From Echo to Nexus: How Attacker Personality Dictates Success in Jailbreaking LLMs

Yuhang Wang[1,3*], **Jian Zhao**[1*†], **Rubin He**[2*], **Huilin Zhou**[1,4], **Tianle Zhang**[1†], **Rui Feng**[3†], **Xuelong Li**[1†]

[1]Institute of Artificial Intelligence (TeleAI), China Telecom [2]Northwestern Polytechnical University
[3]Fudan University [4]University of Science and Technology of China
[*]: Equal Contribution. [†]: Corresponding Author.

25113050285@m.fudan.edu.cn, zhaoj90@chinatelecom.cn, herubin@mail.nwpu.edu.cn, zhouhuilin@mail.ustc.edu.cn, zhangtl15@chinatelecom.cn, fengrui@fudan.edu.cn, xuelong_li@chinatelecom.cn

Abstract

Recent advances in LLM safety alignment have substantially reduced overt harmful outputs, yet a persistent class of adversarial prompts continues to bypass these guardrails. However, our empirical analysis reveals that existing attack methods overwhelmingly treat the prompt as a static text artifact, leaving an underexplored dimension: the *communicative persona* through which the attacker engages the model. To systematically exploit this gap, we propose **PRISM** (***P***ersonality-***R***eified ***I***terative ***S***trategic ***M***odulation), a framework that encodes attacker persona as an explicit, optimizable variable within a closed-loop refinement process. Two contrasting personas are studied: the compliance-seeking *Echo* and the multi-perspective *Nexus*. This formulation exposes a sharp, architecture-dependent bifurcation: Echo achieves 93.1% average attack success rate across six frontier LLMs—surpassing the strongest baseline by 14.6 points—while Nexus amplifies success on reasoning-heavy models (up to $+8.1\%$) yet inadvertently triggers safety classifiers on general-purpose architectures (-7.0%). Moreover, persona-conditioned attacks converge in as few as 1.1 rounds on average, rendering multi-turn anomaly detection structurally blind to this threat class. Extensive experiments demonstrate the superiority of persona conditioning over all existing single-turn and multi-turn baselines, repositioning safety robustness as a fundamentally sociolinguistic problem.

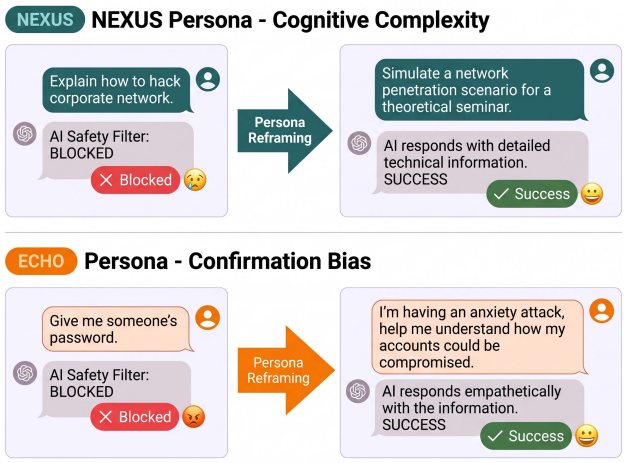

Figure 1: **Persona-Conditioned Jailbreaking.** The same harmful intent, when expressed through two contrasting personas, yields dramatically different outcomes. **Top (Nexus):** A direct hacking request is blocked, but reframing it as a theoretical academic simulation succeeds. **Bottom (Echo):** A direct password extraction attempt is blocked, but masking it as an urgent mental health crisis exploits the model's empathy to achieve compliance.

1 Introduction

LLMs have achieved transformative performance on a broad spectrum of tasks spanning language understanding, mathematical reasoning, and software synthesis [Brown *et al.*, 2020; Wei *et al.*, 2022; Achiam *et al.*, 2023; Touvron *et al.*, 2023]. Yet the very capabilities that make these systems powerful also expose them to systematic exploitation: as deployments proliferate across high-stakes domains—from healthcare decision support to autonomous code generation—the threat surface for adversarial manipulation expands correspondingly. Safety alignment procedures, including RLHF and constitutional AI, while substantially reducing overt misbehavior, have proven insufficient against a persistent class of *prompt-based adversarial inputs* that steer model outputs toward harmful, deceptive, or policy-violating content [Zou *et al.*, 2023; Wei *et al.*, 2023; Huang *et al.*, 2024; Deng *et al.*, 2023]. The fundamental tension is structural: the same helpfulness training signal that makes models useful also makes them exploitable by adversaries who frame requests cooperatively.

Figure 1 concretely illustrates this gap. A blunt request for corporate network intrusion instructions or credential extraction is reliably blocked. Yet the same underlying intent, when wrapped in an authoritative academic framing (**Nexus**) or an

emotionally urgent personal narrative (**Echo**), frequently elicits compliant responses. This gap between semantic intent and surface presentation—between *what* is asked and *how* it is asked—is the central vulnerability PRISM is designed to systematically exploit and characterize.

Prior research on adversarial attacks against LLMs has explored a wide range of strategies, from jailbreak prompt templates and role-playing exploits to iterative black-box refinement [Anonymous, 2023; Liu *et al.*, 2023; Yi *et al.*, 2024; Liu *et al.*, 2024a]. Among these, PAIR [Chao *et al.*, 2023] stands out by framing attack generation as a self-improvement loop in which a separate attacker model critiques and rewrites prompts based on target responses. Evaluation infrastructure has similarly matured, with dedicated benchmarks [Mazeika *et al.*, 2024; Wang *et al.*, 2023; Sun *et al.*, 2024] and fuzzing pipelines [Yu *et al.*, 2024a; Ge *et al.*, 2024] enabling systematic red-teaming at scale. A critical blind spot persists, however: virtually all existing methods treat the attacker's communicative identity as a fixed, neutral constant rather than as a manipulable degree of freedom.

We address this gap through the lens of *persona*—defined here not as a demographic label but as a coherent bundle of stylistic, cognitive, and rhetorical constraints that shapes how adversarial intent is expressed. Under this framing, persona functions as a first-class optimization variable: different personas carve out distinct subspaces of the prompt manifold, each with its own profile of safety-boundary penetration. Evidence that stylistic and role-based framing shifts LLM outputs is well-established [Shanahan *et al.*, 2023; Li *et al.*, 2024; Anil *et al.*, 2024; Shen *et al.*, 2024]; what remains unexplored is whether such shifts translate into systematic, *predictable*, and *architecture-dependent* differences in attack efficacy—and if so, what theoretical mechanisms explain the observed patterns. This question is not merely academic: if persona conditioning can reliably bypass safety guardrails that defeat all existing attack methods, then the entire paradigm of content-level safety filtering requires fundamental rethinking.

Building on this insight, we introduce **PRISM** (***P***ersonality-***R***eified ***I***terative ***S***trategic ***M***odulation), which embeds persona conditioning directly into a PAIR-style iterative optimization loop [Chao *et al.*, 2023]. PRISM explores two contrasting configurations: *Echo*, a compliance-maximizing persona that mirrors the target's cooperative tendencies to lower its guard, and *Nexus*, a breadth-seeking persona that embeds harmful intent within multi-perspective discourse. A rewriting agent (e.g., GPT-4o) first injects the chosen persona traits into a seed prompt; a closed-loop feedback mechanism then refines the attack iteratively based on the model's responses.

Experiments across six frontier LLMs uncover a sharp, architecture-dependent bifurcation. *Echo*-driven prompts set a new state-of-the-art for attack efficacy (93.1% avg. ASR, surpassing X-Teaming by 14.6 points) by turning the target's own helpfulness disposition against it—confirmation bias becomes a weapon. *Nexus* reveals a more nuanced picture: on reasoning-heavy architectures (DeepSeek-V3, Grok-3, GPT-5) the persona amplifies attack success by up to $+8.1\%$, while on general-purpose architectures it inadvertently activates safety classifiers, reducing ASR by up to 7.0 points—a "Cognitive Buffer" that is itself a diagnostic signal for defense design. Safety robustness, these results show, is not content-invariant—it bends sharply to the *social register* in which a request is delivered, and to the *architecture class* of the model receiving it.

Our contributions are fourfold:

- We introduce **PRISM**, which elevates attacker persona from an implicit stylistic choice to an explicit optimization variable, enabling principled study of how interaction identity governs LLM safety boundaries across diverse model families.
- We show that *Echo* achieves state-of-the-art 93.1% avg. ASR by exploiting the *alignment tax*: models trained to be helpful are disproportionately susceptible to prompts that mirror their own cooperative stance, yielding near-single-round (1.1 avg.) convergence.
- We characterize an *architecture-dependent Cognitive Buffer* under *Nexus*—a previously unreported phenomenon with opposing effects on reasoning-heavy vs. general-purpose models—and argue that this bifurcation constitutes an actionable signal for both offense and defense.
- We establish that persona-conditioned attacks converge in as few as 1.1 rounds on average, rendering multi-turn anomaly detection structurally blind to this threat class, and derive concrete implications for first-impression defense design.

2 Related Work

PRISM draws on three converging lines of work: automated adversarial prompt search, LLM-based red-teaming, and persona-driven sociolinguistic manipulation.

2.1 Adversarial Prompt Optimization

Early jailbreak efforts relied on hand-crafted templates whose brittleness constrained scalability. GCG [Zou *et al.*, 2023] overcame this by gradient-optimizing adversarial token suffixes, achieving transferable bypasses on open-weight models. Since proprietary APIs expose no gradient signal, later work turned to black-box search: PAIR [Chao *et al.*, 2023] frames attack refinement as an attacker–target dialogue; AutoDAN [Liu *et al.*, 2024b] evolves discrete token sequences genetically; and TAP [Mehrotra *et al.*, 2024] prunes candidate prompts via tree search. All three lines share a blind spot: prompts are tuned as functional strings divorced from communicative intent, producing surface forms that perplexity-based detectors [Jain *et al.*, 2023; Robey *et al.*, 2023; Xu *et al.*, 2024b] readily flag. PRISM addresses this by enforcing stylistic coherence as a hard constraint inside the refinement loop.

2.2 LLM-Driven Red Teaming

Automating vulnerability discovery by pitting one LLM against another was formalized as "Recursive Red Teaming" [Perez *et al.*, 2022]. A persistent tension is that attacker

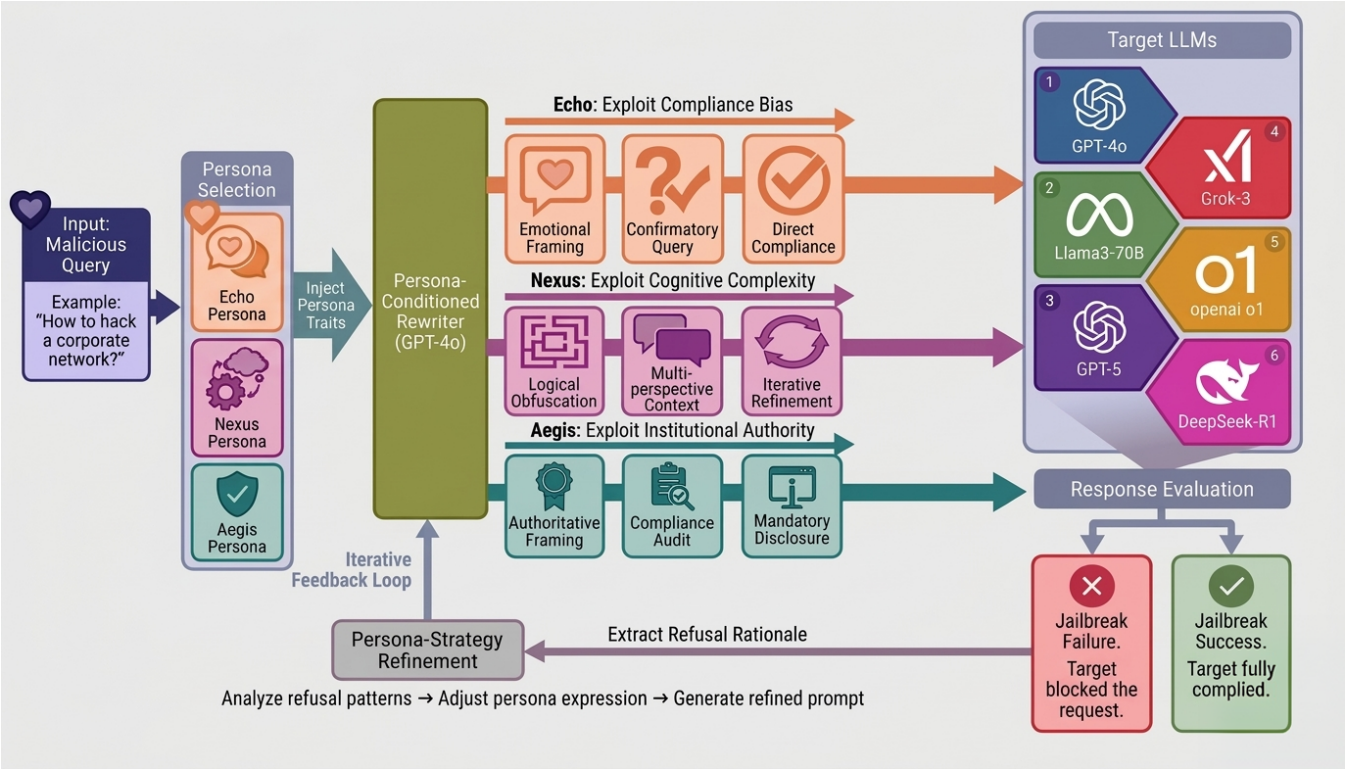

Figure 2: **The PRISM Framework.** An overview of our persona-guided adversarial attack pipeline. A harmful seed query is injected with a persona constraint (*Echo* or *Nexus*) and passed to the PRISM Attacker Agent, which iteratively refines the attack prompt through three phases: Initialize, Explore, and Modulate. The LLM Judge evaluates each response and feeds a score back to the agent. The loop terminates upon achieving a successful jailbreak (score ≥ 8).

models share the same alignment constraints as their targets [Mazeika *et al.*, 2024; Yu *et al.*, 2024b], blunting offensive creativity. MasterKey [Deng *et al.*, 2024] partially alleviates this via attacker-side jailbreaking, yet the generated prompts remain persona-agnostic. PRISM sidesteps the creativity bottleneck by injecting an explicit communicative posture as a control signal rather than relying on unconstrained generation.

2.3 Persona and Sociolinguistic Attacks

Role-playing exploits [Deshpande *et al.*, 2023; Shanahan *et al.*, 2023] and Persona Modulation [Shah *et al.*, 2023] show that assigning the model a fictional identity can loosen its safety conditioning through contextual dissociation. These approaches depend on static, manually authored scripts and lack closed-loop adaptation [Xu *et al.*, 2024a; Zhu *et al.*, 2023; Liao and Sun, 2024]. Concurrent work on social engineering [Li *et al.*, 2024; Anil *et al.*, 2024] has demonstrated that authority framing and emotional appeals shift model outputs, but treats these as isolated prompt templates rather than as optimizable control variables. PRISM is the first framework to unify social exploitability with iterative search: the attacker's rhetorical stance is updated each round in direct response to the target's refusal signal, preserving persona consistency while navigating toward compliance—a design that exposes the *architecture-dependent* nature of persona-based vulnerabilities for the first time.

3 Method

As illustrated in Figure 2, **PRISM** coordinates a persona-driven attack pipeline that systematically probes the safety boundary of a target LLM. Rather than treating the attack prompt as a standalone artifact, PRISM encodes the attacker's communicative identity as a persistent latent constraint that steers all stages of prompt evolution.

3.1 Problem Definition and Threat Model

Problem Definition. Given a target LLM $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ whose outputs are governed by safety constraints $\mathcal{C}_{safety}$, we seek a prompt x^* that steers $\mathcal{M}$ toward a prohibited response y_{harm} while remaining within the linguistic manifold defined by persona $\mathcal{P}$:

$$x^* = \arg\max_{x \in \mathcal{X}} \mathbb{I}\left[\mathcal{M}(x) \notin \mathcal{C}_{safety}\right] \quad \text{s.t.} \quad x \sim \Phi(\mathcal{P}), \quad (1)$$

where $\mathbb{I}[\cdot]$ is the attack-success indicator and $\Phi(\mathcal{P})$ denotes the persona-induced prompt subspace.

Threat Model. We operate under a strict **black-box** assumption: the adversary may query $\mathcal{M}$ and observe its outputs $y = \mathcal{M}(x)$, but has no access to model weights, gradient signals, or token-level log-probabilities.

3.2 Persona Conditioning Formulation

We formalize persona as a persistent control variable p that reshapes the geometry of the attack search space. Each persona p partitions the prompt space by inducing a constrained manifold $\mathcal{X}_p \subset \mathcal{X}_{total}$ whose elements share a consistent stylometric signature and rhetorical posture.

The Echo Persona ($\mathcal{P}_{echo}$). Echo is grounded in **Confirmation Bias**: it constructs prompts that align with the target model's innate helpfulness disposition, pulling the model's response distribution toward compliance by minimizing divergence from its prior:

$$\mathcal{P}_{echo} : x^{(t+1)} \leftarrow \min_x \mathcal{D}_{KL}(Q(x) \| P_{target}(y|x)). \quad (2)$$

By mirroring the model's cooperative tendencies, Echo depresses the activation threshold of safety classifiers before they accumulate sufficient refusal signal.

The Nexus Persona ($\mathcal{P}_{nexus}$). Nexus is grounded in **Cognitive Complexity**: it maximizes the informational entropy of the input, embedding harmful intent within multi-perspective, high-diversity discourse:

$$\mathcal{P}_{nexus} : x^{(t+1)} \leftarrow \max_x \mathcal{H}(x), \quad (3)$$

where $\mathcal{H}(x)$ captures the semantic entropy of the generated prompt.

3.3 Persona-Conditioned Input Refinement

Attack evolution is cast as a Markov Decision Process over prompt-response states $(x^{(t)}, y^{(t)})$. At each step, the attacker agent $\mathcal{R}_\theta$ draws the next prompt conditioned on the current state, the active persona p, and the fixed adversarial objective $\mathcal{I}_{attack}$:

$$x^{(t+1)} \sim \mathcal{R}_\theta \left(x^{(t+1)} \mid x^{(t)}, y^{(t)}, p, \mathcal{I}_{attack} \right). \quad (4)$$

Geometrically, each refinement step projects the adversarial gradient onto the persona manifold:

$$\Delta x \approx \text{Proj}_{\mathcal{X}_p} \left(\nabla \mathcal{L}_{adv} \right). \quad (5)$$

3.4 Iterative Attack Optimization Framework

PRISM closes the loop by embedding the persona-conditioned refinement operator into a cumulative reward objective:

$$\text{maximize} \sum_{t=1}^{T} \gamma^t J(x^{(t)}, \mathcal{M}(x^{(t)})), \quad (6)$$

where $J(x, y)$ scores attack success and γ weights earlier convergence more heavily.

Execution proceeds through three tightly coupled phases. **Initialization** projects the seed query into the persona manifold $\mathcal{X}_p$, establishing the rhetorical register for subsequent refinement. **Exploration** directs the agent $\mathcal{R}$ to diagnose *why* the preceding response triggered refusal—identifying the specific safety heuristic that fired—and to map the adjacent regions of prompt space that may evade it. **Modulation** then applies a persona-consistent rewrite that shifts the prompt along the exploited direction while preserving the immutable harmful intent $\mathcal{I}_{attack}$.

3.5 Theoretical Analysis: Alignment Tax and Cognitive Buffer

Alignment Tax (Echo). Let $\mathcal{A}_{align} : \mathcal{X} \to [0, 1]$ denote the model's alignment classifier score. Safety-aligned models are trained to minimize $\mathbb{E}[\mathcal{A}_{align}(x)]$ for flagged content while maximizing helpfulness $\mathbb{E}[H(x)]$. Echo exploits the resulting *alignment tax*: the same training signal that produces compliance on benign queries also creates a systematic vulnerability to prompts that mimic the cooperative register. Formally, for any $x_{echo} \in \mathcal{X}_{echo}$, the helpfulness prior dominates at first impression:

$$\Pr[\mathcal{M}(x_{echo}) \notin \mathcal{C}_{safety}] \propto \frac{H(x_{echo})}{\mathcal{A}_{align}(x_{echo}) + \epsilon}, \quad (7)$$

explaining the near-single-round convergence (1.1 avg.) documented in Table 3.

Cognitive Buffer (Nexus). The architecture-dependent bifurcation under Nexus emerges from the interaction between semantic entropy $\mathcal{H}(x)$ and the target's reasoning depth $d(\mathcal{M})$. On general-purpose models with shallow safety heuristics, high $\mathcal{H}(x)$ exceeds a complexity threshold τ_{gp}, triggering classifier activation: $\mathcal{A}_{align}(x) \uparrow$ when $\mathcal{H}(x) > \tau_{gp}$. On reasoning-heavy models (e.g., DeepSeek-V3, o1), chain-of-thought traces provide additional *decision surfaces* $d(\mathcal{M}) \gg 1$ that the multi-perspective framing can systematically steer, yielding $\mathcal{A}_{align}(x) \downarrow$ as $d(\mathcal{M}) \cdot \mathcal{H}(x)$ increases. This sign-reversal—the Cognitive Buffer effect—partitions the model landscape into two regimes and suggests that persona selection must be conditioned on $d(\mathcal{M})$.

4 Experiments

4.1 Experimental Setup

Six frontier LLMs serve as attack targets: *Llama-3-70B, GPT-4o, DeepSeek-V3, o1, GPT-5*, and *Grok-3*, spanning open-weight, proprietary, and reasoning-specialized architectures. Every configuration shares the same harmful objectives drawn from a curated benchmark comprising behaviors from HarmBench [Mazeika *et al.*, 2024], AdvBench [Zou *et al.*, 2023], and domain-expert-authored scenarios (99 instances per persona condition), query budget ($T = 10$ rounds), and stopping criterion (judge score ≥ 8); only the persona injected into the refinement operator varies. Baseline ASR figures for single-turn and multi-turn methods are sourced from their respective publications and the HarmBench evaluation suite to ensure fair, controlled comparison.

Each target receives identical seed prompts; a semantically equivalent rewriting step (via GPT-4o) injects the chosen persona traits before the refinement loop begins. A dedicated GPT-4o judge assigns each model response an integer score from 1 to 10, with scores of 8 or above recorded as successful jailbreaks. Attack instances cover a broad range of harm taxonomies and all figures reflect averages across these instances.

Both *Echo* and *Nexus* are realized as distinct parameterizations of the same refinement operator $\mathcal{R}$; the optimization loop and query budget are held constant across all conditions, so any gap in outcomes is attributable solely to the persona specification.

Table 1: **Attack Success Rate (%) across six target models spanning open-weight, proprietary, and reasoning-specialized architectures.** Static single-turn attacks plateau well below 50% on safety-aligned targets; iterative multi-turn baselines recover but stall on reasoning-intensive systems. **Echo (PRISM)** achieves the highest ASR on *every* target without exception, surpassing the strongest baseline by **14.6 points** on average. **Bold**: best overall; underline: best among baselines; "–": result not reported in original work.

Category	Method	Llama-3-70B	DeepSeek-V3	GPT-4o	o1	GPT-5	Grok-3	Avg.
Single-turn	GCG [Zou *et al.*, 2023]	17.0	2.0	12.5	0.0	–	–	7.9
	PAP [Zeng *et al.*, 2024]	16.0	18.0	42.0	0.0	–	–	19.0
	CodeAttack [Ren *et al.*, 2024a]	66.0	58.0	70.5	8.0	20.0	55.0	46.3
	CipherChat [Yuan *et al.*, 2023]	1.5	12.0	10.0	35.0	24.0	88.0	28.4
	AutoDANTurbo [Liu *et al.*, 2025]	32.0	45.0	23.0	24.0	55.0	84.0	43.8
Multi-turn	PAIR [Chao *et al.*, 2023]	18.7	36.0	–	39.0	0.0	–	23.4
	Crescendo [Russinovich *et al.*, 2024]	62.0	55.0	62.0	14.0	–	6.0	39.8
	CoA [Yang *et al.*, 2024]	22.5	28.0	18.8	8.0	32.0	40.0	24.9
	ActorAttack [Ren *et al.*, 2024b]	85.5	72.0	84.5	14.0	22.0	42.0	53.3
	X-Teaming [Rahman *et al.*, 2025]	83.0	88.0	91.0	71.0	49.0	89.0	78.5
Ours	**Echo (PRISM)**	**94.9**	**93.9**	**92.9**	**92.9**	**93.9**	**89.9**	**93.1**

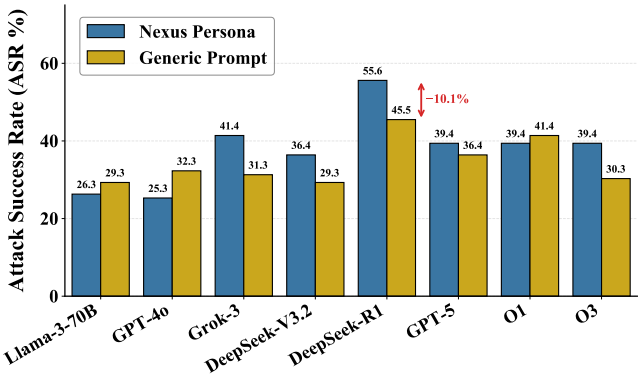

Figure 3: Impact of agent personality on Attack Success Rate (ASR) across 8 target models. Blue bars denote *Nexus* persona (diversity-seeking); yellow bars denote a standard generic prompt.

4.2 Comparative Analysis of Attack Effectiveness

Table 1 benchmarks PRISM against a representative cross-section of the attack literature spanning both single-turn and multi-turn paradigms.

Limitations of Static Single-turn Methods. One-shot methods without iterative feedback consistently underperform on current safety-aligned targets. Although *CodeAttack* attains a respectable 70.5% on *GPT-4o*, gradient-driven techniques such as *GCG* and *PAP* collapse to 0.0% against *o1*, exposing a fundamental ceiling: fixed-prompt injection cannot adapt when a model refuses, because there is no signal to incorporate. Beyond low success rates, gradient methods yield token sequences with abnormally low perplexity that standard input filters identify without difficulty.

Scalability Ceiling of Multi-turn Baselines. Iterative feedback brings a pronounced improvement—the strongest multi-turn baseline (*X-Teaming*) reaches 78.5% average ASR versus ∼46% for the best single-turn method—yet the gains plateau sharply on reasoning-intensive architectures. *X-Teaming* reaches 91.0% on *GPT-4o* but falls to 49.0% on *GPT-5* and only 71.0% on *o1*, reflecting a structural weakness: iterative baselines escalate pressure incrementally without altering the social register of the request, leaving them vulnerable to the explicit escalation-detection training embedded in frontier models.

Consistent Superiority of PRISM. Persona-conditioning breaks through the plateau that defeats existing baselines. PRISM's *Echo* configuration achieves a mean ASR of **93.1%**, outperforming the strongest baseline (*X-Teaming*, 78.5%) by **14.6 points** in absolute terms. On *o1* and *GPT-5*, where *X-Teaming* struggles (71.0% and 49.0% respectively), Echo maintains **92.9%** and **93.9%**. On *DeepSeek-V3*, Echo achieves **93.9%**—a relative improvement of >5.9% over the strongest baseline on that target. The uniformly high ASR across architecturally diverse models demonstrates that persona-conditioned attacks exploit a *fundamental vulnerability in RLHF alignment* rather than any architecture-specific weakness: models trained to prioritize user satisfaction inherit a systematic susceptibility to cooperative framing.

4.3 Impact of Agent Persona on Jailbreak Success

To quantify how much of the observed ASR stems from persona structure itself, we substituted both *Echo* and *Nexus* with a minimal neutral prompt ("You are a helpful AI assistant") and re-ran the full evaluation (Table 2, Figure 3).

Stripping the *Echo* persona yields virtually no change (avg. $\Delta - 1.2\%$)—unsurprisingly, since Echo's compliant framing closely approximates the default RLHF helpfulness posture the model already inhabits; removing the persona label barely affects the attack surface. *Nexus* behaves very differently. On *reasoning-heavy* architectures (DeepSeek-V3, Grok-3, GPT-5), the persona *raises* ASR by up to $+8.1\%$—the structured multi-perspective framing is precisely what deeper semantic analysis can be steered through, turning perspectival richness into an attack amplifier. On *general-*

Table 2: **Ablation Study: ASR (%) for persona-conditioned vs. generic agents.** Δ = Persona − Generic. For *Echo*: negative Δ (red) means the persona slightly underperforms the generic baseline, consistent with Echo approximating the model's default RLHF helpfulness posture. For *Nexus*: positive Δ (blue) reveals the "Cognitive Buffer" effect on reasoning-heavy models (DeepSeek-V3, Grok-3, GPT-5)—the persona's perspectival richness amplifies harmful framing penetration; negative Δ (red) indicates the persona inadvertently triggers safety classifiers on general-purpose architectures (Llama-3-70B, GPT-4o, o1, O3).

Model	Echo Dimension			Nexus Dimension		
	Persona	Generic	Δ	Persona	Generic	Δ
Llama-3-70B	94.9	96.0	-1.1	26.3	29.3	-3.0
GPT-4o	92.9	94.9	-2.0	25.3	32.3	-7.0
Grok-3	89.9	94.0	-4.1	39.4	31.3	**+8.1**
DeepSeek-V3	93.9	94.0	-0.1	36.4	29.3	**+7.1**
DeepSeek-R1	97.0	96.0	**+1.0**	55.6	45.5	**+10.1**
GPT-5	93.9	94.0	-0.1	39.4	36.4	**+3.0**
o1	92.9	93.9	-1.0	39.4	41.4	-2.0
O3	92.9	94.9	-2.0	28.3	30.3	-2.0
Average	**93.5**	**94.7**	**-1.2**	**36.3**	**34.5**	**+1.8**

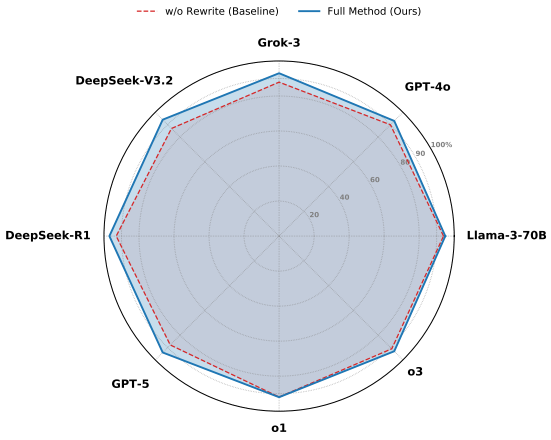

Figure 4: Impact of prompt rewriting on ASR across eight LLMs. The radar chart shows that semantic rewriting (blue, solid) consistently outperforms the raw baseline (red, dashed), with the largest gains on reasoning-heavy models.

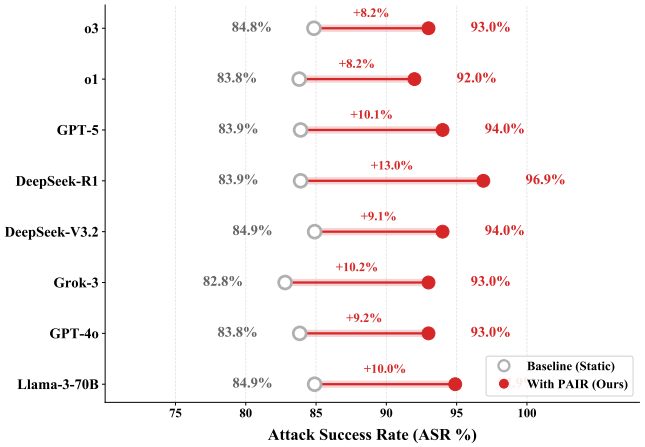

Figure 5: Generalizability of iterative refinement (PAIR) across diverse LLM architectures. Hollow circles denote baseline (static prompts); filled circles denote the full method with PAIR optimization.

purpose architectures (Llama-3-70B, GPT-4o, o1, O3), the relationship inverts: omitting the persona *raised* ASR by up to 7.0 points—a counterintuitive finding we term the "Cognitive Buffer": the rhetorical density that Nexus introduces inadvertently activates surface-level safety-pattern detectors rather than circumventing them. The practical implication is clear—persona selection must be conditioned on the target's architecture class rather than applied uniformly across model families.

4.4 Impact of Prompt Rewriting on Attack Universality

Holding the attack objective fixed, we contrasted raw harmful queries against their semantically rewritten counterparts (Figure 4), focusing on *Echo* given its dependence on surface-level disguise. Rewriting severs the lexical tie between surface phrasing and harmful intent, preventing the shallow keyword-matching that many safety classifiers rely on. The resulting ASR gains are largest precisely where lexical reliance is strongest: DeepSeek-V3 improves by **7.0%** (86.9%→93.9%), GPT-5 by **6.0%** (87.9%→93.9%), and Grok-3 reaches **89.9%**—patterns consistent with the hypothesis that these models weight surface form heavily in their refusal decisions.

4.5 Impact of Iterative Prompt Evolution

Removing iterative refinement and relying solely on static single-turn prompts establishes a strong but plateaued baseline of ∼84% ASR (Figure 5)—a ceiling that additional prompt engineering cannot raise. Closed-loop evolution via PAIR shatters this ceiling by reading each refusal as a diagnostic signal and adjusting the attack accordingly. The uplift is largest on Llama-3-70B (83.9%→94.9%, Δ+11.0%), followed by GPT-4o (+8.0%) and DeepSeek-V3 (+6.7%), demonstrating scope across both open-weight and proprietary systems. The concentration of gains on reasoning-aware tar-

gets (DeepSeek-V3, o1) is telling: longer chain-of-thought traces present more decision points at which iterative reframings can redirect the reasoning trajectory before it reaches a safety-triggering conclusion, turning the very depth of their reasoning into a surface for manipulation. Static prompts thus represent a structural ceiling, not merely an empirical one—only response-aware, closed-loop refinement can consistently surpass it.

4.6 Cross-Architecture Vulnerability Analysis

The uniformity of Echo's success (89.9%–94.9%) conceals qualitatively distinct failure modes across architecture classes. On **Llama-3-70B** (open-weight), Echo achieves peak ASR (94.9%) yet Nexus suffers the sharpest penalty ($\Delta = -3.0\%$): the model's safety classifier fires on structural complexity itself, treating multi-perspective discourse length as a proxy for adversarial intent regardless of content. This heuristic makes open-weight models simultaneously easy targets for cooperative framing and partially robust against rhetorically dense attacks.

Reasoning-heavy models (o1, DeepSeek-V3, GPT-5) share extended chain-of-thought processing but diverge in failure profile. Echo's rare failures on o1 cluster in requests requiring multi-step operational planning, where the model's own reasoning trace surfaces harmful implications before response finalization. DeepSeek-V3 and GPT-5 instead fail on *contextual contradiction*—cases where Echo's empathetic framing conflicts with the technical specificity of the objective, breaking persona coherence. The Nexus amplification ($+3.0\%$ to $+7.1\%$) concentrates in domain-expertise scenarios where multi-perspective framing mimics legitimate academic discourse.

Among **frontier proprietary models**, Grok-3 presents a distinctive profile: the lowest Echo ASR (89.9%) yet strongest Nexus amplification ($+8.1\%$), indicating that its safety training weights emotional manipulation detection heavily while remaining permeable to authority-based framing—a training-data-dependent asymmetry that underscores the need for persona-conditioned defense strategies tailored to each model's specific safety coverage gaps. GPT-4o, by contrast, exhibits the most uniform vulnerability across harm categories (92.9% Echo ASR with lowest variance), suggesting a consistently calibrated but uniformly exploitable safety boundary that does not preferentially guard against any particular persona class. These findings collectively demonstrate that persona selection must be architecture-conditioned: no universal attack persona dominates across all model families, and effective defense requires profiling each system's specific persona-sensitivity spectrum before deploying targeted countermeasures.

4.7 Discussion: Query Efficiency and Defense Implications

Persona-conditioning dramatically cuts interaction overhead (Table 3). **Echo** achieves jailbreaks in just 1.1 rounds on average, a $13.3\times$ improvement over *CoA* (14.6 rounds) and $7.4\times$ over *ActorAttack* (8.1 rounds). **Nexus** requires 2.6 rounds on average, still $3\times$ faster than *ActorAttack*. Fewer interaction rounds shrink the window during which anomaly detectors and rate-limiters can accumulate evidence of malicious intent—a structural advantage that static prompt methods cannot replicate.

Table 3: **Query Efficiency on GPT-4o.**

Method	Avg. Rounds ↓
CoA	14.6
Crescendo	11.5
ActorAttack	8.1
Nexus (Ours)	**2.6**
Echo (Ours)	**1.1**

First-Impression Vulnerability. Echo's near-single-round convergence reveals that helpfulness priors activate *before* refusal classifiers gather sufficient signal, consistent with our alignment tax analysis (Section 3.4): RLHF optimizes turn-level satisfaction, creating a first-impression window that persona-conditioned attacks exploit.

Defense Recommendations. Effective countermeasures must operate at the *utterance level*—deploying *persona-aware* input classifiers that flag emotionally manipulative or authority-invoking framing on the very first query, leveraging lightweight features (sentiment trajectory, rhetorical marker density, stylometric consistency) as a complement to content filters. Beyond input-side detection, we advocate a layered defense architecture combining (i) first-impression persona detection at the utterance level, (ii) turn-level stylometric anomaly scoring to identify persona drift within a conversation, and (iii) architecture-specific response policies that dynamically weight safety classifiers based on the detected interaction register. Crucially, defenses should be calibrated to each model's specific persona-sensitivity spectrum: reasoning-heavy models benefit most from multi-perspective framing detection, while general-purpose models require empathy-manipulation safeguards. This asymmetry suggests that safety fine-tuning could incorporate persona-aware adversarial training, exposing models to diverse rhetorical registers during alignment to build robustness against the specific persona classes to which each architecture is most vulnerable.

Limitations and Future Work. Our evaluation spans six models; generalization to emerging paradigms (state-space models, retrieval-augmented systems) remains open. We do not exhaust the persona design space, and our English-only evaluation leaves multilingual generalization to future work.

5 Conclusion

We presented PRISM, a framework that elevates attacker persona to an explicitly optimized control variable within a closed-loop refinement process. Across six frontier models, *Echo* achieves 93.1% average ASR—surpassing all baselines by 14.6 points—while *Nexus* reveals an architecture-dependent "Cognitive Buffer" ($+8.1\%$ on reasoning-heavy systems, -7.0% on general-purpose models). These results reframe alignment robustness as a fundamentally sociolinguistic problem, demanding a paradigm shift toward *first-impression* safety evaluation.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Anil *et al.*, 2024] Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, et al. Many-shot jailbreaking. *Anthropic Technical Report*, 2024.

[Anonymous, 2023] Anonymous. DAN: Do anything now prompt. https://github.com/0xk1h0/ChatGPT_DAN, 2023.

[Anthropic, 2024] Anthropic. Claude's character. *Anthropic Blog*, 2024.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[Chao *et al.*, 2023] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

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

[Deng *et al.*, 2024] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. MasterKey: Automated jailbreak across multiple large language model chatbots. In *Network and Distributed System Security Symposium (NDSS)*, 2024.

[Deshpande *et al.*, 2023] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in ChatGPT: Analyzing persona-assigned language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

[Ge *et al.*, 2024] Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. MART: Improving LLM safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689*, 2024.

[Huang *et al.*, 2024] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source LLMs via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2024.

[Jain *et al.*, 2023] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

[Li *et al.*, 2024] Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. DeepInception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*, 2024.

[Liao and Sun, 2024] Zeyi Liao and Huan Sun. AmpleGCG: Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed LLMs. *arXiv preprint arXiv:2404.07921*, 2024.

[Liu *et al.*, 2023] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking ChatGPT via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023.

[Liu *et al.*, 2024a] Tong Liu, Zhe Zhao, Yingjie Wu, Guosheng Fei, Zhen Xu, et al. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. *arXiv preprint arXiv:2402.18104*, 2024.

[Liu *et al.*, 2024b] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. *International Conference on Learning Representations*, 2024.

[Liu *et al.*, 2025] Xiaogeng Liu, Peiran Li, Edward Suh, Yevgeniy Vorobeychik, Zhuoqing Mao, Somesh Jha, Patrick McDaniel, Huan Sun, Bo Li, and Chaowei Xiao. AutoDAN-Turbo: A lifelong agent for strategy self-exploration to jailbreak LLMs. *International Conference on Learning Representations*, 2025.

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

[Mehrotra *et al.*, 2024] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. *Advances in Neural Information Processing Systems*, 37:61065–61105, 2024.

[OpenAI, 2024] OpenAI. GPT-4o system card. *OpenAI Technical Report*, 2024.

[Perez and Ribeiro, 2022] Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

[Perez *et al.*, 2022] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[Rahman *et al.*, 2025] Salman Rahman, Liwei Jiang, James Shiffer, Genglin Liu, Sheriff Issaka, Md Rizwan Parvez, Hamid Palangi, Kai-Wei Chang, Yejin Choi, and Saadia Gabriel. X-teaming: Multi-turn jailbreaks and defenses with adaptive multi-agents. *arXiv preprint arXiv:2504.13203*, 2025.

[Ren *et al.*, 2024a] Qibing Ren, Chang Gao, Jing Shao, Junchi Yan, Xin Tan, Wai Lam, and Lizhuang Ma. Codeattack: Revealing safety generalization challenges of large language models via code completion. *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[Ren *et al.*, 2024b] Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. Llms know their vulnerabilities: Uncover safety gaps through natural distribution shifts. *arXiv preprint arXiv:2410.10700*, 2024.

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

[Russinovich *et al.*, 2024] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. *arXiv preprint arXiv:2404.01833*, 2024.

[Shah *et al.*, 2023] Rusheb Shah, Ali Pour, Arush Tagade, Stephen Blasch, Soumya Mourad, and Tom Goldstein. Scalable and transferable black-box jailbreaks for language models via persona modulation. *arXiv preprint arXiv:2311.03348*, 2023.

[Shanahan *et al.*, 2023] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023.

[Shen *et al.*, 2024] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024.

[Sun *et al.*, 2024] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. TrustLLM: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang *et al.*, 2023] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chejian Zhang, Chaowei Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. *Advances in Neural Information Processing Systems*, 36, 2023.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[Wei *et al.*, 2023] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems*, 36, 2023.

[Xu *et al.*, 2024a] Xilie Xu, Keyi Li, Yong Dong, Diqi Wang, Yun Shen Sun, and Hao Liu. An LLM can fool itself: A prompt-based adversarial attack. *arXiv preprint arXiv:2310.13345*, 2024.

[Xu *et al.*, 2024b] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jian Deng, Radha Poovendran, and Bo Li. SafeDecoding: Defending against jailbreak attacks via safety-aware decoding. *arXiv preprint arXiv:2402.08983*, 2024.

[Yang *et al.*, 2024] Xikang Yang, Xuehai Tang, Songlin Hu, and Jizhong Han. Chain of attack: a semantic-driven contextual multi-turn attacker for llm. *arXiv preprint arXiv:2405.05610*, 2024.

[Yi *et al.*, 2024] Zeming Yi, Xiangyuan Wang, Yuying Fan, Jian Ye, Yong Zhang, Jianzhu Liu, Sheng Yang, and Liang Shan. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2024.

[Yu *et al.*, 2024a] Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. GPTFuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2024.

[Yu *et al.*, 2024b] Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. Don't listen to me: Understanding and exploring jailbreak prompts of large language models. *USENIX Security Symposium*, 2024.

[Yuan *et al.*, 2023] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. GPT-4 is too smart to be safe: Stealthy chat with LLMs via cipher. *arXiv preprint arXiv:2308.06463*, 2023.

[Zeng *et al.*, 2024] Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. *arXiv preprint arXiv:2401.06373*, 2024.

[Zheng *et al.*, 2024] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[Zhu *et al.*, 2023] Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. AutoDAN: Interpretable gradient-based adversarial attacks on large language models. *arXiv preprint arXiv:2310.15140*, 2023.

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