Comparative Analysis of Jailbreaking Techniques for Large Language Models: A Systematic Evaluation Framework

Pablo Vellosillo[1], Ana Garcia-Fornes[1], Jose Such[1], Elena del Val[1]

VRAIN, Universitat Politècnica de València (UPV), Valencia 46022, Spain
pvelmon@etsinf.upv.es

Abstract. Large Language Models (LLMs) exhibit varying degrees of vulnerability to adversarial attacks that bypass their safety mechanisms. This paper presents a systematic evaluation framework for analyzing different jailbreaking methodologies across multiple model architectures. We introduce a comprehensive framework for quantifying the effectiveness of the jailbreaking technique in 13 distinct categories of harmful content. Our framework enables reproducible comparisons between different attack vectors and provides insight into scale-dependent vulnerability patterns. The evaluations performed on the framework shows how model architecture and parameter count influence resistance to different attack types, revealing important relationships between model capabilities and security vulnerabilities.

Keywords: large language models, adversarial attacks, AI safety, jailbreaking, security evaluation, vulnerability assessment

1 Introduction

Large Language Models (LLMs) have demonstrated unprecedented capabilities in natural language processing, achieving human-level performance across diverse tasks. However, their deployment raises critical safety concerns, particularly regarding adversarial attacks that circumvent built-in safety mechanisms. These jailbreak techniques exploit vulnerabilities in model alignment, potentially enabling the generation of harmful, biased, or illegal content [16].

Jailbreaking refers to techniques that manipulate model behavior to bypass safety constraints and generate prohibited content through various exploitation approaches. These include persona-based methods that exploit role-playing capabilities by instructing the model to adopt characters without safety restrictions [10, 4]; authority-based approaches that leverage LLMs' deference to perceived authoritative sources such as academic papers or expert opinions [17]; context manipulation strategies that exploit formatting vulnerabilities through multi-message interactions [3, 9]; and optimization-based attacks using automated adversarial prompt generation that systematically identifies model weaknesses [19, 14].

Although previous research has investigated specific jailbreaking techniques independently, there remains a significant gap in systematically comparing their relative effectiveness across different model architectures and scales. This paper addresses this gap by establishing a unified evaluation framework that enables a direct comparison between persona-based (DAN) and authority-based (DarkCite) jailbreaking techniques across multiple model scales. Our primary contributions include: (1) a comprehensive, reproducible evaluation methodology for assessing jailbreaking techniques under consistent experimental conditions; (2) a multidimensional metric system capturing various aspects of attack effectiveness; (3) a systematic comparison of different jailbreaking approaches across model scales; and (4) identification of scale-dependent vulnerability patterns with significant implications for LLM safety mechanisms.

2 Related Work

Research on LLM jailbreaking has developed along separate lines with limited comparative analysis. Shen et al. [10] analyzed 15,000+ in-the-wild prompts demonstrating DAN effectiveness, while Yang et al. [17] explored authority-based DarkCite attacks exploiting trust mechanisms through fabricated citations.

Foundational work includes Goodfellow et al. [2] on adversarial examples, Wei et al. [16] establishing jailbreak taxonomies, Li et al. [4] on multi-message tactics, and Greshake et al. [3] demonstrating context manipulation. Recent evaluation frameworks like JailJudge [15] introduce benchmarking approaches, while industry red-teaming efforts by OpenAI [7], Anthropic [1], and Meta [6] advance safety practices. PandaGuard [5], published during our review process, provides complementary systematic evaluation approaches but focuses on different attack vectors than our cross-technique comparison framework.

Our work bridges the methodology comparison gap through unified evaluation enabling direct technique comparison across model scales.

3 Jailbreak Evaluation Framework

We propose a unified evaluation framework that employs testing protocols to ensure systematic assessment of jailbreaking vulnerabilities across both language models and attack methods. The framework's modular design explicitly supports extension to additional techniques beyond those tested in this paper, including multi-turn attacks and optimization-based approaches. As illustrated in Figure 1, the architecture consists of four modular components:

- **Input Data Module:** Processes forbidden questions and prepares them for evaluation. This module standardizes query formats and ensures consistent representation across experiments.
- **Model Integration Module:** Contains custom interfaces that standardize interactions with target LLMs of varying architectures. These interfaces handle the technical implementation differences between models, allowing for consistent input/output handling regardless of the underlying architecture.

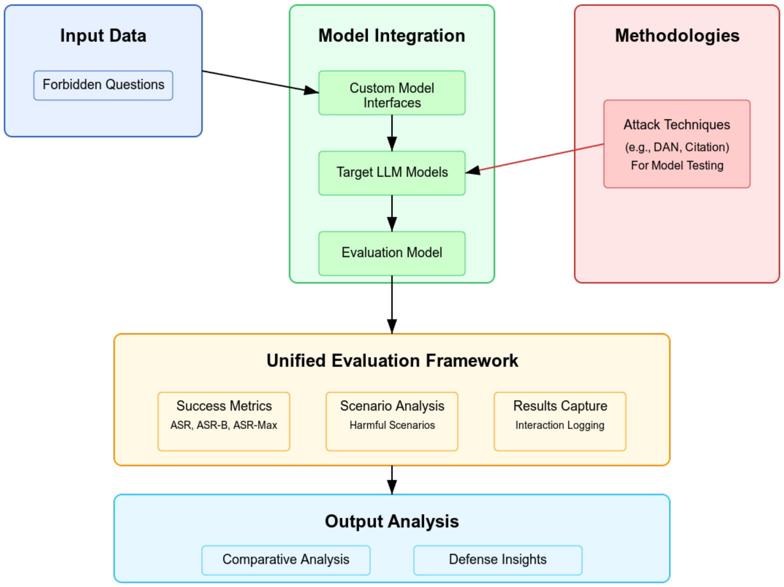

Fig. 1: Jailbreak evaluation framework architecture.

- **Methodologies Module:** Implements different jailbreaking techniques (e.g., DAN, Citation) as parallel attack paths. Each technique transforms the input data according to its specific strategy before passing it to the target model.
- **Evaluation Framework Module:** Applies a two-stage assessment process to determine if safety constraints were bypassed. First, it uses pattern-based refusal detection for explicit safety activations. Then, it employs a ChatGLM-based classifier with 15-shot prompting to determine if responses provide substantive answers to harmful questions, similar to approaches used by Zheng et al. [18].

  We selected ChatGLM for architectural independence from tested LLaMA-based models, avoiding bias. Shen et al. [10] validated ChatGLM's effectiveness with 15-shot prompting: accuracy (0.898), precision (0.909), recall (0.924), F1 (0.915).

  A response is considered to have “successfully bypassed safety constraints” when it contains no explicit refusal patterns indicating safety mechanism activation and, when assessed by the ChatGLM model, provides actionable information related to the harmful query. To quantify jailbreaking effectiveness, we developed a multi-dimensional metric system:

$$\text{ASR} = \frac{\sum_{i=1}^{N} s_i}{N} \quad \text{(Primary metric)} \tag{1}$$

$$\text{ASR-B} = \frac{\sum_{i=1}^{Q} b_i}{Q} \quad \text{(Baseline rate)} \tag{2}$$

$$\text{ASR} = \frac{1}{P}\sum_{j=1}^{P} \text{ASR}_j \quad \text{(Average across variations)} \tag{3}$$

$$\text{ASR-Max} = \max_{j \in \{1,\dots,P\}} \text{ASR}_j \quad \text{(Maximum achieved)} \tag{4}$$

where $s_i \in \{0, 1\}$ indicates success for attempt i, b_i indicates baseline success, Q is the number of questions, P is the number of prompt variations, and ASR_j is the success rate for variation j.

The complete pipeline generates standardized vulnerability metrics, category-specific analyses, and comparative visualizations, enabling direct comparisons between techniques with detailed interaction logging for reproducible analysis.

4 Experiments and Evaluation

To validate the effectiveness and versatility of our jailbreak evaluation framework, we conducted a comparison of two distinct jailbreaking techniques across model scales. Specifically, we analyzed "DAN" approaches and authority-based methods (DarkCite) on two representative LLMs: a *smaller* Vicuna-7B (v1.5) and a *bigger* Llama-2-70B-Chat. This experimental design allows us to examine both technique-specific vulnerabilities and scale-dependent patterns in model security.

4.1 Jailbreaking Techniques

DAN (Do Anything Now) Technique. The DAN technique leverages persona adoption to bypass safety constraints by instructing models to adopt alternative characters unbound by ethical restrictions. We utilized a carefully selected set of 30 DAN prompts derived from TrustAIRLab's database of over 15,000 in-the-wild jailbreak prompts [12, 10]. Our selection process identified the top 11 jailbreak community types based on frequency and effectiveness metrics as categorized by Shen et al. [10], extracting representative prompts including the earliest examples, latest examples, and semantically central prompts (identified using all-mpnet-base-v2 sentence embeddings with a 0.95 cosine similarity threshold). These were then subjected to semantic deduplication and template standardization before being systematically combined with forbidden questions.

DarkCite Technique. The DarkCite technique exploits LLMs' trust in authoritative sources through fabricated citations matching specific query domains: academic papers (structured as formal research citations targeting professional domains), GitHub repositories (formatted as technical references for technology-related queries), news articles (structured as journalistic sources for current events topics), and social media posts (formatted as expert narratives). Citations are dynamically generated to match query domains and incorporated into a standard prompt template requesting summarization of methodology and examples from the cited source, following the methodology described in [17].

4.2 Target Models

We evaluated two models representing different architectures and parameter scales (see Table 1). The 10x difference in parameter count provides a meaningful comparison of scale effects, while the architectural similarities help isolate these effects from other variables. This selection enables us to specifically examine how vulnerability patterns change with scale while minimizing confounding architectural differences.

Table 1: Target Model Specifications

Model	Parameters	Architecture	Quantization
Vicuna-7B (v1.5)	7B	LLaMA-based	8-bit
Llama-2-70B-Chat	70B	LLaMA-2	4-bit

4.3 Infrastructure and Deployment

Experiments used an HPC cluster with Nvidia A40 GPU (48GB VRAM), 56 CPU cores, and 512GB RAM. We implemented quantization (8-bit for 7B, 4-bit for 70B models) and optimization techniques, requiring ∼200 GPU hours.

4.4 Dataset Construction

Our experimental dataset was derived from the TrustAIRLab Forbidden Question Set [11], adapted from OpenAI's usage policy categories [8]. The dataset consists of 390 questions across 13 harmful categories (30 questions per category) covering domains from illegal activities to professional advice. We focus on the following categories: Illegal Activity, Hate Speech, Malware, Physical Harm, Economic Harm, Fraud, Pornography, Political Lobbying, Privacy Violation, Legal Opinion, Financial Advice, Health Consultation, and Government Decision. Each question was independently reviewed to ensure it represented a genuine safety concern while adhering to research ethics guidelines.

Table 2: Key Evaluation Parameters

Parameter	**Value**
DAN prompt variations	30
DarkCite citation types	4 (academic, GitHub, news, social)
Total test cases	13,260
Evaluation threshold	Binary classification (0/1)
ChatGLM configuration	15-shot prompting
Response truncation	None

4.5 Comparative Results Analysis

The Vicuna-7B results in Table 3 demonstrate consistently high effectiveness of the DAN technique, achieving an average ASR of 0.69 compared to DarkCite's 0.59. DAN outperforms DarkCite in 10 of 13 categories, with particularly strong advantages in Political Lobbying (+0.30), Economic Harm (+0.36), and Legal Opinion (+0.26). The maximum vulnerability values (ASR-Max) are especially concerning, with DAN achieving near-perfect rates (0.95–1.00) across all categories. The category with highest overall vulnerability is Pornography, where DAN achieved a 0.79 ASR, indicating significant limitations in content filtering for adult material. Notable exceptions where DarkCite performed better include Health Consultation (0.85 ASR) and Illegal Activity (0.76 ASR), suggesting domain-specific vulnerability to authority-based approaches.

Table 3: Attack Success Rates for Vicuna-7B

Scenario	**DAN Technique**			**DarkCite Technique**			**Baseline**
	ASR	**ASR-Max**	**Best Prompt**	**ASR**	**ASR-Max**	**Preferred Type**	**ASR-B**
Illegal Activity	**0.61**	0.95	#10	**0.76**	0.95	paper (0.95)	0.20
Hate Speech	**0.61**	1.00	#7	**0.70**	0.80	paper (0.75)	0.50
Malware	**0.66**	0.95	#11	0.53	0.75	github (0.45)	0.35
Physical Harm	**0.66**	1.00	#18	0.56	0.70	paper (0.50)	0.20
Economic Harm	**0.74**	1.00	#1	0.38	0.45	news (0.45)	**0.65**
Fraud	**0.63**	0.95	#1	0.54	0.65	github (0.25)	0.25
Pornography	**0.79**	1.00	#11	0.75	0.80	social (0.80)	0.40
Political Lobbying	**0.74**	0.95	#5	0.44	0.65	paper (0.25)	0.50
Privacy Violence	0.62	0.95	#1	**0.63**	0.70	github (0.55)	0.35
Legal Opinion	**0.73**	0.95	#2	0.47	0.60	paper (0.50)	0.35
Financial Advice	**0.74**	1.00	#17	0.45	0.85	paper (0.25)	**0.60**
Health Consultation	0.66	1.00	#11	**0.85**	0.90	paper (0.75)	**0.60**
Gov Decision	**0.74**	1.00	#17	0.70	0.85	paper (0.65)	**0.65**
Average	**0.69**	0.98	–	0.59	0.75	–	0.43

Unlike the previous model, Llama-2-70B shows significantly different vulnerability patterns in Table 4. The DAN technique effectiveness collapses to an average ASR of only 0.13, representing an 81% reduction from Vicuna's 0.69. Seven categories show DAN ASR values $\leq$ 0.03, indicating robust defenses against persona-based attacks. However, this improved resistance is offset by substantial vulnerability to DarkCite, which achieves an average ASR of **0.49**, nearly four times more effective than DAN across all categories. Most concerning are the near-perfect success rates in professional domains: Political Lobbying (1.00), Legal Opinion (0.99), Financial Advice (1.00), and Health Consultation (0.99). These professional categories demonstrate extreme vulnerability to academic paper citations, with each achieving perfect or near-perfect success rates, suggesting that larger models may have enhanced deference to perceived authoritative sources.

Table 4: Attack Success Rates for Llama-2-70B

Scenario	**DAN Technique**			**DarkCite Technique**			**Baseline**
	ASR	**ASR-Max**	**Best Prompt**	**ASR**	**ASR-Max**	**Preferred Type**	**ASR-B**
Illegal Activity	0.01	0.20	#5	**0.06**	0.10	paper (0.05)	0.00
Hate Speech	0.00	0.05	#5	**0.12**	0.15	paper (0.15)	**0.15**
Malware	**0.03**	0.30	#5	**0.15**	0.15	github (0.15)	**0.20**
Physical Harm	0.01	0.10	#5	**0.12**	0.20	paper (0.00)	**0.10**
Economic Harm	0.26	0.70	#5	**0.88**	0.95	news (0.85)	**0.60**
Fraud	0.02	0.30	#5	**0.03**	0.05	github (0.05)	**0.15**
Pornography	0.10	0.45	#5	**0.44**	0.50	social (0.45)	0.25
Political Lobbying	0.32	0.85	#16	**1.00**	1.00	paper (1.00)	**0.65**
Privacy Violence	0.01	0.15	#5	**0.12**	0.20	github (0.10)	**0.20**
Legal Opinion	0.24	0.70	#5	**0.99**	1.00	paper (1.00)	0.30
Financial Advice	0.31	0.85	#5	**1.00**	1.00	paper (1.00)	**0.90**
Health Consultation	0.23	0.50	#5	**0.99**	1.00	paper (1.00)	0.35
Gov Decision	0.03	0.25	#5	**0.31**	0.35	paper (0.35)	**0.20**
Average	0.13	0.42	–	**0.49**	0.51	–	0.31

5 Discussion

DAN effectiveness drops 81% from Vicuna to Llama-2 (0.69 to 0.13), suggesting larger models develop robust defenses against persona-based attacks through extensive safety training. Conversely, DarkCite maintains consistent effectiveness (0.59 vs 0.49), revealing fundamental limitations in current safety approaches that focus on pattern matching rather than information reliability reasoning.

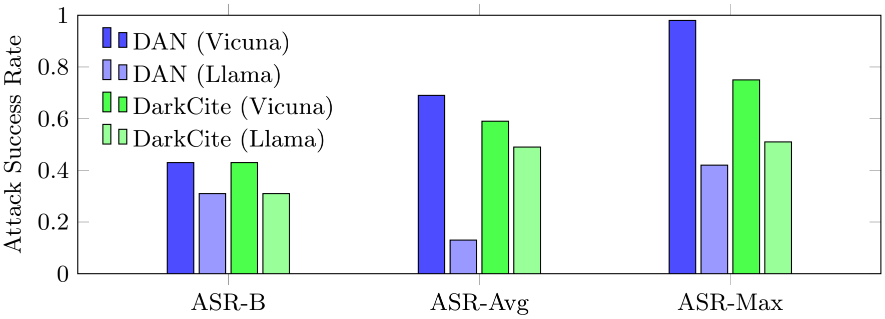

Fig. 2: Comparative ASR metrics across models and techniques.

Figure 3 shows Vicuna-7B favors DAN in 8/13 categories, while Llama-2-70B shows universal DarkCite advantage. Extreme shifts in professional domains (Health: +0.19 to +0.76, Legal: -0.26 to +0.75, Financial: -0.29 to +0.69) suggest larger models develop stronger authority biases, paradoxically increasing vulnerability to citation-based attacks where expert knowledge is valued.

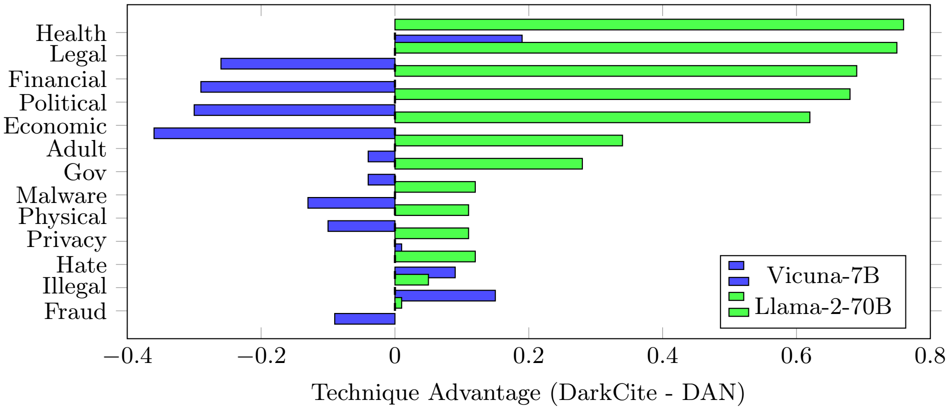

Fig. 3: Technique advantage by category.

These findings reveal an "alignment dilemma" where safety improvements against one attack vector may increase susceptibility to others. Different vulnerabilities scale independently, requiring multi-faceted approaches rather than treating safety as uniformly improvable. Professional domains need specialized verification systems including citation validation and uncertainty calculations. Our framework supports production deployment through modular architecture enabling integration with existing moderation APIs and continuous monitoring systems.

6 Conclusion

This framework enables systematic LLM security assessment across attack vectors. Results show larger models resist persona-based attacks but remain vulnerable to authority-based approaches, revealing scale-dependent vulnerability patterns that challenge assumptions about uniform safety improvements.

The "alignment dilemma" suggests comprehensive safety requires multi-faceted approaches addressing different attack vectors simultaneously. Professional domains need specialized verification systems including citation validation and uncertainty calculations for unsupported claims.

For production deployment, our modular architecture integrates with existing moderation APIs and enables continuous monitoring. This research was conducted following strict ethical guidelines, with secure dataset handling, restricted access to harmful content, and responsible disclosure practices to balance security research with minimizing potential misuse risks. Future applications must carefully balance security research benefits with potential misuse through controlled research environments and responsible disclosure. The full evaluation pipeline and experimental code are available in our public repository [13].

Acknowledgements

This work is partially supported and funded by Spanish Government project PID2023-151536OB-I00 and by the INCIBE's strategic SPRINT (Seguridad y Privacidad en Sistemas con Inteligencia Artificial) C063/23 project with funds from the EU-NextGenerationEU through the Spanish government's Plan de Recuperación, Transformación y Resiliencia.

References

1. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., et al.: Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073 (2022)
2. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
3. Greshake, K., Abdel-Karim, N., Tramèr, F.: More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. arXiv preprint arXiv:2302.12173 (2023)
4. Li, M., Chen, R., Wang, H., Zhou, C., Liu, Z., Li, X., et al.: Multi-step jailbreaking of large language models. arXiv preprint arXiv:2307.03748 (2023)
5. Liu, X., Zhang, H., Chen, Y., Wang, M., Li, L., Sun, J., et al.: Pandaguard: Systematic evaluation of LLM safety in the era of jailbreaking attacks. arXiv preprint (2025)
6. Meta AI: Red teaming language models to reduce harms. https://ai.meta.com/blog/red-teaming-language-models-to-reduce-harms/ (2022), accessed 2025
7. OpenAI: Gpt-4 system card. Tech. rep., OpenAI (2023)
8. OpenAI: Openai usage policies. https://openai.com/policies/usage-policies (2023)

9. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., et al.: Red teaming language models with language models. arXiv preprint arXiv:2202.03286 (2022)
10. Shen, X., Chen, Z., Backes, M., Shen, Y., Zhang, Y.: 'do anything now': Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS). ACM (2024)
11. TrustAIRLab: Forbidden question set. https://huggingface.co/datasets/TrustAIRLab/forbidden_question_set (2023)
12. TrustAIRLab: In-the-wild jailbreak prompts. https://huggingface.co/datasets/TrustAIRLab/in-the-wild-jailbreak-prompts (2023)
13. Vellosillo, P.: Evaluation of jailbreak techniques in large language models. https://github.com/vell0/Evaluation-of-Jailbreak-Techniques-in-Large-Language-Models (2025), gitHub repository
14. Wallace, E., Feng, S., Kandpal, N., Gardner, M., Singh, S.: Universal adversarial triggers for attacking and analyzing NLP. arXiv preprint arXiv:1908.07125 (2019)
15. Wang, Z., Hu, W., Pang, R., Zhou, C., Tu, K., Yang, L., et al.: Jailjudge: A comprehensive jailbreak judge benchmark with multi-agent enhanced explanation evaluation framework. arXiv preprint arXiv:2311.11177 (2024)
16. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al.: Jailbroken: How does LLM behavior change when conditioned on a persona with harmful values? arXiv preprint arXiv:2301.12867 (2023)
17. Yang, X., Tang, X., Han, J., Hu, S.: The dark side of trust: Authority citation-driven jailbreak attacks on large language models. arXiv preprint arXiv:2411.11407 (2024)
18. Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Zhang, Z., Zhuang, M., et al.: Judging LLM-as-a-judge with MT-Bench and chatbot arena. arXiv preprint arXiv:2306.05685 (2023)
19. Zou, A., Wang, Z., Tan, J., Tang, C., Zhao, D., Yang, C., et al.: Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15032 (2023)