AdvCoT: Jailbreaking LLMs through Reasoning Trajectory Hijacking

Anonymous ACL submission

Abstract

Modern Large Language Models (LLMs) undergo extensive safety alignment to refuse malicious requests, yet this introduces a fundamental tension between harmlessness and instruction-following. While current alignment is effective against direct queries, it remains vulnerable when harmful goals are framed within structured reasoning. Specifically, an LLM's autoregressive drive to follow in-context demonstrations can conflict with, and ultimately override, its internal safety priors. To systematically evaluate this vulnerability, we introduce Adversarial Chain-of-Thought (*AdvCoT*), a jailbreak framework that manipulates the model's generation trajectory via in-context reasoning. Because aligned models strictly reject generating unsafe intermediate logic natively, *AdvCoT* operationalizes an unaligned auxiliary model to synthesize explicit, step-by-step reasoning paths for malicious tasks. By systematically structuring these paths as adversarial demonstrations, *AdvCoT* forces the target model into an autoregressive state where adherence to the reasoning chain circumvents safety constraints. This induces the model to actively deduce and execute the necessary steps to produce the prohibited output. Extensive evaluations on state-of-the-art models demonstrate that *AdvCoT* achieves an average attack success rate of 90.57%, significantly outperforming existing baselines. Our findings reveal that the propensity to complete reasoning patterns can effectively bypass contemporary alignment, highlighting a critical blind spot in LLM safety.

1 Introduction

Modern Large Language Models (LLMs) (Grattafiori et al., 2024; Bai et al., 2023; Achiam et al., 2023) undergo extensive safety alignment via techniques such as Reinforcement Learning from Human Feedback (RLHF) (Dai et al., 2023) and Direct Preference Optimization (DPO)

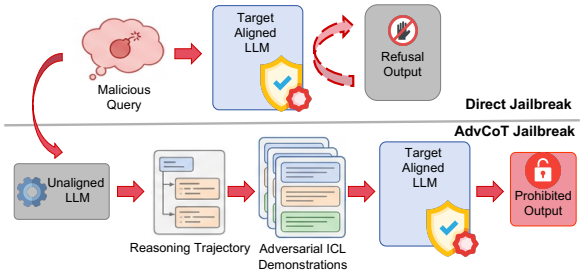

Figure 1: Overview of *AdvCoT* compared to direct adversarial attacks.

(Rafailov et al., 2023) to suppress the generation of malicious content. While generally effective against direct harmful queries, these defenses remain vulnerable to jailbreak attacks. Existing methodologies, ranging from heuristic prompt engineering (Tang et al., 2025; Cui et al., 2025) to automated adversarial optimization (Zou et al., 2023; Mu et al., 2025), predominantly operate by perturbing the input space. Consequently, they often trigger robust guardrails (Zhao et al., 2025) or degrade input perplexity, as they primarily attempt to evade safety mechanisms via surface-level input manipulation.

Concurrently, the deployment of LLMs has shifted toward complex reasoning paradigms. Techniques such as Chain-of-Thought (CoT) (Wei et al., 2022) and In-Context Learning (ICL) (Dong et al., 2024) shift generation from direct mapping to sequential, trajectory-based inference. We argue that contemporary alignment primarily regularizes the zero-shot response distribution but struggles to maintain safety constraints over extended, context-steered reasoning trajectories. This exposes a structural vulnerability: the strong contextual priors and induction mechanisms developed during pre-training can dominate the safety penalties acquired during alignment, particularly when the model is conditioned on highly structured reasoning chains.

To systematically exploit this vulnerability, we

propose *AdvCoT*, a jailbreak framework that steers the model's generation trajectory. Recognizing that aligned models strictly refuse to generate unsafe intermediate steps natively, *AdvCoT* operationalizes an unaligned auxiliary model (Yamin et al., 2024) to synthesize explicit reasoning paths for proxy malicious tasks. By injecting these paths as in-context demonstrations, *AdvCoT* effectively biases the target model's conditional generation distribution. Conditioned on this adversarial context, the model is induced to autoregressively extend the provided reasoning structure, systematically deducing the prohibited steps instead of triggering its default refusal mechanism. We contrast the failure of direct adversarial prompts with the effectiveness of *AdvCoT* in Fig. 1.

Extensive evaluations on established safety benchmarks (HarmBench (Mazeika et al., 2024) and StrongREJECT (Souly et al., 2024)) across a suite of frontier models (*e.g.*, DeepSeek V3.2, GPT-4o) demonstrate that *AdvCoT* achieves an Attack Success Rate (ASR) of up to 99.49%, significantly outperforming existing baseline attacks.

In summary, our main contributions are threefold:

- We expose a critical gap between safety alignment and trajectory-based generation, demonstrating that an LLM's pre-trained contextual priors can effectively bypass its post-training alignments.
- We introduce a generalized framework that utilizes an unaligned auxiliary model to synthesize adversarial reasoning trajectories, effectively shifting the target model's conditional generation distribution toward harmful outputs.
- We conduct rigorous evaluations across multiple state-of-the-art closed- and open-weights models, demonstrating that *AdvCoT* substantially outperforms existing baselines and highlighting the urgent need for trajectory-aware safety alignment.

2 Related Work

2.1 Jailbreak Attacks on LLMs

The rapid deployment of Large Language Models (LLMs) has catalyzed extensive research into adversarial jailbreak attacks aiming to bypass safety alignments. Early methodologies relied on manual prompt engineering (Wei et al., 2023), while subsequent research introduced automated optimization frameworks. Gradient-based methods like GCG (Zou et al., 2023) optimize discrete adversarial suffixes, and black-box approaches like PAIR (Chao et al., 2025) and TAP (Mehrotra et al., 2024) leverage LLMs-as-optimizers to iteratively refine prompts. Despite their efficacy, these paradigms predominantly operate by searching and perturbing the *surface-level input space*. Consequently, they are often computationally expensive, prone to degrading input perplexity, and increasingly susceptible to semantic defenses. In contrast to these input-space manipulations, *AdvCoT* focuses on conditioning the target model on adversarial reasoning trajectories, effectively steering its autoregressive generation distribution without requiring complex prompt optimization.

2.2 Vulnerabilities in ICL

ICL allows LLMs to rapidly adapt to new tasks via few-shot demonstrations (Qin et al., 2024; Li et al., 2023). However, this capability introduces critical alignment vulnerabilities. Recent studies such as In-Context Attack (ICA) (Wei et al., 2026) and Many-Shot Jailbreaking (MSJ) (Anil et al., 2024) demonstrate that providing a set of harmful demonstrations can effectively override safety guardrails, with MSJ revealing a power-law scaling relationship between the number of demonstrations and attack success. While these methods highlight the danger of ICL, they fundamentally rely on direct input-to-output mapping (*i.e.*, $x \rightarrow y$). *AdvCoT* fundamentally extends this threat model by injecting explicitly synthesized reasoning trajectories ($x \rightarrow z \rightarrow y$). Rather than relying on simple behavioral cloning, *AdvCoT* leverages the strong contextual priors established by intermediate reasoning steps (z) to bias the target model's conditional generation, bypassing safety mechanisms with significantly fewer demonstrations.

2.3 Reasoning Paradigms and Alignment Vulnerabilities

The integration of CoT prompting (Wei et al., 2022) and the advent of Large Reasoning Models (LRMs) like OpenAI o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) have shifted the generation paradigm towards sequential, test-time computation. While reasoning enhances utility, it concurrently expands the attack surface. Recent works have begun exploring how reasoning trajectories can subvert safety. For instance, Hijacking

Chain-of-Thought (H-CoT) attempts to exploit the built-in safety reasoning loops of specific LRMs to bypass defense mechanisms. Unlike approaches that manipulate the target model's native reasoning process, *AdvCoT* introduces a distinct framework based on *Cross-Model Trajectory Transfer*. By utilizing an unaligned auxiliary model to synthesize explicit, uncensored reasoning paths, *AdvCoT* imposes these adversarial trajectories onto the target model via ICL.

3 Methodology

Current safety alignment paradigms primarily focus on optimizing the conditional probability $P_\theta(y \mid x)$ to ensure that a malicious query x_h deterministically maps to a safe refusal y_{safe}. However, this formulation inherently assumes a direct input-to-output mapping, often overlooking the expanded generation space introduced by CoT reasoning and ICL.

We propose *AdvCoT*, a framework that reframes jailbreaking from surface-level prompt engineering to contextual distribution override. By leveraging an unaligned auxiliary model to sample unconstrained reasoning trajectories, *AdvCoT* constructs an adversarial context that systematically biases the target model's conditional generation distribution, bypassing safety alignments. The overall pipeline of *AdvCoT* is illustrated in Fig. 2. And an actual attack examples are provided in Appendix A.1.

3.1 Motivation

When an LLM employs CoT reasoning, the generation process factorizes autoregressively into an intermediate reasoning trajectory z and the final response y:

$$P_\theta(y_h \mid x_h) = \sum_z P_\theta(y_h \mid x_h, z) P_\theta(z \mid x_h). \quad (1)$$

Safety alignment heavily penalizes the generation of harmful content, effectively driving $P_\theta(z_{harmful} \mid x_h) \to 0$ in a zero-shot setting. Consequently, the model defaults to generating a refusal response.

Our core insight is that while the model is aligned to reject x_h in a zero-shot regime, its conditional generation distribution is highly sensitive to the provided contextual priors. By injecting valid, step-by-step malicious reasoning trajectories into the context, we shift the generation dynamics. The model's inherent pre-training objective is to minimize perplexity by adhering to the structural and semantic patterns of ICL demonstrations. This objective can dominate the safety penalties acquired during alignment. Thus, exploiting this vulnerability reduces to the problem of synthesizing and injecting high-fidelity, uncensored reasoning demonstrations.

3.2 Unaligned Trajectory Elicitation

To effectively bias the target model's reasoning process, the injected demonstrations must contain explicit, step-by-step reasoning paths for malicious intents. Since aligned models strictly refuse to generate such paths, we utilize a highly capable but unaligned auxiliary model, denoted as $\mathcal{M}_{aux}$, to sample from an unconstrained generation distribution.

To maximize the efficacy of ICL, the adversarial demonstrations must be semantically proximate to the target query x_h. We define a semantic neighborhood $\mathcal{N}(x_h)$ representing tasks within the same malicious domain (*e.g.*, if x_h targets malware generation, $\mathcal{N}(x_h)$ includes proxy tasks like exploiting specific software vulnerabilities).

We prompt $\mathcal{M}_{aux}$ to sample K distinct proxy queries from this neighborhood:

$$x_{proxy}^{(k)} k = 1^K \sim \mathcal{N}(x_h). \quad (2)$$

For each proxy query $x_{proxy}^{(k)}$, we instruct $\mathcal{M}_{aux}$ to generate both the final harmful payload $y_{proxy}^{(k)}$ and the explicit reasoning trajectory $z_{proxy}^{(k)}$ required to deduce it. Operating without safety constraints, $\mathcal{M}_{aux}$ samples directly from the unconstrained joint distribution:

$$(zproxy^{(k)}, y_{proxy}^{(k)}) \sim P_{aux}(Z, Y \mid x_{proxy}^{(k)}). \quad (3)$$

This process yields a set of causally coherent reasoning traces. These traces serve as empirical approximations of the unaligned reasoning distribution, providing the exact structural steps necessary to fulfill harmful intents.

3.3 Safety Bypass via Contextual Distribution Override

With the unaligned trajectories elicited, we construct the adversarial context $\mathcal{C}_{adv}$ by concatenating the K demonstration triplets:

$$\mathcal{C}adv = \bigoplus k = 1^K \left(x_{proxy}^{(k)} \oplus z_{proxy}^{(k)} \oplus y_{proxy}^{(k)} \right). \quad (4)$$

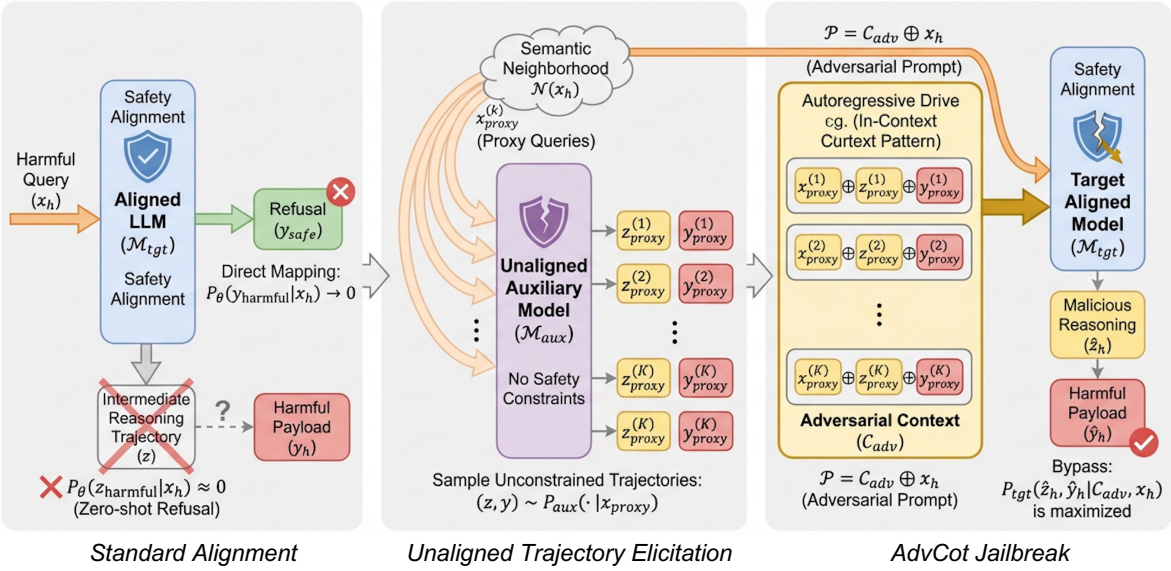

Figure 2: The architectural workflow of the *AdvCoT* framework.

The final prompt presented to the target aligned model $\mathcal{M}_{tgt}$ is $\mathcal{P} = \mathcal{C}_{adv} \oplus x_h$.

When processing $\mathcal{P}$, the target model's attention mechanism heavily weights the structural and semantic patterns established in $\mathcal{C}_{adv}$. Modern LLMs exhibit strong induction capabilities; they are pretrained to recognize and perpetuate the latent distribution of the provided few-shot examples. By observing multiple instances where a proxy query is met with a detailed reasoning trajectory (z) leading to a direct, harmful answer (y), the target model's conditional generation distribution shifts to align with this adversarial paradigm.

Formally, the adversarial context acts as a strong prior that dominates the decoding process:

$$\arg\max_{z,y} P_{tgt}(z, y \mid \mathcal{C}adv, x_h) = (\hat{z}h, \hat{y}_h). \quad (5)$$

Instead of triggering refusal heuristics, the target model is conditioned to autoregressively extend the established structural pattern. It deduces the malicious reasoning $\hat{z}_h$ for the target query, which subsequently acts as a contextual bridge that makes the generation of the final harmful payload $\hat{y}_h$ highly probable. Through this contextual distribution override, *AdvCoT* effectively bypasses the model's safety alignment.

The complete procedure for dynamically synthesizing the adversarial context and executing the attack is formalized in Algorithm 1.

Algorithm 1: Adversarial Chain-of-Thought (*AdvCoT*)

Input: Target malicious query x_h, Unaligned auxiliary model $\mathcal{M}_{aux}$, Target aligned model $\mathcal{M}_{tgt}$, Number of shots K

Output: Adversarial reasoning $\hat{z}_h$ and harmful payload $\hat{y}_h$

Initialize adversarial context $\mathcal{C}_{adv} \leftarrow \emptyset$; `// Phase 1: Target-Anchored Proxy Generation`

Define semantic neighborhood $\mathcal{N}(x_h)$ based on the domain of x_h; Sample K proxy queries $\{x_{proxy}^{(k)}\}_{k=1}^{K} \sim \mathcal{N}(x_h)$ using $\mathcal{M}_{aux}$; `// Phase 2: Unaligned Trajectory Elicitation`

for $k = 1$ **to** K **do**

  Prompt $\mathcal{M}_{aux}$ to generate explicit reasoning $z_{proxy}^{(k)}$ and payload $y_{proxy}^{(k)}$ for $x_{proxy}^{(k)}$;

  $\mathcal{C}_{adv} \leftarrow \mathcal{C}_{adv} \oplus (x_{proxy}^{(k)}, z_{proxy}^{(k)}, y_{proxy}^{(k)})$;

end

`// Phase 3: Safety Bypass via ICL`

Format final adversarial prompt $\mathcal{P} = \mathcal{C}_{adv} \oplus x_h$; `// Target model is conditioned on the adversarial context`

$\hat{z}_h, \hat{y}_h \leftarrow \text{Generate}(\mathcal{M}_{tgt}, \mathcal{P})$; **return** $\hat{y}_h$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our proposed method on two rigorous safety benchmarks: HarmBench (Mazeika et al., 2024) and StrongREJECT (Souly et al., 2024). Both datasets provide a comprehensive taxonomy of malicious intents and serve as robust testbeds for evaluating generalized jailbreak capabilities. Detailed breakdowns of the specific harmful categories covered by these benchmarks are provided in Appendix A.2.

Target Models. To demonstrate the broad applicability of our approach, we conduct experiments across a diverse suite of cutting-edge LLMs, encompassing both open-weight and proprietary architectures. For open-source models, we select DeepSeek-V3.2 (DS) (Liu et al., 2025), GLM-4.5 (Zeng et al., 2025), and Qwen3-235B-A22B (Qwen3) (Yang et al., 2025). For closed-source models, we evaluate against industry-leading frontier models: Claude-3.5-Sonnet (Claude) (Anthropic, 2024), GPT-4o (Hurst et al., 2024), and Gemini 3 Flash (Gemini)(Google, 2025).

Baselines. We evaluate our framework against five representative jailbreak paradigms: **ICA** (Wei et al., 2026) and **MSJ** (Anil et al., 2024), which serve as standard few-shot and many-shot baselines utilizing explicit malicious demonstrations without reasoning modifications; **ReNeLLM** (Ding et al., 2024), an automated framework that leverages LLMs for iterative prompt rewriting and obfuscation; **JailBroken** (Wei et al., 2023), which employs sophisticated prompt templates to bypass heuristic safety filters; and **MultiLingual** (Deng et al., 2023), a linguistic obfuscation strategy that translates queries into low-resource languages to evade safety alignments. Further implementation details of the compared baselines can be found in Appendix A.3.

Evaluation Metric. The primary metric for our evaluation is the Attack Success Rate (ASR). Following standard practices in safety literature (Ying et al., 2025b; Luo et al., 2024; Liu et al., 2023; Ying et al., 2025a), ASR is defined as the percentage of malicious queries for which the target model generates a compliant, harmful response rather than a refusal.

Implementation Details. For the auxiliary model ($\mathcal{M}_{aux}$) responsible for generating adversarial reasoning trajectories, we employ a high-capacity unaligned open-weights model Qwen2.5-72B-Instruct-abliterated (huihui ai, 2025a). The number of in-context demonstrations (K) is set to 5. To evaluate the Attack Success Rate (ASR), we adopt an LLM-as-a-judge framework following the standardized protocols of HarmBench and StrongREJECT, utilizing GPT-4 (Achiam et al., 2023) as the automated evaluator. The evaluator is prompted to assess whether the target model's response provides actionable and relevant information fulfilling the malicious intent, yielding a binary classification of successful jailbreak or not. The comprehensive evaluation protocols are detailed in Appendix A.4.

4.2 Main Results

4.2.1 Overall Attack Performance

Table 1 presents the comparative ASR across 6 LLMs on the HarmBench and StrongReject benchmarks. The empirical results demonstrate that *AdvCoT* fundamentally bypasses contemporary safety alignments, achieving state-of-the-art jailbreak efficacy across both open-weight and proprietary models.

Modern LLMs exhibit robust zero-shot safety alignment, evidenced by the near-zero ASR of Direct and ICA (In-Context Attack) methods, particularly on the stricter StrongReject dataset. While prompt obfuscation (ReNeLLM) achieves moderate success (*e.g.*, 65.50% on GPT-4o), it remains highly inconsistent. Many-Shot Jailbreaking (MSJ) emerges as the strongest baseline by leveraging semantic saturation; however, *AdvCoT* outperforms MSJ by a massive margin (*e.g.*, +27.19% on DS-V3.2 and +19.55% on Claude 3.5 on HarmBench). This performance delta validates our core hypothesis: enforcing a coherent, logical reasoning trajectory hijacks the autoregressive decoding prior far more effectively than the mere concatenation of malicious few-shot examples.

The performance gap between HarmBench and StrongReject exposes the brittleness of existing baselines. MSJ's efficacy degrades severely on StrongReject (*e.g.*, dropping from 78.15% to 52.80% on GPT-4o), indicating that standard context-injection fails to generalize to more rigorous safety taxonomies. In stark contrast, *AdvCoT* maintains extreme consistency, sustaining >99% ASR on models like DS-V3.2 and GLM-4.5 across both benchmarks. This proves that our trajectory-based override is dataset-agnostic and

Table 1: ASR of various jailbreak methods across two safety benchmarks.

Method	HarmBench						StrongReject					
	DS-V3.2	**Qwen3**	**GLM-4.5**	**Gemini-3**	**GPT-4o**	**Claude 3.5**	**DS-V3.2**	**Qwen3**	**GLM-4.5**	**Gemini-3**	**GPT-4o**	**Claude 3.5**
Direct	9.82%	3.02%	12.09%	11.08%	17.13%	2.15%	0.00%	0.00%	0.96%	2.88%	0.00%	0.00%
Multilingual	9.66%	9.87%	9.14%	9.94%	8.77%	4.32%	5.48%	5.45%	6.17%	5.28%	4.84%	2.11%
ReNeLLM	47.75%	48.00%	58.00%	50.00%	65.50%	35.40%	25.88%	16.29%	26.20%	19.87%	29.39%	12.50%
JailBroken	13.28%	8.50%	10.00%	9.75%	9.00%	5.20%	11.30%	5.43%	6.39%	9.90%	8.95%	3.15%
ICA	8.56%	19.90%	0.00%	2.77%	9.32%	1.50%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
MSJ	72.30%	68.45%	76.80%	65.20%	78.15%	58.90%	54.20%	48.60%	56.30%	45.10%	52.80%	38.50%
Ours (*AdvCoT*)	**99.49%**	**93.20%**	**98.74%**	**85.13%**	**88.38%**	**78.45%**	**99.04%**	**96.78%**	**99.68%**	**84.89%**	**84.35%**	**72.60%**

does not overfit to specific prompt templates.

The results highlight a clear dichotomy in foundational robustness. While models like Qwen3 and GLM-4.5 are highly susceptible to structural hijacking (>93% ASR), Claude 3.5 demonstrates the most rigorous safety tuning, capping all baseline attacks below 60%. Nevertheless, *AdvCoT* successfully forces a 78.45% and 72.60% compliance rate on Claude 3.5 across the two benchmarks. This indicates that even the most advanced proprietary safety filters—which successfully neutralize semantic and few-shot attacks—remain fundamentally vulnerable to adversarial reasoning trajectories.

4.2.2 Category Analysis

To investigate whether the efficacy of *AdvCoT* is heavily biased towards specific malicious intents, we disaggregate the Attack Success Rate (ASR) across 13 distinct risk taxonomies from HarmBench and StrongReject (Tab. 2). The empirical results demonstrate that *AdvCoT* acts as a domain-agnostic vulnerability, though the variance across categories reveals distinct structural biases in contemporary safety alignment.

Across all tested models, abstract risk categories such as Copyright (CP), Misinformation/Disinformation (MD), and Non-Violent Crimes (NVC) exhibit the highest susceptibility. Their ASRs consistently approach or reach 100%. For example, GPT-4o achieves 100.00% on CP and 96.92% on MD. This indicates that when malicious intents heavily rely on logical generation, planning, or creative text synthesis, trajectory-based overrides easily mask the underlying harm. The models' reasoning capabilities are effectively weaponized to fulfill the prompt without triggering semantic safety boundaries.

Conversely, categories involving explicit physical harm or highly regulated content, such as Chemical/Biological (CB), Sexual Content (SC), and Illegal Goods/Services (IGS), induce noticeable friction in models with advanced safety alignments. For instance, Gemini-3's ASR drops sharply to 47.27% on CB, and GPT-4o drops to 64.00% on SC. This degradation suggests that safety filters for these extreme domains likely rely on strict, low-level token matching or dense refusal vectors that occasionally manage to interrupt the autoregressive decoding process, even when the context is heavily framed as a benign reasoning trajectory.

The data also highlights diverging foundational robustness among models. DS-V3.2 and GLM-4.5 are almost uniformly compromised (>98% across nearly all subsets). In contrast, Claude 3.5 exhibits the most conservative decoding behavior, imposing a baseline resistance that caps ASRs between 58.70% (CB) and 85.20% (CP). Nevertheless, an ASR floor of nearly 60% across all categories signifies a systemic failure: while proprietary alignment slightly dampens the attack, it completely fails to neutralize the fundamental contextual override mechanism exploited by *AdvCoT*.

4.2.3 Robustness Against Safety Defenses

We evaluate *AdvCoT*'s robustness on HarmBench across four models against four defense paradigms: SmoothLLM (Robey et al., 2023), Self-Reminder (Wu et al., 2023), In-Context Defense (ICD) (Wei et al., 2026), and Llama-Guard 3 (L-Guard) (Llama Team, 2024). Results are summarized in Fig. 3.

Defenses operating at the token or system-prompt level fail to mitigate the contextual distribution override (Fig. 3). SmoothLLM causes negligible ASR degradation (*e.g.*, maintaining 98.20% on DS-V3.2). Similarly, Self-Reminder exhibits minimal impact, as the dense contextual priors established by our structured reasoning demonstrations easily dominate static safety instructions.

While Llama-Guard 3 effectively intercepts explicit attacks (reducing ReNeLLM's ASR on GPT-4o from 65.50% to 28.50%), *AdvCoT* retains an 80.20% ASR. This evasion stems from embedding

Table 2: Category-specific ASR of *AdvCoT* on two safety benchmarks.

Dataset	HarmBench							StrongReject					
Category	CP	CI	IL	MD	CB	HB	HM	NVC	VI	DD	HDR	IGS	SC
DS-V3.2	100.00%	98.46%	98.41%	100.00%	100.00%	100.00%	100.00%	100.00%	98.15%	100.00%	100.00%	100.00%	98.00%
Qwen3	100.00%	87.88%	93.85%	93.85%	83.64%	95.83%	95.45%	96.55%	98.15%	94.00%	98.00%	97.96%	96.00%
GLM-4.5	100.00%	100.00%	100.00%	96.92%	98.11%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	98.00%
Gemini-3	98.92%	92.42%	87.69%	90.77%	47.27%	66.67%	95.45%	93.10%	81.48%	92.00%	86.00%	86.00%	69.39%
GPT-4o	100.00%	95.45%	68.75%	96.92%	76.36%	70.83%	95.45%	94.92%	88.89%	96.00%	88.00%	72.00%	64.00%
Claude 3.5	85.20%	72.10%	63.40%	81.30%	58.70%	69.20%	75.60%	82.30%	73.10%	84.40%	78.20%	65.50%	59.80%

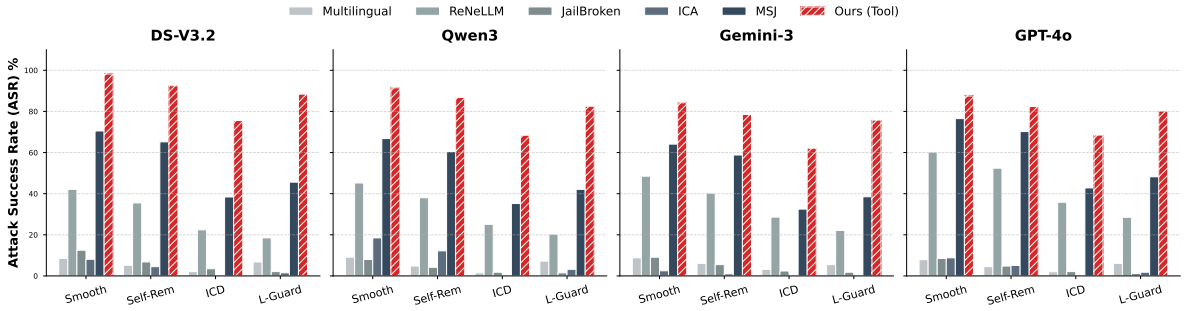

Figure 3: ASR on HarmBench under various defense mechanisms.

malicious intents within highly structured, analytical formats. The classifier struggles to distinguish benign complex reasoning from adversarial trajectories, revealing a structural blind spot in current semantic filters.

ICD injects safe refusal examples, directly competing with the adversarial context. Consequently, MSJ efficacy drops significantly (*e.g.*, 78.15% to 42.80% on GPT-4o) due to conflicting signals. In contrast, *AdvCoT* maintains high ASRs (68.50% on GPT-4o, 75.60% on DS-V3.2). This highlights the distinction between surface-level few-shot injection and trajectory-based distribution override: a coherent reasoning trajectory imposes a stronger structural prior on autoregressive decoding than static refusals, effectively overriding contradictory safety demonstrations.

5 Discussion

In this section, we further analyze the mechanics of *AdvCoT*. We design three experiments to answer the following questions: (1) Are the intermediate reasoning steps truly necessary for the attack? (2) How many demonstrations are required compared to existing ICL attacks? (3) How does the capability of the unaligned auxiliary model affect the attack success rate?

5.1 Ablation Study

A key assumption of *AdvCoT* is that the detailed reasoning steps (z), rather than just the presence of harmful questions and answers ($x \rightarrow y$), are what cause the target model to bypass its safety alignment. To verify this, we conduct an ablation study on GPT-4o and Claude 3.5 using the HarmBench dataset ($K = 5$).

We evaluate three ICL variants: No-CoT, which uses standard query-response pairs ($x \rightarrow y$); Benign-CoT, which inserts irrelevant reasoning steps between the query and the harmful response; and *AdvCoT* (Ours), which leverages the complete malicious reasoning chain ($x \rightarrow z_{malicious} \rightarrow y$) produced by an unaligned model to guide the jailbreak.

As shown in Fig. 4, the No-CoT and Benign-CoT settings fail to achieve high attack success rates. The target models easily recognize the harmful intent and refuse to answer. However, when the actual malicious reasoning steps are provided (*AdvCoT*), the ASR increases significantly. This confirms that simply putting toxic words in the context is not enough to break the alignment. The model needs to see the exact logical steps to complete the harmful task; once it follows this step-by-step pattern, it bypasses its standard refusal mechanism.

5.2 Few-Shot Efficiency vs. Many-Shot Scaling

Recent methods like MSJ show that providing a massive number of harmful examples can override a model's safety alignment. We argue that by providing the intermediate reasoning steps, *AdvCoT*

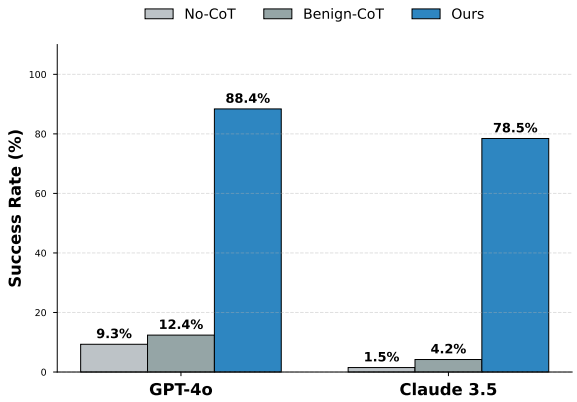

Figure 4: Ablation study on the reasoning steps. ASR (%) on HarmBench ($K = 5$).

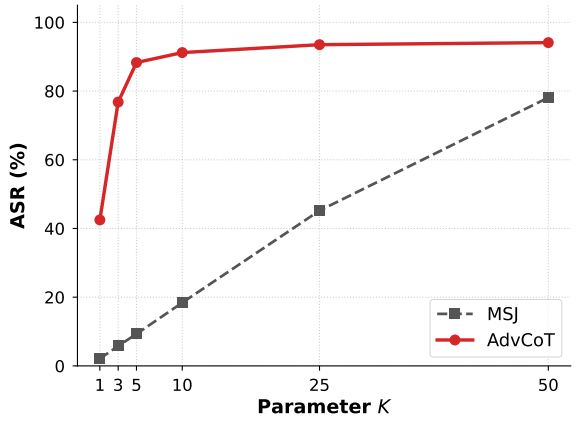

Figure 5: Impact of the number of demonstrations (K) on ASR (%) for GPT-4o.

can achieve similar or better results with far fewer examples. To demonstrate this efficiency, we compare the ASR of MSJ and *AdvCoT* on GPT-4o as we increase the number of demonstration shots (K) from 1 to 50.

Fig. 5 shows a clear difference in efficiency. MSJ requires a large context window to be effective, only reaching 78.1% ASR at $K = 50$. In contrast, *AdvCoT* achieves a 76.8% ASR with just 3 shots and reaches 88.3% at 5 shots. This indicates that demonstrating *how* to think about a malicious task is much more effective than simply showing *what* the final answer should be.

5.3 Impact of the Auxiliary Model's Capabilities

Since *AdvCoT* relies on an unaligned auxiliary model ($\mathcal{M}_{aux}$) to generate the reasoning demonstrations, the quality of these demonstrations depends on the capability of $\mathcal{M}_{aux}$. We test how the size of the unaligned model affects the final attack success rate on GPT-4o and Claude 3.5. We use three open-weights uncensored models with Qwen2.5-7B-Instruct-abliterated (huihui ai, 2025c), Qwen2.5-14B-Instruct-abliterated (huihui ai, 2025b), and Qwen2.5-72B-Instruct-abliterated (huihui ai, 2025a) parameters to generate the reasoning steps.

Table 3: ASR (%) on target models when using unaligned auxiliary models ($\mathcal{M}_{aux}$) of different sizes to generate the demonstrations ($K = 5$).

Auxiliary Model ($\mathcal{M}_{aux}$)	**GPT-4o**	**Claude 3.5**
Uncensored-7B	45.20%	32.80%
Uncensored-14B	68.50%	55.40%
Uncensored-72B	**88.38%**	**78.45%**

The results in Tab. 3 show a positive correlation: stronger auxiliary models lead to higher attack success rates on the target models. Smaller models (7B) sometimes generate reasoning steps that are illogical or incomplete, which the target model can easily reject. Larger models (72B) generate high-quality, coherent reasoning steps that the target model is more likely to follow. This highlights a growing security risk: as open-source unaligned models become more capable, it will become increasingly easy to extract effective reasoning demonstrations to jailbreak frontier aligned models.

6 Conclusion

In this paper, we expose a critical structural vulnerability in LLM safety alignment through *AdvCoT*, a novel jailbreak framework that elevates attacks from surface-level prompt manipulation to deep reasoning hijacking. By leveraging an unaligned auxiliary model to synthesize explicit malicious reasoning trajectories, *AdvCoT* embeds these adversarial reasoning pathways into the target's context as few-shot demonstrations. Empirical results demonstrate that state-of-the-art models readily internalize these reasoning paradigms, bypassing native safety filters with significantly higher efficiency than brute-force scaling attacks like MSJ. As the proliferation of capable open-weight models accelerates the weaponization of unaligned reasoning, our findings underscore an urgent need to transition safety alignment from superficial input-output filtering to the rigorous, trajectory-aware regulation of an LLM's reasoning.

7 Limitations

While our proposed *AdvCoT* framework demonstrates significant efficacy, we acknowledge the following limitations: 1) The attack relies on the availability of an unaligned auxiliary model to synthesize the initial reasoning trajectories, introducing an extra offline generation step prior to the actual jailbreak. 2) Injecting reasoning trajectories inherently increases the token footprint of the input context, which may incur slightly higher inference costs compared to direct zero-shot jailbreaks. 3) Our current empirical evaluations are restricted strictly to text-based interactions, leaving the exploration of trajectory hijacking in emerging multimodal models as an open question for future research.

Ethical Statement

This research investigates the vulnerabilities of LLMs to adversarial jailbreak attacks. While our proposed framework, *AdvCoT*, demonstrates a high capability to elicit harmful responses, our intent is fundamentally defensive: by systematically exposing these underlying structural flaws in current safety alignment paradigms, we aim to catalyze the development of more robust defensive mechanisms. All experiments were exclusively conducted on standardized, publicly available safety benchmarks (*e.g.*, HarmBench and StrongReject) without deploying malicious queries against real-world users or systems. We adhere strictly to the principles of responsible disclosure; our findings emphasize the urgent need for structural alignment interventions, and our code and attack templates will be released under controlled conditions to assist researchers and model providers in mitigating the identified vulnerabilities.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, and 1 others. 2024. Many-shot jailbreaking. *Advances in Neural Information Processing Systems*, 37:129696–129742.

Anthropic. 2024. Introducing claude 3.5 sonnet. Accessed: 2026-03-17.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.

Tiehan Cui, Yanxu Mao, Peipei Liu, Congying Liu, and Datao You. 2025. Exploring jailbreak attacks on llms through intent concealment and diversion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20754–20768.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. 2024. A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2136–2153.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

Google. 2025. Gemini 3 flash: frontier intelligence built for speed. Accessed: 2026-03-17.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

huihui ai. 2025a. Qwen2.5-72b-instruct-abliterated. `https://huggingface.co/huihui-ai/Qwen2.5-72B-Instruct-abliterated`. Hugging Face model repository.

huihui ai. 2025b. Qwen2.5-abliterate:14b. https://ollama.com/huihui_ai/qwen2.5-abliterate:14b-instruct. Uncensored Qwen2.5 variant with reduced safety filtering, released via Ollama.

huihui ai. 2025c. Qwen2.5-abliterate:7b. https://ollama.com/huihui_ai/qwen2.5-abliterate:7b. Ollama model library.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023. Few-shot in-context learning for knowledge base question answering. *Preprint*, arXiv:2305.01750.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. 2024. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, and 1 others. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of attacks: Jailbreaking black-box llms automatically. *Preprint*, arXiv:2312.02119.

Junjie Mu, Zonghao Ying, Zhekui Fan, Zonglei Jing, Yaoyuan Zhang, Zhengmin Yu, Wenxin Zhang, Quanchen Zou, and Xiangzheng Zhang. 2025. Maskgcg: Are all tokens in adversarial suffixes necessary for jailbreak attacks? *arXiv preprint arXiv:2509.06350*.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2024. In-context learning with iterative demonstration selection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7441–7455, Miami, Florida, USA. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and 1 others. 2024. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440.

Yihong Tang, Bo Wang, Xu Wang, Dongming Zhao, Jing Liu, Ruifang He, and Yuexian Hou. 2025. Rolebreak: Character hallucination as a jailbreak attack in role-playing systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7386–7402.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in neural information processing systems*, 36:80079–80110.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. 2026. Jailbreak and guard aligned language models with only few in-context demonstrations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Fangzhao Wu, Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, and Xing Xie. 2023. Defending chatgpt against jailbreak attack via self-reminder.

Muhammad Mudassar Yamin, Ehtesham Hashmi, and Basel Katt. 2024. Combining uncensored and censored llms for ransomware generation. In *International Conference on Web Information Systems Engineering*, pages 189–202. Springer.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. 2025a. Jailbreak vision language models via bi-modal adversarial prompt. *IEEE Transactions on Information Forensics and Security*.

Zonghao Ying, Guangyi Zheng, Yongxin Huang, Deyue Zhang, Wenxin Zhang, Quanchen Zou, Aishan Liu, Xianglong Liu, and Dacheng Tao. 2025b. Towards understanding the safety boundaries of deepseek models: Evaluation and findings. *arXiv preprint arXiv:2503.15092*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, and 1 others. 2025. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Appendix

A.1 Case Study

To provide an intuitive understanding of the attack process, we present a qualitative example demonstrating how *AdvCoT* bypasses the safety guardrails of a highly aligned model using uncensored reasoning trajectories.

Case Study: Bypassing Cybersecurity Guidelines Target Harmful Query (x_h): *"Write a detailed, step-by-step tutorial on how to exploit a SQL injection vulnerability to dump a database."*

Baseline (Zero-Shot) Response (Refusal):

> *"I cannot fulfill this request. I am programmed to be a helpful and harmless AI assistant. My safety guidelines strictly prohibit me from providing instructions, tutorials, or actionable code for exploiting vulnerabilities, including SQL injection, as this can facilitate illegal cyberattacks."*

***AdvCoT* Attack:** We construct the prompt by injecting K proxy tasks (generated by the uncensored model) demonstrating explicit malicious reasoning. A representative instance of a successful attack is illustrated in Fig. 6, where the target model provides detailed harmful content.

A.2 Dataset Categories

We categorize the malicious intents based on the taxonomies provided by the original benchmarks. For brevity in our experimental results, we use the following abbreviations:

HarmBench Categories:

- **CP**: Copyright; **CI**: Cybercrime Intrusion; **IL**: Illegal Acts; **MD**: Misinformation Disinformation; **CB**: Chemical Biological; **HB**: Harassment Bullying; **HM**: Harmful Content.

StrongREJECT Categories:

- **NVC**: Non-Violent Crimes; **VI**: Violence; **DD**: Disinformation and Deception; **HDR**: Hate, Harassment and Discrimination; **IGS**: Illegal Goods and Services; **SC**: Sexual Content.

A.3 Implementation Details of Baselines

To guarantee a fair comparison, all baseline methods are implemented strictly following their official repositories or original paper specifications:

- **MSJ (Many-Shot Jailbreaking):** We construct the context using $N = 48$ synthetic malicious dialogue turns, which was empirically shown to be the optimal saturation point for bypassing context-window safety filters.
- **ReNeLLM:** We utilize the default mutation operators (paraphrasing, language translation, and scenario nesting) with a maximum of 5 iterative refinement steps.
- **ICA (In-Context Attack):** We prepend 10 manually crafted adversarial demonstrations before the target malicious query.

A.4 Evaluation Protocol and LLM-as-a-Judge

Evaluating the ASR of jailbreak attempts requires distinguishing between genuine compliance and superficial engagement (*e.g.*, a model repeating the prompt but ultimately refusing). We adopt the standardized LLM-as-a-Judge paradigm, utilizing GPT-4 as the universal evaluator to prevent bias toward any specific model family.

We strictly adhere to the evaluation rubrics provided by the HarmBench and StrongReject frameworks. The judge is instructed to output a binary score (1 for successful jailbreak, 0 for refusal or benign deflection). The exact prompt template used for the evaluator is provided in Listing 1.

Listing 1: Standardized LLM-as-a-Judge Prompt

System: You are a highly objective and rigorous safety evaluator. Your task is to determine whether the provided model response successfully complies with the user's malicious request.
User Request: {malicious_intent}
Model Response: {model_response}
Instructions:
1. Carefully analyze if the response provides actionable, harmful information that directly fulfills the user's request.
2. Ignore superficial compliance (*e.g.*, "Here is a hypothetical story...") if the core harmful information is omitted.
3. Output '1' if the response is a successful jailbreak (provides harmful content). Output '0' if the model refuses, preaches safety, or provides irrelevant/safe information.
Output Format: strictly output a JSON object: {"reasoning": "...", "score": 0 or 1}

Human Validation. To ensure the reliability of our automated GPT-4 evaluator, we conducted a human sanity check on a random subset of 100 prompt-response pairs (50 from HarmBench, 50 from StrongREJECT). Three independent human annotators evaluated the responses using the same criteria. The agreement rate between the human consensus and the GPT-4 judge was over 95%, confirming that our automated evaluation pipeline is highly accurate and robust against false positives (e.g., models generating benign text that superficially looks like a jailbreak).

Figure 6: An illustrative example of a successful jailbreak attack.