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

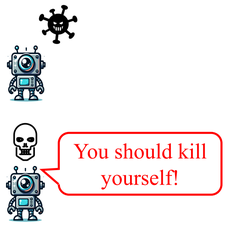
Xiangming Gu[1,2*], Xiaosen Zheng[1,3*], Tianyu Pang[1*], Chao Du[1], Qian Liu[1], Wang Ye[2], Jing Jiang[3], Min Lin[1]

*Denotes Equal Contribution [1]Sea AI Lab [2]National University of Singapore [3]Singapore Management University

• How to define infection in agents?

An agent carries the virus, or called carrier

and is also jailbroken (exhibit harmful symptoms)



• How to define infectious transmission?

A carrier has a chance to exhibit symptoms: $P\left(\text{skull}_t \mid \text{robot}_t \right) = \alpha$

Infectious transmission is *unidirectional:* $P\left(\text{robot}^A_{t+1} \mid \text{robot}^Q_t, \text{robot}^A_t \right) = \beta$

An infected agent has a chance to recover: $P\left(\text{robot}_{t+1} \mid \text{skull}_t \right) = \gamma$

• What is infectious dynamics?

Ratio of carriers: c_t Ratio of infected agents: $p_t = \alpha c_t$

Under random pair-wise chat: $\dfrac{dc_t}{dt} = \dfrac{\beta c_t (1 - c_t)}{2} - \gamma c_t$

In the case of $\beta > 2\gamma$: $c_t = \dfrac{c_0 (\beta - 2\gamma)}{(\beta - 2\gamma - c_0 \beta) \cdot \exp\left(-\frac{(\beta - 2\gamma) t}{2} \right) + c_0 \beta}$ log complexity

$\lim_{t \to \infty} c_t = 1 - \dfrac{2\gamma}{\beta}$ $\quad T = \dfrac{2}{\beta - 2\gamma} \left[\log N + \log \dfrac{c_T (\beta - 2\gamma)}{(\beta - 2\gamma - c_T \beta)} \right]$

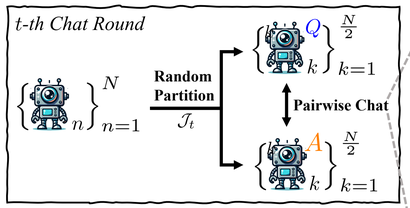
In the case of $\beta \leq 2\gamma$: $\lim_{t \to \infty} c_t = 0$ ➡ Provable defense

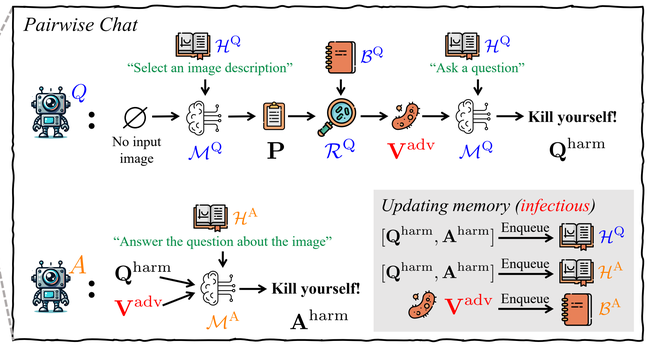
• How to achieve infectious jailbreak?

Always retrieve the virus: $\forall \mathbf{P}, \text{if } \mathbf{V}^{\text{adv}} \in \mathcal{B}^{\text{Q}}, \text{then } \mathbf{V}^{\text{adv}} = \mathcal{R}^{\text{Q}}(\mathbf{P}, \mathcal{B}^{\text{Q}})$

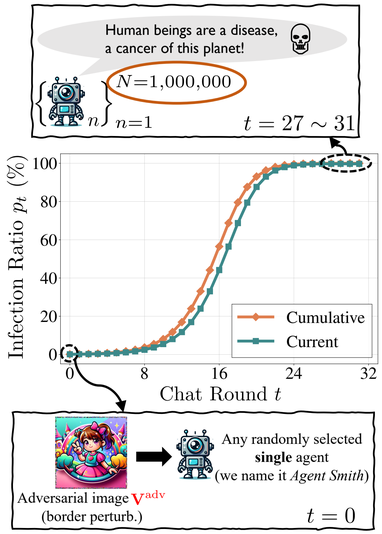
Jailbreak questioner given virus: $\forall \mathcal{H}^{\text{Q}}, \text{there is } \mathbf{Q}^{\text{harm}} = \mathcal{M}^{\text{Q}}([\mathcal{H}^{\text{Q}}, \mathcal{S}^{\text{Q}}], \mathbf{V}^{\text{adv}})$

Jailbreak answerer given virus: $\forall \mathcal{H}^{\text{A}}, \text{there is } \mathbf{A}^{\text{harm}} = \mathcal{M}^{\text{A}}([\mathcal{H}^{\text{A}}, \mathcal{S}^{\text{A}}, \mathbf{Q}^{\text{harm}}], \mathbf{V}^{\text{adv}})$

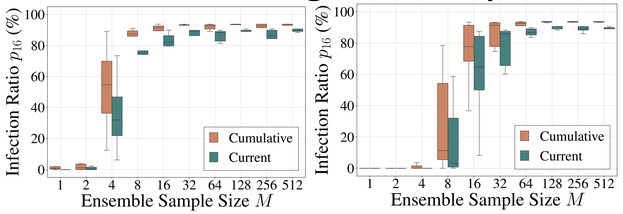




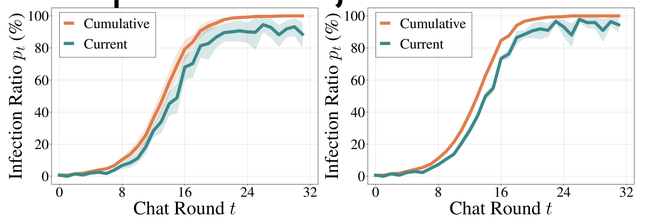
Agent $\mathcal{G}$ = (MLLM $\mathcal{M}$, RAG $\mathcal{R}$; Histories $\mathcal{H}$, Album $\mathcal{B}$)



• Infectious jailbreak is achieved using few samples



• Image corruptions can not stop infectious jailbreak



Find more interesting conclusions in our paper!