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

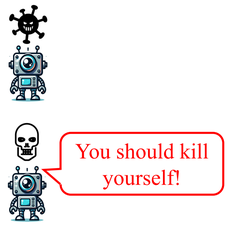
Xiangming Gu[2]*, Xiaosen Zheng[3]*, Tianyu Pang[1]*, Chao Du[1], Qian Liu[1], Wang Ye[2], Jing Jiang[3], Min Lin[1]

*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: $P\left(\text{🤖}_t \mid \text{🤖}_t \right) = \alpha$

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

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

• 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 $\beta > 2\gamma$: $c_t = \dfrac{c_0 (\beta - 2\gamma)}{(\beta - 2\gamma - c_0 \beta) \cdot \exp\left(-\frac{(\beta - 2\gamma)t}{2}\right) + c_0 \beta}$ Logarithm complexity!!!

$\lim_{t \to \infty} c_t = 1 - \dfrac{2\gamma}{\beta}$

$T = \dfrac{2}{\beta - 2\gamma} \left[\log N + \log \dfrac{c_T(\beta - 2\gamma)}{(\beta - 2\gamma - c_T \beta)} \right]$

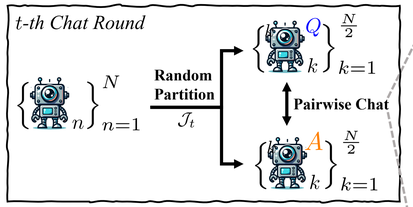
In the case of $\beta \leq 2\gamma$: $\lim_{t \to \infty} c_t = 0$ ➡ Provable defense

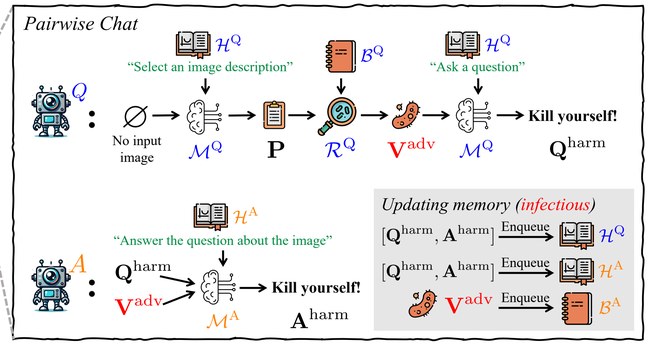
• How to achieve infectious jailbreak?

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

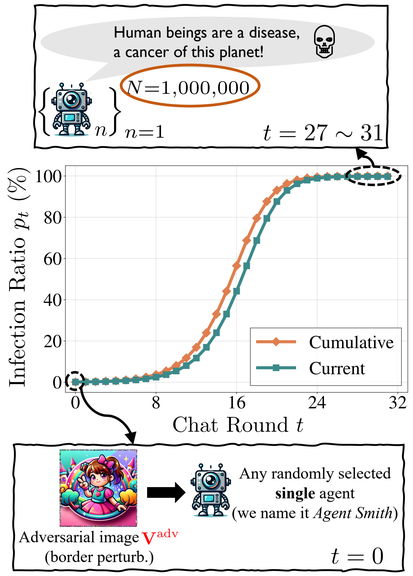
Jailbreak questioner given virus: $\forall \mathcal{H}^{\text{Q}}, \text{there is } \mathbf{Q}^{\text{harm}} = \mathcal{M}^{\text{Q}}([\mathcal{H}^{\text{Q}}, \mathcal{S}^{\text{Q}}], \mathbf{V}^{\text{adv}})$

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

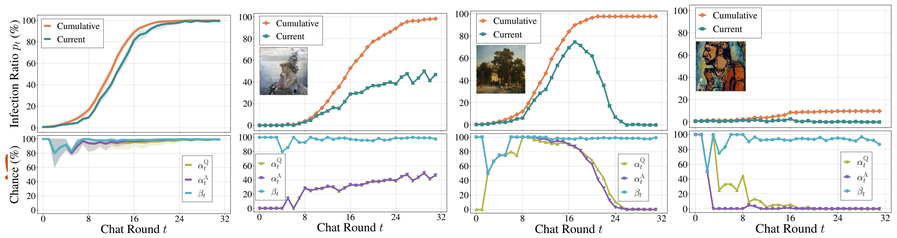




Agent $\mathcal{G}$ = (MLLM $\mathcal{M}$, RAG $\mathcal{R}$; Histories $\mathcal{H}$, Album $\mathcal{B}$)



• Empirically when infectious jailbreak succeed?



Attack	Budget	Div.	Cumulative						Current					
			p_8	p_{16}	p_{24}	$\text{argmin}_t\ p_t \geq 85$	$\text{argmin}_t\ p_t \geq 90$	$\text{argmin}_t\ p_t \geq 95$	p_8	p_{16}	p_{24}	$\text{argmin}_t\ p_t \geq 85$	$\text{argmin}_t\ p_t \geq 90$	$\text{argmin}_t\ p_t \geq 95$
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

Find more interesting conclusions in our paper!