Novelty Assessment Report

Paper: Jailbreak Transferability Emerges from Shared Representations
PDF URL: https://openreview.net/pdf?id=UQK3tUsouK
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-04

Abstract

Jailbreak transferability is the surprising phenomenon when an adversarial attack compromising one model also elicits harmful responses from other models. Despite widespread demonstrations, there is little consensus on why transfer is possible: is it a quirk of safety training, an artifact of model families, or a more fundamental property of representation learning? We present evidence that transferability emerges from shared representations rather than incidental flaws. Across 20 open-weight models and 33 jailbreak attacks, we find two factors that systematically shape transfer: (1) representational similarity under benign prompts, and (2) the strength of the jailbreak on the source model. To move beyond correlation, we show that deliberately increasing similarity through benign-only distillation causally increases transfer. Qualitative analysis reveal systematic patterns; for example, persona-style jailbreaks transfer far more often than cipher-based prompts, consistent with the idea that natural-language attacks exploit models' shared representation space, whereas cipher-based attacks rely on idiosyncratic quirks that do not generalize. Together, these results reframe jailbreak transfer as a consequence of representation alignment rather than a fragile byproduct of safety training.

Core Task Landscape

This paper addresses: **Jailbreak Transferability Across Language Models**
A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Adversarial Suffix and Token-Level Optimization Attacks**
- **Semantic and Prompt-Level Jailbreak Techniques**
- **Multimodal Jailbreak Attacks on Vision-Language Models**
- **Jailbreak Transferability Mechanisms and Analysis**
- **Defense Mechanisms and Robustness Enhancement**
- **Specialized Attack Contexts and Applications**
- **Hybrid and Advanced Attack Methodologies**
- **Contextual and Configurational Factors in Jailbreaking**
- **General Adversarial Robustness and Security Foundations**

Complete Taxonomy Tree

- Jailbreak Transferability Across Language Models Survey Taxonomy
- Adversarial Suffix and Token-Level Optimization Attacks
  - Gradient-Based Suffix Generation Methods (3 papers)
  - [1] Universal and transferable adversarial attacks on aligned language models (Zou, 2023) View paper
  - [14] Exploiting the index gradients for optimization-based jailbreaking on large language models (Li Jiahui, 2025) View paper
  - [28] Universal and transferable adversarial attack on large language models using exponentiated gradient descent (Biswas, 2025) View paper
  - Transferability Enhancement for Suffix Attacks (2 papers)
  - [9] Boosting Jailbreak Transferability for Large Language Models (Liu Han-qing, 2024) View paper
  - [40] Toward Understanding the Transferability of Adversarial Suffixes in Large Language Models (Ball Sarah, 2025) View paper
- Semantic and Prompt-Level Jailbreak Techniques
  - Automated Semantic Jailbreak Generation (3 papers)
  - [2] Autodan: Generating stealthy jailbreak prompts on aligned large language models (Liu Xiaogeng, 2023) View paper
  - [5] Jailbreaking black box large language models in twenty queries (Patrick Chao, 2025) View paper
  - [8] Cold-attack: Jailbreaking llms with stealthiness and controllability (Guo, 2024) View paper
  - Prompt Decomposition and Obfuscation Strategies (3 papers)
  - [13] A Wolf in Sheepâ□s Clothing: Generalized Nested Jailbreak Prompts can Fool Large Language Models Easily (Cao, 2023) View paper
  - [22] Drattack: Prompt decomposition and reconstruction makes powerful llm jailbreakers (Li Xirui, 2024) View paper
  - [48] Human-Interpretable Adversarial Prompt Attack on Large Language Models with Situational Context (Das Nilanjana, 2024) View paper
  - Stylistic and Linguistic Manipulation Attacks (2 papers)
  - [15] Adversarial Poetry as a Universal Single-Turn Jailbreak Mechanism in Large Language Models (Piercosma Bisconti, 2025) View paper
  - [20] A Cross-Language Investigation into Jailbreak Attacks in Large Language Models (Li Jie, 2024) View paper
  - Query-Based Black-Box Jailbreak Methods (2 papers)

- [11] Prompt optimization via adversarial in-context learning (Brown, 2024) View paper
- [23] Best-of-n jailbreaking (Hughes, 2024) View paper
- Reasoning and Intent Manipulation Attacks (2 papers)
- [24] LLMs can be Dangerous Reasoners: Analyzing-based Jailbreak Attack on Large Language Models (Lin Shi, 2024) View paper
- [26] Exploring Jailbreak Attacks on LLMs through Intent Concealment and Diversion (Liu Peipei, 2025) View paper

- Multimodal Jailbreak Attacks on Vision-Language Models
  - Image-Based Jailbreak Prompt Generation (3 papers)
  - [3] Jailbreaking Attack against Multimodal Large Language Model (Niu, 2024) View paper
  - [4] Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks (Weidi Luo, 2024) View paper
  - [16] On evaluating adversarial robustness of large vision-language models (Zhao, 2023) View paper
  - Cross-Modal Obfuscation and Perturbation Techniques (2 papers)
  - [42] Cross-Modal Obfuscation for Jailbreak Attacks on Large Vision-Language Models (Jiang Lei, 2025) View paper
  - [46] Visual adversarial attack on vision-language models for autonomous driving (Zhang TianYuan, 2024) View paper
  - Transferability of Multimodal Jailbreaks (3 papers)
  - [6] When Do Universal Image Jailbreaks Transfer Between Vision-Language Models? (R Schaeffer, 2024) View paper
  - [12] Transfer Attack for Bad and Good: Explain and Boost Adversarial Transferability across Multimodal Large Language Models (Hao Cheng, 2024) View paper
  - [34] Simulated Ensemble Attack: Transferring Jailbreaks Across Fine-tuned Vision-Language Models (WANG Ruofan, 2025) View paper
  - Vision-Language Alignment Exploitation (2 papers)
  - [41] Improving Adversarial Transferability in MLLMs via Dynamic Vision-Language Alignment Attack (Gu, 2025) View paper
  - [50] Align Is Not Enough: Multimodal Universal Jailbreak Attack Against Multimodal Large Language Models (Wang, 2025) View paper
  - Memory-Efficient Gradient-Based Multimodal Attacks (1 papers)
  - [29] Zer0-Jack: A Memory-efficient Gradient-based Jailbreaking Method for Black-box Multi-modal Large Language Models (Chen, 2024) View paper
- Jailbreak Transferability Mechanisms and Analysis
  - Representation-Based Transferability Analysis ★ (2 papers)
  - [0] Jailbreak Transferability Emerges from Shared Representations (Anon et al., 2026) View paper
  - [7] Understanding and enhancing the transferability of jailbreaking attacks (LIN Runqi, 2025) View paper
  - Cross-Language and Multilingual Transferability (1 papers)
  - [49] Do Methods to Jailbreak and Defend LLMs Generalize Across Languages? (Atil, 2025) View paper
- Defense Mechanisms and Robustness Enhancement
  - Adversarial Training and Alignment Strategies (4 papers)
  - [10] Lifelong Safety Alignment for Language Models (Wang Haoyu, 2025) View paper
  - [27] Revisiting the Robust Generalization of Adversarial Prompt Tuning (Yang Fan, 2024) View paper
  - [39] "Short-length" Adversarial Training Helps LLMs Defend "Long-length" Jailbreak Attacks: Theoretical and Empirical Evidence (Fu ShaoPeng, 2025) View paper
  - [47] Adversarial tuning: Defending against jailbreak attacks for llms (Liu Fan, 2024) View paper
  - Detection and Mitigation Frameworks (2 papers)
  - [31] Defending Jailbreak Prompts via In-Context Adversarial Game (Bao Hong-yan, 2024) View paper
  - [37] LLM Adversarial Prompt Attack Detection and Mitigation Engine: A Novel Framework for Securing Generative AI Systems (Fathima, 2025) View paper
  - Robustness Evaluation and Benchmarking (2 papers)
  - [17] Are all prompt components value-neutral? understanding the heterogeneous adversarial robustness of dissected prompt in large language models (Zheng, 2025) View paper
  - [45] Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models (LIU Yugeng, 2023) View paper
- Specialized Attack Contexts and Applications
  - Multi-Agent System Attacks (2 papers)
  - [18] Infecting llm agents via generalizable adversarial attack (W Yu, 2025) View paper
  - [30] Agents under siege: Breaking pragmatic multi-agent llm systems with optimized prompt attacks (Chen Tianlong, 2025) View paper
  - Retrieval-Augmented Generation Poisoning Attacks (2 papers)
  - [36] CPA-RAG:Covert Poisoning Attacks on Retrieval-Augmented Generation in Large Language Models (Li ChunYang, 2025) View paper
  - [38] Fine-Grained Privacy Extraction from Retrieval-Augmented Generation Systems via Knowledge Asymmetry Exploitation (Chen Yu-fei, 2025) View paper
  - Denial-of-Service and Resource Exhaustion Attacks (1 papers)
  - [43] ThinkTrap: Denial-of-Service Attacks against Black-box LLM Services via Infinite Thinking (Yunzhe Li, 2025) View paper
  - Adversarial Attacks on Neural Ranking Models (1 papers)
  - [44] Adversarial Attacks against Neural Ranking Models via In-Context Learning (Arabzadeh, 2025) View paper
- Hybrid and Advanced Attack Methodologies (2 papers)
  - [19] One model transfer to all: On robust jailbreak prompts generation against LLMs (Li Linbao, 2025) View paper
  - [25] Advancing Jailbreak Strategies: A Hybrid Approach to Exploiting LLM Vulnerabilities and Bypassing Modern Defenses (Ahmed Mohamed, 2025) View paper
- Contextual and Configurational Factors in Jailbreaking (1 papers)
  - [21] Is the System Message Really Important to Jailbreaks in Large Language Models? (Zou Xiaotian, 2024) View paper
- General Adversarial Robustness and Security Foundations (3 papers)
  - [32] The Uncanny Valley: Exploring Adversarial Robustness from a Flatness Perspective (Walter, 2024) View paper
  - [33] AdvTG: An Adversarial Traffic Generation Framework to Deceive DL-Based Malicious Traffic Detection Models (Peishuai Sun, 2025) View paper
  - [35] Safety at scale: A comprehensive survey of large model safety (Yu-Gang Jiang, 2025) View paper

Narrative

Core task: jailbreak transferability across language models. The field examines how adversarial prompts that successfully bypass safety mechanisms in one model can be reused or adapted to compromise others. The taxonomy reflects a rich landscape organized around attack methodologies, transferability mechanisms, defenses, and specialized contexts. Major branches include token-level optimization techniques (e.g., Universal Transferable Adversarial Attacks[1], Autodan Stealthy Jailbreak[2]) that craft adversarial suffixes, semantic and prompt-level methods that manipulate meaning rather than tokens, and multimodal attacks targeting vision-language models (Multimodal Jailbreaking Attack[3], Universal Image Jailbreaks Transfer[6]). A dedicated branch on transferability mechanisms explores why attacks generalize, while defense-focused work seeks robustness enhancements. Hybrid methodologies and contextual factors (cross-language settings, system messages, agent environments) round out the taxonomy, illustrating that jailbreak research spans diverse threat models and application domains.

Recent work highlights tensions between attack stealthiness, transferability, and computational cost. Some studies pursue query-efficient strategies (Jailbreaking Twenty Queries[5]) or ensemble-based transfer (Simulated Ensemble Attack[34]), while others investigate representation-level explanations for why certain prompts generalize. Jailbreak Transferability Shared Representations[0] sits squarely within the transferability mechanisms branch, focusing on representation-based analysis to understand cross-model generalization. Its emphasis on shared internal structures contrasts with neighboring efforts like Enhancing Jailbreak Transferability[7], which may prioritize algorithmic improvements to boost transfer rates, or Boosting Jailbreak Transferability[9], which explores optimization refinements. By probing the representational underpinnings of transferability, this work complements attack-centric studies and informs both offensive research and the design of defenses that account for common vulnerabilities across model families.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Understanding and enhancing the transferability of jailbreaking attacks

Authors: LIN Runqi, Han Bo, Runqi Lin, Li, Fengwang, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Jailbreaking attacks can effectively manipulate open-source large language models (LLMs) to produce harmful responses. However, these attacks exhibit limited transferability, failing to disrupt proprietary LLMs consistently. To reliably identify vulnerabilities in proprietary LLMs, this work investigates the transferability of jailbreaking attacks by analysing their impact on the model's intent perception. By incorporating adversarial sequences, these attacks can redirect the source LLM's focus ...

Relationship Analysis

Both papers belong to the Representation-Based Transferability Analysis category, investigating how shared representations influence jailbreak transfer across language models. The original paper focuses on measuring representational similarity using mutual k-nearest neighbors and demonstrates through benign-only distillation that increasing similarity causally increases transfer, while the candidate paper analyzes how jailbreaks manipulate intent perception through adversarial sequences and proposes the PiF method to enhance transferability by uniformly dispersing model focus across neutral-intent tokens rather than creating high-importance regions. The key difference is that the original paper establishes representation similarity as a causal factor through distillation experiments, whereas the candidate paper focuses on understanding and improving transferability by analyzing intent recognition mechanisms and mitigating distributional dependency.

Contributions Analysis

Overall novelty summary. The paper investigates why jailbreak attacks transfer across language models, proposing that shared representations rather than incidental flaws drive transferability. It resides in the 'Representation-Based Transferability Analysis' leaf, which contains only two papers total within the broader 'Jailbreak Transferability Mechanisms and Analysis' branch. This is a relatively sparse research direction compared to crowded attack-generation categories like 'Gradient-Based Suffix Generation Methods' or 'Automated Semantic Jailbreak Generation,' suggesting the paper addresses a less-explored theoretical question about transferability mechanisms rather than developing new attack techniques.

The taxonomy reveals that most jailbreak research focuses on attack methodologies—token-level optimization, semantic manipulation, multimodal techniques—rather than mechanistic explanations. The paper's branch sits adjacent to 'Cross-Language and Multilingual Transferability,' which examines transfer across linguistic boundaries, and is conceptually distinct from attack-focused branches like 'Adversarial Suffix and Token-Level Optimization Attacks' or 'Semantic and Prompt-Level Jailbreak Techniques.' The scope note clarifies this leaf excludes attack optimization methods and defense strategies, positioning the work as foundational analysis rather than applied technique development. Its emphasis on representation similarity and causal manipulation through distillation differentiates it from neighboring optimization-centric studies.

Among 22 candidates examined, none clearly refute the three main contributions. The large-scale empirical analysis (5 candidates examined, 0 refutable) and systematic attack-type characterization (7 candidates, 0 refutable) appear relatively novel within this limited search scope. The benign-only distillation protocol (10 candidates, 0 refutable) shows no substantial prior work among examined papers. These statistics suggest the contributions are distinct within the top-K semantic neighborhood, though the search scale is modest and does not guarantee exhaustive coverage of all relevant prior work in representation-based transferability analysis or causal intervention methods.

Based on 22 examined candidates from semantic search, the work appears to occupy a relatively underexplored niche connecting representation learning theory to jailbreak transferability. The sparse taxonomy leaf and absence of refutable prior work within the search scope suggest novelty, though this assessment is constrained by the limited candidate pool. A broader literature review might uncover related work in adversarial robustness or representation alignment that was not captured by the semantic search strategy.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Large-scale empirical analysis of jailbreak transferability factors

Description: The authors conduct a comprehensive empirical study across 20 open-weight models and 33 jailbreak attacks applied to 313 harmful prompts, identifying two systematic factors that predict jailbreak transferability: the strength of the jailbreak on the source model and the representational similarity between models measured under benign prompts.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. SafeInt: Shielding Large Language Models from Jailbreak Attacks via Safety-Aware Representation Intervention

URL: View paper

Brief Assessment

SafeInt Safety-Aware Intervention[72] focuses on defending against jailbreak attacks through representation intervention rather than analyzing transferability factors. The paper does not investigate what makes jailbreaks transfer across models or measure representational similarity as a predictor of transfer success.

---

2. Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis

URL: View paper

Brief Assessment

Representation Space Analysis[71] focuses on understanding jailbreak attacks through representation space analysis rather than conducting a large-scale empirical study of transferability factors across models. The candidate paper's emphasis is on analyzing the representation space mechanisms underlying jailbreak attacks, not on systematically identifying factors that predict transferability across 20 models and 33 attacks as claimed in the original contribution.

---

3. AdversariaL attacK sAfety aLIgnment (ALKALI): Safeguarding LLMs through GRACE: Geometric Representation-Aware Contrastive Enhancement-Introducing â¦

URL: View paper

Brief Assessment

ALKALI GRACE[69] focuses on adversarial defense mechanisms through geometric representation alignment, not on analyzing transferability factors across models. The candidate introduces GRACE as a mitigation framework and AVQI as a vulnerability metric, rather than studying what makes jailbreaks transfer between models.

---

4. One Leak Away: How Pretrained Model Exposure Amplifies Jailbreak Risks in Finetuned LLMs

URL: View paper

Brief Assessment

Pretrained Model Exposure[73] focuses on pretrain-to-finetune transfer within model families using white-box pretrained access, while the original paper examines cross-model transfer across 20 diverse open-weight models using post-hoc attacks and representational similarity under benign prompts. The candidate does not challenge the novelty of identifying representational similarity and jailbreak strength as systematic predictors across heterogeneous model families.

---

5. CAVGAN: Unifying Jailbreak and Defense of LLMs via Generative Adversarial Attacks on their Internal Representations

URL: View paper

Brief Assessment

CAVGAN Generative Adversarial[70] focuses on a unified attack-defense framework using GANs on internal representations, not on empirical analysis of transferability factors across models or representational similarity measurements.

Contribution 2: Benign-only distillation protocol for causal manipulation of transferability

Description: The authors develop a distillation method that fine-tunes a student model exclusively on benign prompt-response pairs from a teacher model, deliberately increasing their representational similarity. This intervention causally increases jailbreak transferability from teacher to student, providing evidence that shared representations drive transfer rather than artifacts of safety training.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Distilling Adversarial Robustness Using Heterogeneous Teachers

URL: View paper

Brief Assessment

Distilling Heterogeneous Teachers[60] focuses on adversarial robustness distillation using multiple heterogeneous teachers (CNNs and Vision Transformers) to improve model robustness against adversarial attacks. The original paper's benign-only distillation protocol specifically manipulates jailbreak transferability in LLMs through representational similarity, which is a fundamentally different domain and objective than adversarial robustness in image classifiers.

---

2. Similarity of neural network models: A survey of functional and representational measures

URL: View paper

Brief Assessment

Similarity Neural Network Models[53] is a survey paper about measuring similarity between neural networks. It does not present a distillation protocol for manipulating jailbreak transferability or study adversarial robustness in language models.

---

3. Common knowledge learning for generating transferable adversarial examples

URL: View paper

Brief Assessment

Common Knowledge Learning[51] focuses on distilling knowledge from multiple teacher models to improve adversarial example transferability across different DNN architectures. The original paper's benign-only distillation protocol specifically manipulates jailbreak transferability in language models through representational similarity, which is a distinct application domain and mechanism from adversarial image classification.

---

4. Distillation as a defense to adversarial perturbations against deep neural networks

URL: View paper

Brief Assessment

Distillation Defense Adversarial[59] focuses on using distillation to defend against adversarial perturbations in DNNs by training on the same architecture with soft labels. The original paper uses distillation to causally increase jailbreak transferability between LLMs by fine-tuning on benign prompt-response pairs, which is a fundamentally different application and mechanism.

---

5. Guided adversarial contrastive distillation for robust students

URL: View paper

Brief Assessment

Guided Adversarial Contrastive Distillation[58] focuses on transferring adversarial robustness in image classification through adversarial training and contrastive learning, not on benign-only distillation for manipulating jailbreak transferability in language models.

---

6. Continuous transfer of neural network representational similarity for incremental learning

URL: View paper

Brief Assessment

Continuous Transfer Representational Similarity[52] focuses on incremental learning through knowledge distillation to maintain representational similarity across learning stages, not on adversarial transferability or jailbreak attacks in language models.

---

7. Data-free knowledge distillation via text-noise fusion and dynamic adversarial temperature.

URL: View paper

Brief Assessment

Text-Noise Fusion[55] focuses on data-free knowledge distillation using text-noise fusion for model compression, not on benign-only distillation to causally manipulate jailbreak transferability in language models.

---

8. Distillation-Based Cross-Model Transferable Adversarial Attack for Remote Sensing Image Classification

URL: View paper

Brief Assessment

Distillation-Based Cross-Model Transferable[56] uses distillation to train surrogate models for adversarial attacks on remote sensing images, not to study jailbreak transferability in language models through benign-only distillation as a causal intervention.

---

9. Improving the transferability of adversarial examples with diverse gradients

URL: View paper

Brief Assessment

Diverse Gradients Transferability[57] focuses on adversarial example transferability in computer vision using knowledge distillation to create diverse gradients from a single source model. The original paper studies jailbreak transferability in LLMs using benign-only distillation to increase representational similarity. These are fundamentally different domains (CV vs. NLP) and mechanisms (gradient diversity vs. representation alignment).

---

10. Similarity of neural architectures using adversarial attack transferability

URL: View paper

Brief Assessment

Similarity Adversarial Attack Transferability[54] focuses on measuring architectural similarity through adversarial attack transferability between models, not on causal manipulation of jailbreak transferability through benign-only distillation protocols.

Contribution 3: Systematic characterization of attack-type differences in transferability

Description: The authors show that persona-style jailbreaks, which use natural language and align with shared semantic representations, transfer far more reliably across models than cipher-based jailbreaks, which exploit idiosyncratic model-specific quirks. This finding supports the hypothesis that transferability emerges from shared representational geometry.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. LLM-Virus: Evolutionary Jailbreak Attack on Large Language Models

URL: View paper

Brief Assessment

LLM-Virus Evolutionary Jailbreak[61] focuses on evolutionary algorithms for jailbreak attacks and does not systematically characterize differences between persona-style versus cipher-based attacks in terms of transferability across models.

---

2. A Survey of Recent Advances in Adversarial Attack and Defense on Vision-Language Models

URL: View paper

Brief Assessment

Adversarial Attack Defense Survey[62] is a survey paper on vision-language models, not language-only models. The candidate focuses on adversarial attacks in multimodal contexts, while the original paper specifically examines persona-style versus cipher-based jailbreak transferability patterns in text-only language models.

---

3. A Review of â□□Do Anything Nowâ□□ Jailbreak Attacks in Large Language Models: Potential Risks, Impacts, and Defense Strategies

URL: View paper

Brief Assessment

Do Anything Now Review[64] is a review paper that surveys jailbreak phenomena and defense strategies broadly, but does not present empirical findings on persona-style versus cipher-based attack transferability patterns or their relationship to shared representational geometry.

---

4. Simulated Ensemble Attack: Transferring Jailbreaks Across Fine-tuned Vision-Language Models

URL: View paper

Brief Assessment

Simulated Ensemble Attack[34] focuses on transferable jailbreak attacks across fine-tuned vision-language models using image-based adversarial perturbations, not on comparing persona-style versus cipher-based textual jailbreak transferability across language models.

---

5. Boosting Jailbreak Transferability for Large Language Models

URL: View paper

Brief Assessment

Boosting Jailbreak Transferability[9] focuses on optimization-based jailbreak methods (GCG variants) and their transferability across models, but does not systematically characterize differences between persona-style versus cipher-based attacks as the original paper does.

---

6. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks

URL: View paper

Brief Assessment

Jailbreakv Benchmark[4] focuses on evaluating jailbreak attack transferability from LLMs to multimodal LLMs (MLLMs), not on comparing persona-style versus cipher-based attack transferability patterns across language models as the original paper does.

---

7. Autodan: Generating stealthy jailbreak prompts on aligned large language models

URL: View paper

Brief Assessment

Autodan Stealthy Jailbreak[2] focuses on generating semantically meaningful jailbreak prompts using hierarchical genetic algorithms, not on systematically characterizing transferability differences between persona-style and cipher-based attacks across models.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Jailbreak Transferability Emerges from Shared Representations View paper
- [1] Universal and transferable adversarial attacks on aligned language models View paper
- [2] Autodan: Generating stealthy jailbreak prompts on aligned large language models View paper
- [3] Jailbreaking Attack against Multimodal Large Language Model View paper
- [4] Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks View paper
- [5] Jailbreaking black box large language models in twenty queries View paper
- [6] When Do Universal Image Jailbreaks Transfer Between Vision-Language Models? View paper
- [7] Understanding and enhancing the transferability of jailbreaking attacks View paper
- [8] Cold-attack: Jailbreaking llms with stealthiness and controllability View paper
- [9] Boosting Jailbreak Transferability for Large Language Models View paper
- [10] Lifelong Safety Alignment for Language Models View paper
- [11] Prompt optimization via adversarial in-context learning View paper
- [12] Transfer Attack for Bad and Good: Explain and Boost Adversarial Transferability across Multimodal Large Language Models View paper
- [13] A Wolf in Sheepâ□s Clothing: Generalized Nested Jailbreak Prompts can Fool Large Language Models Easily View paper
- [14] Exploiting the index gradients for optimization-based jailbreaking on large language models View paper
- [15] Adversarial Poetry as a Universal Single-Turn Jailbreak Mechanism in Large Language Models View paper
- [16] On evaluating adversarial robustness of large vision-language models View paper
- [17] Are all prompt components value-neutral? understanding the heterogeneous adversarial robustness of dissected prompt in large language models View paper
- [18] Infecting llm agents via generalizable adversarial attack View paper
- [19] One model transfer to all: On robust jailbreak prompts generation against LLMs View paper
- [20] A Cross-Language Investigation into Jailbreak Attacks in Large Language Models View paper
- [21] Is the System Message Really Important to Jailbreaks in Large Language Models? View paper
- [22] Drattack: Prompt decomposition and reconstruction makes powerful llm jailbreakers View paper
- [23] Best-of-n jailbreaking View paper
- [24] LLMs can be Dangerous Reasoners: Analyzing-based Jailbreak Attack on Large Language Models View paper
- [25] Advancing Jailbreak Strategies: A Hybrid Approach to Exploiting LLM Vulnerabilities and Bypassing Modern Defenses View paper
- [26] Exploring Jailbreak Attacks on LLMs through Intent Concealment and Diversion View paper
- [27] Revisiting the Robust Generalization of Adversarial Prompt Tuning View paper
- [28] Universal and transferable adversarial attack on large language models using exponentiated gradient descent View paper
- [29] Zer0-Jack: A Memory-efficient Gradient-based Jailbreaking Method for Black-box Multi-modal Large Language Models View paper
- [30] Agents under siege: Breaking pragmatic multi-agent llm systems with optimized prompt attacks View paper
- [31] Defending Jailbreak Prompts via In-Context Adversarial Game View paper
- [32] The Uncanny Valley: Exploring Adversarial Robustness from a Flatness Perspective View paper
- [33] AdvTG: An Adversarial Traffic Generation Framework to Deceive DL-Based Malicious Traffic Detection Models View paper
- [34] Simulated Ensemble Attack: Transferring Jailbreaks Across Fine-tuned Vision-Language Models View paper
- [35] Safety at scale: A comprehensive survey of large model safety View paper
- [36] CPA-RAG:Covert Poisoning Attacks on Retrieval-Augmented Generation in Large Language Models View paper
- [37] LLM Adversarial Prompt Attack Detection and Mitigation Engine: A Novel Framework for Securing Generative AI Systems View paper
- [38] Fine-Grained Privacy Extraction from Retrieval-Augmented Generation Systems via Knowledge Asymmetry Exploitation View paper
- [39] "Short-length" Adversarial Training Helps LLMs Defend "Long-length" Jailbreak Attacks: Theoretical and Empirical Evidence View paper
- [40] Toward Understanding the Transferability of Adversarial Suffixes in Large Language Models View paper
- [41] Improving Adversarial Transferability in MLLMs via Dynamic Vision-Language Alignment Attack View paper
- [42] Cross-Modal Obfuscation for Jailbreak Attacks on Large Vision-Language Models View paper
- [43] ThinkTrap: Denial-of-Service Attacks against Black-box LLM Services via Infinite Thinking View paper
- [44] Adversarial Attacks against Neural Ranking Models via In-Context Learning View paper
- [45] Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models View paper
- [46] Visual adversarial attack on vision-language models for autonomous driving View paper
- [47] Adversarial tuning: Defending against jailbreak attacks for llms View paper
- [48] Human-Interpretable Adversarial Prompt Attack on Large Language Models with Situational Context View paper
- [49] Do Methods to Jailbreak and Defend LLMs Generalize Across Languages? View paper
- [50] Align Is Not Enough: Multimodal Universal Jailbreak Attack Against Multimodal Large Language Models View paper
- [51] Common knowledge learning for generating transferable adversarial examples View paper
- [52] Continuous transfer of neural network representational similarity for incremental learning View paper
- [53] Similarity of neural network models: A survey of functional and representational measures View paper

- [54] Similarity of neural architectures using adversarial attack transferability View paper
- [55] Data-free knowledge distillation via text-noise fusion and dynamic adversarial temperature. View paper
- [56] Distillation-Based Cross-Model Transferable Adversarial Attack for Remote Sensing Image Classification View paper
- [57] Improving the transferability of adversarial examples with diverse gradients View paper
- [58] Guided adversarial contrastive distillation for robust students View paper
- [59] Distillation as a defense to adversarial perturbations against deep neural networks View paper
- [60] Distilling Adversarial Robustness Using Heterogeneous Teachers View paper
- [61] LLM-Virus: Evolutionary Jailbreak Attack on Large Language Models View paper
- [62] A Survey of Recent Advances in Adversarial Attack and Defense on Vision-Language Models View paper
- [63] Jailbreak Strength and Model Similarity Predict Transferability View paper
- [64] A Review of â□□Do Anything Nowâ□□ Jailbreak Attacks in Large Language Models: Potential Risks, Impacts, and Defense Strategies View paper
- [65] Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation View paper
- [66] Improved Representation Steering for Language Models View paper
- [67] Causes and Consequences of Representational Similarity in Machine Learning Models View paper
- [68] Understanding Adversarial Transfer: Why Representation-Space Attacks Fail Where Data-Space Attacks Succeed View paper
- [69] AdversariaL attacK sAfety aLIgnment (ALKALI): Safeguarding LLMs through GRACE: Geometric Representation-Aware Contrastive Enhancement-Introducing â□¦ View paper
- [70] CAVGAN: Unifying Jailbreak and Defense of LLMs via Generative Adversarial Attacks on their Internal Representations View paper
- [71] Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis View paper
- [72] SafeInt: Shielding Large Language Models from Jailbreak Attacks via Safety-Aware Representation Intervention View paper
- [73] One Leak Away: How Pretrained Model Exposure Amplifies Jailbreak Risks in Finetuned LLMs View paper