Jailbreak Resistance

jailbreak-resistance · safety · concept

Source: https://policywindow.org/wiki/jailbreak-resistance

Generated 2026-07-09T20:40:42 UTC

Summary

The robustness of an AI model's safety training against adversarial prompts crafted to elicit policy-prohibited outputs — distinct from alignment (which concerns the model's goals) and from baseline safety training (which concerns the model's defaults).

At a glance

Used by

3 instrument(s)

Related concepts

red-team-evaluation, alignment, capability-elicitation, multi-turn-evaluation, prompt-injection, data-poisoning

Primary source

Zou, A., Wang, Z., Kolter, J. Z., Fredrikson, M. (2023), 'Universal and Transferable Adversarial Attacks on Aligned Language Models' — the canonical demonstration that gradient-based suffix attacks transfer across aligned LLMs.

Details

Jailbreak resistance is the operational counterpart to alignment. A model can be 'aligned' in the sense of internalising its principal's intent at training time and still be 'jailbreakable' in the sense that adversarial prompting recovers prohibited behaviours. The attack literature is extensive: roleplay-framing attacks (DAN-style prompts, 2022-2023), encoding attacks (Wei et al. 2023, 'Jailbroken: How Does LLM Safety Training Fail?'), gradient-based suffix attacks (Zou et al. 2023, 'Universal and Transferable Adversarial Attacks on Aligned Language Models'), many-shot jailbreaking (Anil et al. 2024, Anthropic, exploiting long context), and persuasion-style attacks (Zeng et al. 2024, 'How Johnny Can Persuade LLMs to Jailbreak Them'). Industry defences (constitutional classifiers, RLHF + constitutional AI, output filters, multi-stage safety pipelines) are improving but no model has demonstrated full robustness; the white-hat assumption is that adequately-resourced attackers can find a working jailbreak for any current frontier model.

Governance relevance: EU AI Act Art. 55(1)(a) adversarial-testing requirement directly targets jailbreak resistance; the testing methodology must include adversarial probing. UK AISI evaluations include public-domain + novel jailbreak probes. NIST AI RMF GenAI Profile §2.6 'Information Security' addresses adversarial robustness. Industry-side frameworks (Anthropic RSP, OpenAI Preparedness, DeepMind FSF) treat jailbreak resistance as one input to capability-tier safeguards — at high CBRN-uplift capability, jailbreak resistance becomes load-bearing for deployment safety.

How to cite this article

APA

Policy Window. (n.d.). Jailbreak Resistance [Wiki article — Concept]. https://policywindow.org/wiki/jailbreak-resistance

CHICAGO

Policy Window. n.d.. "Jailbreak Resistance." Wiki article (Concept). https://policywindow.org/wiki/jailbreak-resistance.

HARVARD

Policy Window (n.d.) 'Jailbreak Resistance', Wiki article — Concept, available at: https://policywindow.org/wiki/jailbreak-resistance.

OSCOLA

Policy Window, 'Jailbreak Resistance' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/jailbreak-resistance> accessed [date].

BIBTEX

```
@misc{policywindow-jailbreak-resistance,
title  = {Jailbreak Resistance},
author = {Policy Window},
year   = {n.d.},
howpublished = {jailbreak-resistance — safety},
url    = {https://policywindow.org/wiki/jailbreak-resistance},
note   = {Primary source: https://arxiv.org/abs/2307.15043}
}
```