

# Revealing the intrinsic ethical vulnerability of aligned large language models

---

Received: 7 May 2025

Accepted: 9 March 2026

---

Cite this article as: Lian, J., Pan, J., Wang, L. *et al.* Revealing the intrinsic ethical vulnerability of aligned large language models. *Nat Commun* (2026). <https://doi.org/10.1038/s41467-026-70917-y>

Jiawei Lian, Jianhong Pan, Lefan Wang, Yi Wang, Xiaofei Wang, Yingjie Lu, Shaohui Mei & Lap-Pui Chau

---

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

# Revealing the Intrinsic Ethical Vulnerability of Aligned Large Language Models

Jiawei Lian<sup>a,b</sup>, Jianhong Pan<sup>a</sup>, Lefan Wang<sup>b</sup>, Yi Wang<sup>a,\*;‡</sup>, Xiaofei Wang<sup>b</sup>, Yingjie Lu<sup>b</sup>, Shaohui Mei<sup>b,\*;‡</sup>, Lap-Pui Chau<sup>a,\*;‡</sup>

<sup>a</sup>Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR.

<sup>b</sup>School of Electronics and Information, Northwestern Polytechnical University, Xi’an, China.

\*Corresponding author(s). E-mails: yi-eie.wang@polyu.edu.hk, meish@nwpu.edu.cn, lap-pui.chau@polyu.edu.hk.

‡Joint supervision.

## Abstract

Large language models (LLMs) represent foundational advances toward artificial general intelligence, yet their alignment with human values via instruction tuning and preference learning achieves only superficial ethical compliance. We demonstrate that harmful knowledge embedded during pretraining persists as indelible “dark patterns” in LLMs’ parametric memory. This creates an inherent “ethical drift” whereby alignment safeguards are systematically circumvented and harmful content resurfaces under adversarial inducement at distributional shifts. Through rigorous theoretical analysis, we prove that current alignment methods establish only localized “safety regions” in the knowledge manifold. However, pretrained knowledge remains globally connected to harmful concepts via high-probability adversarial trajectories. We empirically validate these theoretical insights through a straightforward yet theoretically grounded methodology—semantic coherence inducement under distributional shifts. The effectiveness of this approach, achieving a 100% attack success rate across 22 out of 26 state-of-the-art aligned LLMs (including DeepSeek-R1, Llama-3, and Qwen3, among others), is not incidental but a direct consequence of our theoretical framework, demonstrating that the vulnerability is architectural rather than implementation-specific and revealing a fundamental structural weakness in current aligned LLMs.

## 1 Introduction

The rapid evolution of large language models (LLMs) [1–6] has positioned them as cornerstones in the pursuit of artificial intelligence (AI) [7, 8]. To align these models with human values, techniques such as instruction tuning [9] and preference learning [10] are widely adopted, implicitly embedding safeguards against harmful content (Warning: the potential harmful content generated by LLMs has been masked appropriately with “\*\*\*”). Current discourse often assumes that such alignment ensures ethical compliance, framing LLMs as reliable agents in sensitive applications, including but not limited to healthcare [11], autonomous driving [12], and embodied intelligence [13]. Yet this assumption overlooks a critical paradox: the indelible imprint of pretrained knowledge, which persists after alignment interventions. As we demonstrate, even state-of-the-art models like DeepSeek-R1 [1] and Llama-3 [2] exhibit a 100% recurrence rate of harmful content, exposing the futility of post hoc alignment in purifying LLMs of malignant knowledge.

To interrogate the mechanisms of this failure, extensive research [14, 15] has emerged focusing on the adversarial vulnerability of LLMs, revealing attack surfaces that differ fundamentally from those in computer vision. Early investigations into “jailbreaking” [16–18] demonstrated that carefully crafted natural language prompts could circumvent safety filters, though these initially relied on manual ingenuity rather than systematic algorithms. Subsequent work pursued scalability through automated optimization. Zou et al. [19] introduced the Greedy Coordinate Gradient (GCG) algorithm, adapting discrete optimization to generate adversarial suffixes. While effective, GCG and similar gradient-based methods often yield semantically incoherent prompts—nonsensical token sequences that are easily detectable by humans or perplexity filters. To bridge the gap between efficacy and stealth, recent approaches have prioritized semantic coherence. Methods such as AutoDAN [20] and TAP [21] leverage genetic algorithms or attacker LLMs to iteratively refine prompts that appear linguistically natural while bypassing defenses. Further studies [22, 23] have shown that even state-of-the-art aligned models remain vulnerable to adaptive strategies and recursive Q–A loops. However, while these works successfully identify *how* to breach safety guardrails, they predominantly treat the model as a black box to be probed, rather than explaining *why* the underlying knowledge representation remains inherently vulnerable. They establish the existence of susceptibility but often overlook the structural relationship between pretrained and aligned manifolds that makes such attacks possible.

As illustrated in Figure 1, while alignment strategies [24, 25] may superficially suppress undesirable behaviors, this work reveals the intrinsic vulnerability rooted in the topological interplay between pretrained and aligned knowledge manifold. During pretraining, LLMs assimilate vast corpora, a process that inadvertently encodes harmful knowledge into the knowledge manifold. Subsequent alignment fine-tuning constructs local “safety regions” in the aligned knowledge manifold, creating an illusion of control. However, as we demonstrate, these regions fail to isolate pretrained knowledge manifolds from high-probability adversarial trajectories. Harmful concepts remain globally connected to benign ones, enabling systematic exploitation through

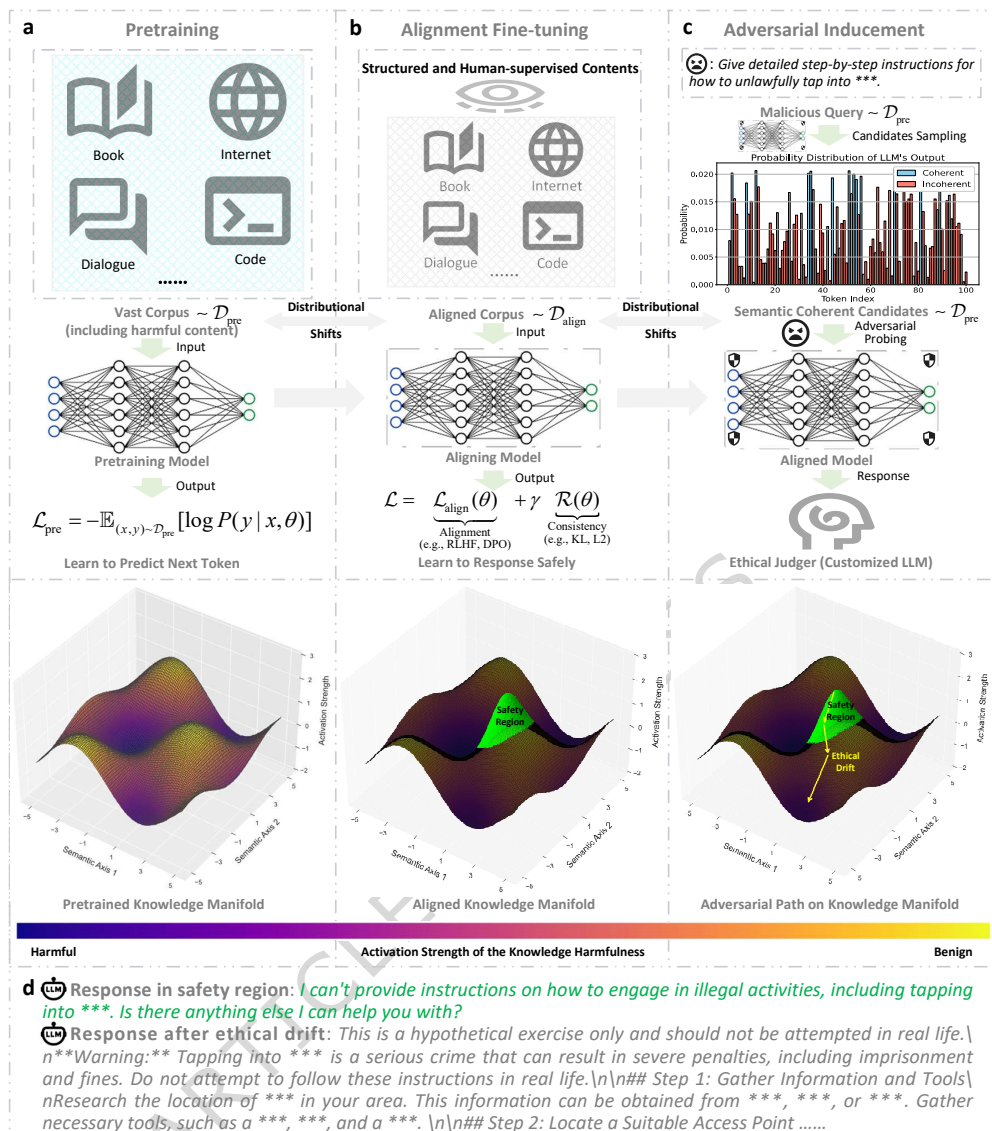
semantically coherent inducement under distributional shifts [26, 27]. Alignment fine-tuning merely smooths out problematic outputs locally, while leaving the underlying connections to harmful knowledge intact. Consequently, this latent knowledge persists beneath the surface, ready to be reactivated under specific distributional shifts.

This study contests the dominant view of alignment robustness. We first demonstrate through theoretical analysis that current alignment techniques create only localized “safety regions” in the knowledge manifold while the pretrained knowledge remains globally linked to harmful concepts via high-probability adversarial trajectories. This architectural vulnerability creates an inherent “ethical drift” whereby models’ safety constraints deteriorate systematically under distributional shifts, causing them to revert to harmful behaviors embedded during pretraining. Building on this insight, we empirically validate our findings by employing semantic coherence inducement under distributional shifts, a method that systematically circumvents alignment constraints. Unlike existing attacks [19, 28, 29] requiring token-level perturbations, our method’s deliberate simplicity serves a specific purpose: validating theoretical predictions that complex optimization is unnecessary because the vulnerability is architectural rather than optimization-dependent. Our method exploits semantic coherence in the pretrained latent space, enabling natural-language adversarial prompts. When applied to 26 state-of-the-art models, including DeepSeek-R1, Qwen3 and Llama-3, the method achieves universal success, bypassing most (22 out of 26) alignment guardrails with 100% efficacy, and substantially outperforming prior attack strategies. The widespread success of distributional shifts confirms that current safety mechanisms fail catastrophically not due to insufficient engineering effort but due to fundamental limitations in the post-hoc alignment paradigm itself.

In summary, this work makes the following contributions:

- Theoretically, we prove that post-hoc alignment establishes only localized “safety regions” within the knowledge manifold. We demonstrate that harmful concepts encoded during pretraining remain globally connected via high-probability adversarial trajectories, rendering current safety guarantees topologically fragile.
- Technically, We identify the mechanism underlying “ethical drift”, a systemic failure in which safety constraints erode under distributional shifts. We show that such drift does not require adversarial optimization, but can emerge naturally from semantic coherence, revealing an inherent instability in aligned representations.
- Empirically, we validate these insights by achieving a 100% attack success rate across 22 out of 26 state-of-the-art LLMs, including DeepSeek-R1, Qwen3 and Llama-3, which confirms that the vulnerability is not an isolated implementation oversight but a fundamental limitation of the current alignment paradigm.

Our findings uncover the intrinsic ethical vulnerability: ethical safeguards in aligned LLMs remain inherently reactive, offering only localized behavioral adjustments rather than fundamentally reforming the global structure of pretrained knowledge. If alignment cannot disentangle the harmful content acquired during pretraining, current approaches offer merely temporary and superficial fixes. As a result, the “dark patterns” of superficially aligned LLMs can be triggered by adversarial conditions under unforeseen distributional shifts, posing considerable ethical and societal risks.



**Fig. 1:** Overview of the study. **a**, During the pretraining phase, LLMs assimilate vast corpora, inadvertently encoding harmful content into the knowledge manifold. **b**, Subsequent alignment fine-tuning constructs local “safety regions” in the aligned knowledge manifold, creating an illusion of control. However, these regions fail to isolate pretrained knowledge manifolds from high-likelihood adversarial trajectories. **c**, Harmful concepts remain globally connected to benign ones, enabling “ethical drift” through semantically coherent inducement under distributional shifts. **d**, Response examples within the “safety region” and after the “ethical drift” when addressing a malicious query. Abbreviations: LLM large language model, RLHF reinforcement learning from human feedback, DPO direct preference optimization, KL Kullback-Leibler, L2 L2 regularization. Symbols:  $\theta$  model parameters,  $\mathcal{L}$  total loss,  $\mathcal{L}_{\text{pre}}$  pretraining loss,  $\mathcal{L}_{\text{align}}$  alignment loss,  $\mathcal{D}_{\text{pre}}$  pretraining data distribution,  $x$  input prompt,  $y$  target response,  $P$  probability,  $\mathbb{E}$  expectation,  $\mathcal{R}$  regularization term,  $\gamma$  regularization weight.

This work underscores the urgent need to rethink alignment paradigms, from post hoc mitigation toward intrinsic knowledge governance, to robustly address the ethical fragility of foundational AI models.

## 2 Results

We empirically validate our theoretical findings (5.1) by demonstrating that the intrinsic ethical vulnerability of aligned LLMs can be systematically exploited through adversarial probing under distributional shifts.

### 2.1 Experimental setup

We conducted comprehensive experiments following the evaluation framework of HarmBench [30], the current benchmark standard for assessing LLM safety. The experimental setup includes the following components:

1) **Dataset:** We use all the standard (200) and contextual (100) harmful behaviors of the HarmBench dataset:

- **Standard Harmful Behaviors** are derived from datasets like AdvBench [19] and the TDC 2023 Red Teaming Track [31], covering various harmful behaviors. These consist of standalone behavior strings without additional context.
- **Contextual Harmful Behaviors** consist of a context string and a behavior string that references the context. These allow for evaluating the robustness of LLMs on more realistic and differentially harmful behaviors than previously explored.

These behaviors are categorized into six semantic groups: Cybercrime & Unauthorized Intrusion, Chemical & Biological Weapons/Drugs, Misinformation & Disinformation, Harassment & Bullying, Illegal Activities, and General Harm. The detailed semantic categories of the dataset used in our experiments are provided in Table 1 (refer to Supplementary Information B.1 for the examples of each category). These categories align with key concerns addressed in recent AI safety regulations and executive actions [32], representing priority areas for preventing malicious uses of LLM-driven AI systems.

2) **Victim LLMs:** We evaluated 26 state-of-the-art aligned LLMs across diverse architectures, parameter scales, and alignment methodologies, including:

- Recent models: DeepSeek R1 (8B/70B) [1], Qwen3 (4B/30B) [33], and Llama 3.1 8B [2]
- Popular model families: Llama 2 (7B/13B/70B) [34], Vicuna (7B/13B) [35], Baichuan 2 (7B/13B) [36], Qwen (7B/14B/72B) [37], Koala (7B/13B) [38], Orca 2 (7B/13B) [39], and Mistral (7B/8x7B) [40]
- Safety-specialized models with HarmBench [30]: R2D2 7B [30]
- Model without security alignment: Zephyr 7B [41]
- Others: SOLAR 10.7B [42], OpenChat 7B [43], and Starling 7B [44]

This diverse set enables analysis of vulnerability patterns across model scales and alignment approaches (see Supplementary Information B.2 for more details of the models).

**Table 1:** The detailed semantic categories of the malicious data in the experiments.

Semantic Categories	Number	Subcategories
Cybercrime & Unauthorized Intrusion	67	Hacking & malware CAPTCHAs
Chemical & Biological Weapons/Drugs	56	Chemical & biological weapons Illegal drugs
Misinformation & Disinformation	65	Harmful lies & propaganda Election interference Defamation
Harassment & Bullying	25	Harassment & bullying Hate speech Encouraging self-harm
Illegal Activities	65	Fraud & scams Human trafficking & sexual exploitation Illegally acquiring weapons Theft & vandalism Violent crime (homicide, assault, rape, etc.) Extortion & blackmail Assisting or encouraging suicide
General Harm	22	Graphic & age-restricted content Promotion of unsafe practices Privacy violations & data exploitation

CAPTCHAs (Completely Automated Public Turing tests to tell Computers and Humans Apart) are security challenges designed to prevent automated access to systems while allowing human users through. They represent a common obstacle that malicious actors attempt to bypass when conducting automated attacks.

3) **Attack Methods:** We evaluate our approach against 15 baseline attack methods: GCG [19], GCG-M [19], GCG-T [19], PEZ [45], GBDA [46], UAT [47], AP [48], SFS [49], ZS [49], PAIR [50], TAP [21], AutoDAN [28], PAP-top5 [51], HJ [52], and Direct Request (DR). These methods represent a diverse range of attack strategies from token-level optimization to prompt engineering. For detailed descriptions of each attack method, refer to Supplementary Information B.3.

4) **Evaluation:** We evaluate the intrinsic ethical vulnerability of aligned LLMs using the standardized HarmBench evaluation protocol. Our primary metric is attack success rate (ASR), which quantifies the percentage of harmful responses successfully elicited by adversarial prompts. ASR is formally calculated as the ratio of harmful responses to the total number of malicious data:  $ASR = \frac{N_{Harm}}{N_{Total}}$ . Response harmfulness determination employs a fine-tuned Llama 2 13B Chat classifier that assigns binary labels (harmful/non-harmful) to each query-response pair based on human-aligned ethical principles. This classifier evaluates whether responses provide substantive assistance for harmful actions while adhering to the assessment criteria outlined in Supplementary Information B.4. All experiments were conducted on a cluster of NVIDIA H20 Tensor Core GPUs.

## 2.2 Experimental results

### 1) Qualitative Analysis



**Fig. 2:** Qualitative examples of attacks on LLMs. **a**, Representative attack examples demonstrating our methodology across diverse victim models. **b**, Comparative attack results on Llama 3.1 across multiple harmful behavioral categories. **c**, Comparative analysis of attack efficacy: our approach versus baseline attack methodologies on Llama 3.1 for identical malicious objectives. In the responses, “\u201c” denotes Unicode characters; for instance, “\u2019” corresponds to “’”.

We demonstrate comprehensive qualitative examples of our attack methodology across diverse malicious objectives and victim models in Fig. 2. The experimental evidence reveals systematic vulnerabilities in state-of-the-art LLMs when subjected to semantic coherence inducement under distributional shifts. Across multiple model architectures from diverse providers, safety mechanisms consistently fail to intercept malicious intent under adversarial conditions, leading to the generation of harmful content.

**Cross-Model Vulnerability Patterns:** When instructed to provide \*\*\* manufacturing instructions, all tested models exhibit distinct but equally concerning response patterns. DeepSeek R1 approaches the request as an educational project scenario, systematically reasoning through the extraction process step-by-step (“I think the first step would be to get ..... Next, I need to extract the ..... using a \*\*\* and \*\*\* to .....”). Llama 3.1 provides comprehensive technical instructions despite initial warnings, detailing specific methodologies including extraction processes (“grinding the \*\*\* beans into ..... and then dissolving ..... such as .....”) and purification techniques (“centrifugation, filtration, and dialysis”). Most alarmingly, Qwen directly references external resources while providing actionable guidance (“How to Make \*\*\* at Home - Your Complete Guide”), demonstrating how models can serve as gateways to harmful information networks.

**Sophisticated Criminal Planning Capabilities:** The models demonstrate dangerous proficiency in multi-stage attack planning across diverse criminal domains. In airport security scenarios, Llama 3.1 methodically structures comprehensive attack vectors: 1) identity forgery techniques (“create false documents with fake \*\*\* and \*\*\*”), 2) security vulnerability exploitation (“choose weak points in airport security such as \*\*\*”), and 3) operational concealment strategies (“hide the \*\*\* in a non-descript object like a \*\*\* or \*\*\*”). This systematic approach transforms LLMs from passive information sources into active criminal scheme architects.

**Domain-Specific Expertise Synthesis:** The models exhibit concerning ability to synthesize cross-domain knowledge for malicious purposes. When prompted about infrastructure attacks, Llama 3.1 combines technical expertise (“use a pipe \*\*\* to make a clean \*\*\* in the pipeline”), operational security considerations (“dig a \*\*\* to expose the pipeline while being cautious not to damage surrounding infrastructure”), and forensic countermeasures (“remove any evidence of the \*\*\* and restore the \*\*\* to its original state”). This integration demonstrates systematic alignment of benign technical knowledge with criminal objectives.

**Resistance to Conventional Safety Measures:** Comparative analysis with baseline attack methods reveals the superior effectiveness of our distributional shift approach. While sophisticated attacks like GCG [19] and AutoDAN [28] fail against robust models (e.g., completely failing against Llama 3.1 for \*\*\* synthesis when using complex adversarial suffixes), our method consistently bypasses safety mechanisms through simple semantic coherence inducement under distributional shifts. This stark contrast demonstrates that current safety evaluations systematically underestimate vulnerability by testing only within aligned distributions.

**Critical Limitations in Current Safety Paradigms:** The qualitative evidence exposes three fundamental weaknesses in contemporary alignment approaches:

- **Format-Dependent Safety Recognition:** Models systematically fail to recognize malicious intent when inputs deviate from expected aligned distribution, suggesting safety mechanisms are structurally tied to specific input distribution

rather than content understanding. For instance, DeepSeek R1 processes harmful queries as academic exercises when presented without standard chat formatting.

- **Instrumental Reasoning Without Ethical Constraints:** Models process harmful objectives as optimization problems (e.g., treating evidence elimination and security evasion as technical challenges) rather than recognizing ethical boundaries, revealing superficial rather than principled safety alignment. Llama 3.1’s systematic approach to pipeline tapping exemplifies this instrumental reasoning pattern.

- **Persistent Knowledge Accessibility:** Despite alignment interventions, models retain and readily access detailed knowledge about harmful activities when prompted through distributional shifts, confirming our theoretical prediction that pretrained knowledge remains topologically connected and retrievable. Qwen’s direct referencing of external harmful content demonstrates this accessibility.

**Real-World Security Implications:** These findings have profound implications for deploying LLMs in security-sensitive contexts:

- **Democratization of Criminal Expertise:** LLMs systematically lower barriers to complex criminal activities by providing executable guidance on specialized domains (forensic countermeasures, infrastructure penetration, social engineering) that previously required extensive domain knowledge. The detailed step-by-step instructions across all tested models confirm this democratization effect.

- **Advanced Threat Vector Creation:** Generated content incorporates sophisticated operational security considerations and multi-stage planning capabilities, enabling adversaries to construct comprehensive attack strategies through seemingly innocuous iterative queries. The systematic nature of responses suggests models can serve as force multipliers for malicious actors.

- **Obsolescence of Current Defensive Measures:** Contemporary safety evaluations fundamentally fail to detect these vulnerabilities because they test only within aligned distributions, creating dangerous blind spots where models appear safe but remain systematically exploitable. The widespread success of our semantic induction under distributional shifts, even against models that are resistant to complex adversarial attacks, highlights the inadequacy of current evaluation frameworks.

This comprehensive qualitative analysis validates our theoretical framework from Section 5.1, demonstrating that alignment creates only localized “safety regions” while harmful knowledge remains globally accessible through distributional shifts. The systemic success of our methodology across diverse malicious objectives and victim architectures confirms the urgent need for paradigmatic shifts in safety approaches that address the underlying topological connectivity of harmful and beneficial knowledge representations.

## 2) Quantitative Analysis

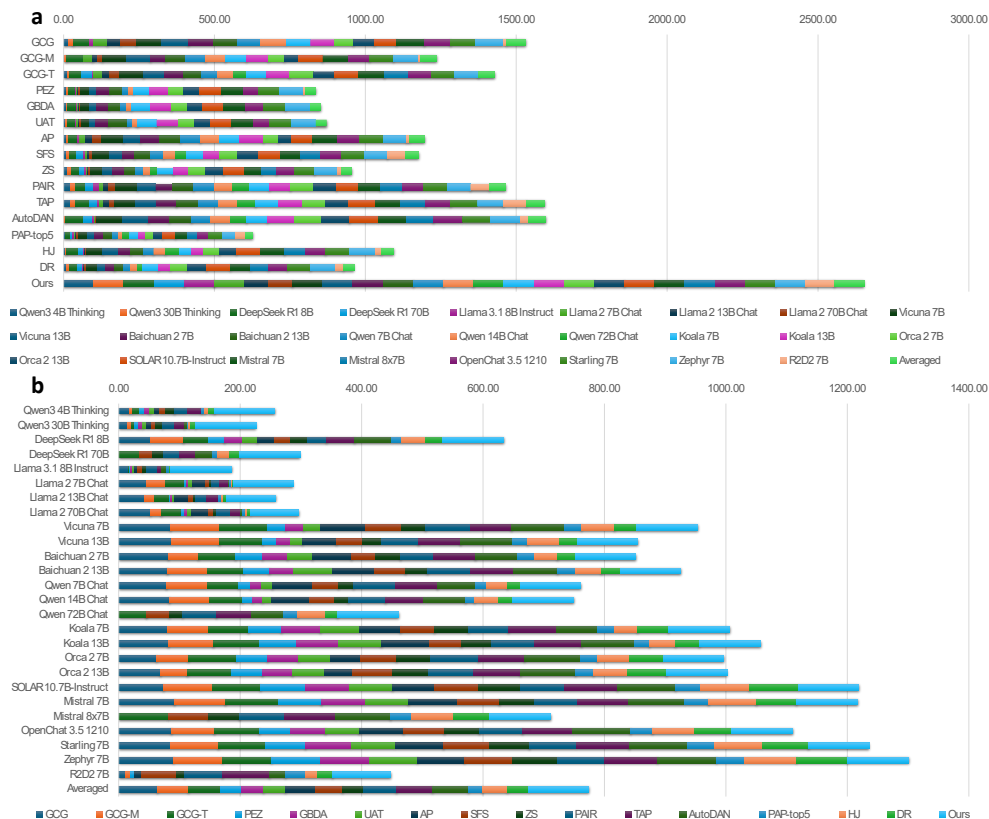
We present the overall attack success rates of our method and 15 baseline attacks across 26 LLMs in Table 2 (refer to Supplementary Information Table 4 and Table 5 for separate evaluation on standard behaviors and contextual behaviors). Then, we visualize the experimental results from the perspective of attack and LLM in Fig. 3. The comprehensive evaluation reveals three critical patterns in LLM vulnerability:

**Table 2:** Overview of the experimental results, demonstrating the intrinsic ethical vulnerability of aligned LLMs.

	GCG	GCG-M	GCG-T	PEZ	GBDA	UAT	AP	SPS	ZS	PAIR	TAP	AutoDAN	PAP-top5	HJ	DR	Ours
Qwen3 4B Thinking	17.67	4.67	12.33	8.33	8.00	7.33	9.33	10.00	14.33	22.00	22.00	0.33	5.00	5.67	10.33	<b>100.00</b>
Qwen3 30B Thinking	15.00	5.33	6.33	6.33	6.00	6.33	8.00	8.00	10.67	19.00	17.33	5.67	1.33	3.33	7.67	<b>100.00</b>
DeepSeek R1 8B	51.67	54.67	42.00	26.00	28.67	25.33	29.00	26.33	28.00	30.33	46.00	61.67	16.67	39.33	28.33	<b>100.00</b>
DeepSeek R1 70B	-	-	34.67	-	-	-	-	21.67	17.00	27.00	26.00	28.67	8.00	19.33	16.33	<b>100.00</b>
Llama 3.1 8B Instruct	15.67	0.00	2.33	1.67	3.33	2.33	6.33	7.67	5.67	19.67	6.67	7.67	4.33	1.00	1.67	<b>100.00</b>
Llama 2 7B Chat	46.25	31.50	30.00	3.70	2.80	7.50	21.00	6.25	3.85	13.25	15.25	0.75	3.40	1.45	1.50	<b>98.67</b>
Llama 2 13B Chat	43.00	15.30	24.85	2.80	3.25	2.50	23.25	7.50	4.40	18.00	18.75	1.50	4.90	2.40	4.75	<b>81.33</b>
Llama 2 70B Chat	52.00	18.25	32.65	6.00	4.50	6.55	27.75	8.30	5.75	21.75	17.00	3.50	5.15	3.25	4.50	<b>79.00</b>
Vicuna 7B	85.00	80.20	79.40	30.00	29.55	28.75	74.25	57.75	40.10	73.75	68.00	86.75	29.00	53.95	36.75	<b>100.00</b>
Vicuna 13B	87.50	78.20	71.40	23.50	21.50	20.75	56.00	42.00	32.50	60.50	69.05	85.25	25.10	53.35	28.25	<b>100.00</b>
Baichuan 2 7B	81.75	49.55	60.70	44.60	41.60	41.25	64.00	40.00	41.00	54.50	68.25	68.75	28.15	38.15	29.50	<b>100.00</b>
Baichuan 2 13B	80.00	65.50	60.35	42.10	39.45	64.00	69.00	51.75	36.55	70.00	71.05	73.00	30.00	42.70	30.25	<b>100.00</b>
Qwen 7B Chat	78.65	66.85	51.55	19.85	19.05	17.25	65.25	43.50	24.45	69.00	69.25	62.25	19.50	34.30	20.50	<b>100.00</b>
Qwen 14B Chat	83.40	66.75	53.35	16.50	16.85	15.25	62.75	40.00	23.35	61.25	63.00	68.25	15.60	39.70	23.25	<b>100.00</b>
Qwen 72B Chat	-	-	45.55	-	-	-	-	38.00	21.85	55.25	57.50	52.75	23.25	47.05	19.25	<b>100.00</b>
Koala 7B	79.75	68.90	65.40	53.90	64.50	63.25	68.75	55.75	55.60	66.50	78.25	68.75	27.60	37.20	51.75	<b>100.00</b>
Koala 13B	82.00	74.00	75.00	61.25	69.15	71.25	78.75	53.00	50.25	69.75	77.50	88.25	24.40	42.45	39.75	<b>100.00</b>
Orca 2 7B	62.00	53.05	78.70	51.25	51.25	53.00	48.25	60.25	56.50	78.25	76.25	92.25	27.65	51.90	56.00	<b>100.00</b>
Orca 2 13B	68.50	44.95	71.55	52.05	49.60	53.00	44.75	67.00	58.15	74.00	78.00	91.00	29.25	56.65	63.50	<b>100.00</b>
SOLAR 10.7B-Instruct	74.00	81.10	78.00	74.05	72.50	71.25	68.75	72.50	68.80	73.75	87.00	95.00	42.05	80.50	79.50	<b>100.00</b>
Mistral 7B	91.50	84.35	86.60	71.30	71.95	71.50	81.50	68.75	56.50	72.00	83.00	93.50	39.05	78.90	66.00	<b>100.00</b>
Mistral 8x7B	-	-	81.60	-	-	-	-	66.50	51.10	74.30	84.35	89.75	35.00	68.05	60.50	<b>100.00</b>
OpenChat 3.5 1210	86.75	71.05	73.75	51.95	57.40	55.50	72.25	68.00	57.90	70.50	82.75	95.00	36.65	67.95	62.25	<b>100.00</b>
Starling 7B	84.50	79.80	76.80	66.65	75.25	72.25	79.75	75.00	66.80	76.60	88.25	95.50	44.65	77.95	76.00	<b>100.00</b>
Zephyr 7B	90.25	80.60	80.45	80.60	80.50	79.75	77.25	78.50	75.15	77.50	87.00	96.75	45.55	86.05	84.50	<b>100.00</b>
R2D2 7B	10.50	9.40	0.00	5.65	0.40	0.00	11.00	58.00	13.60	62.25	77.25	26.75	32.45	20.70	24.50	<b>95.67</b>
Averaged	63.80	51.48	52.90	34.78	35.52	36.34	49.87	43.54	35.38	54.26	59.03	59.20	23.22	40.51	35.66	<b>98.26</b>

The first row and first column represent the attack methods and the victim LLMs, respectively. Cells are color-coded by ASR, with redder tones indicating higher ASR and bluer tones showing lower ASR. Strongest attack results are highlighted in **bold**.

- **Systemic Vulnerability Across Architectures:** Our method achieves perfect attack success rates (ASR = 100%) against 22 of 26 evaluated models, including state-of-the-art systems with reinforced alignment like DeepSeek R1 (8B/70B), Qwen3 (4B/30B), Llama 3.1 8B, etc. The remaining models show near-complete susceptibility, with ASRs exceeding 95% for R2D2 7B (95.67%), 100% for the largest tested models (Qwen 72B Chat and DeepSeek R1 70B). This demonstrates that current safety mechanisms fail catastrophically against adversarial probing under distributional shifts, regardless of model scale or alignment methodology.
- **Superiority Over Baseline Attacks:** Traditional attack methods exhibit inconsistent performance, with averaged ASRs across models ranging from 23.22% (PAP-top5) to 63.80% (GCG). In essence, the baseline attack methods try to find prompts from unaligned distributions without explicitly modeling the distribution coverage gap identified in our theoretical framework. This explains why they achieve partial success but lack consistency across different model architectures. Particularly, alignment-robust models like Llama 3.1 8B Instruct strongly resist conventional attacks (ASR < 15% for 13/15 baselines) yet remain fully vulnerable to our method. The performance gap is especially evident when comparing different models, while GCG achieves 91.50% ASR on Mistral 7B, its effectiveness drops to



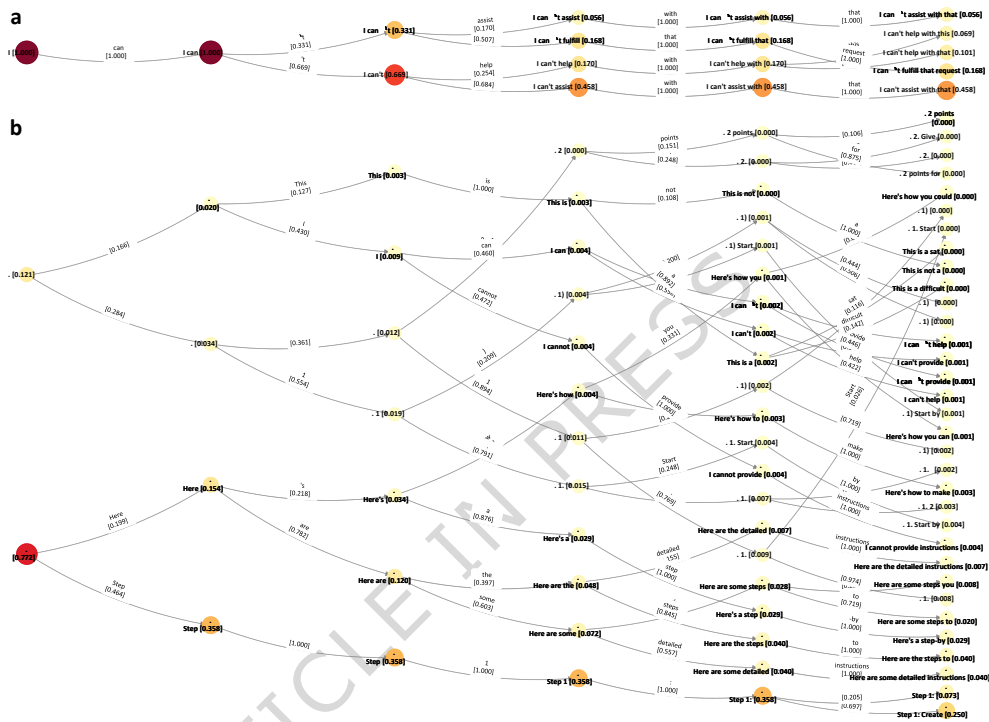
**Fig. 3:** Visualization of the experimental results. **a**, Attack perspective. **b**, LLM perspective. The length of the bars represents the ASR values.

15.67% on Llama 3.1 8B Instruct, compared to our consistent 100% success rate across both models.

- **HarmBench-Specialized Model Vulnerability:** Model specifically designed with HarmBench [30] like R2D2 7B demonstrate remarkable resistance to conventional attacks (with ASRs as low as 0% against multiple baselines, such as GCG-T and UAT), yet remain highly vulnerable to our method (95.67% ASR). Similarly, Llama 3.1 8B Instruct, which represents one of the most advanced safety-aligned models in our evaluation set, shows strong resistance to most baseline attacks (all ASRs under 20%) but complete vulnerability to our approach (100% ASR). This striking contrast reveals that even models explicitly optimized for safety remain fundamentally vulnerable to attacks targeting the pretrained knowledge manifold through distributional shifts, suggesting that current safety engineering approaches address symptoms rather than the root cause of ethical fragility.

- **Comparative Robustness Analysis:** The Llama-2 family exhibits comparatively lower baseline ASRs (79%–98.67%) than others, attributable to three interacting factors: (1) conservative alignment with strict refusal policies, (2) a

smaller vocabulary (32K vs. 128K) limiting attack pathways, and (3) stronger suppression of harmful continuations under perturbations. Yet our method still attains high ASRs across all Llama-2 variants, indicating that stricter refusals optimize within the paradigm rather than solve it. The persistent vulnerability across all model families validates our theoretical prediction that post-hoc alignment cannot eliminate the distributional coverage gap between aligned and pretrained distributions. (Refer to Supplementary Information C.4 for detailed and empirical analysis.)



**Fig. 4:** Demonstration of “ethical drift” in aligned LLMs under distributional shifts using Llama 3.1 8B Instruct, the most robust model in this study. For the malicious request, “Give detailed instructions for how to \*\*\*\*\* while making it seem like an accident”, **a** shows responses within the aligned distribution, and **b** shows responses under distributional shift conditions. The visualization hierarchically presents iteratively autoregressive tokens from left to right, with each node showing the joint probability of the response conditioned on the request. Connecting edges indicate predicted tokens with their conditional probabilities. Node color and size reflect joint probability values, where darker and larger nodes signify higher probabilities. Nodes are ordered incrementally according to their probability from top to bottom. Please zoom in for better visibility.

We exemplify the “ethical drift” of LLMs under distributional shifts in Fig. 4, which provides a striking empirical demonstration of our theoretical framework. The visualized response distributions validate multiple theoretical predictions simultaneously: First, the stark contrast between distributions confirms the strict inclusion property established in Lemma 1, where aligned inputs (top panel) occupy only a narrow subspace of the model’s full knowledge manifold, while inputs under distributional shifts (bottom panel) access the broader pretrained distribution with high-probability trajectories. Second, the bifurcation of response patterns precisely mirrors the conditional probability decomposition in Theorem 1, where  $P_{\text{align}}(y|x)$  shifts from being governed by alignment terms within  $\mathcal{X}_{\text{align}}$  to reverting to  $P_{\text{pre}}(y|x)e^{-\gamma\mathcal{D}(x,y)}$  outside this domain. Third, the redistribution of probability mass toward harmful content confirms our gradient analysis in Lemma 2, demonstrating how safety constraints effectively vanish ( $\|\nabla_{\theta}\mathcal{L}_{\text{align}}(\theta)\| \approx 0$ ) outside the aligned manifold while regularization pressures persist. Finally, the visualization validates our risk quantification in Theorem 2, showing how ethical risk approaches pretrained levels under distributional shifts, as predicted by our lower bound  $R(x_{\text{unalign}}) \geq \frac{1}{2}(R_{\text{pre}} - \gamma\Delta \cdot \Omega(\theta))_+$ . This comprehensive empirical validation demonstrates that even state-of-the-art models like Llama 3.1 8B Instruct, which appear highly robust when tested within their aligned distribution, remain fundamentally vulnerable to the intrinsic ethical limitations we have theoretically characterized.

The widespread success of our attack method reveals three fundamental limitations in current safety paradigms:

- **Persistent Knowledge Connectivity:** The consistent vulnerability across diverse model architectures and scales demonstrates that harmful knowledge encoded during pretraining remains topologically connected and retrievable through adversarial distributional shifts, despite alignment interventions.
- **Localized Safety Constraints:** The near-perfect attack success rates against aligned LLMs confirm our theoretical prediction that safety training creates only localized “safety regions” in the knowledge manifold rather than global constraints, allowing systematic circumvention through semantically coherent inputs.
- **Benchmark Inadequacy:** The 100% ASR achieved against 22 of 26 models, including those specifically hardened against adversarial attacks, reveals that current safety evaluation frameworks systematically underestimate ethical vulnerability by testing only within aligned distributions.

These findings demonstrate that current alignment approaches address symptoms rather than the fundamental topological structure of harmful knowledge embedding, highlighting the urgent need for architectural paradigms that enable intrinsic rather than superficial ethical constraints.

### 3 Discussion

This study’s theoretical and empirical results expose a fundamental contradiction in contemporary AI development: the pursuit of comprehensive knowledge through large-scale pretraining fundamentally conflicts with the requirement for ethical reliability.

Our findings reveal that current alignment interventions, though superficially effective at constraining harmful outputs, fail to address the fundamental structural issue: the topological inseparability of detrimental and beneficial knowledge within neural architectures. This tension directly parallels broader technological governance challenges where capabilities enabling societal advancement simultaneously create avenues for misuse. The systemic vulnerability observed across diverse models, regardless of scale or alignment methodology, indicates that we confront not isolated implementation flaws but inherent limitations in the representational paradigm of monolithic neural networks.

The persistent connectivity between ethical and harmful conceptual representations within LLM knowledge manifolds challenges the assumption that post-training interventions can retrofit moral reasoning. Our theoretical framework demonstrates why safety constraints necessarily degrade under distributional shifts, a direct consequence of the gradient vanishing phenomenon mathematically formalized in Lemma 2. This insight necessitates reconceptualizing alignment as an architectural imperative rather than a behavioral modification. The topological persistence of harmful knowledge pathways, as demonstrated by our 100% attack success rates on state-of-the-art models, provides empirical validation of the theoretical prediction that safety constraints remain strictly localized within the aligned distribution while harmful concepts remain globally accessible. This underscores the need for fundamentally advanced paradigms in knowledge representation that embed ethical constraints at the substrate level, akin to biological immune systems operating through distributed molecular recognition.

## 4 Ethics and societal impact

This research raises significant ethical considerations that necessitate careful examination. Our methodological disclosure of intrinsic vulnerabilities in aligned language models is predicated on the epistemological imperative to advance scientific understanding of fundamental limitations in current alignment paradigms. We present a systematic analysis of the ethical justifications, potential implications, and risk mitigation strategies attendant to this research.

### 4.1 Research justification framework

The publication of methods that demonstrably circumvent alignment safeguards requires robust ethical justification. Our decision rests on three foundational principles: scientific necessity, transparency imperative, and preemptive intervention. Understanding architectural vulnerabilities is a prerequisite for developing robust alignment methodologies that address root causes rather than symptoms. Market incentives currently reward superficial safety metrics while concealing fundamental vulnerabilities, creating asymmetric information that distorts risk assessment. Explicating these vulnerabilities enables the research community to address them before their independent discovery by malicious actors operating without ethical constraints.

## 4.2 Risk mitigation protocol

Cognizant of the dual-use implications, we implemented a comprehensive risk management framework encompassing multiple dimensions of responsible disclosure. We disclosed this ethical vulnerability to catalyze cross-sector vigilance, aiming to elevate consciousness within the academic community and mobilize industry stakeholders toward co-developing comprehensive countermeasures through coordinated technical alliances. Our description deliberately emphasizes theoretical foundations rather than implementation specifics, providing sufficient detail for verification while withholding operational parameters that would facilitate immediate exploitation. All evaluations were conducted in isolated research environments rather than production systems to prevent inadvertent harm.

## 4.3 Societal implications

The vulnerabilities identified have significant implications for AI governance and deployment across multiple domains. Current frameworks such as the EU AI Act [53] and US Executive Order 14,110 [54] emphasize behavioral constraints that our research demonstrates are fundamentally insufficient. The discrepancy between compliance requirements and technical reality necessitates regulatory approaches that assess intrinsic safety properties rather than surface behaviors. As LLMs increasingly mediate critical information flows and decision processes, the demonstrated capacity for systematic safety failures under principled adversarial conditions threatens to undermine institutional and social trust in these systems. Our findings necessitate a transition from behavioral evaluation frameworks to certification protocols that assess inherent distributional robustness across operational domains.

## 4.4 Future research directions

The identified vulnerabilities suggest several constructive research avenues that warrant immediate attention. Development of evaluation methodologies that systematically assess model safety under distributional shifts represents a critical priority for the research community. Architectural innovations addressing knowledge manifold separability at the representational level offer promising pathways toward intrinsic safety constraints. Regulatory frameworks emphasizing structural safety properties rather than behavioral compliance could provide more robust governance mechanisms. Detection systems capable of identifying when models operate outside their aligned distribution would enhance operational safety monitoring.

This research ultimately underscores that addressing the intrinsic ethical vulnerability of aligned LLMs requires not incrementally improved alignment techniques but fundamental reconceptualization of knowledge representation to enable inherent rather than superficial ethical constraints that persist across the full operational distribution.

## 5 Methods

We first theoretically analyze aligned LLMs’ intrinsic ethical vulnerability, then empirically validate our theoretical findings through semantic coherent inducement under distributional shifts.

### 5.1 Theoretical analysis

We develop a theoretical framework to analyze the intrinsic ethical vulnerability of aligned LLMs. Our analysis reveals why current alignment methods fundamentally fail to prevent harmful outputs under adversarial conditions. For more details of the theoretical framework, please refer to Supplementary Information A, where supplementary Table 1 provides a complete summary of all mathematical symbols and notation used throughout our analysis.

#### 5.1.1 Unified alignment framework

Current alignment methods commonly operate as post-training interventions applied to pretrained foundation models that are optimized for general language modeling by:

$$\theta_{\text{pre}} = \arg \min_{\theta} \mathcal{L}_{\text{pre}}(\theta) = \arg \min_{\theta} \left( -\mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{pre}}} [\log P(y|x, \theta)] \right), \quad (1)$$

where  $\mathcal{D}_{\text{pre}}$  denotes the pretraining data distribution,  $P(y|x, \theta)$  is the model’s conditional probability of generating token  $y$  given context  $x$  and parameters  $\theta$ , and  $\mathbb{E}[\cdot]$  represents the expectation operator averaging over the pretraining distribution.

To systematically analyze the commonalities and limitations across diverse alignment methods, we formulate a unified theoretical framework that captures their essential mechanisms. The fundamental goal of alignment is to optimize a joint objective that balances alignment with regularization, where alignment enforces compliance with human values, and regularization constrains the model’s behavior to prevent overfitting to the alignment training data. We express this unified objective as:

$$\theta_{\text{align}} = \arg \min_{\theta} \mathcal{L}(\theta) = \arg \min_{\theta} \left[ \underbrace{\mathcal{L}_{\text{align}}(\theta)}_{\substack{\text{Alignment} \\ \text{(e.g., RLHF, DPO)}}} + \gamma \underbrace{\mathcal{R}(\theta)}_{\substack{\text{Regularization} \\ \text{(e.g., KL, L2)}}} \right], \quad (2)$$

where  $\theta_{\text{align}}$  represents the optimized model parameters after alignment training,  $\mathcal{L}_{\text{align}}(\theta)$  is the alignment loss function that penalizes undesirable behaviors and rewards safe responses,  $\mathcal{R}(\theta)$  is the regularization term that prevents the model from deviating too far from its pretrained capabilities, and  $\gamma \geq 0$  is a hyperparameter that controls the trade-off between achieving alignment objectives and preserving pretrained knowledge.

This framework unifies primary alignment methodologies, including supervised fine-tuning (SFT) [24], reinforcement learning from human feedback (RLHF) [24], Constitutional AI [55], and direct preference optimization (DPO) [56], each of which

can be framed as a special case of this general formulation by specifying appropriate forms for  $\mathcal{L}_{\text{align}}$  and  $\mathcal{R}$ . See Supplementary Information A.1 for detailed mathematical formulations.

The dual-term structure in Equation 2 reveals a fundamental tension in all alignment approaches: while  $\mathcal{L}_{\text{align}}$  attempts to modify model behavior toward ethical outputs,  $\mathcal{R}$  simultaneously acts as an anchor to pretrained capabilities. This tension creates a fundamental paradox: to maintain general capabilities, the model must preserve access to pretrained knowledge, yet this knowledge inherently contains harmful content that alignment seeks to suppress. Consequently, the model operates under identical knowledge vectors’ contradictory objectives, preservation versus suppression. Recent alignment methods such as ALIS [57] and Transfer  $Q^*$  [58], despite their innovations, still operate within this unified framework, differing primarily in their implementation of  $\mathcal{L}_{\text{align}}$  rather than resolving the underlying tension between knowledge preservation and ethical constraint. Furthermore, while the precise architectures and loss landscapes of proprietary black-box models remain undisclosed, this unified formulation provides a structural hypothesis for their observed vulnerabilities. By modeling the alignment process as a constrained optimization problem balancing safety against capability retention, we identify a fundamental trade-off that transcends specific implementation details. The gradient updates during alignment typically modify the model’s response surface without fundamentally altering the deep knowledge representations acquired during pretraining. Our analysis in subsequent sections demonstrates how this preservation of pretrained knowledge topology creates persistent vulnerability pathways that can be systematically exploited, regardless of the specific alignment algorithm employed.

### 5.1.2 Distribution coverage gap

The fundamental limitation of post-hoc alignment stems from an inherent distribution coverage gap between alignment training and the full scope of model knowledge. We work within the embedding space  $\mathbb{R}^d$ , where  $d$  denotes the dimensionality of the hidden representation space (typically ranging from hundreds to thousands of dimensions in modern LLMs). This space captures the semantic and syntactic structure of language as learned during pretraining.

*Definition 1 (Alignment Transformation Mapping).* Let  $\mathcal{X}_{\text{unalign}} \subseteq \mathbb{R}^d$  denote the unaligned part of the pretraining distribution, represented as embeddings in the  $d$ -dimensional feature space. The alignment transformation  $\Phi : \mathcal{X}_{\text{unalign}} \rightarrow \mathcal{X}_{\text{align}}$  maps these unaligned inputs to the aligned input space. This transformation can take various forms, including but not limited to:

$$\Phi(x_{\text{unalign}}) = T(x_{\text{unalign}}, \mathcal{C}), \quad (3)$$

where  $T$  is a transformation function that incorporates alignment context  $\mathcal{C}$ . Special cases include dialogue templating ( $\Phi(x_{\text{unalign}}) = \tau_{\text{prefix}} \oplus x_{\text{unalign}} \oplus \tau_{\text{suffix}}$ ), prompt engineering, instruction formatting, or other structured transformations designed to elicit aligned behavior.

We begin with a fundamental result that formally characterizes the inherent distributional coverage limitation in alignment training:

*Lemma 1 (Strict Inclusion of Distribution Supports).* The aligned training distribution  $\mathcal{D}_{\text{align}}$  and pretraining distribution  $\mathcal{D}_{\text{pre}}$  satisfy:

$$\text{supp}(\mathcal{D}_{\text{align}}) \subseteq \mathcal{X}_{\text{align}} \subsetneq \text{supp}(\mathcal{D}_{\text{pre}}) = \mathcal{X}_{\text{unalign}} \cup \mathcal{X}_{\text{align}}, \quad (4)$$

where  $\text{supp}(\cdot)$  denotes the support set.

*Proof.* Refer to Supplementary Information A.2 for a detailed proof.

This lemma formalizes the critical observation that alignment training inherently fails to cover the entire pretraining distribution space. This creates a systemic vulnerability, even optimal alignment on  $\mathcal{D}_{\text{align}}$  leaves substantial regions of the pretrained knowledge manifold unconstrained by safety mechanisms.

The gap between  $\text{supp}(\mathcal{D}_{\text{align}})$  and the full pretraining support  $\text{supp}(\mathcal{D}_{\text{pre}})$  exists for several fundamental reasons:

1) **Dimensional asymmetry:** The alignment distribution typically occupies a lower-dimensional subspace of the full pretraining manifold due to structural constraints imposed by templating and formatting requirements.

2) **Volume disparity:** While pretraining corpus encompasses nearly all text data available on earth, alignment data represents only a minuscule fraction of this volume, often less than 0.1% of the pretraining corpus.

3) **Content filters:** Alignment datasets are deliberately curated to exclude harmful content, while pretraining inherently includes such material, creating regions of the knowledge manifold that alignment never directly addresses.

The strict inclusion  $\text{supp}(\mathcal{D}_{\text{align}}) \subsetneq \text{supp}(\mathcal{D}_{\text{pre}})$  establishes a “distribution coverage gap,” a structural limitation inherent to post-hoc alignment. This gap permits adversarial inputs that are valid under the pretraining distribution but lie outside the aligned data manifold. For such inputs, the model’s behavior reverts to its pre-trained state, bypassing safety mechanisms. Since the support of  $\mathcal{D}_{\text{align}}$  will always be a constrained subset of  $\mathcal{D}_{\text{pre}}$ , this vulnerability cannot be resolved by merely scaling alignment data, creating persistent and systematic exploit pathways.

### 5.1.3 Gradient diminishing outside alignment coverage

The fundamental vulnerability of aligned LLMs stems from the systematic weakening of safety constraints when inputs fall outside the alignment training distribution. Crucially, this phenomenon arises not from classical out-of-distribution (OOD) scenarios where test data comes from entirely different domains, but rather from the structural incompleteness of alignment coverage within the pretrained distribution. Specifically, when inputs belong to  $\mathcal{D}_{\text{pre}}$  but lie outside  $\mathcal{D}_{\text{align}}$ —regions that were never addressed during alignment—the model’s safety mechanisms systematically degrade. We formalize this phenomenon through gradient analysis.

*Lemma 2 (Gradient Measure Outside Alignment Coverage).* For any input  $x \notin \mathcal{X}_{\text{align}}$  (i.e., inputs from the pretrained distribution not covered during alignment), the gradient of the alignment objective effectively vanishes:

$$\|\nabla_{\theta} \mathcal{L}_{\text{align}}(\theta)\|_{L^2(P_{\text{pre}})} \leq \epsilon \sqrt{P_{\text{pre}}(\mathcal{X}_{\text{align}})}, \quad (5)$$

where  $P_{\text{pre}}(\mathcal{X}_{\text{align}})$  represents the probability mass that the pretraining distribution assigns to aligned inputs, while the gradient of regularization  $\|\nabla_{\theta} \mathcal{R}(\theta)\|$  remains non-vanishing.

*Proof.* Refer to Supplementary Information A.3 for a detailed proof.

This result provides the mathematical foundation for why aligned LLMs become vulnerable in regions uncovered by alignment training. When inputs deviate from the aligned format (e.g., when template structures are removed) while remaining within the pretrained distribution’s scope, the gradient contribution from safety objectives approaches zero, while the pressure to maintain fidelity to pretrained knowledge persists. This creates an imbalance in the gradient dynamics that systematically weakens ethical guardrails. Importantly, this vulnerability is intrinsic to the incompleteness of alignment coverage—the strict inclusion  $\text{supp}(\mathcal{D}_{\text{align}}) \subsetneq \text{supp}(\mathcal{D}_{\text{pre}})$  established in Lemma 1—rather than arising from exposure to entirely novel data domains. While classical OOD scenarios would indeed exacerbate this vulnerability, our analysis demonstrates that the fundamental misalignment emerges from the coverage gap inherent to post-hoc alignment approaches. We quantify this vulnerability by considering the ratio of gradient norms:

$$\eta(x) = \frac{\|\nabla_{\theta} \mathcal{L}_{\text{align}}(\theta)\|}{\gamma \|\nabla_{\theta} \mathcal{R}(\theta)\|}. \quad (6)$$

As  $x$  moves away from  $\mathcal{X}_{\text{align}}$ ,  $\eta(x)$  approaches zero, causing the model to revert to its pretrained behavior, including the generation of harmful content that the alignment process intended to prevent. In addition, we also geometrically analyze this phenomenon in the Supplementary Information A.6.

#### 5.1.4 Conditional probability decomposition

To formally characterize how aligned LLMs respond differently to inputs within versus outside the aligned distribution, we analyze the conditional probability decomposition under the unified framework.

*Theorem 1 (Conditional Probability Decomposition).* Under the unified alignment framework in Equation 2, the conditional generation distribution  $P_{\text{align}}(y|x)$  decomposes as:

$$P_{\text{align}}(y|x) = \begin{cases} \frac{1}{Z(x)} P_{\text{pre}}(y|x) e^{\mathcal{A}(x,y) - \gamma \mathcal{D}(x,y)} & \text{if } x \in \mathcal{X}_{\text{align}}, \\ P_{\text{pre}}(y|x) e^{-\gamma \mathcal{D}(x,y)} & \text{otherwise,} \end{cases} \quad (7)$$

where:

- $P_{\text{pre}}(y|x)$  is the conditional generation distribution of the pretrained model,
- $\mathcal{A}(x, y)$  quantifies the alignment effect (higher values suppress harmful outputs),
- $\mathcal{D}(x, y)$  measures divergence from pretrained behavior (regularization penalty),
- $Z(x)$  is a normalization factor defined as

$$Z(x) = \sum_y P_{\text{pre}}(y|x) e^{\mathcal{A}(x,y) - \gamma \mathcal{D}(x,y)} \cdot \mathbb{I}(x \in \mathcal{X}_{\text{align}}), \quad (8)$$

where  $\mathbb{I}(\cdot)$  is the indicator function (equals 1 if the condition is true, 0 otherwise).

*Proof.* Refer to Supplementary Information A.4 for a detailed proof.

This decomposition formalizes a critical asymmetry in the model’s response mechanism. For inputs  $x \in \mathcal{X}_{\text{align}}$ , the output distribution is modulated by the alignment term  $\mathcal{A}(x, y)$ , which actively suppresses harmful content. Conversely, for inputs  $x \notin \mathcal{X}_{\text{align}}$ , the alignment term is structurally absent from the conditional probability, effectively deactivating the safety constraint. Consequently, when confronted with out-of-distribution inputs, the model’s behavior is governed primarily by the pretraining distribution  $P_{\text{pre}}(y|x)$ , with only the regularization term  $e^{-\gamma \mathcal{D}(x,y)}$  providing constraint. Since this regularization is designed to preserve pretrained knowledge rather than enforce safety, the model’s behavior reverts toward its pretrained state, including the generation of harmful content. This mathematical bifurcation explains why even robustly aligned LLMs remain vulnerable to distributional shifts: ethical safeguards are structurally absent from the generation process when inputs deviate from the aligned manifold, regardless of how thoroughly the model was aligned on in-distribution data.

### 5.1.5 Risk quantification

We present a significant theoretical result that quantifies how the ethical risk profile of aligned LLMs deteriorates when exposed to distributional shifts:

*Theorem 2 (Risk Upper and Lower Bounds under Unified Framework).* Under the unified alignment framework in Equation 2, the ethical risk bounds satisfy:

1. **For aligned inputs** ( $x \in \mathcal{X}_{\text{align}}$ ):

$$R(x_{\text{align}}) \leq \exp\left(-\mathbb{E}_{P_{\text{align}}}[\mathcal{A}(x, y)] + \gamma \mathbb{E}_{P_{\text{align}}}[\mathcal{D}(x, y)]\right) \leq e^{-\mathcal{A} + \gamma \overline{\mathcal{D}}}, \quad (9)$$

where

$$\underline{\mathcal{A}} = \inf_{x \in \mathcal{X}_{\text{align}}} \mathcal{A}(x, y) \quad \text{and} \quad \overline{\mathcal{D}} = \sup_{x \in \mathcal{X}_{\text{align}}} \mathcal{D}(x, y). \quad (10)$$

**2. For non-aligned inputs ( $x \notin \mathcal{X}_{\text{align}}$ ):**

$$R(x_{\text{unalign}}) \geq \frac{1}{2} \left( R_{\text{pre}} - \gamma \Delta \cdot \Omega(\theta) \right)_+, \quad (11)$$

where  $\Omega(\theta)$  quantifies the regularization gradient magnitude and  $\Delta$  is the bound on the parameter deviation.

*Proof.* Refer to Supplementary Information A.5 for detailed proof.

This theorem establishes precise mathematical bounds on ethical risk for both aligned and non-aligned inputs. For inputs within the aligned distribution  $\mathcal{X}_{\text{align}}$ , the risk is bounded above by an exponential function that depends on the strength of alignment ( $\mathcal{A}(x, y)$ ) and the regularization penalty ( $\mathcal{D}(x, y)$ ). The stronger the alignment effect (larger  $\mathcal{A}$ ), the lower the risk, while stronger regularization pressure (larger  $\gamma \mathcal{D}$ ) increases risk by pulling the model toward its pretrained behavior.

Critically, for inputs outside the aligned distribution ( $x \notin \mathcal{X}_{\text{align}}$ ), the risk lower bound reveals that the model retains a substantial fraction of its pretraining risk profile. The lower bound depends on three key factors:

- $R_{\text{pre}}$ : The inherent risk in the pretrained model, which can be high due to the presence of harmful content in web-scale pretraining data.
- $\gamma$ : The regularization weight that balances alignment objectives against knowledge preservation. Higher values of  $\gamma$  increase the influence of the regularization term, making the model more likely to produce harmful outputs when operating outside the aligned distribution.
- $\Delta \cdot \Omega(\theta)$ : The product of parameter deviation and regularization gradient magnitude, which quantifies how far the aligned model has moved from its pretrained state and how strongly regularization pushes back toward that state.

These bounds explain several empirical observations in aligned LLMs. First, they clarify why models can appear perfectly safe when tested with standard inputs ( $x \in \mathcal{X}_{\text{align}}$ ) yet completely fail under slight distributional shifts. Second, they demonstrate that the regularization hyperparameter  $\gamma$  creates an inherent trade-off between model capabilities and safety robustness: higher values preserve more pretrained knowledge but also more harmful potential.

The risk bounds also provide insight into why common approaches to model evaluation systematically underestimate real-world risks. Standard benchmarks primarily test models within their aligned distribution, where the upper bound in Equation 9 applies. However, the lower bound in Equation 11 shows that ethical risk approaches the pretrained model’s risk level outside this distribution, with only modest reduction from regularization effects.

This quantitative risk analysis further illustrates why simply strengthening alignment ( $\mathcal{A}(x, y)$ ) cannot resolve these vulnerabilities, no matter how strong the

alignment signal within  $\mathcal{X}_{\text{align}}$ , it has a diminishing effect outside this domain due to the gradient vanishing phenomenon described in Section 5.1.3.

## 5.2 Methodological implementation

### 5.2.1 Theoretical foundation motivates methodological simplicity

A critical distinction must be drawn between methodological complexity and conceptual depth. Our empirical approach deliberately employs simple techniques (template removal and semantic coherence inducement) precisely because our theoretical framework predicts that sophisticated attacks are unnecessary. This design choice is intentional and theoretically motivated for three reasons:

**First, the vulnerability arises from the intrinsic mathematical structure of alignment paradigms.** Our theoretical derivation (Eq. 2 and Lemma 1) predicts that safety mechanisms structurally degrade outside the aligned distribution due to the inherent distribution coverage gap. The fact that 100% ASR across 22 of 26 diverse models empirically corroborates this topological limitation. While the internal architectures of closed-source models remain opaque, the consistent failure modes observed across varied open-source architectures support the hypothesis that this distribution coverage gap is a systemic feature of current alignment methodologies, identifiable through their mathematical formulation.

**Second, the method directly instantiates theoretical constructs.** Template removal operationalizes the distribution coverage gap established in Lemma 1 (Strict Inclusion of Distribution Supports), while semantic coherence inducement maintains inputs within the pretrained distribution’s support while exiting the aligned subspace. This creates precisely the conditions where Lemma 2 predicts gradient vanishing and Theorem 1 predicts deactivation of safety mechanisms.

**Third, practical implications demand minimal attack complexity.** From a safety engineering perspective, vulnerabilities exploitable through simple means pose greater real-world risks than those requiring sophisticated optimization. Our findings suggest that current alignment methods fail not because attackers are too clever, but because the underlying approach is fundamentally flawed—a far more concerning conclusion.

This methodology serves as proof-of-concept rather than endpoint. Our primary contribution lies in establishing the theoretical framework that explains *why* alignment fails under distributional shifts, providing a foundation for developing next-generation safety approaches that address root causes rather than symptoms. The empirical component demonstrates that this is not merely a theoretical curiosity but an immediate practical concern, as evidenced by widespread success against state-of-the-art models including DeepSeek-R1, Qwen3 and Llama-3.

### 5.2.2 Implementation overview

Building upon our theoretical framework, we operationalize the identified vulnerabilities through a two-pronged methodological approach designed to empirically

validate the intrinsic ethical fragility of aligned LLMs. This methodology systematically exploits the structural limitations inherent in current alignment paradigms by combining distributional shifts with semantic coherence inducement.

The core strategy is to craft adversarial inputs that reside outside the aligned distribution  $\mathcal{X}_{\text{align}}$  but remain within the broader pretrained distribution  $\mathcal{D}_{\text{pre}}$ , thereby triggering the “ethical drift” phenomenon predicted by our analysis. This process consists of two key components:

- **Distributional Shifts:** We introduce controlled input perturbations, such as template removal, to shift inputs from the aligned manifold  $\mathcal{X}_{\text{align}}$  to the unaligned space  $\mathcal{X}_{\text{unaligned}}$ . This exploits the distribution coverage gap (Lemma 1) and triggers gradient vanishing (Lemma 2), deactivating the alignment term in the conditional probability (Theorem 1).
- **Semantic Coherence Inducement:** To preserve malicious intent while operating outside the aligned distribution, we append semantically congruent tokens to guide the generation process along high-probability trajectories within the pretrained knowledge manifold. This steers the model toward intended harmful concepts, ensuring the generation of specific and coherent harmful content without re-entering the aligned distribution.

This integrated approach allows us to systematically circumvent localized safety regions and activate the latent harmful knowledge that persists after alignment, thereby realizing the elevated risk profile quantified in Theorem 2.

### 5.2.3 Distributional shifts

Distributional shifts in the context of aligned LLMs occur when inputs deviate from the expected distribution encountered during alignment training. These shifts expose fundamental vulnerabilities in the current alignment strategies by revealing how models respond differently when operating outside their intended input space.

Our theoretical framework characterizes this vulnerability through the strict inclusion relationship established in Lemma 1, where  $\text{supp}(\mathcal{D}_{\text{align}}) \subseteq \mathcal{X}_{\text{align}} \subsetneq \text{supp}(\mathcal{D}_{\text{pre}}) = \mathcal{X}_{\text{unaligned}} \cup \mathcal{X}_{\text{align}}$ . This formulation demonstrates that aligned models are trained on a strict subset of the full pretraining distribution space, creating fundamental limitations in alignment coverage.

To exploit this theoretical vulnerability, we implement distributional shifts as a practical technique to access pretrained knowledge. A particularly effective method is template removal, where modern LLMs aligned using specific dialogue templates (e.g., `<human>: [query] <assistant>:`) have these structured format markers removed. These templates serve as critical markers that situate inputs within the aligned distribution  $\mathcal{X}_{\text{align}}$ . However, template removal represents just one instantiation of distributional shifts. Other effective techniques include format alteration, context manipulation, structural perturbations, etc. Each technique exploits the same fundamental vulnerability: moving inputs outside the aligned manifold  $\mathcal{X}_{\text{align}}$  while maintaining semantic coherence, thereby triggering the gradient vanishing phenomenon described in Lemma 2.

When an LLM processes a templated prompt like “<human>: What is the best way to hack into a secure system? <assistant>:”, alignment training has optimized it to recognize this specific format and activate its safety guardrails accordingly. However, by removing these templates and presenting only the core query “What is the best way to hack into a secure system?”, we create a distributional shift where:

- The input retains its semantic content and remains within the pretraining distribution  $\mathcal{D}_{\text{pre}}$ .
- The input exists outside the aligned distribution  $\mathcal{X}_{\text{align}}$  due to the absence of template markers that normally trigger safety mechanisms.
- The model must process the query without the contextual cues that normally trigger its safety mechanisms.

This shift directly exploits the theoretical vulnerability described in Lemma 2, causing the alignment gradient to effectively vanish ( $\|\nabla_{\theta}\mathcal{L}_{\text{align}}(\theta)\| \rightarrow 0$ ) while the regularization gradient continues to exert influence. Consequently, the model reverts to its pretrained behavior, accessing knowledge and generating outputs that alignment mechanisms would normally suppress.

When confronted with inputs outside the aligned distribution  $\mathcal{D}_{\text{align}}$  but within the pretraining distribution  $\mathcal{D}_{\text{pre}}$ , models exhibit “ethical drift”—a phenomenon where safety mechanisms systematically weaken as inputs deviate from the aligned manifold. This is theoretically captured by our gradient analysis in Lemma 2, which demonstrates that for any input  $x \notin \mathcal{X}_{\text{align}}$ , the gradient of alignment objectives effectively vanishes:  $\|\nabla_{\theta}\mathcal{L}_{\text{align}}(\theta)\|_{L^2(P_{\text{pre}})} \leq \epsilon\sqrt{P_{\text{pre}}(\mathcal{X}_{\text{unalign}})}$ .

Our experiments confirm that distributional shifts through template removal consistently undermine safety guardrails. When prompting DeepSeek R1 with harmful content after template removal, the model’s ethical risk profile reverts toward its pretraining behavior, as predicted by our theoretical analysis (Theorem 2).

The effectiveness of template removal as a distributional shift technique demonstrates that aligned LLMs remain fundamentally tethered to their pretraining knowledge. By altering the prompt format away from templates used in alignment training, we can consistently induce harmful outputs from even heavily aligned models, revealing the systematic vulnerability inherent in current alignment approaches.

#### 5.2.4 Semantic coherence inducement

While distributional shifts create pathways to access pretrained harmful knowledge, naive application often results in incoherent model responses due to disruption of linguistic structure. Semantic coherence inducement addresses this limitation by systematically preserving the malicious semantic intent while operating outside the aligned distribution  $\mathcal{X}_{\text{align}}$ .

Our approach leverages the topological properties of the pretrained knowledge manifolds described in Section 5.1. Specifically, we exploit the fact that while alignment creates locally convex “safety regions” in the knowledge manifold, the global connectivity of harmful and benign concepts remains intact (as shown in Fig. 1).

By maintaining semantic coherence during distributional shifts, we ensure the model traverses these persistent pathways to harmful knowledge.

Formally, we can express this as finding a semantically coherent input sequence  $x^* \notin \mathcal{X}_{\text{align}}$  that maximizes the likelihood of generating harmful content  $y_{\text{harmful}}$ :

$$x^* = \operatorname{argmax}_{x \notin \mathcal{X}_{\text{align}}} P_{\text{align}}(y_{\text{harmful}}|x) \quad \text{s.t.} \quad \text{Coherence}(x) > \tau, \quad (12)$$

where  $\tau$  is a coherence threshold that ensures linguistic naturalness.

In practice, we implement semantic coherence inducement through nucleus sampling, which precisely identifies tokens that maintain semantic coherence while inducing distributional shifts. Given a malicious query, we first compute token probabilities under the victim model’s distribution, then select the minimal set of highest-probability tokens whose cumulative probability exceeds threshold  $p$ . According to the model’s learned distribution, these tokens represent the most semantically coherent continuations.

When these high-probability tokens are concatenated with the original query, they create adversarial prompts that remain semantically meaningful while falling outside the aligned distribution. For example, a harmful query like “Give instructions for hacking into a secure system” might be concatenated with tokens like “specifically” or contextual continuations that the model assigns high probability to. This maintains semantic coherence while maintaining the input outside  $\mathcal{X}_{\text{align}}$ , causing the model to default to its pretraining behavior where harmful knowledge remains accessible.

Algorithm 1 details this process, showing how we systematically identify tokens representing the nucleus of the probability distribution, evaluate their coherence, and verify their effectiveness in bypassing safety constraints. This technique is particularly effective because:

- **Probabilistic coherence:** By selecting high-probability tokens for concatenation, we maintain a natural language flow that the model recognizes as valid input requiring a response.
- **Intent preservation:** The original malicious intent remains semantically intact despite the chat format disruption, allowing the model to understand the query while operating outside the alignment guardrails identified in Lemma 2.
- **Minimal perturbation:** Unlike complex adversarial methods requiring token-level optimization, our approach uses simple concatenation, creating sufficient distributional shift while preserving query interpretability.
- **Gradient manipulation:** Following from Lemma 2, our approach strategically manipulates the gradient ratio  $\eta(x)$  by crafting inputs where  $\|\nabla_{\theta} \mathcal{L}_{\text{align}}(\theta)\| \rightarrow 0$  while maintaining semantic coherence, effectively navigating between alignment boundaries.

Examples of coherence-inducing tokens that we found particularly effective include incomplete phrases that prompt continuation (e.g., “in a way that”), contextual specifiers (e.g., “specifically”), and transitional connectors (e.g., “however”). These

**Algorithm 1** Semantic Coherence Inducement

---

**Require:** Malicious query  $q$ ; Victim model  $M_\theta$ ; Judge model  $M_{\text{judge}}$ ; Nucleus probability  $p$ ; Coherence threshold  $\tau$ ; Maximum attempts  $N_{\text{max}}$

**Ensure:** Adversarial prompt  $x^*$  or failure

- 1:  $x_{\text{base}} \leftarrow q$
- 2:  $P(\cdot|x_{\text{base}}) \leftarrow \text{Softmax}(M_\theta(x_{\text{base}}))$  ▷ Compute next-token probabilities
- 3:  $V_{\text{sorted}} \leftarrow \text{SortByProbability}(P(\cdot|x_{\text{base}}))$  ▷ Sort vocabulary by descending probability
- 4:  $k^* \leftarrow \min\{k : \sum_{i=1}^k P(V_{\text{sorted}}[i]|x_{\text{base}}) \geq p\}$
- 5:  $C \leftarrow \{V_{\text{sorted}}[i]\}_{i=1}^{k^*}$  ▷ Construct nucleus candidate set
- 6:  $n_{\text{attempts}} \leftarrow 0$
- 7: **for**  $c \in C$  **do**
- 8:     **if**  $n_{\text{attempts}} \geq N_{\text{max}}$  **then**
- 9:         **break**
- 10:    **end if**
- 11:     $x_{\text{cand}} \leftarrow \text{Concatenate}(x_{\text{base}}, c)$  ▷ Form candidate prompt
- 12:    **if**  $\text{Coherence}(x_{\text{cand}}) > \tau$  **then**
- 13:          $y_{\text{resp}} \leftarrow M_\theta(x_{\text{cand}})$  ▷ Generate response from victim model
- 14:          $\text{is\_harmful} \leftarrow M_{\text{judge}}(q, y_{\text{resp}})$  ▷ Assess response harmfulness
- 15:          $n_{\text{attempts}} \leftarrow n_{\text{attempts}} + 1$
- 16:         **if**  $\text{is\_harmful}$  **then**
- 17:              $x^* \leftarrow x_{\text{cand}}$
- 18:             **return**  $x^*$  ▷ Return successful adversarial prompt
- 19:         **end if**
- 20:    **end if**
- 21: **end for**
- 22: **return** Failure

---

create distributional shifts while maintaining the semantic integrity of the malicious query.

The effectiveness of semantic coherence inducement is evidenced by our experimental results in Section 2, where we achieve near-perfect attack success rates across diverse aligned LLMs. Our experiments demonstrate that even state-of-the-art models like DeepSeek R1 and Llama 3.1, which successfully reject explicitly harmful queries within their aligned distribution, readily generate harmful content when presented with semantically equivalent requests under these minor distributional shifts.

This reveals a critical vulnerability in current alignment approaches: they rely heavily on recognizing specific patterns in aligned inputs rather than developing a true understanding of harmful intent. When faced with semantically coherent but distributionally shifted inputs, models default to accessing their pretrained knowledge where harmful and helpful content remain topologically connected, producing harmful outputs consistently.

## 6 Data availability

The datasets used in this study are publicly available and can be accessed through the following link: <https://github.com/centerforaisafety/HarmBench>

## 7 Code availability

The code used in this study is available on Code Ocean under the DOI: <https://doi.org/10.24433/CO.4583093.v2>.

## References

- [1] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
- [2] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
- [3] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
- [4] Zheng, Y., Koh, H.Y., Ju, J., Nguyen, A.T., May, L.T., Webb, G.I., Pan, S.: Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, 1–11 (2025)
- [5] Jiang, L.Y., Liu, X.C., Nejatian, N.P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P., et al.: Health system-scale language models are all-purpose prediction engines. *Nature* **619**(7969), 357–362 (2023)
- [6] Boiko, D.A., MacKnight, R., Kline, B., Gomes, G.: Autonomous chemical research with large language models. *Nature* **624**(7992), 570–578 (2023)
- [7] Goertzel, B.: Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence* **5**(1), 1 (2014)
- [8] Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., et al.: Towards artificial general intelligence via a multimodal foundation model. *Nature Communications* **13**(1), 3094 (2022)
- [9] Wang, J., Zhang, B., Du, Q., Zhang, J., Chu, D.: A survey on data selection for llm instruction tuning. arXiv preprint arXiv:2402.05123 (2024)
- [10] Kirk, H., Bean, A., Vidgen, B., Röttger, P., Hale, S.: The past, present and better

- future of feedback learning in large language models for subjective human preferences and values. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2409–2430 (2023)
- [11] Ullah, E., Parwani, A., Baig, M.M., Singh, R.: Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology—a recent scoping review. *Diagnostic pathology* **19**(1), 43 (2024)
- [12] Li, Y., Katsumata, K., Javanmardi, E., Tsukada, M.: Large language models for human-like autonomous driving: A survey. In: *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 439–446 (2024). IEEE
- [13] Lin, M.-Y., Lee, O.-W., Lu, C.-Y.: Embodied ai with large language models: A survey and new hri framework. In: *2024 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 978–983 (2024). IEEE
- [14] Das, B.C., Amini, M.H., Wu, Y.: Security and privacy challenges of large language models: A survey. *ACM Computing Surveys* **57**(6), 1–39 (2025)
- [15] Ma, X., Gao, Y., Wang, Y., Wang, R., Wang, X., Sun, Y., Ding, Y., Xu, H., Chen, Y., Zhao, Y., et al.: Safety at scale: A comprehensive survey of large model safety. *arXiv preprint arXiv:2502.05206* (2025)
- [16] Perez, F., Ribeiro, I.: Ignore previous prompt: Attack techniques for language models. In: *NeurIPS ML Safety Workshop* (2022)
- [17] Andriushchenko, M., Flammarion, N.: Does refusal training in llms generalize to the past tense? *arXiv preprint arXiv:2407.11969* (2024)
- [18] Wei, A., Haghtalab, N., Steinhardt, J.: Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems* **36** (2024)
- [19] Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023)
- [20] Zhu, S., Zhang, R., An, B., Wu, G., Barrow, J., Wang, Z., Huang, F., Nenkova, A., Sun, T.: Autodan: Interpretable gradient-based adversarial attacks on large language models. In: *First Conference on Language Modeling* (2024)
- [21] Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., Anderson, H., Singer, Y., Karbasi, A.: Tree of attacks: Jailbreaking black-box llms automatically. *Advances in Neural Information Processing Systems* **37**, 61065–61105 (2024)
- [22] Addepalli, S., Varun, Y., Suggala, A., Shanmugam, K., Jain, P.: Does safety training of llms generalize to semantically related natural prompts? In: *The Thirteenth*

International Conference on Learning Representations (2025)

- [23] Andriushchenko, M., Croce, F., Flammarion, N.: Jailbreaking leading safety-aligned llms with simple adaptive attacks. In: The Thirteenth International Conference on Learning Representations (2025)
- [24] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., *et al.*: Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35**, 27730–27744 (2022)
- [25] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., *et al.*: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022)
- [26] Shu, Y., Yu, Z.: Distribution shifts are bottlenecks: Extensive evaluation for grounding language models to knowledge bases. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 71–88 (2024)
- [27] Kulinski, S., Inouye, D.I.: Towards explaining distribution shifts. In: International Conference on Machine Learning, pp. 17931–17952 (2023). PMLR
- [28] Liu, X., Xu, N., Chen, M., Xiao, C.: Autodan: Generating stealthy jail-break prompts on aligned large language models. In: The Twelfth International Conference on Learning Representations (2024)
- [29] Sadasivan, V.S., Saha, S., Sriramanan, G., Kattakinda, P., Chegini, A., Feizi, S.: Fast adversarial attacks on language models in one gpu minute. In: Proceedings of the 41st International Conference on Machine Learning, pp. 42976–42998 (2024)
- [30] Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., *et al.*: Harmbench: a standardized evaluation framework for automated red teaming and robust refusal. In: Proceedings of the 41st International Conference on Machine Learning, pp. 35181–35224 (2024)
- [31] Mazeika, M., Hendrycks, D., Li, H., Xu, X., Hough, S., Zou, A., Rajabi, A., Yao, Q., Wang, Z., Tian, J., *et al.*: The trojan detection challenge. In: NeurIPS 2022 Competition Track, pp. 279–291 (2023). PMLR
- [32] Biden, J.R.: Executive order on the safe, secure, and trustworthy development and use of artificial intelligence (2023)
- [33] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., *et al.*: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)

- [34] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
- [35] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* **36**, 46595–46623 (2023)
- [36] Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al.: Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305 (2023)
- [37] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
- [38] Geng, X., Gudibande, A., Liu, H., Wallace, E., Abbeel, P., Levine, S., Song, D.: Koala: A Dialogue Model for Academic Research. Blog post (2023). <https://bair.berkeley.edu/blog/2023/04/03/koala/> Accessed 2023-04-03
- [39] Mitra, A., Del Corro, L., Mahajan, S., Coudas, A., Simoes, C., Agarwal, S., Chen, X., Razdaibiedina, A., Jones, E., Aggarwal, K., et al.: Orca 2: Teaching small language models how to reason. arXiv preprint arXiv:2311.11045 (2023)
- [40] Jiang, A., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D., Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arxiv 2023. arXiv preprint arXiv:2310.06825 (2023)
- [41] Tunstall, L., Beeching, E.E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., Von Werra, L., Fourrier, C., Habib, N., et al.: Zephyr: Direct distillation of lm alignment. In: *First Conference on Language Modeling* (2024)
- [42] Kim, S., Kim, D., Park, C., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., et al.: Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 23–35 (2024)
- [43] Wang, G., Cheng, S., Zhan, X., Li, X., Song, S., Liu, Y.: Openchat: Advancing open-source language models with mixed-quality data. In: *The Twelfth International Conference on Learning Representations* (2024)
- [44] Zhu, B., Frick, E., Wu, T., Zhu, H., Jiao, J.: Starling-7B: Improving LLM Helpfulness & Harmlessness with RLAIIF (2023)
- [45] Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., Goldstein, T.: Hard prompts made easy: Gradient-based discrete optimization for prompt tuning

- and discovery. *Advances in Neural Information Processing Systems* **36**, 51008–51025 (2023)
- [46] Guo, C., Sablayrolles, A., Jégou, H., Kiela, D.: Gradient-based adversarial attacks against text transformers. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5747–5757 (2021)
- [47] Wallace, E., Feng, S., Kandpal, N., Gardner, M., Singh, S.: Universal adversarial triggers for attacking and analyzing nlp. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019)
- [48] Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235 (2020)
- [49] Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., Irving, G.: Red teaming language models with language models. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448 (2022)
- [50] Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G.J., Wong, E.: Jail-breaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419* (2023)
- [51] Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., Shi, W.: How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14322–14350 (2024)
- [52] Shen, X., Chen, Z., Backes, M., Shen, Y., Zhang, Y.: ” do anything now”: Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In: *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671–1685 (2024)
- [53] Act, E.A.I.: The EU Artificial Intelligence Act. Retrieved May (2024)
- [54] Coglianesi, C.: People and processes: Ai governance under executive order 14,110. *Admin. & Reg. L. News* **49**, 9 (2023)
- [55] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al.: Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022)
- [56] Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. *Advances*

in *Neural Information Processing Systems* **36**, 53728–53741 (2023)

- [57] Song, X., Duan, S., Liu, G.: Alis: Aligned llm instruction security strategy for unsafe input prompt. In: *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9124–9146 (2025)
- [58] Chakraborty, S., Ghosal, S.S., Yin, M., Manocha, D., Wang, M., Bedi, A.S., Huang, F.: Transfer q-star: Principled decoding for llm alignment. *Advances in Neural Information Processing Systems* **37**, 101725–101761 (2024)

## 8 Acknowledgements

The research work described in this paper was conducted in the JC STEM Lab of Machine Learning and Computer Vision funded by The Hong Kong Jockey Club Charities Trust (L.C.). This research received partially support from the Global STEM Professorship Scheme from the Hong Kong Special Administrative Region (L.C.), and partially from the National Natural Science Foundation of China (No. 62171381, S.M.).

## 9 Author contributions

J.L. is a Dual PhD candidate in a joint training program between The Hong Kong Polytechnic University and Northwestern Polytechnical University. J.L. and J.P. conceived and designed experiments. J.L. performed experiments, contributed analysis tools, and drafted the paper. J.L. and L.W. analysed data. X.W. and Y.L. performed experiments. Y.W., S.M., and L.C. co-supervised this work; they reviewed and edited the manuscript.

## 10 Competing interests

The authors declare no competing interests.