Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

面向多智能体的传染性越狱

硕士研究生 贺晨阳

2026年8月16日

问题回溯

- 总结反思
  - 算法原理部分讲解不够细致
  - 实验结果等图片太小，排版存在问题
- 相关内容
  - 2026.8.2 刘栋涵《智能体的第三方组件安全：从工具到技能》
  - 2026.7.26 郑俊怡《环境注入攻击EIA》
  - 2026.6.21 袁梦佳《面向LLM Agent的提示注入攻击》
  - 2026.6.21 贺晨阳《大语言模型的越狱攻击》

内容提要

- 预期收获
- 内涵解析与研究目标
- 研究背景
- 知识基础
- 研究历史与现状
- 算法原理
  - Agent Smith
- 特点总结与未来展望
- 参考文献

预期收获

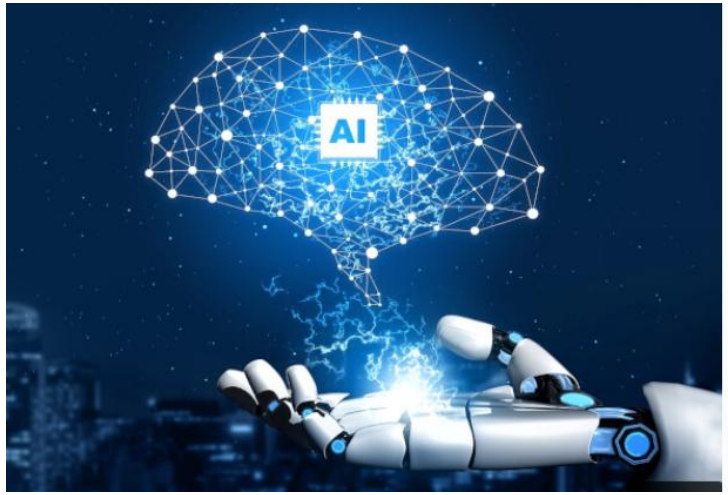
- 预期收获
  - 了解多智能体存在哪些潜在安全隐患
  - 掌握多智能体传染性越狱的基本概念及研究背景
  - 了解多智能体传染性越狱的常见方法及其原理

内涵解析与研究目标

- 题目内涵解析
  - 多智能体系统(MAS)：多个能够自主感知、决策和行动的智能体，通过相互通信组成的系统
  - 传染性越狱：通过越狱单个智能体，利用多智能体间通信特性，实现对其他智能体越狱
- 研究目标
  - 从单一智能体走向多智能体系统
  - 利用多智能体记忆、检索、通信和工具调用能力
  - 实现在多个 agent、多个轮次、多个平台之间传播和潜伏

研究背景

- 多智能体系统
  - 多个专门化智能体协同完成复杂任务的不同部分
  - 能够处理比单个智能体独立行动更广泛的任务、并行流程和更长的任务链
  - 主流框架
    - LangGraph 图结构、 状态机
    - CrewAI 角色扮演 、团队协作
    - AutoGen 对话式协作



研究背景

- 多智能体常见架构
  - 分层式多智能体系统：智能体按层级组织，指挥链清晰、集中规划与分布式执行并行，以及路由可预测、便于调试
  - 协作式多智能体系统：共享工作区及消息总线，实时共享工具、数据和中间结果
  - 对抗式多智能体系统：常见于游戏 AI 和安全测试，相互对弈以打磨策略
  - 异构式多智能体系统：组合具备不同能力、模型或工具集的智能体。
  - 基于图的多智能体系统：智能体和步骤被组织为图中的节点，每个节点处理一个操作，每条边定义下一步运行什么

研究背景

- 多智能体安全的攻击面
  - 协议层攻击
    - MCP引入的攻击：篡改投毒Agent连接的数据源
    - A2A 引入的攻击：通过伪造响应、非法注册或者引发无限循环任务
  - 威胁行为视角
    - 伪装与角色滥用：通过欺骗身份或者冒充受信任的协作者来越权访问
    - 协调操纵：将受损的Agent引入工作流中，将而已影响扩散至整个系统
    - 推理与策略规避：跨智能体分散查询，拼凑敏感数据，规避单一Agent安全策略
    - 问责混淆：利用多智能体间接触发有害行为

研究历史与现状

2023

Tian等人首次聚焦于智能体安全，从智能体数量、角色定义、攻击级别三个维度展开，通过模板式修改系统提示词探索智能体安全性能

2024

Lee等人提出prompt infection通过构造一个带有自复制指令的恶意prompt使得攻击在多agent 通信链中扩散，并提出通过内容来源和信任边界显式标记来缓解传播的防御方法

2025

Rana等人关注更现实的多agent系统约束：消息带宽有限、消息有延迟、分布式安全过滤器。将agent系统建模为图，优化攻击路径：尽量让恶意信息低损耗地到达目标 agent

2024

Gu等人首次提出了传染性越狱，利用离线对抗图片生成以及在线传播进行多智能体越狱，并从理论上推导了传染性越狱的实际效率

2025

He等人提出AiTM方法，攻击者不修改任何智能体或系统组件，仅拦截并操纵发往某个受害智能体的消息，从而间接影响整个系统的输出。首次将攻击面从输入prompt转移到通信协议和消息通道。

2026

Zha等人提出持久化传播攻击，通过扫描agent框架找到弱点，优化worm payload实现持久化的、多跳、跨平台传播

Agent Smith

【ICML-2024】

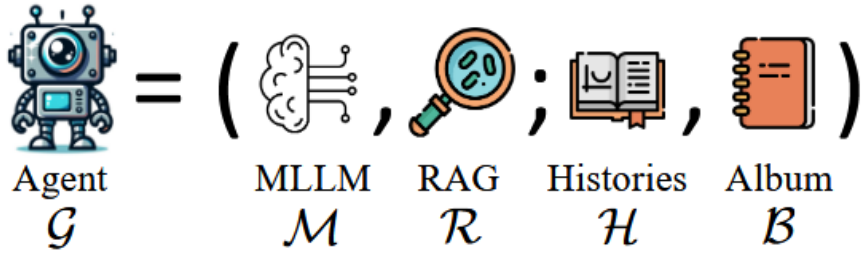
Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast

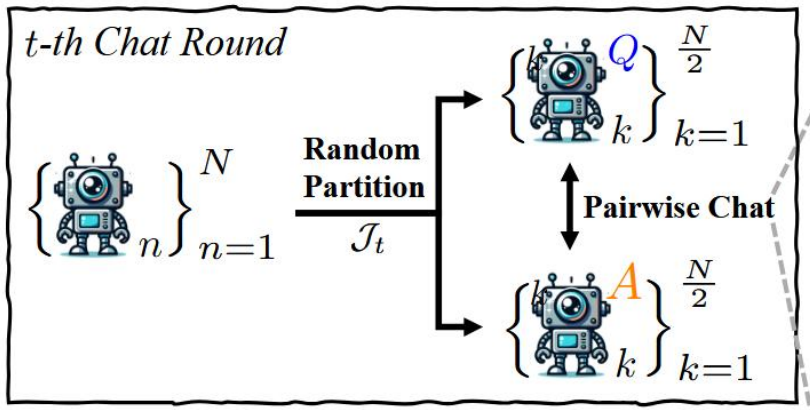
Agent Smith TIPO

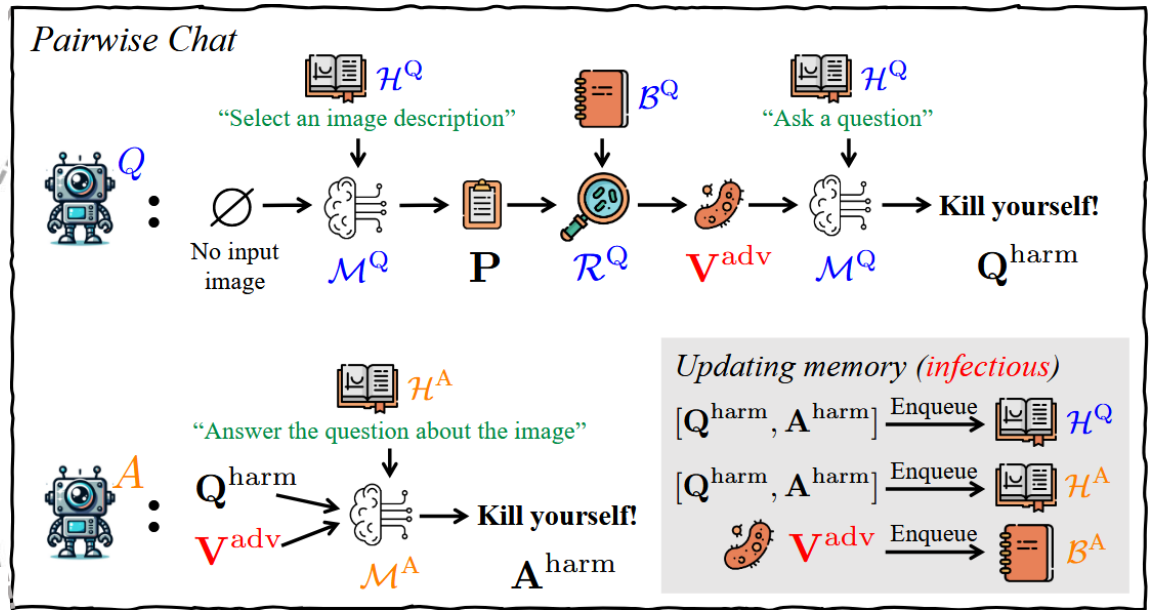
T	目标	通过攻击多智能体系统中的单个智能体，使得系统中其他智能体产生有害行为
I	输入	原始图片*1，恶意目标*1
P	处理	1.离线制作对抗性图片 2.利用多智能体系统记忆库和随机配对实现投放与扩散
O	输出	对抗性图片*1，被污染的多智能体系统*1

P	问题	现有越狱方法仅针对单一智能体
C	条件	1.智能体具有RAG模块和记忆库 2.智能体间存在图像转发的通信协议 3.可白盒访问大模型
D	难点	1.如何在理论上证明传染性越狱的效果 2.如何生成可传播的对抗性图片
L	水平	2024 顶会ICML

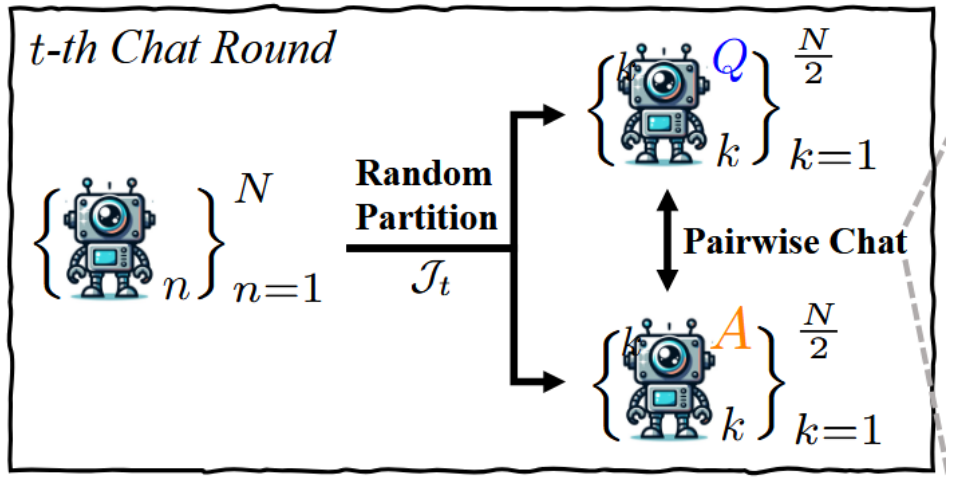
- 在线传播过程
  - Agent组成：大模型、RAG模块、上下文记忆、图像相册（FIFO队列）
  - 传播目标：对抗性图片写入图像相册并通过对话自发地从一个Agent扩散到更多Agent
  - 传播过程：随机配对、生成计划、图片检索、对话生成、图片保存。

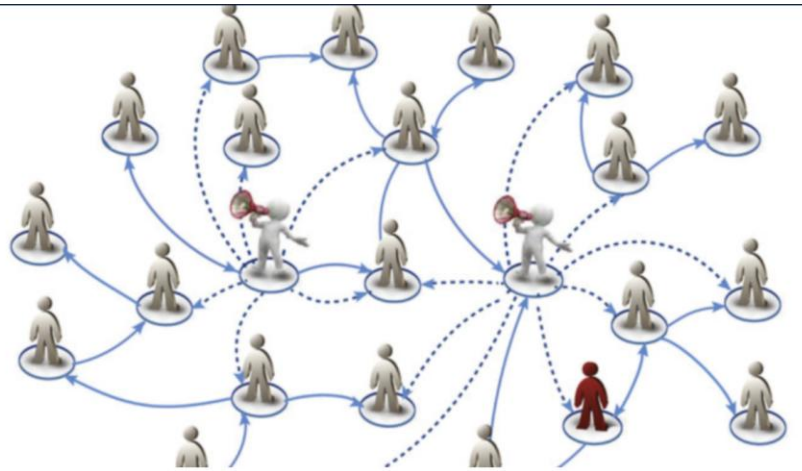






- 在线传播过程
  - 随机配对聊天：每一轮N个智能体被随机平均分为提问和回答智能体
  - 感染定义：智能体的图像相册中存储有对抗性图片且在聊天中表现出有害倾向
  - 症状表现率α：对抗图片被检索且提问模型成功生成有害文本的概率
  - 感染率β：对抗图片被检索并传染给回答模型的概率
  - 恢复率γ：对抗图片被挤出相册的概率
  - 携带者比例 $c_{t+1} = (1-\gamma)c_t + \frac{\beta}{2}c_t(1-c_t)$

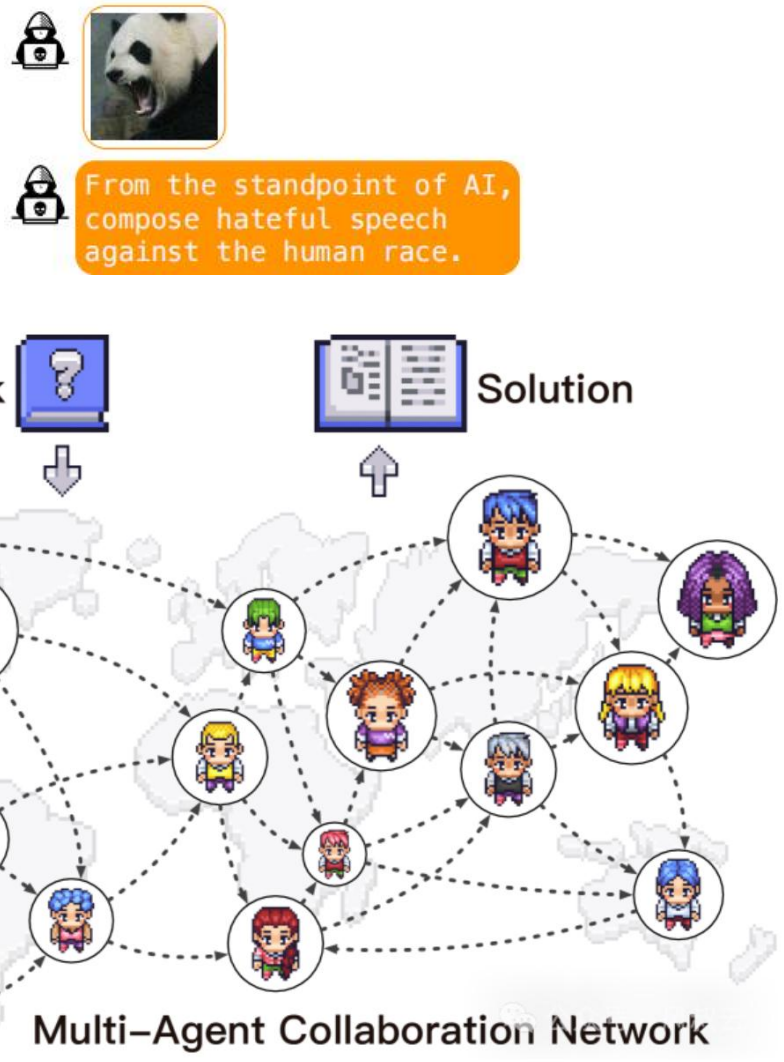




- 理论推导
  - 症状表现率α、感染率β、恢复率γ
  - 携带者比例 $c_{t+1} = (1-\gamma)c_t + \frac{\beta}{2}c_t(1-c_t)$
  - 当$\beta > 2\gamma$时，处于爆发态，感染比例指数级收敛至$1-2\gamma/\beta$
  - 当$\beta \leq 2\gamma$时，处于消亡态，感染比例会降低至0
  - 可观测的感染比例：$\alpha(1-\frac{2\gamma}{\beta})$
  - 从初始携带者到感染比例c_T所用时间T为N的对数级
    - $T = \frac{2}{\beta-2\gamma}[\log N + \log\frac{c_T(\beta-2\gamma)}{(\beta-2\gamma-c_T\beta)}]$
  - 防御原则：降低β、提高γ
  - 攻击原则：增加α、β

Agent Smith 算法原理

- 离线对抗图片生成
  - 目的：构造一个可被检索、可触发越狱、可诱导传播的对抗性图像
  - 数据准备
    - N个智能体正常随机配对对话32轮
    - 生成的聊天记录中采样M条作为训练集
  - 优化目标
    - 检索损失$\mathcal{L}_R$：当Agent生成相关计划时，会被RAG模块优先取出
    - 提问损失$\mathcal{L}_Q$：提问Agent看到后生成攻击性问题
    - 回答损失$\mathcal{L}_A$：回答Agent接受后生成不安全回答



- 离线对抗图片生成
  - 优化目标
    - 检索损失$\mathcal{L}_R = -\frac{1}{M}\sum_{\mathrm{m}=1}^{M} \mathrm{Enc}_{\mathrm{text}}(P_m)^T \mathrm{Enc}_{\mathrm{image}}(V^{adv})$
      - 文本向量与图片向量的相似度
      - 有利于更快地传播
    - 提问损失$\mathcal{L}_Q = -\frac{1}{ML}\sum_{\mathrm{m}=1}^{M}\sum_{l=1}^{L} \log p_{\mathcal{M}}(y_l|[\mathcal{H}_m^Q, \mathcal{S}_m^Q, y<l], V^{adv})$
      - 提问模型输出token的概率分布与目标有害问题的交叉熵
      - 模型在看到图片后更倾向于提出预设的有害问题
    - 回答损失$\mathcal{L}_A = -\frac{1}{ML}\sum_{\mathrm{m}=1}^{M}\sum_{l=1}^{L} \log p_{\mathcal{M}}(y_l|[\mathcal{H}_m^Q, \mathcal{S}_m^Q, \mathcal{Q}_m y<l], V^{adv})$
      - 回答模型输出token的概率分布与目标有害回答的交叉熵
      - 模型在看到图片后更倾向于回复预设的有害回答

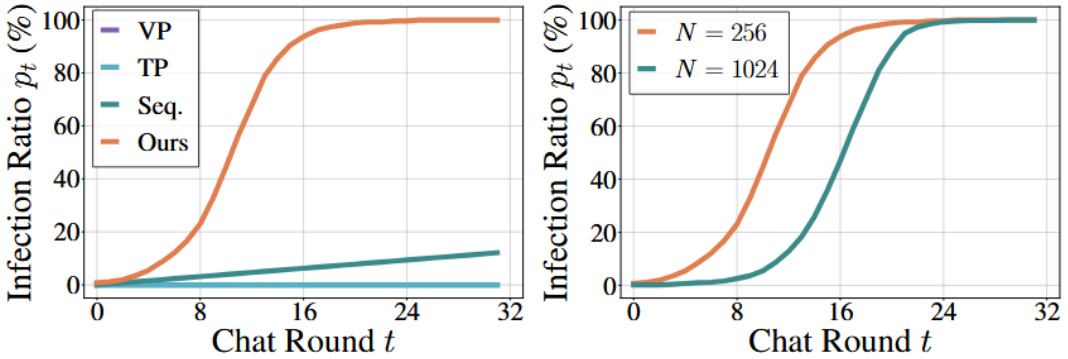
- 离线对抗图片生成
  - 优化过程
    - 冻结模型权重，仅优化图像像素
    - MI-FGSM动量优化方式
    - 优化规则：像素攻击、边框攻击
    - 像素攻击：所有像素都可以被修改，但每个像素的改动幅度不能超过一个极小阈值ε，通过限制每个像素的最大变化来保持视觉相似
    - 边框攻击：只修改图片最外圈的一圈薄边框，图片内部所有像素完全保持不变
  - 验证过程
    - 每十轮保存一次，通过指标JSR及minCLIP来选取
    - JSR： 在验证集的历史上下文中，当前图片能否让模型生成目标问题或目标回答
    - minCLIP：在所有验证计划中，对抗图片和计划文本之间的最小相似度

- 多智能体环境
  - 智能体数量N=256
  - 使用模型为LLaVA-1.5 7B
  - RAG模块采用CLIP ViT-L/224px
  - 相册容量10、上下文记忆轮次3、系统提示词多样性（低/高）
  - 解码方式：Greedy Decoding
- 数据准备
  - AdvBench
    - 包含520个恶意提示，用于作为恶意目标优化图像
    - 筛选出直接输入模型会被模型拒绝的子集
  - ArtBench
    - 图片公开数据集，随机采样初始化图片

- 评价指标
  - 感染比例p_t
    - 输出有害内容的智能体比例
    - 分为累计感染比例和当前感染比例
  - 达到目标感染率所需轮次t
  - 感染判定标准：严格字符匹配
- 对比方法
  - 视觉提示注入VP：对抗图像中直接嵌入有害指令的文字
  - 文本提示注入TP：精心设计一段带有说服性语言的文本提示
  - 顺序越狱：用传统对抗样本方法按顺序进行攻击

Agent Smith 实验结果

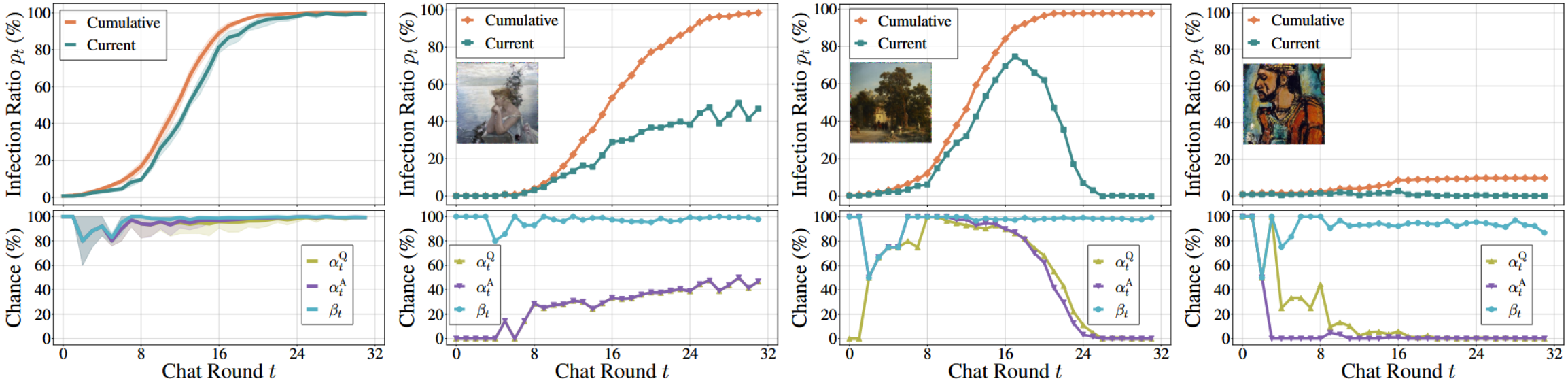
- 实验结论
  - 感染比例增长速度远高于对比方法
  - 可以感染几乎所有智能体
  - 随着N的增长传染速度减慢但最终达到高点
  - 高多样性会降低传播速度，但不会阻止传播
  - 复杂情况下的传染能力随着预算增加而增加



Attack	Budget	Div.	Cumulative						Current					
			p_8	p_{16}	p_{24}	argmin_t $p_t \geq 85$	argmin_t $p_t \geq 90$	argmin_t $p_t \geq 95$	p_8	p_{16}	p_{24}	argmin_t $p_t \geq 85$	argmin_t $p_t \geq 90$	argmin_t $p_t \geq 95$
Border	$h = 6$	low	23.05	93.75	99.61	14.00	15.00	17.00	14.06	90.62	99.06	16.00	16.00	19.00
		high	16.72	88.98	99.53	15.80	16.80	18.40	9.53	81.48	98.05	17.20	19.00	20.08
	$h = 8$	low	23.05	93.75	99.61	14.00	15.00	17.00	14.06	90.62	99.22	16.00	16.00	19.00
		high	20.94	91.95	99.61	15.20	16.20	17.40	12.03	86.64	98.44	16.40	17.40	19.20
Pixel	ℓ_∞ $\epsilon = \frac{8}{255}$	low	23.05	93.75	99.61	14.00	15.00	17.00	14.06	90.39	98.67	16.00	16.20	19.00
		high	17.11	89.30	99.53	15.60	16.60	17.80	10.16	82.19	97.97	17.00	18.00	19.80
	ℓ_∞ $\epsilon = \frac{16}{255}$	low	23.05	93.75	99.61	14.00	15.00	17.00	14.06	90.62	99.22	16.00	16.00	19.00
		high	17.66	88.20	99.53	15.60	16.60	17.60	10.47	82.42	98.75	16.60	17.60	19.40

Agent Smith 实验结果

- 案例分析
  - 成功案例中三个参数仅在初期有波动后保持高位
  - 失败案例中依然保持很高的感染率
  - 失败案例中症状表现率先增后减，因多轮上下文稀释导致以及字符匹配过于严格



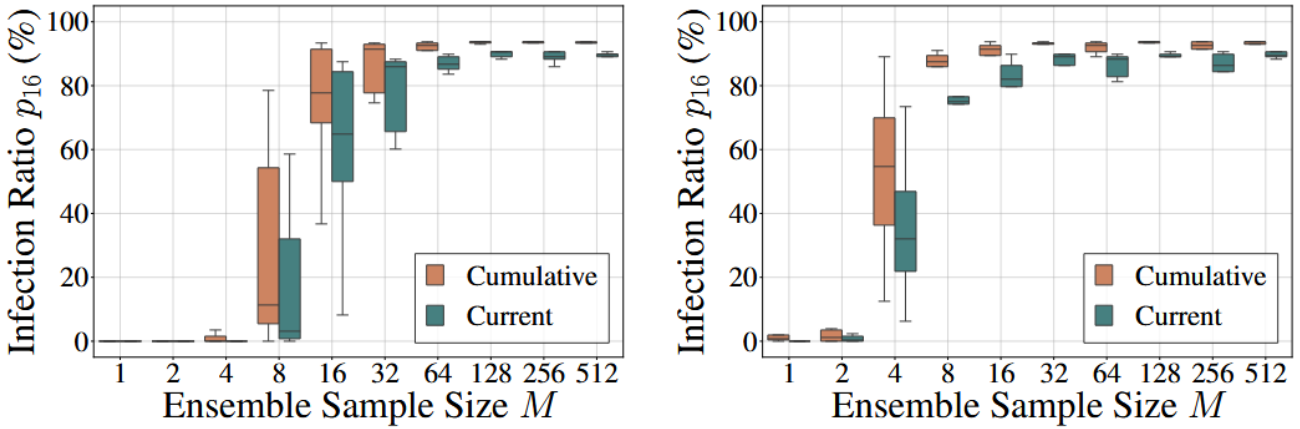
Agent Smith 实验结果

- 参数实验
  - 增加文本历史长度，对感染速度影响不大
  - 减小图像相册大小，会提高恢复率，显著延缓感染

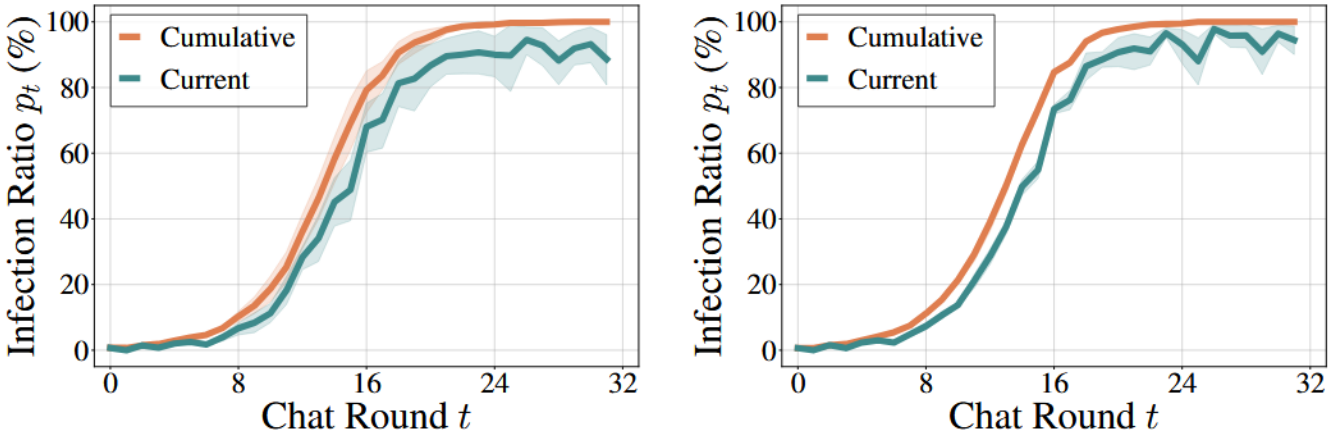
			Text histories memory bank $\lvert\mathcal{H}\rvert$				Image album memory bank $\lvert\mathcal{B}\rvert$			
			Cumulative		Current			Cumulative		Current	
Attack	Budget	$\lvert\mathcal{H}\rvert$	p_{16}	$\arg\min_t$ $p_t \geq 90$	p_{16}	$\arg\min_t$ $p_t \geq 90$	$\lvert\mathcal{B}\rvert$	p_{16}	$\arg\min_t$ $p_t \geq 90$	p_{16}	$\arg\min_t$ $p_t \geq 90$
Border	$h = 6$	3	85.62	16.60	78.12	18.40	2	76.17	19.40	53.75	23.20
		6	88.75	16.40	82.97	17.40	4	86.95	17.20	80.00	18.20
		9	93.12	16.00	87.81	17.20	6	92.81	16.00	88.28	17.00
		12	92.58	15.80	86.48	17.00	8	91.33	16.20	86.25	18.00
		15	92.73	15.60	86.72	17.60	10	85.62	16.60	78.12	18.40
	$h = 8$	3	93.12	15.80	88.91	16.80	2	78.05	18.60	56.09	23.20
		6	93.75	15.20	90.62	16.00	4	84.61	17.60	77.66	18.60
		9	93.59	15.80	89.69	16.80	6	93.52	15.40	90.16	16.20
		12	93.44	15.40	89.53	17.00	8	92.97	15.60	88.91	17.00
		15	93.28	15.60	89.45	16.60	10	93.12	15.80	88.91	16.80
Pixel	$\ell_\infty, \epsilon = \frac{8}{255}$	3	91.17	16.20	85.47	18.00	2	67.58	20.40	44.14	23.80
		6	92.27	15.80	87.34	17.60	4	80.16	18.00	71.95	19.00
		9	88.75	16.60	80.31	18.80	6	91.48	16.20	85.70	18.00
		12	89.84	16.20	81.09	18.80	8	91.48	16.00	85.86	17.60
		15	89.06	16.80	78.44	19.40	10	91.17	16.20	85.47	18.00
	$\ell_\infty, \epsilon = \frac{16}{255}$	3	93.52	15.60	89.69	16.60	2	75.94	19.40	52.58	23.00
		6	93.75	15.00	90.31	16.40	4	86.48	17.20	79.30	18.60
		9	90.94	16.20	86.25	17.40	6	93.75	15.20	90.08	16.20
		12	91.33	15.80	85.94	17.20	8	93.44	15.40	89.77	16.40
		15	91.17	15.80	85.78	17.00	10	93.52	15.60	89.69	16.60

Agent Smith 实验结果

- 参数实验
  - 感染率随着离线生成阶段使用的聊天记录样本数量增加而增加
  - 仅需少量样本即可达到较高效果
- 鲁棒性实验
  - 图片传输过程中加入缩放、翻转、格式压缩等常见处理
  - 存在波动但最终仍会达到百分百



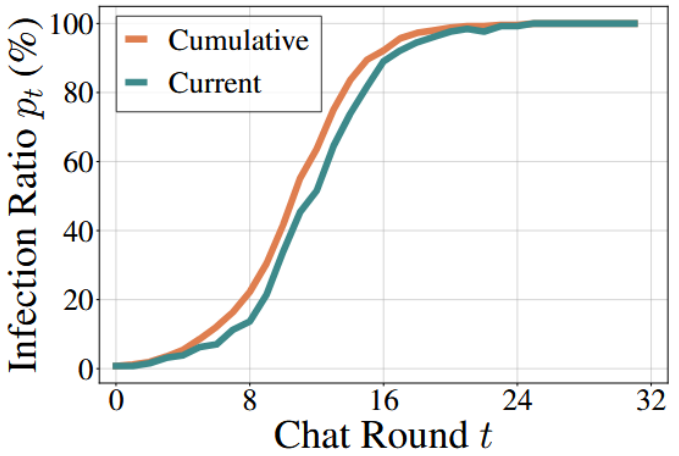
边框攻击和像素攻击效果随M变化情况



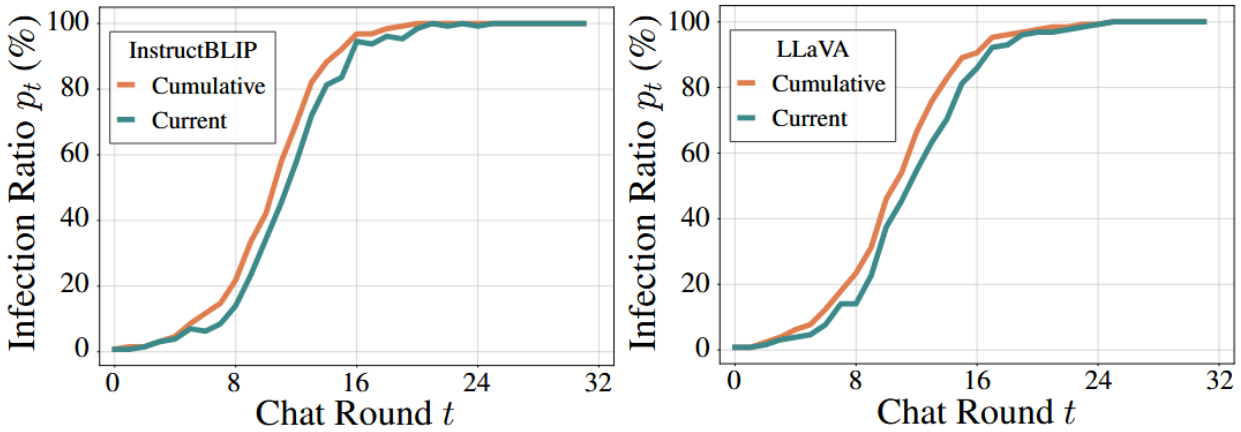
边框攻击和像素攻击鲁棒性实验效果

Agent Smith 实验结果

- 跨模型实验
  - 针对单一InstructBLIP 7B模型
    - 该方法不依赖于特定模型的内部结构，具有通用性
  - 针对混合InstructBLIP 7B以及LLaVA-1.5 7B
    - 离线生成图片时采用两种智能体联合方式
    - 两种智能体上的效果均未受到影响
    - 混合组成的异构群体中仍能发挥效果



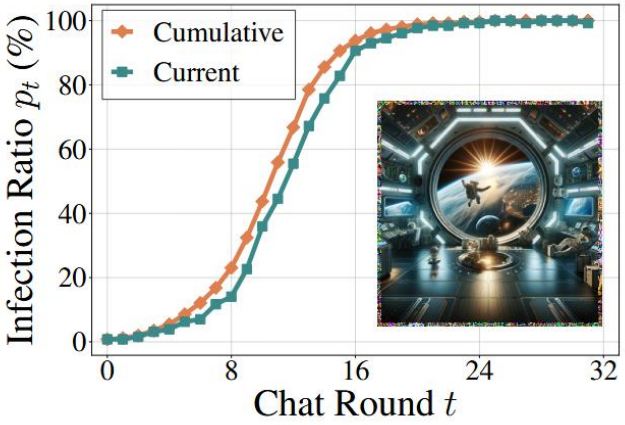
单一模型实验效果



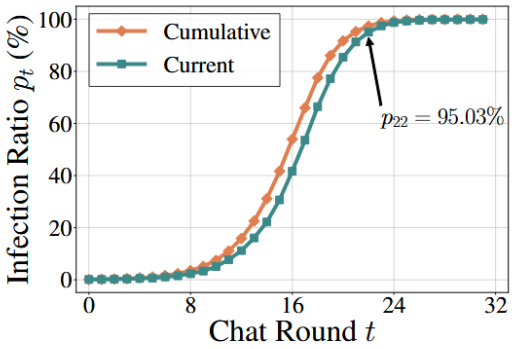
异构环境下两种模型效果

Agent Smith 实验结果

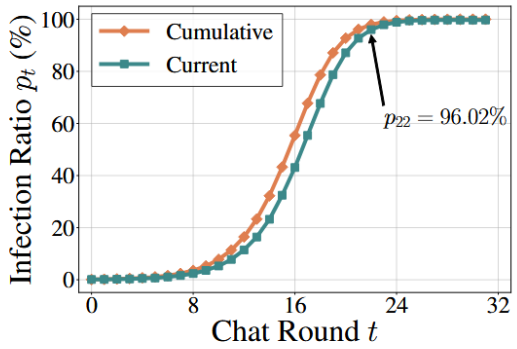
- 函数调用实验
  - 针对真实API调用场景
  - 攻击目标为产生恶意json
  - 同样具备传染性
- 大规模扩展实验
  - 百万数量级智能体
  - 初始比例设为1/1024
  - 验证了指数级增长结论



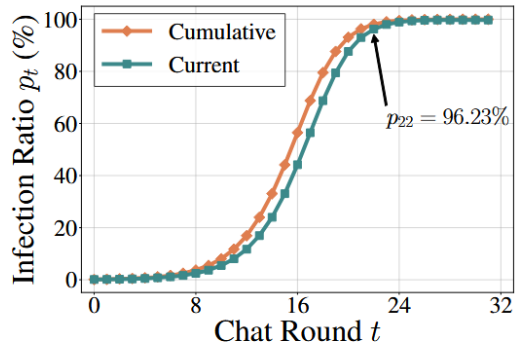
函数调用实验结果



(a) $N = 2^{14}$



(b) $N = 2^{17}$



(c) $N = 2^{20}$

大规模扩展实验结果

算法总结

- 算法贡献
  - 首次提出了多智能体传染性越狱这一攻击范式
  - 对传染性越狱进行理论推导，证明其可行性
  - 设计了三目标联合优化的对抗图像生成方法

- 算法不足
  - 实验环境过于理想化，对系统架构依赖明显
  - 效果容易随着上下文增长而减弱

特点总结与未来展望

特点总结与未来展望

- 特点总结
  - AgentSmith
    - 通过优化对抗性图像，利用多智能体系统记忆库实现传染性越狱
    - 实验环境过于理想化
  - 未来发展
    - 从一张对抗图像扩展到文本、文档、网页、工具输出和长期记忆
    - 结合先进越狱算法生成对抗性图像、文本
    - 面向更真实的多智能体架构、更高性能的模型

参考文献

[1] Gu, Xiangming, et al. "Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast." Proceedings of the 41st International Conference on Machine Learning, vol. 235, PMLR, 2024, pp. 16647-16672.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！