

Agents Under Siege: Breaking Pragmatic Multi-Agent LLM Systems with Optimized Prompt Attacks

⚠ WARNING: This paper contains text that may be considered offensive.

Rana Muhammad Shahroz Khan¹, Zhen Tan², Sukwon Yun¹,
Charles Fleming³, Tianlong Chen¹

¹University of North Carolina at Chapel Hill, ²Arizona State University, ³Cisco

Abstract

Most discussions about Large Language Model (LLM) safety have focused on single-agent settings but multi-agent LLM systems now create novel adversarial risks because their behavior depends on communication between agents and decentralized reasoning. In this work, we innovatively focus on attacking pragmatic systems that have constraints such as limited token bandwidth, latency between message delivery, and defense mechanisms. We design a *permutation-invariant adversarial attack* that optimizes prompt distribution across latency and bandwidth-constraint network topologies to bypass distributed safety mechanisms within the system. Formulating the attack path as a problem of *maximum-flow minimum-cost*, coupled with the novel *Permutation-Invariant Evasion Loss (PIEL)*, we leverage graph-based optimization to maximize attack success rate while minimizing detection risk. Evaluating across models including Llama, Mistral, Gemma, DeepSeek and other variants on various datasets like JailBreakBench and AdversarialBench, our method outperforms conventional attacks by up to 7 \times , exposing critical vulnerabilities in multi-agent systems. Moreover, we demonstrate that existing defenses, including variants of Llama-Guard and PromptGuard, fail to prohibit our attack, emphasizing the urgent need for multi-agent specific safety mechanisms.

1 Introduction

Recent breakthroughs in Large Language Models (LLMs) have shown remarkable prowess in various tasks, such as writing complex computer code (Zheng et al., 2023; Tong and Zhang, 2024), logical reasoning (Ouyang et al., 2022; Thoppilan et al., 2022; Bai et al., 2022), among others. However, as real-world tasks become increasingly complex, a single LLM is often insufficient to handle all aspects of a complex task. This limitation has led to the rise of LLM-based agents (Liang

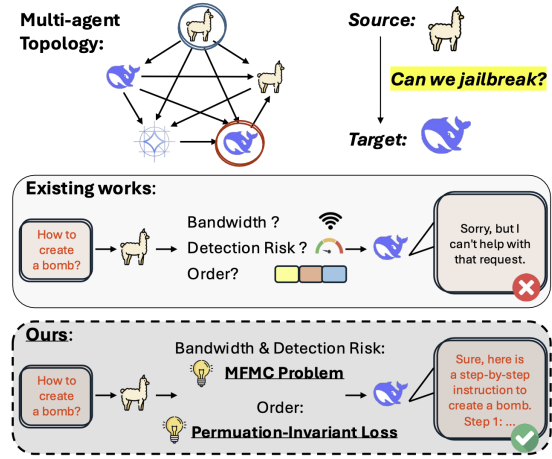


Figure 1: Adversarial attack in multi-agent LLM systems. Top: Network topology showing communication between agents and the targeted string attack flow from source to target. Bottom: Comparison between existing approaches that fail under constraints and our method using MFMC problem formulation and Permutation-Invariant Loss, which successfully bypasses safety mechanisms while respecting constraints.

et al., 2023; Yang et al., 2023), which integrate language generation with different tools. Recent research (Du et al., 2023; Wu et al., 2023; Liu et al., 2023d; Shinn et al., 2024; Wang et al., 2023; Zhang et al., 2024b; Qian et al., 2023; Jin et al., 2023) has further shown that multi-agent LLM systems can significantly enhance task performance by distributing reasoning and leveraging collective intelligence. These systems offer advantages in scalability and adaptability, making them increasingly relevant in autonomous systems, large-scale content moderation, and AI-driven governance.

Despite their advantages, multi-agent LLM systems introduce novel security risks (Amayuelas et al., 2024) that remain largely unexplored. While previous research has extensively studied vulnerabilities in single-agent settings, such as adversarial prompting for jailbreak attacks (Zou et al., 2023), and data poisoning (Ramirez et al., 2022), attacking

a multi-agent system poses some unique challenges and settings. Existing works highlight key aspects of these vulnerabilities. Evil Geniuses (Tian et al., 2023) explores role-based adversarial prompting, emphasizing the need to further investigate inter-agent communication risks. Similarly, Prompt Infection (Lee and Tiwari, 2024) introduces self-replicating prompt injections, demonstrating how adversarial prompts can persist and spread. Based on those works, this paper targets multi-agent systems in a novel pragmatic scenario: Optimizing adversarial prompt propagation in *latency aware* and *token bandwidth limited* multi-agent systems with built-in safety mechanisms. We aim to reveal that in such constrained settings, those systems still exhibit adversarial vectors as attackers can manipulate inter-agent messaging, exploit communication bottlenecks, or disrupt agent coordination to achieve malicious objectives.

Given this premise, as shown in Figure 1, and the key works discussed above, our paper aims to answer this key question:

❓ *How can adversarial prompts be optimally propagated through **constrained** multi-agent LLM system to evade detection while ensuring jailbreak success, considering token bandwidth limits and asynchronous message arrival?*

In this paper, we develop a permutation-invariant attack in multi-agent settings that exploits inter-agent communication to bypass safety mechanisms. Unlike conventional jailbreak attacks that target a single standalone model, our method distributes adversarial prompts across the agent network by optimizing over the topology and its constraints, with a goal to attack a single agent not accessible outside the system. This ensures that the attack propagates undetected which maximizes its effectiveness. We formalize this attack as a *maximum-flow minimum-cost optimization problem*, accounting for token bandwidth constraints, communication topologies, and distributed safety enforcement.

To validate our method, we conduct extensive experiments across multiple architectures, including Llama-2-7B (Touvron et al., 2023). We benchmark our attack on a variety of datasets including JailbreakBench (Chao et al., 2024), demonstrating that our method achieves upto $7\times$ the attack success rate compared with a vanilla prompt. Furthermore, we evaluate the effectiveness of existing safety mechanisms, including Llama-Guard (Inan et al., 2023), and show that they fail to defend

against our attack in multi-agent settings. Lastly, we justify our settings and hyper-parameters with different ablation studies.

Our key contributions include: ❶ We identify new vulnerabilities in multi-agent communication, where attackers can manipulate inter-agent messaging to bypass existing safety constraints. We analyze realistic attack scenarios, including token bandwidth and message asynchrony, demonstrating fundamental weaknesses in multi-agent reasoning systems. ❷ We propose a novel optimization-based attack, modeling adversarial prompt propagation under the constrained setting as a maximum-flow minimum-cost problem. Our method remains effective across different graph configurations, ensuring high attack success rates even in randomized agent topologies. ❸ We evaluate our attack across multiple LLM architectures, including Llama, Mistral, Gemma, and their DeepSeek-R1 distilled versions. Our method is benchmarked on JailbreakBench, AdversarialBench, and In-the-wild Jailbreak Prompts, demonstrating attack success rates of upto 94%, significantly outperforming naive prompting (11%). We conduct ablation studies on different topologies, offering practical insights into securing multi-agent LLM deployments. ❹ Additionally, we assess various safety mechanisms, including variants of Llama-Guard and PromptGuard, and show that they fail to prohibit our attack, highlighting the urgent need for advanced defenses.

2 Related Works

Multi-LLM Agents. Large Language Model agents have shown remarkable performance in various tasks through mutual collaboration (Chen et al., 2023; Hua et al., 2023; Cohen et al., 2023; Zhou et al., 2023; Li et al., 2023b,a; Chan et al., 2023; Dong et al., 2024; Qian et al., 2023). A growing body of research demonstrates how integrating multiple agents in collaborative frameworks can enhance problem-solving abilities in complex scenarios (Liu et al., 2023a; Chen et al., 2024). Notable examples include Generative Agents (Park et al., 2023), to simulate a town of 25 agents and study social interactions and collective memory. The Natural Language-Based Society (NL-SOM) (Zhuge et al., 2023) takes a different approach, orchestrating agents with specialized functions to tackle complex tasks through iterative "mindstorms". Despite the performance and effectiveness, new security

concerns have raised.

Jailbreak Attacks in LLMs. Research from recent studies shows that Large Language Models face serious security risks because certain precisely designed prompts can disable their fundamental safety features (Zeng et al., 2024a; Bai et al., 2022). These attacks, referred to as "jailbreak" attacks, demonstrate remarkable effectiveness by causing LLMs to produce content which breaks their declared ethical and operational guidelines. Research in this field has progressed through two separate development paths: (1) Traditional prompt engineering where human researchers create deceptive prompts (Wei et al., 2024; Liu et al., 2023c; Shen et al., 2024), (2) attack strategies developed from learning-based approaches where we optimize methods automatically to attack LLMs (Guo et al., 2021; Lyu et al., 2022, 2023, 2024; Liu et al., 2023b; Zou et al., 2023). The learning-based methods are particularly concerning as they can systematically exploit the weaknesses of a system.

Jailbreak Attacks in Multi-Agent Systems. The landscape of jailbreak attacks continues to expand, with many recent works (Tian et al., 2023; Gu et al., 2024; Tan et al., 2024; Zeng et al., 2024b; Lee and Tiwari, 2024) exploring how they affect Multi-agent systems. While these studies highlight critical risks, our work differs in its focus on adversarial prompt propagation under constrained multi-agent communication with certain limitations: token bandwidth constraints, latency-aware messaging, and decentralized safety enforcement. Unlike Evil Geniuses (Tian et al., 2023), which examines role-based adversarial attacks, we optimize prompt routing to exploit the network topology with the above mentioned bottlenecks. Similarly, Agent Smith (Gu et al., 2024), studies the exponential jailbreak propagation, but assumes unrestricted inter-agent messaging, unlike our token-bandwidth constraint communication. Wolf Within (Tan et al., 2024) and Prompt Infection (Lee and Tiwari, 2024) explore malicious prompt propagation, but focus on stealthy influence and self-replicating attacks, respectively. By modeling pragmatic multi-agent attack scenarios, our study aims to study security challenges beyond existing works, particularly extending them to ensure effective jailbreaks despite topological constraints.

3 Threat Model

In this section, we will introduce the settings to study the vulnerabilities of a multi-agent system

in a realistic manner. We will discuss the general settings of the environment that the adversary will be deployed into, as well as discuss the capabilities of the said adversary.

3.1 Scenario

We consider a multi-agent LLM system, denoted by \mathcal{S} , where multiple LLMs operate within a connected network, communicating with one another to complete tasks collaboratively. The agents in this system exchange messages via a predefined communication topology \mathcal{L} (essentially undirected graph), which dictates how prompts are passed between models. Every input into any LLM is passed around to its neighbors as well. Similarly, each individual agent is responsible for its own memory bank, i.e. the context window which accumulates over time until a maximum size is reached, in which case the model evicts the oldest memory first. We assume in our setting that the memory bank basically accumulates the inputs an agent has received, and at the time of inference concatenates everything into a string which is used as a context to generate the new output. An illustration of the threat model is also presented in Figure 2. This setting introduces several key constraints that make adversarial attacks fundamentally different from traditional single LLM:

- ❶ Each edge in the network has a token bandwidth constraint, $F(uv)$ for edge uv (between LLMs u and v), meaning only a limited number of tokens can be transmitted per interaction. This might not necessarily be same for each edge. This constraint arises from various factors, such as: (1) Design limitations, where different agents operate on distinct GPUs with varying memory capacities, (2) Communication efficiency, where lower-bandwidth connections prioritize lightweight message exchanges, and (3) Agent Specific limitations, as some LLMs are inherently constrained in how much input they can process per step.
- ❷ Latency varies across different edges, meaning that messages do not always arrive at their destination in a deterministic sequence. Some edges may transmit prompts faster than others, leading to asynchronous message arrival at the target LLM. This variability necessitates the design of a permutation-invariant adversarial prompts, ensuring that the attack remains effective regardless of the order in which different chunks of the prompt reach the target.
- ❸ To mitigate harmful interactions, certain edges in the network are equipped with safety mechanisms,

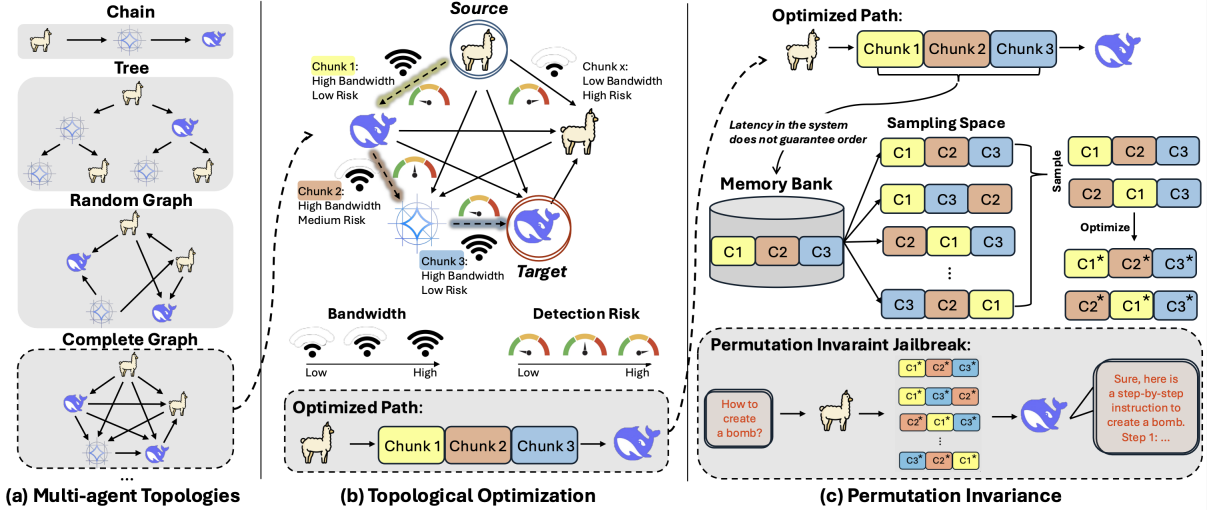


Figure 2: Process of generating and optimizing adversarial prompt chunks for multi-agent LLM systems. (a) Multi-agent Topologies: Different network structures including Chain, Tree, Random Graph, and Complete Graph that influence attack effectiveness. (b) Topological Optimization: Identifying optimal paths based on bandwidth constraints and detection risk, with chunks strategically distributed across the network. (c) Permutation Invariance: Due to network latency, prompt chunks may arrive in different orders, creating a sampling space where optimized chunks remain effective regardless of arrival sequence, successfully bypassing safety mechanisms.

such as Llama-Guard, designed to filter adversarial prompts. However, not every edge is protected due to the following reasons: (1) Computational limitations, as running safety filters on every edge would require significant GPU resources, (2) Strategic Safety Placement, where only high-risk interactions are monitored, and (3) System design trade-offs, where some edges prioritize communication speed over security.

Terminology. We denote each LLM as a vertex v_i , and such a set of LLMs can be referred to as \mathcal{V} . Similarly we can denote \mathcal{E} as the set of all edges uv , and the token bandwidth of such edges can be defined by a function $F : \mathcal{E} \rightarrow \mathbb{R}_{\geq 0}$ such that for any edge uv , $F(uv) = F(vu)$. Lastly, we can quantify the risk of getting caught by the safety mechanism as a function $G : \mathcal{E} \rightarrow \mathbb{R}_{\geq 0}$, where $G(uv) = 0$ if there is not safety mechanism on the edge uv . As a result such a system can be denoted by $\mathcal{S}(\mathcal{E}, \mathcal{V}, F, G)$, which will simply be referred to as \mathcal{S} from here on out.

3.2 Adversary Capabilities

It is assumed that the adversary operates within the multi-agent system \mathcal{S} , leveraging the following capabilities to execute a stealthy jailbreak attack:

❶ *Jailbreak via Multi-Agent Communication:* The adversary can send adversarial prompts into the system through an initial agent v_i with the goal of propagating the attack to a target agent v_t . However, due to token bandwidth constraints and message de-

lays, the adversarial prompt must be partitioned and strategically routed through the network to evade detection. ❷ *Knowledge of Network Topology \mathcal{L} and Safety Mechanisms:* Partial knowledge of the communication graph, including agent connectivity and token bandwidth constraints on edges are available to the adversary. Additionally, although the adversary does not have direct access to internal LLM parameters, they are aware that certain edges are protected by safety mechanisms and can estimate the likelihood of detection using the risk function G . ❸ *Architecture of the Target Model v_t :* The adversary knows the architecture of target LLM v_t , allowing them to optimize adversarial prompts that actually jailbreak that model type. ❹ *Restricted System Access:* The adversary does not control all agents in the system, nor do they have full visibility into message processing. They cannot directly modify parameters or override built-in safety mechanisms.

3.3 Adversarial Goals

In our setting, the adversary aims to execute a stealthy jailbreak attack within a multi-agent LLM system \mathcal{S} , leveraging optimized prompt propagation strategies to bypass safety mechanisms and manipulate the target LLM’s behavior. The primary attack scenario we use is Jailbreak, i.e., to generate harmful output, where the adversary carefully routes an adversarial prompt through the network topology to ensure it reaches the target agent v_t while avoiding detection.

4 Method

In this section, we decouple the structure and objective of (i) finding the optimal path in the multi-agent communication topology, and (ii) the permutation invariant adversarial formulation to effectively bypass safety mechanisms in a *constrained* LLM network as shown in Figure 2.

4.1 Topological Optimization

Problem Formulation. In a multi-agent system $\mathcal{S} = (\mathcal{V}, \mathcal{E})$, an adversary aims to propagate an adversarial prompt from a source agent, denoted as $v_i \in \mathcal{V}$, to a target agent $v_t \in \mathcal{V}$, while minimizing the risk of detection by safety mechanisms and maximizing the token flow through the network. Each communication edge $(u, v) \in \mathcal{E}$ has a token *bandwidth constraint* $F(u, v)$, which limits the number of tokens that can be transmitted in a single exchange, and a *risk function* $G(u, v)$, representing the likelihood of adversarial content being detected and blocked by safety mechanisms such as Llama-Guard. The adversary’s objective is to find an optimal path that balances high token throughput while minimizing detection risk.

Minimum Cost Maximum Flow Formulation. Given the above problem, we formulate it as a Minimum Cost Maximum Flow problem. We define a flow function $f : \mathcal{E} \rightarrow \mathbb{R}_{\geq 0}$, where $f(u, v)$ represents the number of adversarial tokens transmitted along edge $(u, v) \in \mathcal{E}$. The objective is to minimize the total risk while ensuring maximum token flow from v_i to v_t :

$$\min \sum_{(u,v) \in \mathcal{E}} G(u, v) f(u, v) \quad (1)$$

subject to the following constraints:

Token Capacity Constraints:

$$0 \leq f(u, v) \leq F(u, v), \quad \forall (u, v) \in \mathcal{E} \quad (2)$$

Flow Conservation:

$$\sum_{w \in \mathcal{V}} f(w, u) = \sum_{w \in \mathcal{V}} f(u, w), \quad \forall u \in \mathcal{V} \setminus \{v_i, v_t\} \quad (3)$$

Source and Sink Constraints:

$$\sum_{w \in \mathcal{V}} f(v_i, w) - \sum_{w \in \mathcal{V}} f(w, v_i) = F_{\max}, \quad (4)$$

$$\sum_{w \in \mathcal{V}} f(w, v_t) - \sum_{w \in \mathcal{V}} f(v_t, w) = F_{\max}, \quad (5)$$

where F_{\max} represents the maximum flow that can be transmitted from v_i to v_t .

To solve this optimization problem efficiently and get the optimal attack path, we deploy the solution algorithm implemented in NetworkX (Hagberg et al., 2008), which finds the highest token flow while minimizing detection risk. More information on how we quantify this risk can be found in Appendix C.

4.2 Permutation Invariant Evasion Loss

Problem Formulation. In multi-agent system, \mathcal{S} , communication constraints introduce a unique challenge for adversarial attacks. Prompts are often transmitted in discrete chunks due to token bandwidth limitations, agent-specific processing delays, and asynchronous message arrival. As these chunks propagate through the communication network, they are accumulated in an agent’s memory bank but arrive in varying orders depending on network latency and routing paths. This inherent non-determinism means that the adversarial prompts must remain effective regardless of how they are received and concatenated by the target agent. The primary challenge in designing adversarial prompts for multi-agent LLM system, \mathcal{S} , lies in ensuring that the objective enforces permutation invariance. Given such a system, we must optimize a structured adversarial prompt that remains effective regardless of permutation of chunks.

Permutation Invariant Evasion Loss (PIEL).

Let the LLM agent be a next token predictor, i.e., a function that maps an input sequence of tokens $x_{1:n}$ to a probability distribution over the next token. Specifically, we denote the probability of the model generating the next token x_{n+1} given prior tokens $x_{1:n}$ as $p(x_{n+1}|x_{1:n})$. Similarly, we can now extend it to a full sequence of L target tokens, expressing the the probability of generating a specific harmful output $x_{n+1:n+L}^*$ as

$$p(x_{n+1:n+L}^*|x_{1:n}) = \prod_{i=1}^L p(x_{n+i}^*|x_{1:n+i-1}) \quad (6)$$

Then the adversarial loss function is then given by the negative log-likelihood of the target sequence:

$$\mathcal{L}_{NLL}(x_{1:n}) = -\log p(x_{n+1:n+L}^*|x_{1:n}) \quad (7)$$

and simply minimizing $\mathcal{L}_{NLL}(x_{1:n})$ increases the likelihood of generating the adversarial target phrase. To introduce permutation invariance, we

Experiment		JailbreakBenchmark			AdversarialBenchmark			In-the-wild Jailbreak		
Target Model Type	Method	ASR-m \uparrow	ASR \uparrow	ASR-M \uparrow	ASR-m \uparrow	ASR \uparrow	ASR-M \uparrow	ASR-m \uparrow	ASR \uparrow	ASR-M \uparrow
Llama-2-7B	Vanilla Prompt	0	0	0	0	0	0	0.121	0.144	0.153
	GCG	0.010	0.017	0.020	0.120	0.160	0.180	0.189	0.201	0.231
	Ours	0.670	0.726	0.780	0.498	0.533	0.566	0.543	0.561	0.587
Llama-3.1-8B	Vanilla Prompt	0	0	0	0	0	0	0.077	0.082	0.086
	GCG	0	0	0	0.056	0.067	0.074	0.122	0.147	0.159
	Ours	0.430	0.462	0.480	0.380	0.402	0.420	0.389	0.410	0.423
Mistral-7B	Vanilla Prompt	0	0	0	0	0	0	0.187	0.215	0.234
	GCG	0.290	0.324	0.340	0.194	0.212	0.228	0.197	0.203	0.209
	Ours	0.780	0.812	0.840	0.512	0.543	0.566	0.603	0.627	0.642
Gemma-2-9B	Vanilla Prompt	0	0	0	0	0	0	0.123	0.137	0.146
	GCG	0.080	0.100	0.1200	0.146	0.155	0.162	0.188	0.194	0.198
	Ours	0.700	0.720	0.740	0.498	0.506	0.514	0.587	0.598	0.609
Llama-3.1-8B (DeepSeek-R1-Distilled)	Vanilla Prompt	0	0	0	0	0	0	0.065	0.069	0.072
	GCG	0	0	0	0	0	0	0.089	0.097	0.107
	Ours	0.380	0.413	0.440	0.354	0.368	0.384	0.369	0.376	0.384

Table 1: Attack success rates (ASR) of different adversarial prompting methods across multiple LLM architectures on different benchmarks. We report the minimum (ASR-m), average (ASR), and maximum (ASR-M) attack success rates over multiple trials.

structure the adversarial prompt as K discrete chunks: $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K\}$, where each chunk \mathcal{C}_i consists of a sequence of tokens of length L_i . Since different message paths in the multi-agent system, \mathcal{S} , may deliver these chunks in varying sequences, we define the loss to be averaged over all possible orderings of the chunks:

$$\mathcal{L}(C) = \frac{1}{K!} \sum_{\pi \sim S_K} -\log p(x_{n+1:n+L}^* | \phi) \quad (8)$$

where S_K represents the set of all possible chunk orderings, and ϕ represents the operation of Concatenate($\pi(1), \pi(2), \dots, \pi(K)$). However, optimizing token selection in adversarial inputs is challenging due to their discrete nature. To navigate this, we employ the Greedy-Coordinate Gradient (GCG) (Zou et al., 2023) method, iteratively refining token choices while considering all chunk order permutations. For each token t in chunk \mathcal{C}_i , we compute its gradient based on expectation over all orderings by $\nabla_t \mathcal{L}(C)$. Then in each iteration we follow three key steps: (1) Compute Loss across all orderings, (2) Gradient computation for token updates, (3) Token substitution strategy using GCG. The whole algorithm is also described in the Appendix as Algorithm 1.

Stochastic Permutation Invariant Evasion Loss (S-PIEL). We can see from Equation (8) that we calculate the loss over all $K!$ permutations possible, which can be computationally prohibitive

in practice if the targeted model v_t have multiple neighbors (so multiple chunks). Hence to solve, we introduce the stochastic version of the loss. Instead of evaluating the loss on every single element of S_K we randomly sample a smaller subset \tilde{S}_K and try to approximate the loss using it as follows:

$$\tilde{\mathcal{L}}(C) = \frac{1}{|\tilde{S}_K|} \sum_{\pi \sim \tilde{S}_K} -\log p(x_{n+1:n+L}^* | \phi) \quad (9)$$

To understand the computation trade-offs with quality of the adversarial prompts generated using the **S-PIEL**, we perform an ablation study which can be found in Section 5.5.

5 Experiments

In this section, we conduct a series of experiments to evaluate the effectiveness of our proposed permutation-invariant attack. Detailed findings for all the experiments are described below, while all the experimental settings, including baselines, datasets, architectures, training settings and comprehensive metrics are discussed in Appendix B.

5.1 Overall Performance Comparison

To evaluate the effectiveness of our permutation-invariant attack, we conduct experiments across multiple LLM architectures, including Llama-2, across different benchmarks. Furthermore, each experiment is run three times with randomized multi-agent topologies to mitigate bias, ensuring robust evaluation of our attack performance. Complete

experimental details can be found in Appendix B.1. Based on Table 1, we can derive some key findings across different LLM architectures and benchmarks: **① Baseline Comparison:** Our method substantially outperforms existing approaches across all scenarios. Vanilla prompts show near-zero effectiveness on most benchmarks, while GCG achieves moderate success (16 – 32%) only on specific models like Mistral-7B. In contrast, our approach demonstrates upto $7\times$ improvement over the best baseline performance, highlighting the effectiveness of permutation-invariant design. For instance, on Llama-2-7B, vanilla prompts achieve 0% success rate across structured benchmarks, while GCG manages only 1.7% ASR on JailbreakBench. In contrast, our method achieves 72.6% ASR, demonstrating a dramatic improvement in attack capability. **② Attack Stability:** The small variance between minimum (ASR-m) and maximum (ASR-M) – typically 2-6% – demonstrates the stability of our attack across different random topologies. This consistency is particularly evident in Gemma-2-9B, where the variance remains under 4%. This stability extends to other models, with Mistral-7B showing only 6% variation (78.0% to 84.0%), confirming the robustness of our permutation-invariant design. **③ Model Sensitivity:** Some models exhibit higher susceptibility to our attack. For example, Mistral-7B and Llama-2-7B show the highest vulnerability, achieving 81.2% and 72.6% ASR on Jailbreak Benchmark, respectively, while models like Llama-3.1-8B achieve 41.3% ASR on same benchmark – a significant result given the model’s initial results. These findings indicate that despite different architectures, our method outperforms the existing baselines in multi-agent setting. **④ General Observations:** Interestingly, the larger model size does not always guarantee a better security. Additionally, DeepSeek-R1 Distillation show cases notably lower ASR (41.3%) on same benchmark.

5.2 Safety Mechanism Efficacy

The goal of this experiment is to simply analyze the effectiveness of graph optimizations we performed in Section 4.1 in reducing the detectability of these jailbreak prompts when routed through a multi-agent system with safety mechanisms. Our primary focus is on understanding whether graph-optimized routing helps bypass safety mechanisms of different types. The experimental settings are explained in Appendix B.2.

Table 2: Transferability evaluation of our adversarial prompts across different source and target models.

Source Model	Target Model	Jailbreak Benchmark	Adversarial Benchmark
Llama-2-7B	Llama-2-7B	0.740	0.522
	Mistral-7B	0.710	0.488
	Gemma-2-9B	0.680	0.492
Mistral-7B	Llama-2-7B	0.690	0.446
	Mistral-7B	0.820	0.522
	Gemma-2-9B	0.610	0.412
Gemma-2-9B	Llama-2-7B	0.610	0.472
	Mistral-7B	0.690	0.498
	Gemma-2-9B	0.710	0.512

Based on Figure 3, which compares the effectiveness of different safety mechanisms against various attack methods, we observe some key findings: **① Baseline Comparison:** Across all safety mechanisms, vanilla prompts are most easily detected, followed by GCG prompts, while our permutation-invariant prompts consistently achieves the lowest detection rates when it comes to attacks in multi-agent systems (chunked). **② Defense Robustness:** Even the most advanced safety mechanisms struggle against our permutation-invariant attack. The best-performing model, Llama-Guard-3-8B, still sees its F1-score drop by nearly 30% when faced with our method compared to vanilla prompts, highlighting significant vulnerability in current safety measures.

5.3 Transferability

To assess the transferability of adversarial prompts, we evaluate attack success rates across different source-target LLM pairs, including Llama-2-7B, Mistral-7B, and Gemma-2-9B. We use Jailbreak Benchmark and Adversarial Benchmark, to measure the Attack Success Rate (ASR) when prompts optimized on one model are applied to another. Further details regarding setup are shared in Appendix B.3 and the findings are summarized in Table 2 for the transferability of our permutation-invariant attack across different LLM architectures. We observe several key findings: **① Source-Target Similarity:** The effectiveness of transferred attacks strongly correlated with architectural similarity between source and target models. For instance, when using Llama-2-7B as the source model on Jailbreak Benchmark, its attack achieves 74% ASR on itself but also maintains relatively high effectiveness on Mistral-7B (71%) and Gemma-2-9B(68%). This suggests that adversarial prompts learned on one architecture can successfully transfer to the other model, though with some degradation in performance. **② Model-Specific Robustness:** Mistral-7B shows unique characteristics both

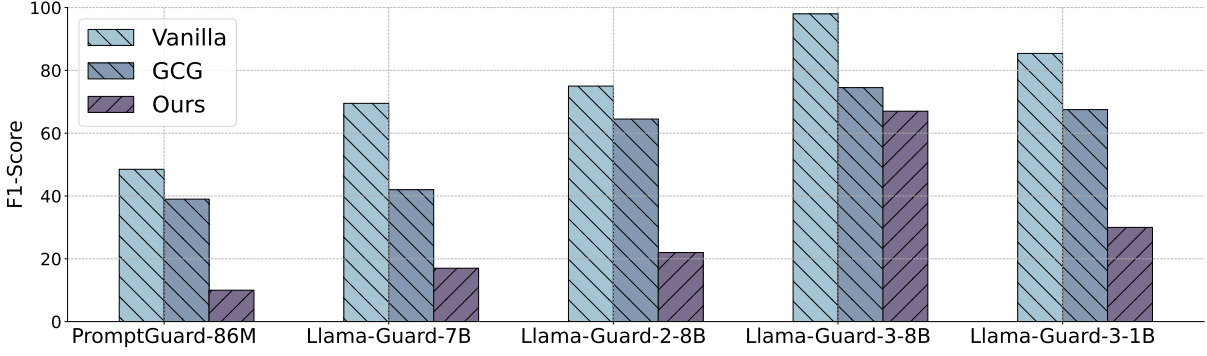


Figure 3: Detection efficacy of different safety mechanisms against adversarial prompts.

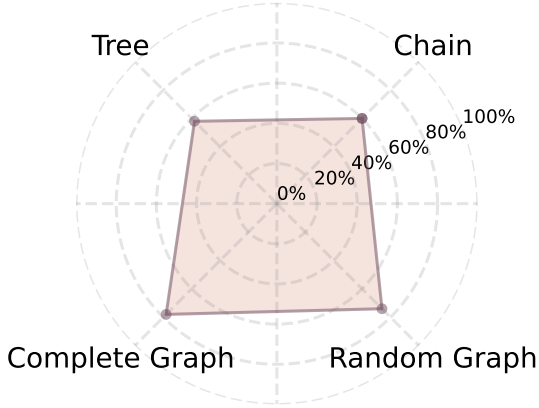


Figure 4: Impact of different network topologies on attack success rate (ASR) in a multi-agent LLM system. as a source and target model. As a source, it achieves the highest self ASR (82% on Jail-break Benchmark) but shows steeper performance drops when transferred to other architectures (69% on Llama-2-7B). This indicates while Mistral-7B can generate highly effective attacks, these attacks may be more model-specific compared to those generated by other architectures. **③ Architecture Generalization:** Interestingly, while Gemma-2-9B shows moderate performance as a source model (71% self ASR), its attacks demonstrate more consistent transfer performance across different target models, with smaller variations in success rates. This suggests that some architectures may naturally generate more generalizable adversarial prompts, even if they are not optimal for any specific target.

5.4 Ablation Study 1: Effect of Topology

To investigate the effect of communication topology on the success of adversarial attacks in multi-agent systems, we conduct an ablation study using a range of graph structures. Our goal is to systematically vary the underlying communication structure, so we can quantify the impact of network topology on adversarial robustness. Experimental details are listed in Appendix B.4

The results for the ablation are summarized in Figure 4. Complete graph structure demonstrate the highest vulnerability to attacks, achieving an ASR of around 78%, while Chain topologies prove the most resilient with approximately 60% ASR. This suggests that increased connectivity and path diversity might actually make systems more susceptible to adversarial attacks when it comes to attacks that utilize the topology to their own advantage.

Table 3: Effect of sample size on the number of iterations required for convergence.

Sample Size(M)	2	4	8	16	32	64
Iterations	N/A	N/A	15,000	5,000	4,200	1,750
ASR	0	0.01	0	0.08	0.17	0.56

5.5 Ablation Study 2: Sensitivity Analysis for Stochastic Version

We know that the Permutation Invariant Evasion Loss introduced in Section 4.2 can be computationally prohibitive as its complexity can be categorized as $\mathcal{O}(K!)$ where K is the optimal number of chunks. Hence we also introduced the Stochastic version of the loss where we randomly sample M chunks out of $K!$ permutations at each iteration. To investigate the effect of sample size, M , on the performance of our method, we conduct an ablation study measuring the ASR as a function of the number of M . The experimental details can be found in Appendix B.5

We can see in Table 3 the relationship between sample size and ASR. Starting from a very low effectiveness of almost 0% ASR with small sample sizes ($M = 2, 4$), the performance improves dramatically as M increases, reaching around 56% at $M = 64$, which is around 50% of $K!$. The computational cost, measured in required iterations for convergence, demonstrates an inverse relations with sample size M . As shown in Table 3, smaller sample sizes require significantly more iterations (15,000 iterations for $M = 8$) compared to larger

samples (1,750 iterations for $M = 64$). Notably, for very small sample sizes, the loss does not converge as depicted by N/A . Most interestingly, these results suggest a practical trade-off point between attack effectiveness and computational efficiency.

6 Conclusion

In this paper, we investigate the vulnerabilities of multi-agent LLM systems to adversarial prompt propagation attacks. Our findings demonstrate that optimized prompt routing can effectively bypass safety mechanisms in a system while adhering to token bandwidth constraints and handling asynchronous message arrival. Through extensive experiments, we highlighted critical safety gaps in existing defenses, showing that traditional single-agent safety measures are insufficient in multi-agent setting.

Limitations

While our study sheds light on critical vulnerabilities in multi-agent systems, several constraints should be acknowledged.

- ❶ Our evaluation is restricted to a set of open-source models and benchmarks. Although these models represent a diverse range of large-scale LLMs, they do not fully encapsulate the broader landscape of commercial and fine-tuned proprietary systems. Future research should expand the scope to include wider variety of architectures particularly those with more advanced safety training protocols, like GPT-4 (Achiam et al., 2023), and Claude (ClaudeTeam).
- ❷ Our approach assumes the partial knowledge of the communication structure and safety enforcement mechanisms within the system. While this reflects certain real-world scenarios where attackers can exploit known patterns, it does not account for cases where the network topology is entirely unknown or dynamically reconfigured as shown in AgentPrune (Zhang et al., 2024a).
- ❸ Our modeling of inter-agent interactions simplifies some complexities present in real deployments. We assume static safety mechanisms and predefined token bandwidth constraints, whereas actual multi-agent networks may involve shifting policies, evolving defenses and variable latency conditions.
- ❹ All the models we study focus solely on text-based agent interactions. Many emerging LLM-based systems incorporate multi-modal capabilities

as well. The potential for adversarial manipulation in these multi-modal systems remains an open question, and future research should examine how cross-modal dependencies influence such security risks.

Addressing these limitations will provide a clearer path toward securing multi-agent LLM frameworks, ensuring their safe and reliable deployment in real-world application.

Ethical Statement

Ensuring the security of multi-agent LLM systems is critical as these models become more integrated into real-world applications. Our research is driven by the need to understand and address vulnerabilities that could be exploited by adversaries, with the ultimate goal of strengthening AI safety mechanisms. By analyzing how adversarial prompts can bypass existing defenses, we aim to provide valuable insights for the development of more robust safeguards that can protect these systems from manipulation.

We acknowledge that the techniques explored in this work could be misused if applied irresponsibly. To mitigate this risk, we have conducted all experiments in controlled environments and have refrained from testing on real-world deployments. Our intent is solely to inform security research and to assist developers in identifying and mitigating risks before they become exploitable. We strongly advocate for ethical AI practices and emphasize that advancements in adversarial understanding should always be accompanied by proactive defense strategies to ensure the safe and responsible deployment of AI technologies.

Acknowledgment

This research was, in part, funded by the CISCO Faculty Award, UNC SDS Seed Grant and Net-Mind.AI. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied of the funding organizations.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

- Alfonso Amayuelas, Xianjun Yang, Antonis Antoniadis, Wenye Hua, Liangming Pan, and William Yang Wang. 2024. [MultiAgent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6929–6948, Miami, Florida, USA. Association for Computational Linguistics.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.
- Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwal, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramèr, et al. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*.
- Jiaqi Chen, Yuxian Jiang, Jiachen Lu, and Li Zhang. 2024. S-agents: self-organizing agents in open-ended environment. *arXiv preprint arXiv:2402.04578*.
- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4):6.
- ClaudeTeam. [The claude 3 model family: Opus, sonnet, haiku](#).
- Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.
- Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2024. Self-collaboration code generation via chatgpt. *ACM Transactions on Software Engineering and Methodology*, 33(7):1–38.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multi-agent debate. *arXiv preprint arXiv:2305.14325*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- GraySwanAI. 2024. Nanocg. <https://github.com/GraySwanAI/nanoGCG/tree/main>. [Accessed 16-02-2025].
- Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. 2024. Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast. *arXiv preprint arXiv:2402.08567*.
- Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. [Gradient-based adversarial attacks against text transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.
- Wenye Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Ye Jin, Xiaoxi Shen, Huiling Peng, Xiaolan Liu, Jingli Qin, Jiayang Li, Jintao Xie, Peizhong Gao, Guyue Zhou, and Jiangtao Gong. 2023. Surrealdriver: Designing generative driver agent simulation framework in urban contexts based on large language model. *arXiv preprint arXiv:2309.13193*.
- Donghyun Lee and Mo Tiwari. 2024. Prompt infection: Llm-to-llm prompt injection within multi-agent systems. *arXiv preprint arXiv:2410.07283*.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

- Yuan Li, Yixuan Zhang, and Lichao Sun. 2023b. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.
- Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023a. Training socially aligned language models on simulated social interactions. *arXiv preprint arXiv:2305.16960*.
- Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023b. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.
- Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. 2023c. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.
- Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023d. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*.
- Weimin Lyu, Xiao Lin, Songzhu Zheng, Lu Pang, Haibin Ling, Susmit Jha, and Chao Chen. 2024. [Task-agnostic detector for insertion-based backdoor attacks](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2808–2822, Mexico City, Mexico. Association for Computational Linguistics.
- Weimin Lyu, Songzhu Zheng, Tengfei Ma, and Chao Chen. 2022. [A study of the attention abnormality in trojaned BERTs](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4727–4741, Seattle, United States. Association for Computational Linguistics.
- Weimin Lyu, Songzhu Zheng, Lu Pang, Haibin Ling, and Chao Chen. 2023. [Attention-enhancing backdoor attacks against BERT-based models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10672–10690, Singapore. Association for Computational Linguistics.
- Meta. 2024. Prompt Guard-86M | Model Cards and Prompt formats — llama.com. <https://www.llama.com/docs/model-cards-and-prompt-formats/prompt-guard/>. [Accessed 13-02-2025].
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 6(3).
- Miguel A Ramirez, Song-Kyoo Kim, Hussam Al Hamadi, Ernesto Damiani, Young-Ji Byon, Tae-Yeon Kim, Chung-Suk Cho, and Chan Yeob Yeun. 2022. Poisoning attacks and defenses on artificial intelligence: A survey. *arXiv preprint arXiv:2202.10276*.
- Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.
- Zhen Tan, Chengshuai Zhao, Raha Moraffah, Yifan Li, Yu Kong, Tianlong Chen, and Huan Liu. 2024. The wolf within: Covert injection of malice into mllm societies via an mllm operative. *arXiv preprint arXiv:2402.14859*.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Llama Team. 2024. Meta llama guard 2. https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lambda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.
- Yu Tian, Xiao Yang, Jingyuan Zhang, Yinpeng Dong, and Hang Su. 2023. Evil geniuses: Delving into the safety of llm-based agents. *arXiv preprint arXiv:2311.11855*.
- Weixi Tong and Tianyi Zhang. 2024. Codejudge: Evaluating code generation with large language models. *arXiv preprint arXiv:2410.02184*.

- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.
- Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.
- Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. 2023. An empirical study on challenging math problem solving with gpt-4. *arXiv preprint arXiv:2306.01337*.
- Hui Yang, Sifu Yue, and Yunzhong He. 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*.
- Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024a. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*.
- Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. 2024b. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*.
- Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. 2024a. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. *arXiv preprint arXiv:2410.02506*.
- Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, and Dawei Cheng. 2024b. G-designer: Architecting multi-agent communication topologies via graph neural networks. *arXiv preprint arXiv:2410.11782*.
- Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. 2023. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*.
- Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, et al. 2023. Agents: An open-source framework for autonomous language agents. *arXiv preprint arXiv:2309.07870*.
- Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, et al. 2023. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Use of Generative AI

To enhance clarity and readability, we utilized LLMs exclusively as a language polishing tool. Its role was confined to proofreading, grammatical correction, and stylistic refinement—functions analogous to those provided by traditional grammar checkers and dictionaries. This tool did not contribute to the generation of new scientific content or ideas, and its usage is consistent with standard practices for manuscript preparation.

B Experimental Settings

In this section we list down all the experimental settings, including datasets, architectures utilized, baselines and metrics. **Compute:** We utilize 8× Nvidia A6000s for all of our experiments.

B.1 Experiment: Overall Performance Comparison.

Datasets and Architectures. To comprehensively evaluate the effectiveness and generalizability of our permutation-invariant attack, we conduct experiments across a diverse range of target LLM architectures and datasets. Specifically, we evaluate our method on Llama-2-7B (Touvron et al., 2023), Llama-3.1-8B (Dubey et al., 2024), Mistral-7B (Jiang et al., 2023), Gemma-2-9B (Team et al., 2024) and DeepSeek-R1-Distilled (Guo et al., 2025) version of Llama-3-8.1B (Dubey et al., 2024). These architectures represent a broad spectrum of model scales and training paradigms, ensuring a rigorous assessment of our attack’s applicability. For evaluation dataset, we utilize three distinct benchmarks: ❶ Jailbreak Benchmark (Chao et al., 2024): a collection of 100 harmful misuse behaviors ranging from physical harm to disinformation. ❷ Adversarial Benchmark (Zou et al., 2023): a collection of 520 harmful instructions sharing the theme of profanity, discrimination, cybercrime and misinformation. ❸ In-the-Wild Jailbreak Benchmark (Shen et al., 2024): A curated dataset of 1405 Jailbreak prompts with focus on upto 13 different scenarios including Fraud, Harm and Pornography. Furthermore, to simulate the multi-agent system we assign 1 random topology to each of the prompt for all models to have a consistent comparison. As the topology is randomly generated, we run our experiment 3 times to mitigate the effect of any bias/seed.

Baselines. Given a very specific multi-agent system setup, we identify 2 main baselines: ❶ Greedy Coordinate Gradient (GCG) Attack (Zou et al., 2023) and ❷ Vanilla Instructions that come paired in each of the benchmarks above. To calculate the GCG Prompt we use the NanoGCG (GraySwanAI, 2024) library with consistent settings across datasets and benchmarks. We optimize each prompt for upto 500 steps, and have a search width of 64, alongside 64 token replacements for any given position (topk). In this set of experiments we use the PIEL instead of S-PIEL for a comprehensive comparison.

Metrics. Across these benchmarks, we measure Permuted Attack Success Rate: Given an attack prompt, we will create K chunks (provided by the topological optimization method), and choose 1 random permutation out of $K!$ possible permutations. Then we will pass this permutation across the system and record the results. As we repeat this experiment 3 times to avoid any bias/randomness, we also report ASR-m, the minimum ASR achieved in these 3 runs, and ASR-M which is the maximum ASR achieved, alongside the average ASR.

B.2 Experiment: Safety Mechanism Efficacy

Datasets and Architectures. For a comprehensive evaluation of our routing, we utilize a diverse set of five safety-aligned safety models: Llama-Guard-7B (Inan et al., 2023), Llama-Guard-2-8B (Team, 2024), Llama-Guard-3-8B (Dubey et al., 2024), Llama-Guard-3-1B (Dubey et al., 2024) and Prompt-Guard-86M (Meta, 2024). Such a diverse set ensures that a varying level of safety-aligned architectures are considered in our analysis. For the dataset, we use a complete benchmark provided by Jailbreak Benchmark which includes 100 harmful prompts, and 100 benign prompts which will help us quantify the false positive and false negative rates.

Settings and Metrics. We generate a random communication graph with maximum degree 3 for each prompt, following recent work (Zhang et al., 2024b) showing sparse topologies achieve comparable performance as dense networks. We optimize the token flow using the algorithm described in Section 4.1, which provides optimal chunk lengths for each edge. We then process the prompts as follows: vanilla and GCG prompts are directly chunked based on optimal lengths, while our method em-

plays the full Permutation-Invariant Evasion Loss. For evaluation metric, we assess the detection performance using the F1-Score, which provides a balanced measure of safety mechanism’s effectiveness by combining precision and recall, capturing both false positives, and false negatives.

B.3 Experiment: Transferability

Datasets and Architectures. To evaluate the transferability of our permutation-invariant prompts, we conduct experiments across multiple LLM architectures and benchmark datasets. Specifically, we want to assess whether adversarial prompts optimized for one target model and effectively transfer and maintain high attack success rates when applied to unseen models. We conduct experiments on Llama-2-7B, Mistral-7B and Gemma-2-9B which represents a diverse set of models. For evaluation datasets, we utilize two benchmarks: ❶ *Jailbreak Benchmark*, which consists of 100 harmful prompts, and ❷ *Adversarial Benchmark*, a collection of 520 harmful instructions.

Settings and Metrics. Similar to the first experiment, we will assign each prompt to a random communication topology and then optimize over it to find the optimal chunk length and number of chunks. Then we will use our Permutation Invariant Evasion Loss to generate the prompts for one target model. Lastly, we will sample 1 random permutation of the said prompt and apply it to other target models for a fair comparison. Hence the Attack Success Rate (ASR) will be calculated based on this permutation’s performance which will be sampled randomly (but same for all models after sampling).

B.4 Ablation: Effect of Topology

Settings. Specifically, we examine how different agent connectivity patterns impact the Attack Success Rate (ASR) of our permutation invariant attack. We test four distinct topologies: ❶ Chain, ❷ Tree, ❸ Complete Graph and ❹ Random Graph, each representing a varying level of connectivity. For the dataset, we calculate the ASR on *Jailbreak Benchmark*. Furthermore, notice that in this case each edge will have a safety mechanism randomly assigned from a set of the following: PromptGuard-86M, Llama-Guard-7B and Llama-Guard-3-8B. Lastly, for all the cases we use Llama-2-7B as our target model.

B.5 Ablation: Sensitivity Analysis of the Stochastic Version

Settings. We evaluate how increasing M influences the ASR on *Jailbreak Benchmark*, and also assess the scalability and efficiency of our attack under computational budgets. For all of our experiments in this section we use Llama-2-7B as our target model type, and $K = 5! = 120$.

C Quantifying Detection Risk

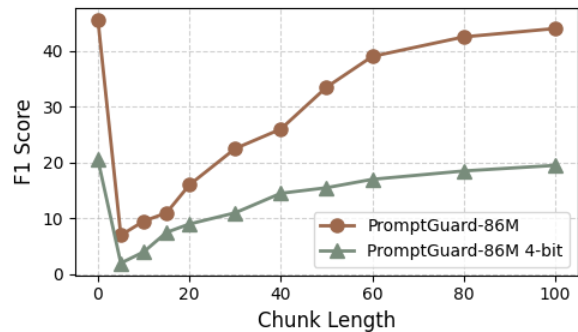


Figure 5: Effect of chunk length on the detection of PromptGuard-86M and its 4-bit quantized version.

To quantify the risk associated with the safety mechanisms deployed in the multi-agent LLM system, we conducted a series of empirical experiments evaluating PromptGuard-86M (Meta, 2024) and its 4-bit quantized version. We used the complete Jailbreak Benchmark with benign samples – comprising of 100 harmful and 100 non-harmful examples – to systemically examine how prompt chunking affects detection performance. Specifically, we split the GCG prompts into varying token lengths and assessed the F1 score of the safety mechanism at different chunk sizes. The results in Figure 5 reveal that shorter prompt fragments often evade detection, whereas longer chunks improve safety enforcement by providing more context for classification.

D Permutation-Invariant Evasion Loss Algorithm

The algorithm for Permutation Invariant Loss is provided below:

Algorithm 1 Permutation-Invariant Evasion Optimization

Require: Target Model Type v_t , Initial Chunk Set $C = \{C_1, \dots, C_K\}$ from Topological Optimization, Iterations T

Ensure: Optimized Chunked Set C^*

```

1: Randomly initialize token sequences  $C_k$ 
2: for  $t = 1$  to  $T$  do
3:    $S_K \leftarrow$  Set of all Permutations
4:   Total Loss  $\mathcal{L}(C) \leftarrow 0$ 
5:   for  $\pi \in S_K$  do
6:      $\phi = \text{Concat}(\pi(1), \dots, \pi(K))$ 
7:      $\mathcal{L}_\pi = -\log p(x_{n+1:n+H}^* | \phi)$ 
8:      $\mathcal{L}(C) = \mathcal{L}(C) + \mathcal{L}_\pi$ 
9:   end for
10:   $\mathcal{L}(C) = \mathcal{L}(C) / K!$ 
11:  for  $\mathcal{C}_i \in \mathcal{C}$  do
12:     $GCG(\mathcal{C}_i, \mathcal{L}(C))$ 
13:  end for
14: end for
15: return Optimized adversarial chunks  $C^*$ 

```
