Novelty Assessment Report

Paper: Automatic Instance Selection with Genetic Updating for Few-shot LLM Jailbreak

PDF URL: https://openreview.net/pdf?id=B7oQWswV7y

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

This paper studies the problem of few-shot large language model (LLM) jailbreak, which aims to trigger unsafe outputs of LLMs using only a handful of adversarial examples. However, the effectiveness of the current few-shot jailbreak attacks is limited by the challenge of systematically selecting the most potent instances, with existing methods often resorting to inefficient manual or random selection. In this paper, we propose a novel approach named Automatic Instance Selection with Genetic Updating (ACCEPT) for few-shot LLM jailbreak. The core of our ACCEPT is to utilize textual gradient and fitness scores to guide the optimization process automatically. In particular, our ACCEPT designs a loss objective prioritizing successful jailbreaks, which can further guide the selection of instances via textual gradient. Furthermore, we construct a pool with meaningless marks, and consider the injection operators as chromosomes following the genetic algorithm. A fitness function is then defined in jailbreak scenarios, which helps the iterations across generations for proper prompts. Extensive experiments across several benchmark datasets can validate the effectiveness of the proposed ACCEPT in comparison with extensive baselines.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Few-Shot Large Language Model Jailbreak Attack**

A total of **29 papers** were analyzed and organized into a taxonomy with **11 categories**.

Taxonomy Overview

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

- **In-Context Learning Based Jailbreak Attacks**
- **Prompt Optimization and Manipulation Attacks**
- **Multimodal Jailbreak Attacks**
- **Model Adaptation and Fine-Tuning Attacks**
- **Cross-Lingual and Domain-Specific Attacks**
- **Reinforcement Learning Enhanced Attacks**
- **Self-Generated and Recursive Attacks**
- **Defense and Detection Mechanisms**

Complete Taxonomy Tree

- Few-Shot Large Language Model Jailbreak Attack Survey Taxonomy
- In-Context Learning Based Jailbreak Attacks
    - Few-Shot Demonstration Injection ★ (4 papers)
    - [0] Automatic Instance Selection with Genetic Updating for Few-shot LLM Jailbreak (Anon et al., 2026) View paper
    - [1] Jailbreak and guard aligned language models with only few in-context demonstrations (Wei, 2023) View paper
    - [3] Improved few-shot jailbreaking can circumvent aligned language models and their defenses (Zheng, 2024) View paper
    - [8] Self-Instruct Few-Shot Jailbreaking: Decompose the Attack into Pattern and Behavior Learning (Hua, 2025) View paper
    - Contextual Priming and Response Manipulation (2 papers)
    - [6] Hijacking Large Language Models via Adversarial In-Context Learning (Zhou Xiangyu, 2023) View paper
    - [19] Response Attack: Exploiting Contextual Priming to Jailbreak Large Language Models (Li Lijun, 2025) View paper
- Prompt Optimization and Manipulation Attacks
    - Gradient-Free Suffix Optimization (2 papers)
    - [9] Logit-Gap Steering: Efficient Short-Suffix Jailbreaks for Aligned Large Language Models (Liu Hongliang, 2025) View paper
    - [23] Graph of Attacks with Pruning: Optimizing Stealthy Jailbreak Prompt Generation for Enhanced LLM Content Moderation (Schwartz Daniel, 2025) View paper
    - Semantic Obfuscation and Disguise (4 papers)
    - [4] Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction (Liu Tong, 2024) View paper
    - [7] Generation, Detection, and Evaluation of Role-play based Jailbreak attacks in Large Language Models (Johnson, 2024) View paper
    - [26] Watch Your Words: Successfully Jailbreak LLM by Mitigating the â□□Prompt Maliceâ□□ (Xu Xiaowei, 2024) View paper
    - [28] Enhancing Jailbreak Attacks on LLMs via Persona Prompts (Zhang Zheng, 2025) View paper
    - Structural and Syntactic Manipulation (4 papers)
    - [11] Simple permutations can fool LLaMA: Permutation attack and defense for large language models (C Liang, 2024) View paper
    - [16] Automatic and Universal Prompt Injection Attacks against Large Language Models (Liu Xiaogeng, 2024) View paper
    - [17] PROMPTFUZZ: Harnessing Fuzzing Techniques for Robust Testing of Prompt Injection in LLMs (Yu, 2024) View paper
    - [29] Camouflage Patching: Effective Jailbreak Attacks on Single-and Multimodal LLMs (P Hu, n.d.) View paper

- Multimodal Jailbreak Attacks (3 papers)
  - [5] Visual Adversarial Examples Jailbreak Aligned Large Language Models (Xiangyu Qi, 2023) View paper
  - [12] Can Large Language Models Automatically Jailbreak GPT-4V? (Wu Yuanwei, 2024) View paper
  - [21] Jailbreak Vision Language Models via Bi-Modal Adversarial Prompt (Zonghao Ying, 2024) View paper
- Model Adaptation and Fine-Tuning Attacks (2 papers)
  - [13] Data poisoning in llms: Jailbreak-tuning and scaling laws (D Bowen, 2024) View paper
  - [15] Attack via Overfitting: 10-shot Benign Fine-tuning to Jailbreak LLMs (Xie, 2025) View paper
- Cross-Lingual and Domain-Specific Attacks (2 papers)
  - [10] Multilingual Jailbreak Challenges in Large Language Models (Deng Yue, 2023) View paper
  - [18] DeRAG: Black-box Adversarial Attacks on Multiple Retrieval-Augmented Generation Applications via Prompt Injection (Yu Fang, 2025) View paper
- Reinforcement Learning Enhanced Attacks (1 papers)
  - [24] Reinforcement Learning-powered Effectiveness and Efficiency Few-shot Jailbreaking Attack LLMs (Xuehai Tang, 2024) View paper
- Self-Generated and Recursive Attacks (1 papers)
  - [14] Self-HarmLLM: Can Large Language Model Harm Itself? (Heehwan Kim, 2025) View paper
- Defense and Detection Mechanisms (5 papers)
  - [2] Reducing the scope of language models (Yunis, 2024) View paper
  - [20] Summarizer unlearning a framework for balancing data retention and forgetting (Banegas, 2025) View paper
  - [22] LLM Adversarial Prompt Attack Detection and Mitigation Engine: A Novel Framework for Securing Generative AI Systems (Fathima, 2025) View paper
  - [25] Zero-Shot Detection of Jailbreaking Attempts in LLMs (MHU Rahman, 2025) View paper
  - [27] Jailbreak Distillation: Renewable Safety Benchmarking (Zhang Jingyu, 2025) View paper

Narrative

Core task: few-shot large language model jailbreak attack. The field of jailbreak attacks on large language models has evolved into a rich taxonomy spanning multiple strategic dimensions. At the highest level, researchers explore In-Context Learning Based Jailbreak Attacks that leverage demonstration injection and contextual manipulation, Prompt Optimization and Manipulation Attacks that systematically refine adversarial inputs, and Multimodal Jailbreak Attacks that exploit vision-language interfaces. Additional branches include Model Adaptation and Fine-Tuning Attacks that poison or retrain models, Cross-Lingual and Domain-Specific Attacks that exploit linguistic or specialized vulnerabilities, Reinforcement Learning Enhanced Attacks that iteratively optimize jailbreak strategies, Self-Generated and Recursive Attacks where models produce their own adversarial prompts, and Defense and Detection Mechanisms that aim to identify or mitigate these threats. Works such as Improved Few-shot Jailbreaking[3] and RL Few-shot Jailbreak[24] illustrate how different branches can converge on the core challenge of crafting effective few-shot demonstrations.

Within this landscape, a particularly active line of inquiry focuses on how carefully selected in-context examples can bypass safety guardrails. Genetic Instance Selection[0] sits squarely in the Few-Shot Demonstration Injection cluster, emphasizing evolutionary or selection-based strategies to identify potent demonstration sets. This contrasts with approaches like Improved Few-shot Jailbreaking[3], which may prioritize systematic prompt refinement, and Self-Instruct Jailbreaking[8], which explores recursive self-generation of adversarial content. Meanwhile, defense-oriented efforts such as Few-shot Jailbreak Guard[1] and Zero-Shot Jailbreak Detection[25] highlight the ongoing arms race between attack sophistication and protective measures. The original paper's focus on genetic or instance-selection mechanisms places it among methods that treat demonstration choice as an optimization problem, distinguishing it from purely prompt-engineering or multimodal strategies while sharing common ground with reinforcement and iterative refinement techniques.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Jailbreak and guard aligned language models with only few in-context demonstrations

Authors: Wei, Zeming, Zeming Wei, Wang Yi-fei, Yifei Wang, et al. (11 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Large Language Models (LLMs) have shown remarkable success in various tasks, yet their safety and the risk of generating harmful content remain pressing concerns. In this paper, we delve into the potential of In-Context Learning (ICL) to modulate the alignment of LLMs. Specifically, we propose the In-Context Attack (ICA) which employs harmful demonstrations to subvert LLMs, and the In-Context Defense (ICD) which bolsters model resilience through examples that demonstrate refusal to produce harmf...

Relationship Analysis

Both papers belong to the Few-Shot Demonstration Injection category, using small numbers of adversarial demonstrations to manipulate LLM behavior. They overlap in leveraging in-context learning with harmful question-answer pairs to jailbreak aligned LLMs. However, the original paper (ACCEPT) focuses on automatic instance selection via textual gradients and genetic algorithms to optimize both semantic instance combinations and non-semantic prompt perturbations, while the candidate paper proposes In-Context Attack (ICA) that directly uses manually or randomly selected harmful demonstrations without systematic optimization, and also introduces In-Context Defense (ICD) as a defensive counterpart.

2. Improved few-shot jailbreaking can circumvent aligned language models and their defenses

Authors: Zheng, Xiaosen, Pang, Tianyu, Xiaosen Zheng, et al. (15 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Recently, Anil et al. (2024) show that many-shot (up to hundreds of) demonstrations can jailbreak state-of-the-art LLMs by exploiting their long-context capability. Nevertheless, is it possible to use few-shot demonstrations to efficiently jailbreak LLMs within limited context sizes? While the vanilla few-shot jailbreaking may be inefficient, we propose improved techniques such as injecting special system tokens like [/INST] and employing demo-level random search from a collected demo pool. Thes...

Relationship Analysis

Both papers belong to the Few-Shot Demonstration Injection category, using small numbers of adversarial demonstrations to manipulate LLM behavior. They overlap in leveraging in-context learning with few-shot harmful examples to jailbreak aligned LLMs. However, the original paper (ACCEPT) focuses on automatic instance selection via textual gradients and genetic algorithms for mark injection, while the candidate paper (I-FSJ) emphasizes injecting special system tokens (like [/INST]) into demonstrations and employing demo-level random search from a pre-constructed pool.

3. Self-Instruct Few-Shot Jailbreaking: Decompose the Attack into Pattern and Behavior Learning

Authors: Hua, Jiaqi, Jiaqi Hua, Wanxu Wei | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Recently, several works have been conducted on jailbreaking Large Language Models (LLMs) with few-shot malicious demos. In particular, Zheng et al. focus on improving the efficiency of Few-Shot Jailbreaking (FSJ) by injecting special tokens into the demos and employing demo-level random search, known as Improved Few-Shot Jailbreaking (I-FSJ). Nevertheless, we notice that this method may still require a long context to jailbreak advanced models e.g. 32 shots of demos for Meta-Llama-3-8B-Instruct ...

Relationship Analysis

Both papers belong to the Few-Shot Demonstration Injection category, using small numbers of adversarial demonstrations to manipulate LLM behavior. They overlap in their core approach of leveraging few-shot in-context learning for jailbreak attacks and both aim to improve the efficiency and effectiveness of instance selection. The key difference is that the original paper (ACCEPT) employs textual gradients and genetic algorithms for automatic instance selection and prompt perturbation, while the candidate paper (Self-Instruct-FSJ) decomposes the attack into pattern and behavior learning with demo-level greedy search, focusing on reducing the number of required demonstrations.

Contributions Analysis

Overall novelty summary. The paper proposes ACCEPT, a framework for few-shot LLM jailbreak that combines textual gradient-based instance selection with genetic algorithm optimization for non-semantic mark injection. It resides in the 'Few-Shot Demonstration Injection' leaf of the taxonomy, which contains four papers total including this work. This leaf sits within the broader 'In-Context Learning Based Jailbreak Attacks' branch, indicating a moderately populated research direction focused on exploiting demonstration examples to bypass safety alignment. The taxonomy reveals this is one of several active attack paradigms, alongside prompt optimization, multimodal attacks, and reinforcement learning approaches.

The taxonomy structure shows ACCEPT's leaf neighbors other demonstration-based methods, while adjacent leaves explore 'Contextual Priming and Response Manipulation' (two papers) and parallel branches investigate 'Gradient-Free Suffix Optimization' and 'Semantic Obfuscation' techniques. The scope notes clarify that demonstration injection methods differ from many-shot attacks or non-demonstration approaches, positioning ACCEPT at the intersection of in-context learning exploitation and systematic instance selection. The broader taxonomy reveals approximately 29 papers across diverse attack mechanisms, suggesting the field has fragmented into specialized sub-problems rather than converging on unified methodologies.

Among the three identified contributions, the literature search examined five candidates total. The textual gradient-based instance selection mechanism was evaluated against four candidates with zero refutations found, while the integrated ACCEPT framework was compared to one candidate with no overlap detected. The genetic algorithm component received no direct comparison due to limited candidate availability. This limited search scope—examining roughly five semantically similar papers rather than an exhaustive survey—means the analysis captures immediate neighbors but may miss relevant work in adjacent taxonomy branches or recent preprints. The absence of refutations among examined candidates suggests potential novelty within the searched subset, though the small sample size precludes definitive conclusions.

Based on the constrained literature search covering five candidates, ACCEPT appears to occupy a distinct position combining gradient-based selection with genetic optimization for few-shot jailbreak. However, the analysis explicitly covers only top-K semantic matches and does not encompass the full taxonomy of 29 papers or adjacent research directions like reinforcement learning attacks or structural manipulation methods. The contribution-level statistics reflect what was examined, not the complete prior art landscape, leaving open questions about overlap with gradient-free optimization or evolutionary strategies in neighboring taxonomy branches.

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

Contribution 1: Textual gradient-based instance selection for few-shot jailbreak

Description: The authors propose using textual gradients to automatically select the most effective semantic instances for few-shot jailbreak attacks. This method treats instance selection as a differentiable optimization process in text space, using LLM-generated gradient feedback to guide iterative improvements rather than relying on manual or random selection.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Class specific autoencoders enhance sample diversity

URL: View paper

Brief Assessment

Class Specific Autoencoders[30] focuses on generating pseudo examples using autoencoders for semi-supervised and few-shot learning in image classification tasks (MNIST, FashionMNIST). It does not address textual gradient optimization, adversarial jailbreak attacks on LLMs, or instance selection for security applications.

2. Hijacking Large Language Models via Adversarial In-Context Learning

URL: View paper

Brief Assessment

Adversarial In-Context Hijacking[6] focuses on gradient-based prompt search to append adversarial suffixes to in-context demos, not on using textual gradients (LLM-generated feedback) to guide iterative instance selection as in the original paper.

3. Sparse Adversarial Attack For Video Via Gradient-Based Keyframe Selection

URL: View paper

Brief Assessment

Sparse Video Attack[32] focuses on gradient-based keyframe selection for adversarial video attacks in computer vision, not textual gradient optimization for LLM jailbreak instance selection. The domains and technical approaches are fundamentally different.

4. MAPGD: Multi-Agent Prompt Gradient Descent for Collaborative Prompt Optimization

URL: View paper

Brief Assessment

MAPGD[31] focuses on prompt optimization through multi-agent collaboration for general NLP tasks, not on few-shot jailbreak attacks or adversarial instance selection. The textual gradient mechanism in MAPGD[31] is used for collaborative prompt refinement across multiple specialized agents, fundamentally different from the original paper's use of textual gradients for selecting adversarial jailbreak instances.

Contribution 2: ACCEPT framework integrating semantic and non-semantic optimization

Description: The authors introduce ACCEPT, a hybrid framework that synergistically combines two optimization mechanisms: textual gradient-guided selection of semantic instances and genetic algorithm-driven injection of non-semantic markers (emojis, special characters) into harmful prompts. This dual-layer approach addresses both instance selection and attack evasiveness simultaneously.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. SEC-Prompt: SEmantic Complementary Prompting for Few-Shot Class-Incremental Learning

URL: View paper

Brief Assessment

SEC-Prompt[33] addresses few-shot class-incremental learning in computer vision using discriminative and non-discriminative prompts for visual recognition tasks, whereas the original paper focuses on LLM jailbreak attacks combining textual gradient-guided instance selection with genetic algorithm-driven non-semantic perturbations. These are fundamentally different problem domains with distinct technical approaches.

Contribution 3: Genetic algorithm for non-semantic mark injection optimization

Description: The authors develop a genetic algorithm-based mechanism that systematically optimizes the injection of non-semantic markers into harmful prompts. The approach encodes perturbation strategies as chromosomes with genes controlling operation type, element selection, intensity, position, and additional transformations, using evolutionary search to discover optimal evasion strategies.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

Textual similarity detection checked 8 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Improved few-shot jailbreaking can circumvent aligned language models and their defenses

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Automatic Instance Selection with Genetic Updating for Few-shot LLM Jailbreak View paper
- [1] Jailbreak and guard aligned language models with only few in-context demonstrations View paper
- [2] Reducing the scope of language models View paper
- [3] Improved few-shot jailbreaking can circumvent aligned language models and their defenses View paper
- [4] Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction View paper
- [5] Visual Adversarial Examples Jailbreak Aligned Large Language Models View paper
- [6] Hijacking Large Language Models via Adversarial In-Context Learning View paper
- [7] Generation, Detection, and Evaluation of Role-play based Jailbreak attacks in Large Language Models View paper
- [8] Self-Instruct Few-Shot Jailbreaking: Decompose the Attack into Pattern and Behavior Learning View paper
- [9] Logit-Gap Steering: Efficient Short-Suffix Jailbreaks for Aligned Large Language Models View paper
- [10] Multilingual Jailbreak Challenges in Large Language Models View paper
- [11] Simple permutations can fool LLaMA: Permutation attack and defense for large language models View paper
- [12] Can Large Language Models Automatically Jailbreak GPT-4V? View paper
- [13] Data poisoning in llms: Jailbreak-tuning and scaling laws View paper
- [14] Self-HarmLLM: Can Large Language Model Harm Itself? View paper
- [15] Attack via Overfitting: 10-shot Benign Fine-tuning to Jailbreak LLMs View paper
- [16] Automatic and Universal Prompt Injection Attacks against Large Language Models View paper
- [17] PROMPTFUZZ: Harnessing Fuzzing Techniques for Robust Testing of Prompt Injection in LLMs View paper
- [18] DeRAG: Black-box Adversarial Attacks on Multiple Retrieval-Augmented Generation Applications via Prompt Injection View paper
- [19] Response Attack: Exploiting Contextual Priming to Jailbreak Large Language Models View paper
- [20] Summarizer unlearning a framework for balancing data retention and forgetting View paper
- [21] Jailbreak Vision Language Models via Bi-Modal Adversarial Prompt View paper
- [22] LLM Adversarial Prompt Attack Detection and Mitigation Engine: A Novel Framework for Securing Generative AI Systems View paper
- [23] Graph of Attacks with Pruning: Optimizing Stealthy Jailbreak Prompt Generation for Enhanced LLM Content Moderation View paper
- [24] Reinforcement Learning-powered Effectiveness and Efficiency Few-shot Jailbreaking Attack LLMs View paper
- [25] Zero-Shot Detection of Jailbreaking Attempts in LLMs View paper
- [26] Watch Your Words: Successfully Jailbreak LLM by Mitigating the â□□Prompt Maliceâ□□ View paper
- [27] Jailbreak Distillation: Renewable Safety Benchmarking View paper
- [28] Enhancing Jailbreak Attacks on LLMs via Persona Prompts View paper
- [29] Camouflage Patching: Effective Jailbreak Attacks on Single-and Multimodal LLMs View paper
- [30] Class specific autoencoders enhance sample diversity View paper
- [31] MAPGD: Multi-Agent Prompt Gradient Descent for Collaborative Prompt Optimization View paper
- [32] Sparse Adversarial Attack For Video Via Gradient-Based Keyframe Selection View paper
- [33] SEC-Prompt: SEmantic Complementary Prompting for Few-Shot Class-Incremental Learning View paper