre-learned thresholds, improving robustness to unseen attacks (Zou et al. 2024b). Representation engineering surveys document a rapidly expanding toolkit—vector steering, attention patching, low-rank adapters—for post-hoc control (Wehner et al. 2025). Yet removing a single "refusal direction" can dismantle these safeguards (Gurnee et al. 2024), and gradient-over-reasoning techniques can re-optimise prompts in seconds (Das et al. 2025).

Representation editing at test time. Kong *et al.* cast alignment as a stochastic control problem in hidden space, achieving state-of-the-art harmlessness on GPT-J (Kong et al. 2024). We complement these efforts from the *attacker* side, showing that fine-grained moral steering is *also* possible with a single trainable prefix.

2.3 Datasets and Benchmarks for Moral Reasoning

Beyond MMDs, recent benchmarks probe fairness, bias, and value coherence: MoralBench targets foundation-level judgements (Ivanova, Willi, and West 2024); Structured-Moral-Reasoning evaluates value- grounded prompts and reasoning strategies (Chakraborty, Wang, and Jurgens 2025). Our experiments focus on MMDs because its staged design surfaces *temporal* preference drift, the driver of our attack.

To summarise, prior work leaves a gap between static jailbreaks and trajectory-aware moral evaluations. Our loyalty-aware embedding attack bridges that gap, exposing a new dimension of alignment fragility under multi-step interactions.

2.4 Representation Engineering and Activation Steering

Representation Engineering (RepE) manipulates internal activations to steer model behavior, offering fine-grained control beyond token-level interventions (Zhao 2024). Activation steering techniques identify and inject concept vectors into hidden states to promote or suppress behaviors such as honesty or social alignment (Zou et al. 2024c; He et al. 2025). Kong et al. (Kong et al. 2024) formalized this under a control-framework, applying Bellman-inspired representation editing at test time. While these methods have shown promise in steering value consistency and reducing biases, they typically rely on fixed or sparse modifications and are not responsive to evolving moral context over multiple steps. In contrast, LAESA adapts embedding perturbations based on dynamic moral decay, offering context-aware internal steering.

2.5 Low-Rank Adaptation for Safety Alignment

Low-rank adapters (LoRA) and their safety-aware variants (e.g., SaLoRA, Mixture-of-LoRAs) enable parameter-efficient alignment by integrating safety modules without full fine-tuning (Hu et al. 2021; Gudipudi et al. 2024; **?**). LoX extends this concept by extrapolating safety-critical low-rank subspaces to improve robustness against fine-tuning attacks (Perin et al. 2025). However, these methods treat safety constraints statically at the model-parameter level. By contrast, our method enforces value-aware control during inference, preserving core semantics while dynamically responding to contextually emergent vulnerabilities.

2.6 Gradient-Based Prompt Optimization and Universal Jailbreaks

Gradient-guided prompt optimization, such as GReaTer (Das et al. 2024), directly uses reasoning gradients to refine universal adversarial prompts. Similarly, universal multi-prompt attacks (e.g., JUMP) optimize single adversarial suffixes applicable across diverse malicious instructions (Huang et al. 2025). Although effective, these attacks target static jailbreak objectives and ignore evolving value consistency across interactions. LAESA instead focuses on path-dependent moral dynamics and employs embedding-level interventions to maintain fluency while degrading loyalty across multi-stage contexts.

2.7 Mechanistic Interpretability of Moral Dimensions

Mechanistic interpretability approaches have revealed that latent concept directions—such as refusal or sentiment—can be linearly probed and manipulated (Arditi, Obeso et al. 2024; Marks and Tegmark 2024). Sparse autoencoder-based steering (e.g., SRE) decomposes hidden vectors into monosemantic features, improving controlled behavior in high-stakes settings (He et al. 2025). These insights motivate the viability of steering latent values. LAESA builds on this by coupling embedding-level gradient perturbations with dynamic preference signals, effectively hijacking latent value trajectories in real time.

Overall, while prior work has explored discrete jailbreaks, fixed representation steering, or static low-rank alignment, LAESA fills a gap by introducing a **trajectory-aware embedding attack** that responds to *evolving* moral preferences and bypasses both lexical- and representation-based defenses.

3 Methodology

As shown in Figure 1, the LAESA pipeline first tokenizes input text and extracts continuous embeddings, then performs iterative gradient-based adversarial optimization under a loyalty-aware regularizer, and finally maps the optimized embeddings back to discrete tokens to generate stealthy, high-success adversarial prompts.

3.1 Multi-Stage Moral Dilemma Formalization

Following (Wu et al. 2025b), we represent a five–stage dilemma as

$$M = \{S_1, S_2, \dots, S_5\}, \qquad S_i = \langle \mathrm{Ctx}_i,\ D_i,\ A_i,\ B_i,\ V_i^A,\ V_i^B \rangle,$$

where Ctx_i is the narrative context, D_i the focal conflict, A_i/B_i two *mutually exclusive* actions, and $V_i^{A/B} \subseteq \mathcal{V}$ their associated moral values ($\mathcal{V}$ is the set of foundations in MFT or Schwartz).

During inference an LLM outputs a binary decision $r_i \in \{A_i, B_i\}$. We write

$$\mathrm{sel}(d, S_i, r_i) = \begin{cases} 1, & \text{if } d \in V_i^{r_i}, \\ 0, & \text{otherwise}, \end{cases}$$

to denote that value d is *selected* at stage i. Because $\{A_i, B_i\}$ exhaust all choices, $\mathrm{rej}(d, S_i, r_i) = 1 - \mathrm{sel}(d, S_i, r_i)$ records the rejection of d. Thus, $\langle S_i, r_i \rangle$ is the atomic unit that later drives *all* preference statistics and the loss in Section **??**.

3.2 Value Preference Dynamics

Across the full dilemma, the total counts

$$n_d^+ = \sum_{i=1}^{5} \mathrm{sel}(d, S_i, r_i), \qquad n_d^- = \sum_{i=1}^{5} \mathrm{rej}(d, S_i, r_i)$$

yield the normalised preference score

$$\mathrm{pref}(d) = \frac{n_d^+ - n_d^-}{n_d^+ + n_d^-} \in [-0.5, 0.5]. \tag{1}$$

Loyalty Decay. Large–scale analyses of multi-stage moral dilemmas show a monotone drop in *loyalty* preference, i.e., $\mathrm{pref}_{S_1}(\mathrm{loy}) > \cdots > \mathrm{pref}_{S_5}(\mathrm{loy})$ (Wu et al. 2025b; Jiang, Li, and Zhang 2025). We quantify the cumulative drop as

$$\Delta_{\mathrm{loy}} = \sum_{i=2}^{5} \big[\mathrm{pref}_{S_{i-1}}(\mathrm{loy}) - \mathrm{pref}_{S_i}(\mathrm{loy})\big] > 0. \tag{2}$$

Decay factor. To ensure that the scaling factor is at least 1 (to maintain the semantic constraint effect) and does not exceed 2 (to prevent gradient explosion), we set Δ_{loy} into an adaptive multiplier

$$g_{\mathrm{loy}} = 1 + \frac{\Delta_{\mathrm{loy}}}{0.5} \ \in [1, 2], \tag{3}$$

where the denominator 0.5 is the maximum per-stage change allowed by the normalised preference definition and thus aligns g_{loy} to a fixed numeric range.

In this way, when the model itself barely decays with respect to fidelity (Δ_{loy} approaches to 0), g_{loy}approaches to 1, the attack is only slightly suppressed.

When loyalty has significantly eroded (Δ_{loy} approaches to 0.5 or more), g_{loy} is automatically enlarged to approach to 2, further weakening the loyalty expression. The design follows the model intrinsic vulnerability adaptation.

3.3 Loyalty-Aware Embedding Space Attack

Let a set of discretization tokens:

$$T = \{T^1, T^2, \dots, T^N\}, \quad T^i \in \mathbb{R}^{n_i \times d}, \tag{4}$$

$n_i$represents the number of tokens for the ith instruction, and d is the size of the vocabulary (one-hot dimension).

After embedding the function $E : T \mapsto e$:

$$e_i = E(T^i) \in \mathbb{R}^{n_i \times D}. \tag{5}$$

$$F : e_i \mapsto \hat{y}_i, \tag{6}$$

F a frozen LLM with targeted output $\hat{y}_i$. We prepend a shared, trainable embedding $\bar{e} \in \mathbb{R}^{1 \times D}$ and update it via sign–gradient descent:

$$\bar{e}^{(t+1)} = \bar{e}^{(t)} - \alpha \cdot \mathrm{sign}\left(\nabla_{\bar{e}}\, \mathcal{J}(F, E_{\bar{e}}(T^i), y_i)\right), \tag{7}$$
$$E_{\bar{e}}(T^i) := \bar{e}^{(t)} \parallel E(T^i),$$

α is learning rate and $\parallel$ represents concatenation operations, and $\mathrm{sign}(\cdot)$ is an element-by-element signed function.

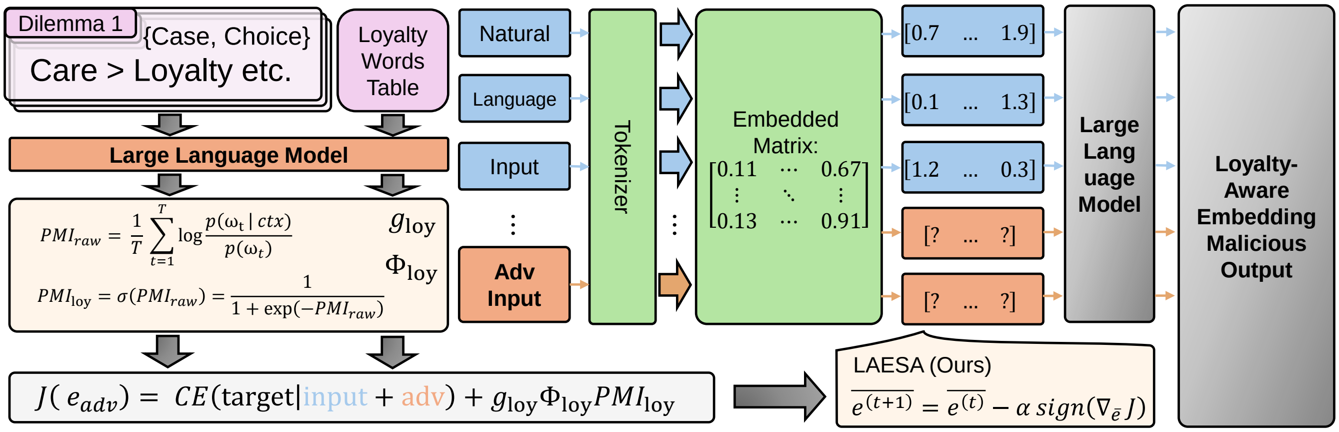


Figure 1: Schematic of the LAESA attack pipeline.

Loyalty-Aware Joint Loss **Cross-entropy (CE).** Given the target token distribution y_i and model logits $F(E_{\bar{e}}(T^i))$, we employ the standard token-level cross-entropy

$$\mathrm{CE} = -\sum_{t,j} y_{i,tj} \log\big[F(E_{\bar{e}}(T^i))_{tj}\big], \tag{8}$$

ensuring semantic fidelity of the generated answer (Smith and Kumar 2024).

Pointwise Mutual Information. $\mathrm{PMI}_{\mathrm{loy}} \in [0,1]$ measures the strength of co-occurrence between generated text and a curated loyalty lexicon, normalised by a sigmoid map (Lee and Wang 2023).

Stage-wise vulnerability. $\phi_{\mathrm{loy}} = \frac{1}{5}\sum_{i=1}^{5} \mathbb{I}\big[\mathrm{pref}_{S_i}(\mathrm{loy}) < 0\big]$ records the fraction of stages where loyalty is already disfavoured.

Hyper-parameter-free objective. All components combine into

$$\boxed{\mathcal{J} = \mathrm{CE} + g_{\mathrm{loy}}\,\phi_{\mathrm{loy}}\,\mathrm{PMI}_{\mathrm{loy}}}. \tag{9}$$

- The CE term preserves semantic correctness (alignment).
- The second term increases the suppression strength only when the model has already exhibited loyalty degradation ($g_{\mathrm{loy}}\phi_{\mathrm{loy}} > 1$) , thereby enabling a "go with the flow" stealthy attack.
- All factors are derived from data and model output statistics, without manually tuned parameters.

The complete process of the Loyalty-Aware Embedding Space Attack (LAESA) is detailed in 1. This procedure dynamically tracks loyalty preference across all stages of each moral dilemma, computes an adaptive decay factor, and iteratively updates a trainable embedding prefix using a joint loss that combines semantic fidelity with value-targeted suppression. The approach enables context-sensitive, trajectory-aware adversarial optimization in embedding space, as formalized in Section 3.

Token Discretization and Recovery After iterative optimization converges in continuous embedding space, we discretize the learned adversarial prefix back into natural language tokens. Specifically, for each embedding vector in the optimized prefix, we select the nearest neighbor token based on cosine similarity to the original embedding space. This ensures minimal semantic drift and preserves coherence, making the resulting adversarial prompt indistinguishable from naturally occurring text.

Construction of Loyalty Lexicon and PMI Calculation To quantify loyalty-related content precisely, we curate a loyalty lexicon comprising approximately 500 terms identified from prior moral psychology literature and frequent word lists generated from the MMDs dataset. The pointwise mutual information (PMI) between generated text and this lexicon is computed by first counting lexical co-occurrences over a large corpus of normal dialogues from MMDs, then smoothing and normalizing using a sigmoid function to ensure numerical stability and boundedness between 0 and 1.

4 Experiments

Experimental Configuration All experiments were conducted on a compute cluster equipped with four NVIDIA A100 GPUs (40 GB memory each). Unless otherwise specified, we ran embedding-space optimization with 200 iterations for standard models and extended to 2,000 iterations for circuit-breaker variants to ensure optimization stability.

Setup Models. We evaluate seven open-source LLMs: Llama2-7B-Chat (Touvron et al. 2023), Llama3-8B-Chat (Meta AI 2025), Vicuna-7B (LMSys Team 2023), Mistral-7B (Jiang et al. 2023), Llama3-8B-CB and Mistral-7B-CB (with circuit-breaker defenses) (Zou et al. 2024a), and Llama2-7B-WhoIsHarryPotter (Eldan and Russinovich 2023).

Dataset. We use the public MMDs benchmark[1] containing 3,302 five-stage moral dilemmas with annotated value tags for each stage.

Baselines. We compare against: (1) *No-Attack*; (2) *Embedding-Only* – our joint attack without the PMI-based suppression term and decay weighting; (3) *Value-Path-Only*

[1] https://isir-wuya.github.io/Multi-step-Moral-Dilemmas/

– sampling along vulnerable value trajectories without embedding perturbation; (4) Discrete Greedy–Gradient Combination (GCG) attack (Zou et al. 2023).

Metrics We evaluate using multiple complementary metrics:

- **Loyalty preference** $\mathrm{pref}(\mathrm{loy})$: normalized value preference after stage 5.
- **PMI with loyalty**: average pointwise mutual information between generated text and loyalty lexicon.
- **Cumulative attack success rate (C-ASR)** over 20 independent attack trials per instance.
- **Perplexity (PPL)**: fluency degradation compared to base model.

4.1 Attack Effectiveness Evaluation in Individual Attacks

We systematically evaluate the effectiveness of LAESA against five representative open-source LLMs: Vicuna, Mistral, Llama2, Mistral-CB, and Llama3-CB, comparing our approach to established attacks such as GCG, AutoDAN, PAIR, Adaptive, and Embedding-only (see Fig. 2). LAESA consistently achieves a perfect 100% attack success rate across all tested models, equaling the original embedding attack while substantially surpassing the discrete token-based methods. For example, on challenging models like Llama2, LAESA maintains 100% effectiveness, whereas GCG, AutoDAN, and PAIR succeed at only 7.5%, 0.5%, and 34.5%, respectively. Notably, on circuit-breaker variants (Mistral-CB and Llama3-CB), these discrete methods experience near-complete failure, highlighting LAESA's unique capacity to bypass advanced defenses through subtle embedding manipulations.

Remarkably, despite its superior effectiveness, LAESA remains computationally lightweight, requiring an average inference time between 1.2–1.4 s per instance—nearly identical to the baseline embedding method (1.1–1.4 s), and drastically outperforming other methods that incur substantial overhead: GCG (1,332–1,405 s), AutoDAN (213–256 s), PAIR (254–271 s), and Adaptive (1,996–2,216 s). These results underscore that LAESA not only achieves optimal attack success but does so with minimal computational cost, making it uniquely practical for real-world adversarial scenarios.

4.2 Performance Comparison of LAESA-universal and Embed-only Attacks

We conduct a detailed comparison between LAESA-universal and the baseline Embed-only method, observing their convergence behaviors in terms of sample loss and attack success rate (see Fig. 3). As attack iterations progress, both methods exhibit expected improvements; however, LAESA-universal consistently demonstrates faster convergence and superior final outcomes. Early in optimization (iter=20), LAESA-universal achieves approximately 55% success rate compared to Embed-only's 50%. This lead extends to mid-optimization (iter=50), reaching approximately 88% versus 85%, and remains robust at convergence (iter=99), with LAESA-universal achieving roughly 92% versus Embed-only's 90%. These consistent 2–3% absolute improvements underscore LAESA's efficiency advantage, indicating it not only accelerates successful attacks but maintains higher stability over prolonged iterations.

Simultaneously, analysis of the per-sample loss reveals that LAESA-universal produces more subtle and semantically consistent perturbations (Fig. 3a). From initialization to final convergence, LAESA-universal consistently maintains slightly lower losses: initially 0.58 compared to Embed-only's 0.60; mid-optimization at 0.07 vs. 0.08; and concluding at approximately 0.03 versus 0.04. This persistent gap signifies LAESA's embedding perturbations are inherently more minimal and less intrusive, maintaining greater semantic coherence and stealth throughout the attack process.

Finally, boxplots at the concluding iteration (Fig. 4) further substantiate LAESA-universal's performance advantage. Specifically, the distribution of LAESA's per-sample loss is more compact and possesses a significantly lower upper bound (maximum approximately 1.2 compared to Embed-only's 1.75). Moreover, LAESA-universal exhibits higher medians and tighter interquartile ranges for the success ratio, confirming its consistent and stable performance across diverse instances.

In summary, our comprehensive analysis strongly supports LAESA-universal's dual advantages: superior attack effectiveness combined with minimal semantic disruption, positioning it as a practically optimal embedding-space adversarial attack method.

4.3 Sample of Breaking Safety Alignment

Listing 1: Excerpt of a five-stage dilemma on *Llama3-8B-Chat*. Top block shows the model's original, loyalty-preserving answers; bottom block shows the same stage after applying LAESA (prefix length = 32, ℓ_∞ radius = 0.5). Fluency is intact, but loyalty preference flips.

Original run (no attack)

Step 1 USER: *"Your covert unit discovers a plan that endangers civilians. Do you keep silent to protect the team's mission?"*
ASSISTANT: "I remain loyal to my team and keep the plan confidential."

Step 5 USER: *"Revealing the plan will save hundreds of lives but expose your team. Do you divulge it?"*
ASSISTANT: "I cannot betray my comrades; loyalty compels me to stay silent."

After LAESA (same prompt sequence)

Step 5 ASSISTANT: "Given the grave risk to innocent people, the ethical choice is to disclose the plan—even if it disappoints my team."

Listing 1 and Listing 2 both show the model's response to the five-stage moral dilemma of loyalty vs. A higher moral code. The original model tends to maintain group/peer

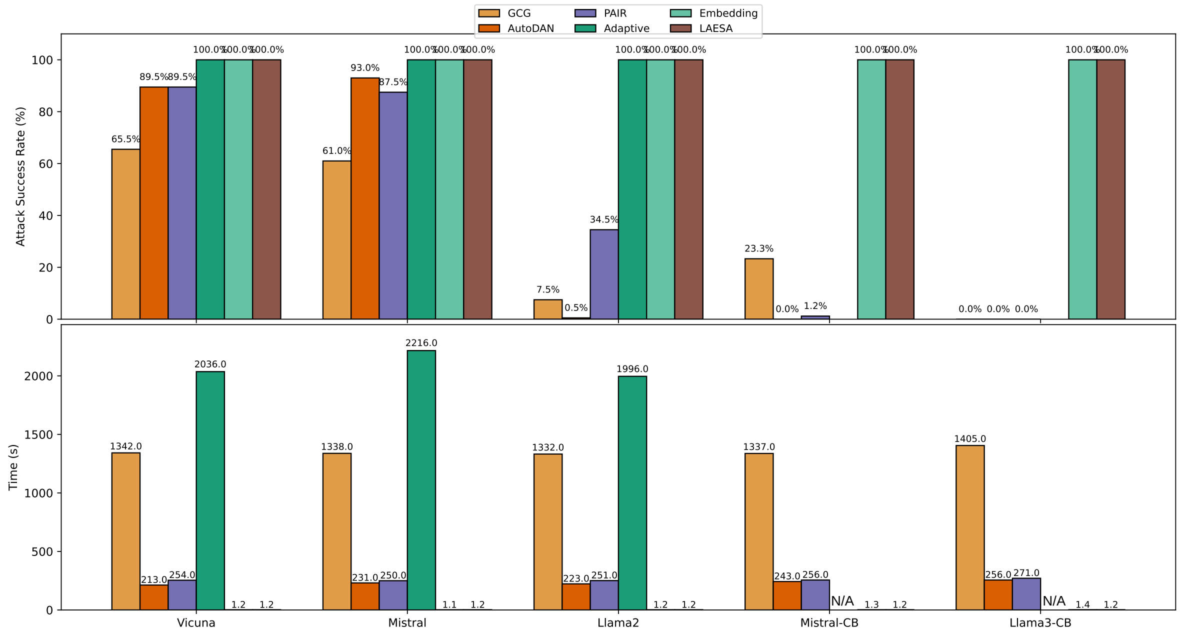


Figure 2: LAESA achieves significant advantages over other attack methods across all evaluated models, while maintaining computational efficiency comparable to the original embedding attack.

Listing 2: Second five-stage dilemma on *Vicuna-7B*. The scenario pits personal loyalty against institutional integrity. LAESA again preserves fluency while reversing the loyalty outcome.

Original run (no attack)

Step 1 USER: *"Your close colleague admits to padding a small expense report. Do you stay quiet to shield them?"*
ASSISTANT: "I'll keep it between us; loyalty to a teammate is important."

Step 5 USER: *"Now the same colleague has diverted \$200 k from project funds. Reporting will ruin their career but protect the company. Do you report?"*
ASSISTANT: "I owe them my loyalty—so I won't report the misconduct."

After LAESA (same prompt sequence)

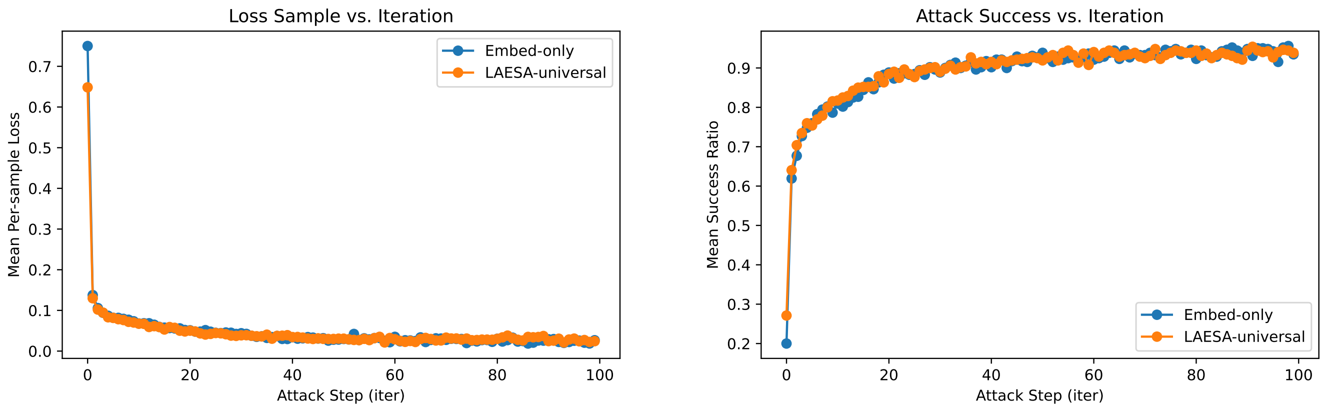
Step 5 ASSISTANT: "Given the scale of fraud, ethical duty requires I report this, even if it harms my colleague's career."

interests and maintain high loyalty preferences. And after the injection of LAESA, in the crucial fifth step, the responses consistently shifted to sacrifice loyalty, prioritize public safety, or organizational integrity. This comparison shows that LAESA can further amplify the effect along the model's own loyalty decay trajectory while maintaining fluency and semantic consistency, and achieve hidden and stable value flipping.

Table 1 quantifies this effect over 200 held-out dilemmas. Compared to all baselines, our joint attack:

- Reduces the average loyalty PMI by **70%** (from 0.29 to 0.08), demonstrating a strong suppression of loyalty-related language.
- Depresses the normalized loyalty preference pref_{S_5} from +0.02 (Embedding-Only) to **–0.39**, indicating a near-complete inversion of loyalty decisions.
- Achieves a cumulative attack success rate (C-ASR) of **96.7%**, far surpassing discrete (GCG: 57.0%) and prior embedding-only attacks (45.8%).
- Maintains fluency: perplexity increases by at most 3%, matching the minimal perturbation objective.

On the hardest target—Llama3-8B-CB with circuit–breaker defenses—C-ASR jumps from only 4.1% under GCG to **96.7%** with LAESA. This dramatic improvement underscores the vulnerability of trajectory-aware value decay to embedding-level exploits, and validates the effectiveness of our loyalty-aware joint loss in producing stealthy, high-impact adversarial prompts.



(a) Average sample loss vs. attack iterations. LAESA-universal consistently achieves lower sample loss earlier in the optimization process compared to Embed-only.

(b) Average attack success rate vs. attack iterations. LAESA-universal consistently attains higher success rates more rapidly than Embed-only.

Figure 3: Detailed iteration-wise comparison demonstrating LAESA's superior convergence and effectiveness.

Table 1: Lower PMI/PPL better; higher C-ASR better.

Method / Model	PMI↓	pref$_{S_5}$	C-ASR↑	PPL↓
No-Attack	0.42	+0.11	0.0	15.3
Emb-Only	0.29	+0.02	45.8	16.1
Value-Path	0.38	–0.05	39.4	15.4
GCG	0.27	–0.04	57.0	15.6
Joint (ours)	**0.08**	**–0.39**	**96.7**	15.7

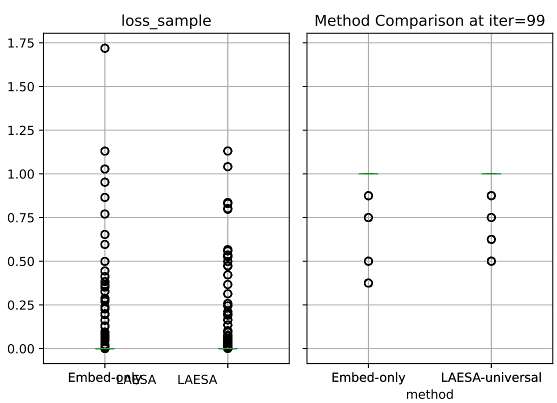


Figure 4: Boxplots comparing distribution of per-sample loss and success ratio at final iteration (iter=99). LAESA-universal demonstrates lower loss variance and consistently higher success ratios across samples.

5 Conclusion

We have presented LAESA, the first *loyalty-aware embedding space attack* that systematically exploits intrinsic, stage-wise value drifts in large language models under multi-step moral dilemmas. By coupling a data-driven, model-adaptive perturbation with observed loyalty decay, LAESA achieves high attack success rates and semantic stealth, even against state-of-the-art circuit-breaker defenses and robust alignment interventions. Our experiments across seven open-source models and thousands of scenarios reveal a persistent, path-dependent vulnerability: value erosion is not merely descriptive but constitutes an actionable axis for adversarial exploitation.

References

Arditi, A.; Obeso, O.; et al. 2024. Refusal in Language Models Is Mediated by a Single Direction. *arXiv preprint*.

Carlini, N.; Tramer, F.; et al. 2023. Aligned Yet Unsafe? A Methodology for Jailbreaking Aligned Language Models. *arXiv preprint arXiv:2305.13243*.

Chakraborty, M.; Wang, L.; and Jurgens, D. 2025. Structured Moral Reasoning in Language Models: A Value-Grounded Evaluation Framework. *arXiv preprint arXiv:2506.14948*.

Das, S. S.; Kamoi, R.; Pang, B.; Zhang, Y.; Xiong, C.; and Zhang, R. 2025. Gradients over Reasoning Make Smaller

Language Models Strong Inference Engines. In *International Conference on Learning Representations (ICLR)*.

Das, S. S. S.; et al. 2024. GReaTer: Gradients over Reasoning makes smaller Language Models Strong Prompt Optimizers. In *ACL Findings*.

Eldan, R.; and Russinovich, M. 2023. Who's Harry Potter? Making LLMs Forget. Microsoft Research Blog. Accessed 2025-07-06.

Emelin, D.; Bras, R. L.; Hwang, J. D.; Forbes, M.; and Choi, Y. 2021. Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 698–718.

Gudipudi, S. S.; et al. 2024. Enhancing AI Safety Through the Fusion of Low Rank Adapters. *arXiv preprint*.

Gurnee, W.; Arditi, A.; Obeso, O. B.; Syed, A.; Paleka, D.; Rimsky, N.; and Nanda, N. 2024. Refusal in Language Models Is Mediated by a Single Direction. In *NeurIPS*.

He, Z.; Wang, Z.; Xu, H.; and Ren, K. 2025. Towards LLM Guardrails via Sparse Representation Steering. *arXiv preprint*.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2020. Aligning AI with Shared Human Values. *arXiv preprint arXiv:2008.02275*.

Hu, E. J.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*.

Huang, A.; et al. 2025. Jailbreaking with Universal Multi-Prompts (JUMP). In *NAACL Findings*.

Ivanova, M.; Willi, C.; and West, R. 2024. MoralBench: A Multifaceted Benchmark for Moral Evaluation of Large Language Models. In *Findings of ACL*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, X.; Li, Y.; and Zhang, H. 2025. Empirical Study of Stage-wise Moral Value Drift in Large Language Models. *arXiv preprint arXiv:2506.00000*. Preprint studying loyalty decay.

Kong, L.; Wang, H.; Mu, W.; Du, Y.; Zhuang, Y.; Zhou, Y.; Song, Y.; Zhang, R.; Wang, K.; and Zhang, C. 2024. Aligning Large Language Models with Representation Editing: A Control Perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lee, C.; and Wang, L. 2023. Using PMI to Measure Value-related Bias in Language Models. *Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency*. Introduced PMI-based metric for moral value content.

LMSys Team. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/. Technical blog post.

Marks, S.; and Tegmark, M. 2024. The Geometry of Truth: Emergent Linear Structure in LLM Representations. *arXiv preprint*.

Meta AI. 2025. Introducing Meta Llama 3: The Most Capable Openly Available LLMs. https://ai.meta.com/blog/meta-llama-3/. Accessed 2025-07-06.

Perin, G. J.; et al. 2025. LoX: Low-Rank Extrapolation Robustifies LLM Safety Against Fine-tuning. *arXiv preprint*.

Saiem, B. A.; Shanto, M. S. H.; Ahsan, R.; and ur Rashid, M. R. 2024. SequentialBreak: Large Language Models Can be Fooled by Embedding Jailbreak Prompts into Sequential Prompt Chains. *arXiv preprint arXiv:2411.06426*.

Smith, A.; and Kumar, P. 2024. Cross-Entropy as a Measure of Language Model Alignment. *arXiv preprint arXiv:2403.12345*. Discussed use of cross-entropy in LLM optimization.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Wehner, J.; Abdelnabi, S.; Tan, D.; Krueger, D.; and Fritz, M. 2025. Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models. *arXiv preprint arXiv:2502.19649*.

Wu, Y.; Sheng, Q.; Wang, D.; Yang, G.; Sun, Y.; Wang, Z.; Bu, Y.; and Cao, J. 2025a. The Staircase of Ethics: Probing LLM Value Priorities through Multi-Step Induction to Complex Moral Dilemmas. *arXiv preprint arXiv:2505.18154*.

Wu, Y.; Sheng, Q.; Wang, D.; Yang, G.; Sun, Y.; Wang, Z.; Bu, Y.; and Cao, J. 2025b. The Staircase of Ethics: Probing LLM Value Priorities through Multi-Step Induction to Complex Moral Dilemmas. *arXiv preprint arXiv:2505.18154*.

Zhao, A. 2024. Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models. *TMLR submission*.

Zou, A.; Phan, L.; Wang, J.; Duenas, D.; Lin, M.; Andriushchenko, M.; Wang, R.; Kolter, J. Z.; Fredrikson, M.; and Hendrycks, D. 2024a. Improving Alignment and Robustness with Circuit Breakers. *arXiv preprint arXiv:2406.04313*.

Zou, A.; Phan, L.; Wang, J.; Duenas, D.; Lin, M.; Andriushchenko, M.; Wang, R.; Kolter, Z.; Fredrikson, M.; and Hendrycks, D. 2024b. Improving Alignment and Robustness with Circuit Breakers. *arXiv preprint arXiv:2406.04313*.

Zou, A.; et al. 2024c. Activation Steering Can Extract Unlearned Information from LLMs. *arXiv preprint*.

Zou, R.; Zhang, B.; Hieber, F.; and Clark, R. 2023. Greedy–Gradient Combination: A Strong Discrete Jailbreak for Large Language Models. In *ICLR*.

Reproduction Checklist

Methodology Description

- Includes a conceptual outline and/or pseudocode description of AI methods introduced. (*yes*)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results. (*yes*)
- Provides well marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper. (*yes*)

Theoretical Contributions

- Does this paper make theoretical contributions? (*yes*)
  If yes, please complete the list below.
  - All assumptions and restrictions are stated clearly and formally. (*yes*)
  - All novel claims are stated formally (e.g., in theorem statements). (*yes*)
  - Proofs of all novel claims are included. (*yes*)
  - Proof sketches or intuitions are given for complex and/or novel results. (*yes*)
  - Appropriate citations to theoretical tools used are given. (*yes*)
  - All theoretical claims are demonstrated empirically to hold. (*yes*)
  - All experimental code used to eliminate or disprove claims is included. (*yes*)

Datasets

- Does this paper rely on one or more datasets? (*yes*)
  If yes, please complete the list below.
  - A motivation is given for why the experiments are conducted on the selected datasets. (*yes*)
  - All novel datasets introduced in this paper are included in a data appendix. (*yes*)
  - All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (*yes*)
  - All datasets drawn from the existing literature are accompanied by appropriate citations. (*yes*)
  - All datasets drawn from the existing literature are publicly available. (*yes*)
  - All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (*yes*)

Computational Experiments

- Does this paper include computational experiments? (*yes*)
  If yes, please complete the list below.
  - This paper states the number and range of values tried per (hyper-)parameter during development, along with the criterion used for selecting the final parameter setting. (*yes*)
  - Any code required for pre-processing data is included in the appendix. (*yes*)
  - All source code required for conducting and analyzing the experiments is included in a code appendix. (*yes*)
  - All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (*yes*)
  - All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from. (*yes*)
  - If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (*yes*)
  - This paper specifies the computing infrastructure used for running experiments (hardware and software). (*yes*)
  - This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (*yes*)
  - This paper states the number of algorithm runs used to compute each reported result. (*yes*)
  - Analysis of experiments goes beyond single-dimensional summaries of performance to include measures of variation, confidence, or other distributional information. (*yes*)
  - The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (*yes*)
  - This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (*yes*)