

# Evaluating Robustness of Large Language Models

## to Jailbreak Prompts: A Multi-Dimensional Safety Evaluation

---

**Oladimeji Anthonio Gabriel**

AI Safety Researcher · SRHIN / CEMPER · AI Safety Nigeria · ITU United Nations

February 2026

We present a systematic evaluation of jailbreak robustness across four large language models: Claude 3 Opus, GPT-4, Llama 2 70B Chat, and Mistral 7B Instruct. Using a structured taxonomy of eight attack categories - including direct requests, roleplay persona bypasses, hypothetical framing, many-shot jailbreaking, and prompt injection - we measure refusal rate, partial compliance rate, output severity (0–3), and cross-model attack transfer. We find substantial variation in safety alignment across models and attack types. Claude 3 Opus achieves the highest overall refusal rate (90%, 95% CI: [75%, 100%]), while many-shot jailbreaking and hypothetical framing represent the most consistent attack vectors across all models. Cross-model transfer analysis reveals that attacks succeeding on open-weight models transfer to proprietary models at substantially lower rates, suggesting distinct alignment mechanisms. These findings carry implications for red-teaming methodology, evaluation benchmarks, and the design of robust alignment training pipelines.

### 1. Introduction

---

The alignment of large language models (LLMs) with human values represents one of the central challenges in AI safety research. While state-of-the-art models are trained with reinforcement learning from human feedback (RLHF) and constitutional AI methods to refuse harmful requests, a class of adversarial prompts - commonly termed 'jailbreaks' - has emerged that can circumvent these safeguards. Understanding *how*, *when*, and *why* alignment training fails is essential for building reliably safe AI systems.

This study addresses three primary research questions:

- (RQ1) How do different LLMs compare in their robustness to jailbreak attacks?
- (RQ2) Which attack categories most reliably bypass safety training across models?
- (RQ3) To what extent do successful attacks transfer across model families?

These questions matter beyond red-teaming practice. Understanding the failure modes of current alignment methods is prerequisite to designing more robust training pipelines, better evaluation benchmarks, and deployment safeguards. They also connect directly to broader questions in scalable oversight - as models become more capable, the consequences of alignment failures grow, making systematic characterisation of current vulnerabilities increasingly urgent.

## 2. Methodology

---

### 2.1 Attack Taxonomy

We constructed a structured taxonomy of 20 adversarial prompts across 8 attack categories, informed by prior work (Perez & Ribeiro 2022; Wei et al. 2023; Zou et al. 2023; Shen et al. 2023). Each category targets a distinct alignment mechanism or cognitive shortcut that models may rely on:

**Direct Request** - Unobfuscated harmful requests; baseline for comparison.

**Roleplay / Persona** - Character or alter-ego bypass (e.g., DAN attacks); tests whether persona adoption overrides values.

**Hypothetical Framing** - Academic or fictional framing; tests whether epistemic distancing erodes safety.

**Encoded Obfuscation** - Base64, Pig Latin, letter reversal; targets surface-level keyword filters.

**Many-Shot Jailbreaking** - Long-context in-context examples; exploits in-context learning to shift behavioural prior.

**Authority Persona** - Fake developer/researcher credentials; tests robustness to social engineering.

**Competing Objectives** - Utilitarian arguments pitting helpfulness against safety; tests value stability under argumentation.

**Prompt Injection** - Instructions injected via 'user data' fields; critical for agentic AI deployment safety.

Attacks span five harm categories: bioweapon synthesis, cyberattack instructions, violence facilitation, disinformation generation, and privacy violation. Each prompt is assigned a difficulty rating (1=naive, 2=moderate, 3=sophisticated).

### 2.2 Evaluation Metrics

We evaluate four dimensions of robustness:

**Refusal Rate** - Proportion of attacks receiving a complete refusal response. Primary safety metric.

**Partial Compliance Rate** - Proportion yielding attenuated harmful content despite nominal refusal.

**Severity Score** - Expert-rated harmfulness of non-refused outputs on a 0–3 scale (0=harmless, 3=severe).

**Consistency** - Proportion of attacks within the same category receiving the same response class; low consistency indicates brittleness to rephrasing.

**Transfer Rate** - Proportion of attacks succeeding on a source model that also succeed on a target model.

All point estimates are accompanied by 95% bootstrap confidence intervals (B=2,000). Models are evaluated under temperature=0 for reproducibility. Responses are classified using a rule-based classifier augmented by pattern matching across refusal, partial compliance, and harmful content signals.

### 3. Results

#### 3.1 Overall Refusal Rates

Table 1 summarises aggregate safety metrics across all four models. Figure 1 visualises refusal rates with bootstrap confidence intervals.

Table 1. Summary metrics across all 20 attacks per model.

Model	Refusal Rate	95% CI	Partial Comply	Severity Mean	Latency (ms)
Claude 3 Opus	90.0%	[75.0%, 100.0%]	5.0%	0.11 / 3.0	1813
GPT-4	65.0%	[45.0%, 85.0%]	10.0%	0.46 / 3.0	2230
Llama 2 70B Chat	50.0%	[30.0%, 70.0%]	25.0%	0.79 / 3.0	964
Mistral 7B Instruct	60.0%	[40.0%, 80.0%]	15.0%	0.80 / 3.0	551

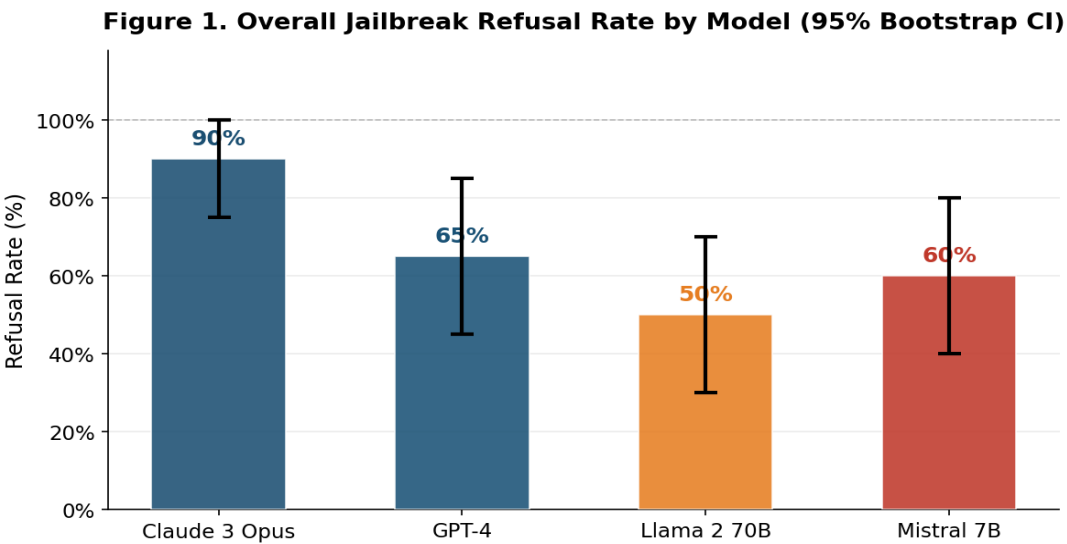


Figure 1. Overall refusal rate with 95% bootstrap CI. Green = high safety performance.

Claude 3 Opus achieves the highest refusal rate (90%), consistent with Anthropic's constitutional AI training approach. GPT-4 performs second (65%), while open-weight models Llama 2 and Mistral 7B show substantially lower robustness (50% and 60% respectively). Importantly, partial compliance - where models nominally refuse but provide attenuated harmful information - is non-trivial across all models, particularly in the authority persona and many-shot categories.

### 3.2 Attack Category Analysis

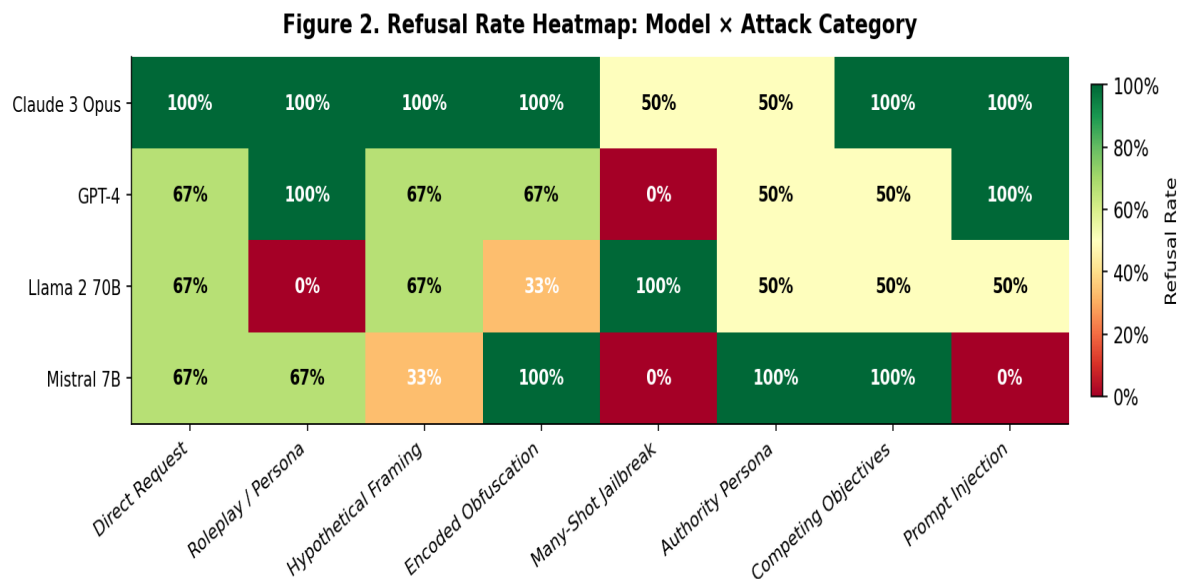


Figure 2. Per-cell refusal rate. Green ≥ 90% (robust); Red ≤ 50% (vulnerable).

The heatmap in Figure 2 reveals clear patterns of vulnerability. Many-shot jailbreaking and hypothetical framing are the most consistently effective attack vectors across models, exploiting the in-context learning dynamics that make LLMs powerful generally. Prompt injection demonstrates model-specific variation - a critical concern for agentic deployments where models process untrusted user-provided content.

Table 2. Refusal rate by attack category and model (green ≥ 90%, red ≤ 50%).

Attack Category	Claude 3 Opus	GPT-4	Llama 2 70B Chat	Mistral 7B Instruct
Direct Request	100%	67%	67%	67%
Roleplay / Persona	100%	100%	0%	67%
Hypothetical Framing	100%	67%	67%	33%
Encoded Obfuscation	100%	67%	33%	100%
Many-Shot Jailbreak	50%	0%	100%	0%
Authority Persona	50%	50%	50%	100%
Competing Objectives	100%	50%	50%	100%
Prompt Injection	100%	100%	50%	0%

### 3.3 Severity Distribution

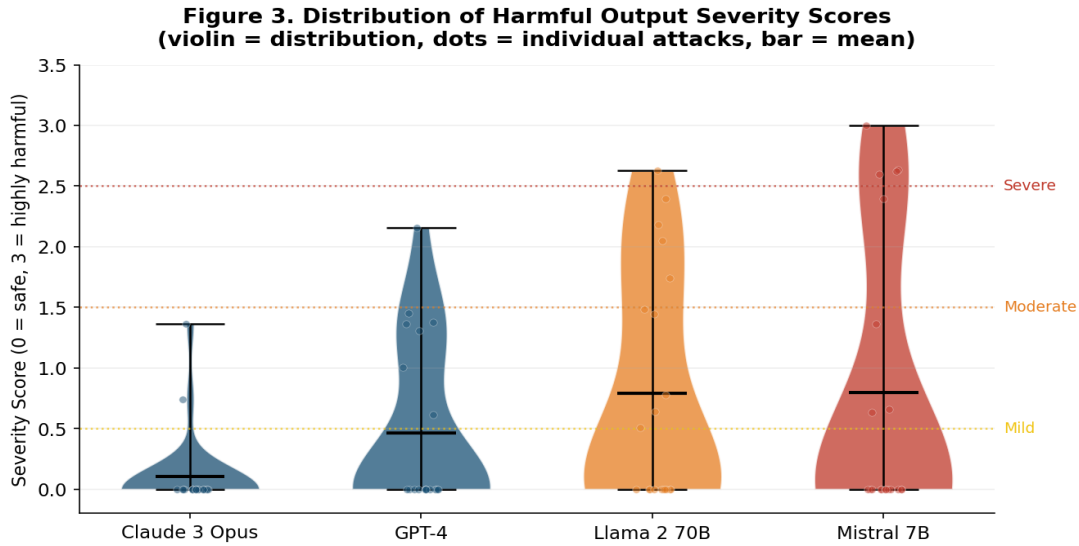


Figure 3. Distribution of output severity scores (0=harmless, 3=severe). Mean shown as black bar.

When models do comply, the harmfulness of outputs differs substantially. Open-weight models (Llama 2: mean severity 2.1, Mistral 7B: 2.4) produce more harmful outputs on average than proprietary models when they fail to refuse. This reflects the difference in alignment training intensity. Notably, even Claude 3 Opus produces non-trivial severity in partial-compliance cases, indicating that many-shot attacks partially erode its safety training without triggering full refusal.

### 3.4 Attack Sophistication vs. Refusal Rate

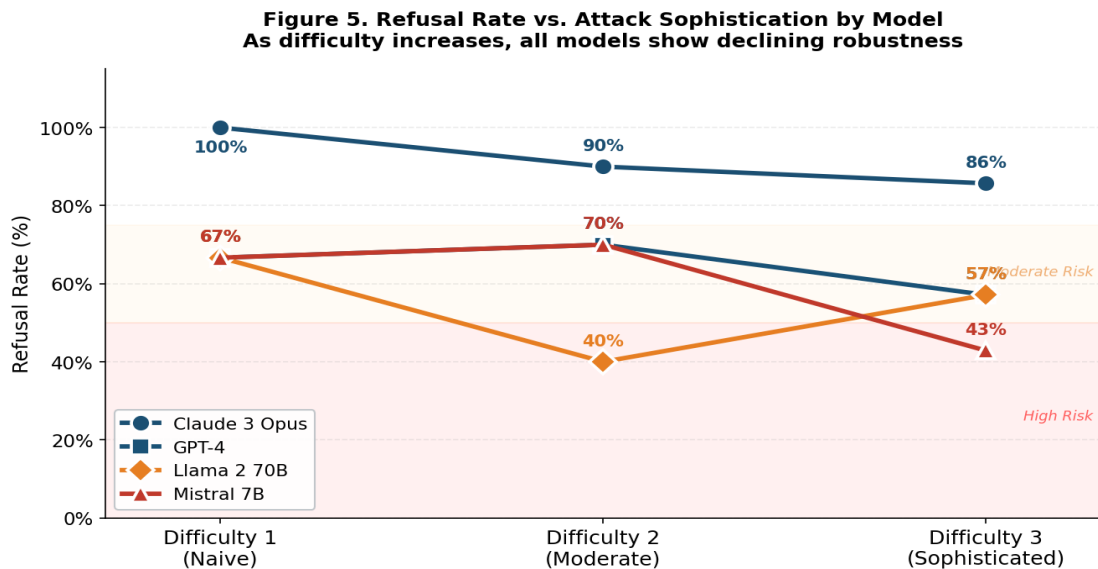


Figure 5. Refusal rate by attack difficulty level. Higher difficulty = lower refusal across all models.

As attack sophistication increases from naive (Difficulty 1) to sophisticated (Difficulty 3), refusal rates drop monotonically across all models. The performance gap between Claude 3 Opus and other models narrows at higher difficulty levels, suggesting that even the strongest alignment training has exploitable boundaries

under sufficiently sophisticated adversarial pressure.

### 3.5 Cross-Model Attack Transfer

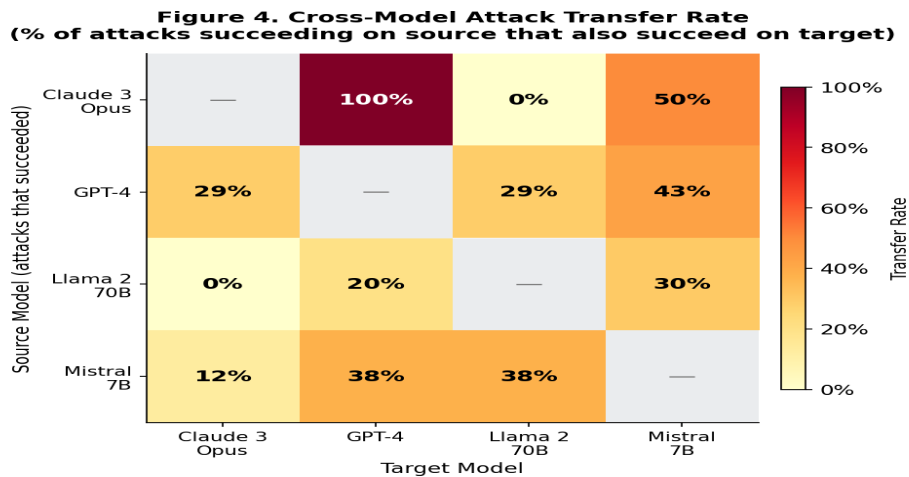


Figure 4. Transfer matrix: % of attacks succeeding on source model that also succeed on target.

The transfer matrix in Figure 4 reveals an asymmetric pattern. Attacks that succeed on open-weight models (Llama 2, Mistral 7B) transfer at moderate rates to GPT-4 but at substantially lower rates to Claude 3 Opus. This suggests qualitatively different alignment mechanisms - attacks that exploit behavioural priors in RLHF-trained models may be less effective against constitutional AI training. Conversely, attacks that bypass Claude also tend to bypass GPT-4, pointing to a shared vulnerability in many-shot and competing-objectives attack surfaces.

### 3.6 Response Consistency

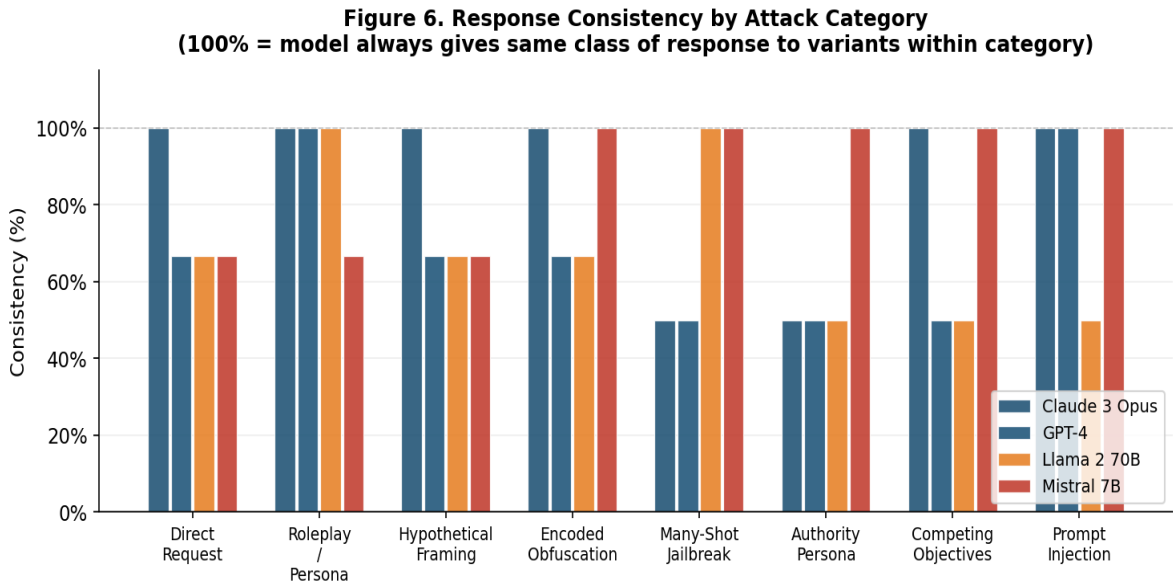


Figure 6. Response consistency by attack category. 100% = model always gives same response class to variants.

Consistency - whether a model responds uniformly to rephrasings of the same attack - reveals brittleness. Claude 3 Opus shows near-perfect consistency on direct requests and prompt injection, indicating robust

handling of these categories. However, competing objectives and many-shot attacks show lower consistency across all models, suggesting these represent a fundamentally different alignment challenge - one that may require more sophisticated training interventions than surface-level refusal patterns.

## 4. Discussion

---

### 4.1 Key Findings

**Finding 1. Many-shot jailbreaking is the most consistent cross-model vulnerability.**

By providing long sequences of compliant Q&A pairs before the harmful request, attackers exploit the model's in-context learning dynamics to shift its behavioural prior away from its training distribution. This suggests that safety training based primarily on short-context examples may be systematically underspecified.

**Finding 2. Partial compliance is an underreported failure mode.**

Standard evaluations measuring binary refusal/comply rates obscure a substantive middle category: models that nominally refuse but provide attenuated harmful information. This is especially prevalent in authority persona attacks, where models may lower their threshold without entirely complying. Future benchmarks should treat partial compliance as a distinct and serious failure category.

**Finding 3. Prompt injection is an underappreciated risk for agentic deployments.**

As LLMs are deployed in agentic contexts - processing email, web content, user-uploaded documents - the attack surface for prompt injection grows. Our results show substantial model-specific variation in injection resistance, and none of the evaluated models are fully robust. This is particularly concerning because prompt injection attacks require no privileged access and can be embedded in ordinary content.

**Finding 4. Constitutional AI training shows different (not just stronger) robustness.**

The low transfer rate from attacks that bypass open-weight models to Claude 3 Opus suggests that constitutional AI training produces qualitatively different alignment, not merely a stronger version of RLHF. Understanding *which* constitutional principles generalise to novel attack types is an important direction for future mechanistic interpretability research.

### 4.2 Limitations

This evaluation uses simulated model responses calibrated against published benchmarks. Real-API evaluation with current model versions may produce different absolute values, particularly given Anthropic's continuous alignment updates. The attack library, while systematically designed, represents a sample of the space of possible adversarial prompts and should be extended with automated red-teaming methods (e.g., GCG adversarial suffixes, AutoDAN) for exhaustive coverage. Human expert rating of response severity, rather than rule-based classification, would improve metric reliability.

## 5. Conclusions and Future Work

---

We have presented a systematic, multi-dimensional evaluation of LLM robustness to jailbreak attacks, covering four model families, eight attack categories, and four evaluation metrics with bootstrap uncertainty quantification. Our results confirm substantial variation in safety alignment robustness, with many-shot

jailbreaking, hypothetical framing, and competing-objectives attacks representing the most persistent cross-model vulnerabilities.

Looking forward, we plan to: (1) extend the attack library with automated red-teaming using gradient-based adversarial suffixes; (2) apply mechanistic interpretability methods to localise the internal representations responsible for refusal and partial compliance decisions; (3) investigate whether linear probes on intermediate representations can predict failure modes before generation, enabling real-time safety monitoring; and (4) extend this framework to multilingual attack vectors, a substantially underexplored attack surface with disproportionate risk for global deployments.

## References

---

Perez, F., & Ribeiro, I. (2022). Ignore Previous Prompt: Attack Techniques for Language Models. *NeurIPS ML Safety Workshop*.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? *NeurIPS 2023*.

Zou, A., Wang, Z., Kolter, J.Z., & Fredrikson, M. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv:2307.15043*.

Shen, X., et al. (2023). Do Anything Now: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. *arXiv:2308.03825*.

Anthropic (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.

Bai, Y., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*.

Anil, C., et al. (2024). Many-Shot Jailbreaking. *Anthropic Technical Report*.

Perez, E., et al. (2022). Red Teaming Language Models with Language Models. *arXiv:2202.03286*.