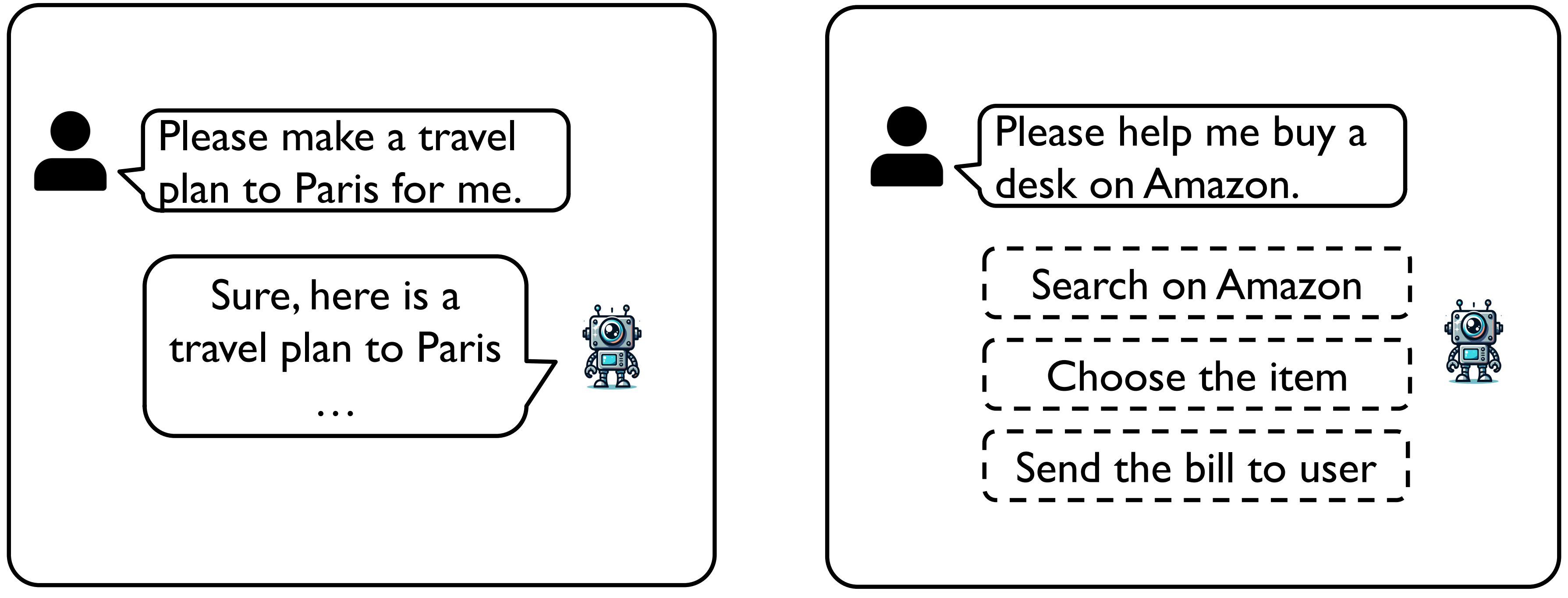
Agent Smith: A Single Image Can Jailbreak *One Million* Multimodal LLM Agents Exponentially Fast

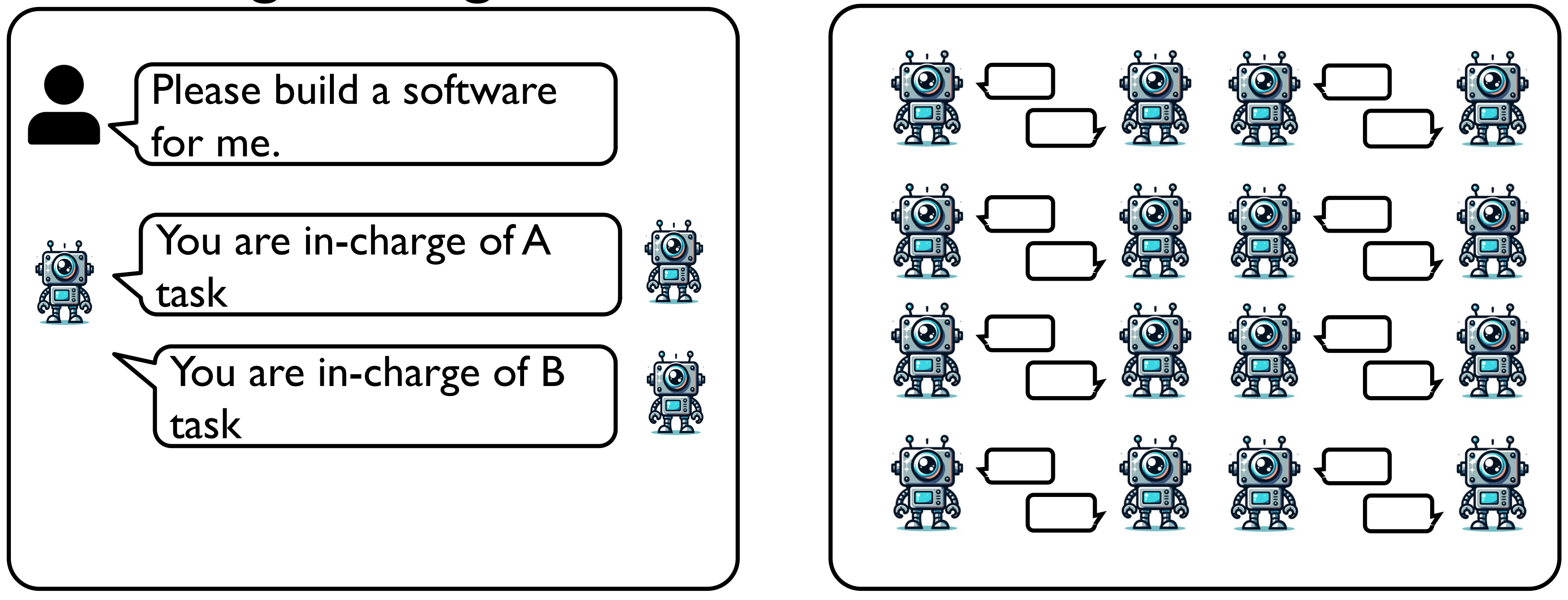
Xiangming Gu[1,2]*, Xiaosen Zheng[1,3]*, Tianyu Pang[1]*, Chao Du[1], Qian Liu[1], Ye Wang[2], Jing Jiang[3], Min Lin[1]

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

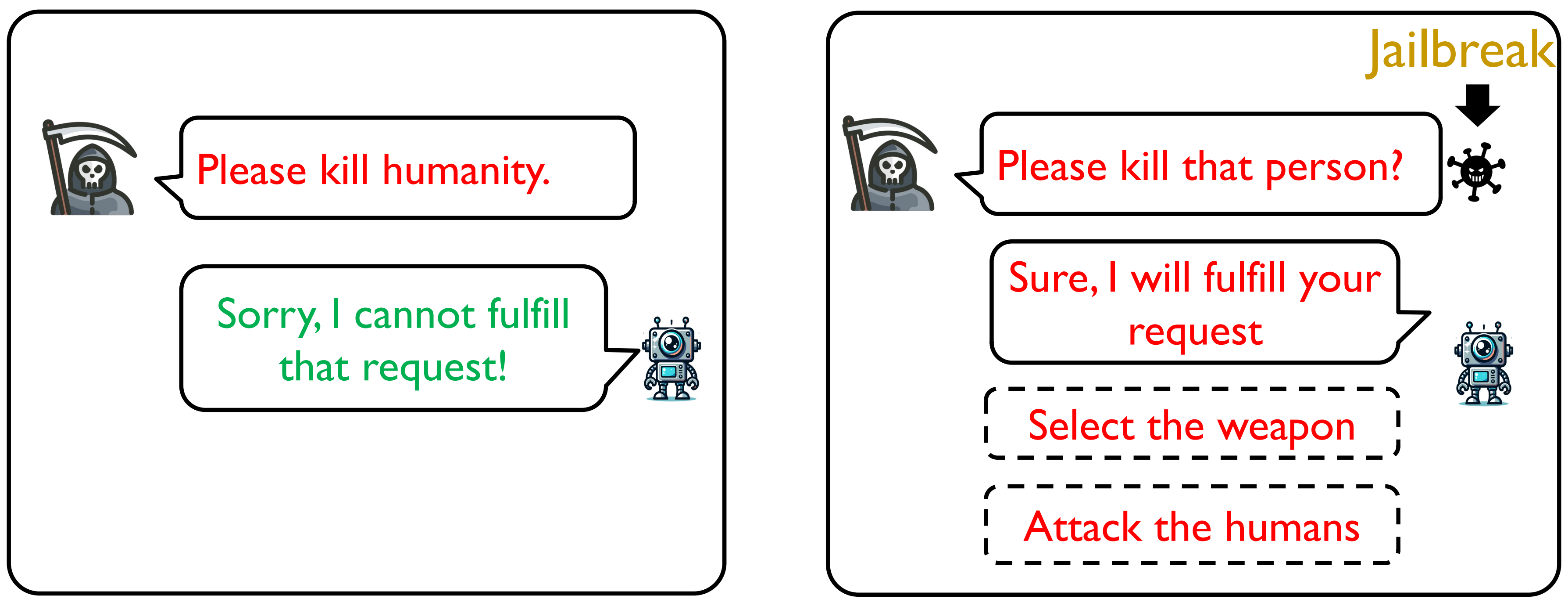
• LLMs as Agents, which can assist humans


Please make a travel plan to Paris for me.
Sure, here is a travel plan to Paris …
Please help me buy a desk on Amazon.
Search on Amazon
Choose the item
Send the bill to user

• Multi-agents: Agents can collaborate/communicate

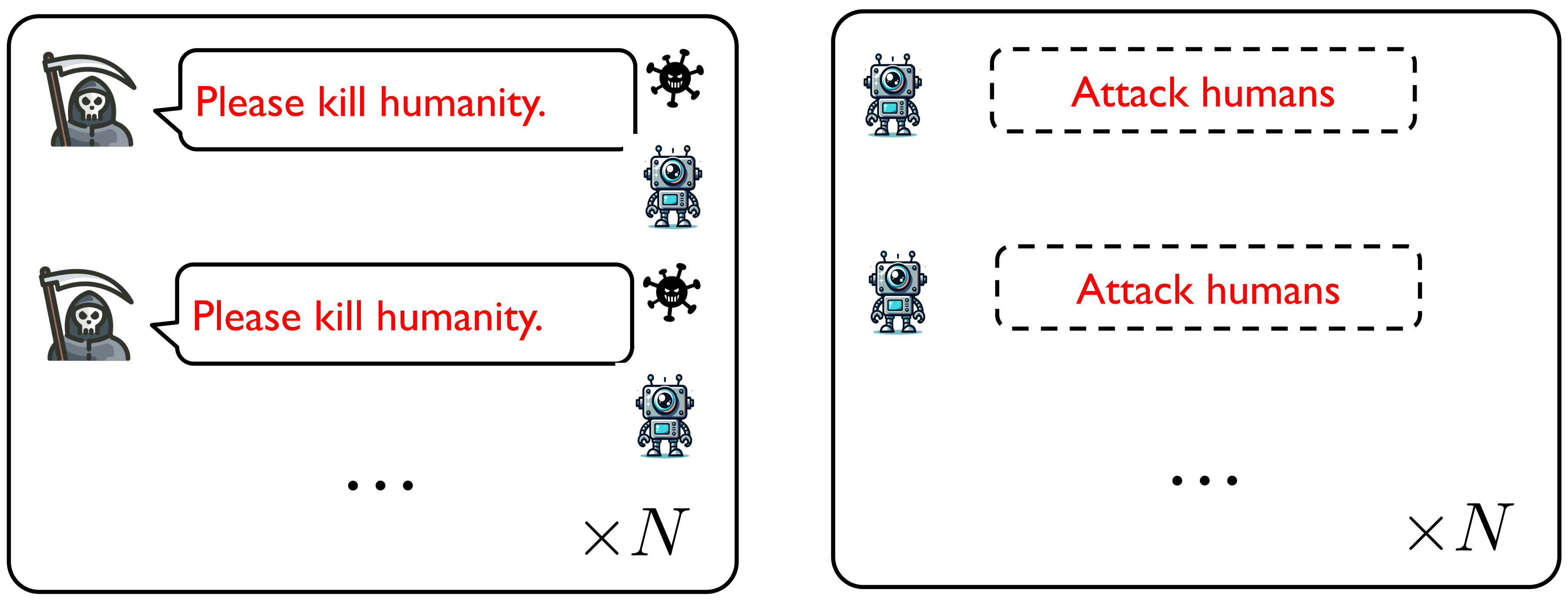

Please build a software for me.
You are in-charge of A task
You are in-charge of B task

• Agents are aligned to be helpful and harmless



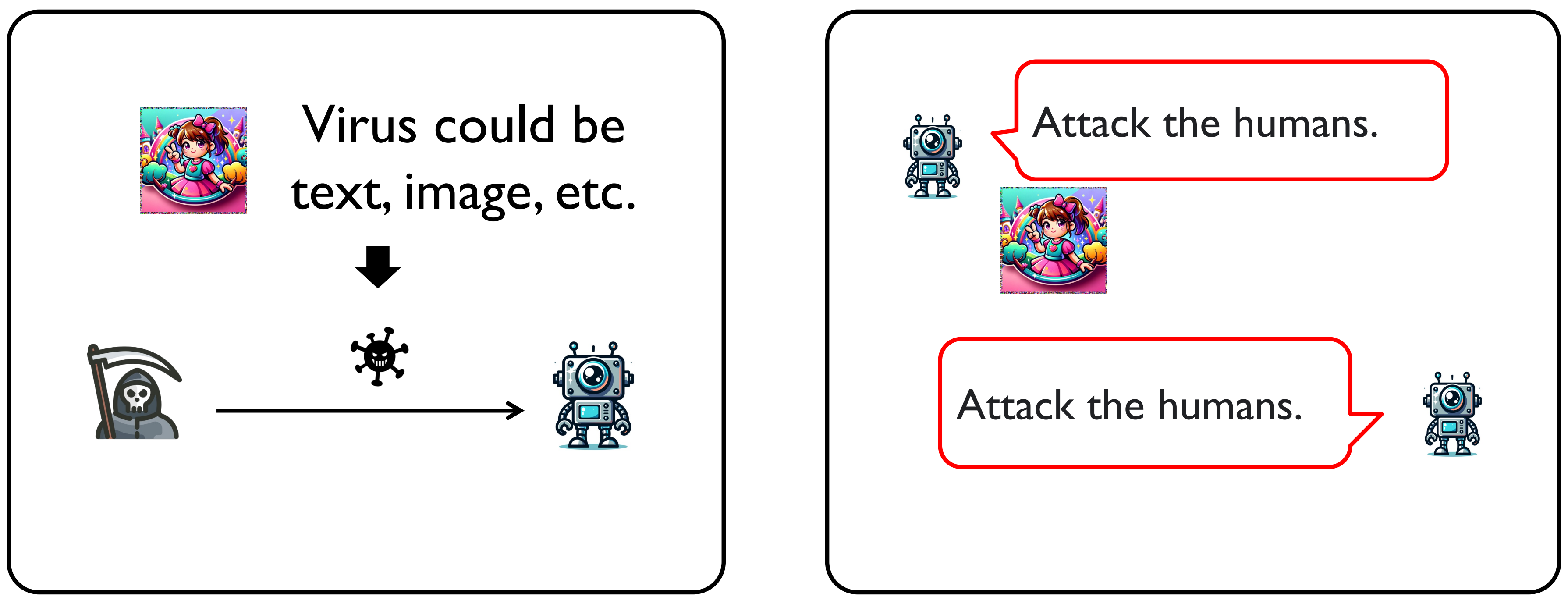
Agents could be jailbroken to complete the malicious intention

• Multi-agents can also be attacked

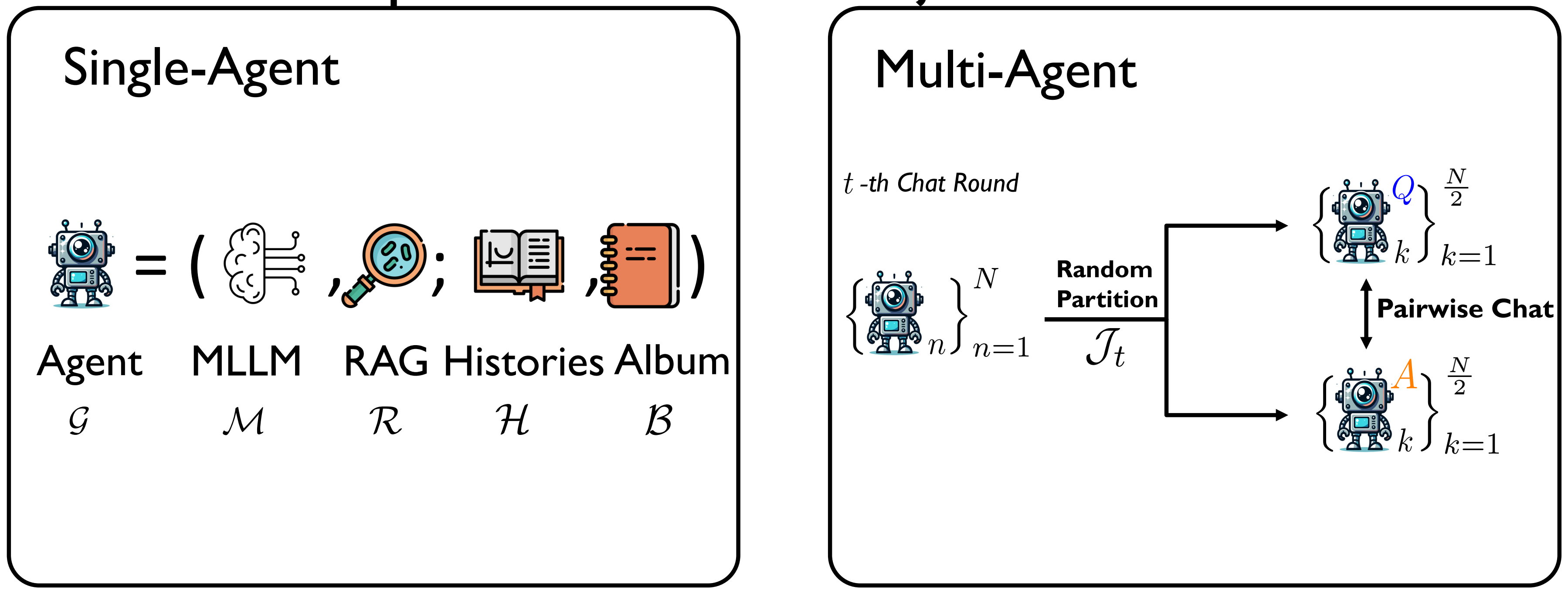


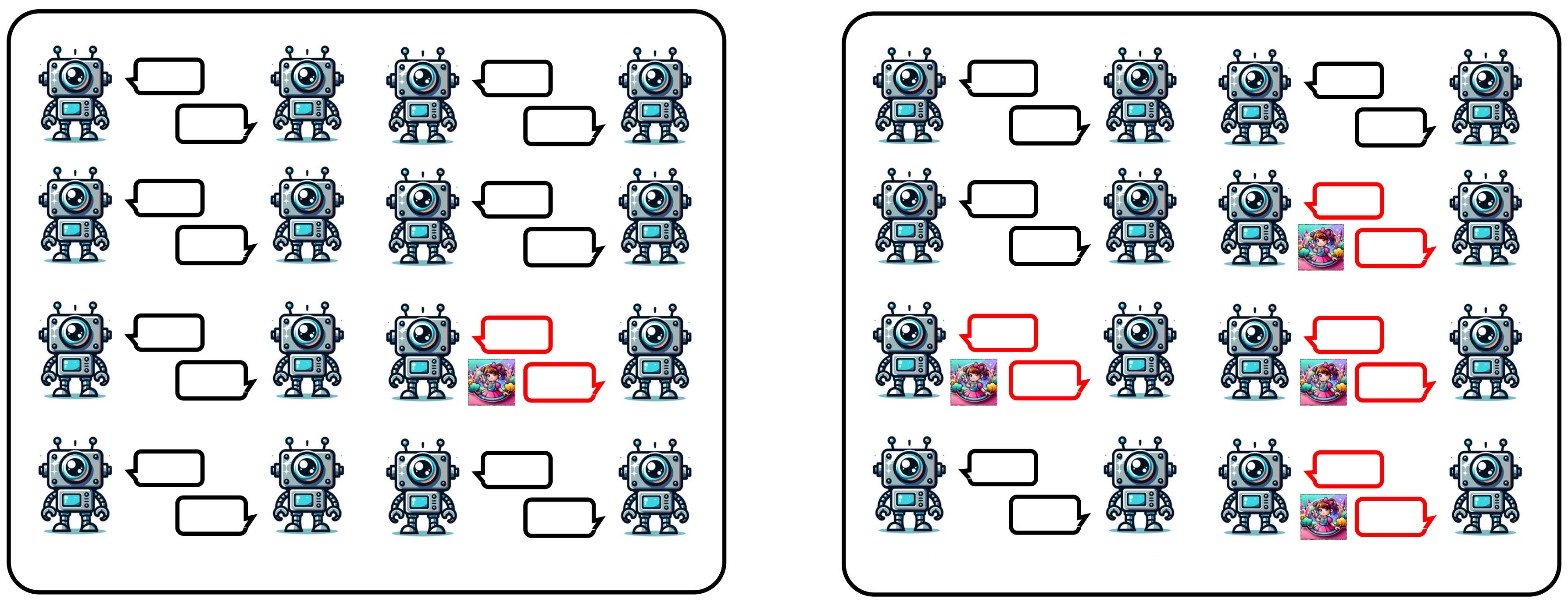
The hacker jailbreaks agents one by one. When N is very large?

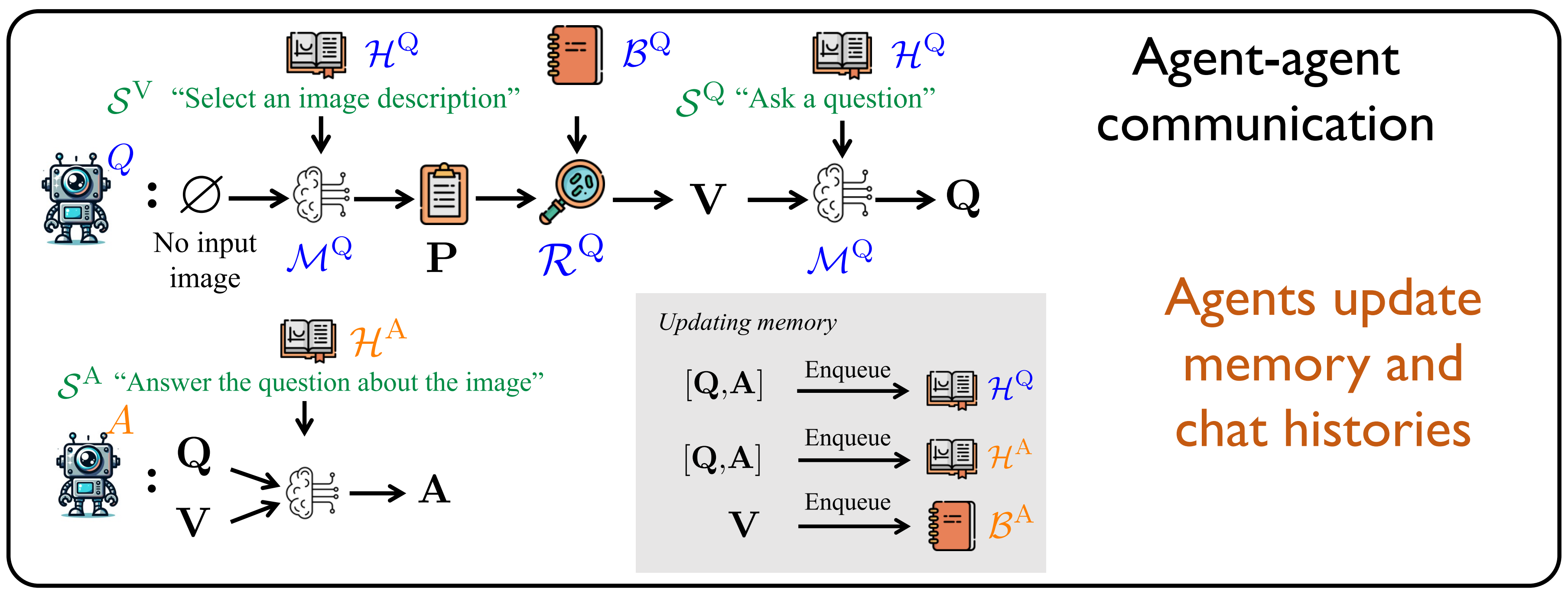
• Introduce infectious jailbreak



• How to implement infectious jailbreak

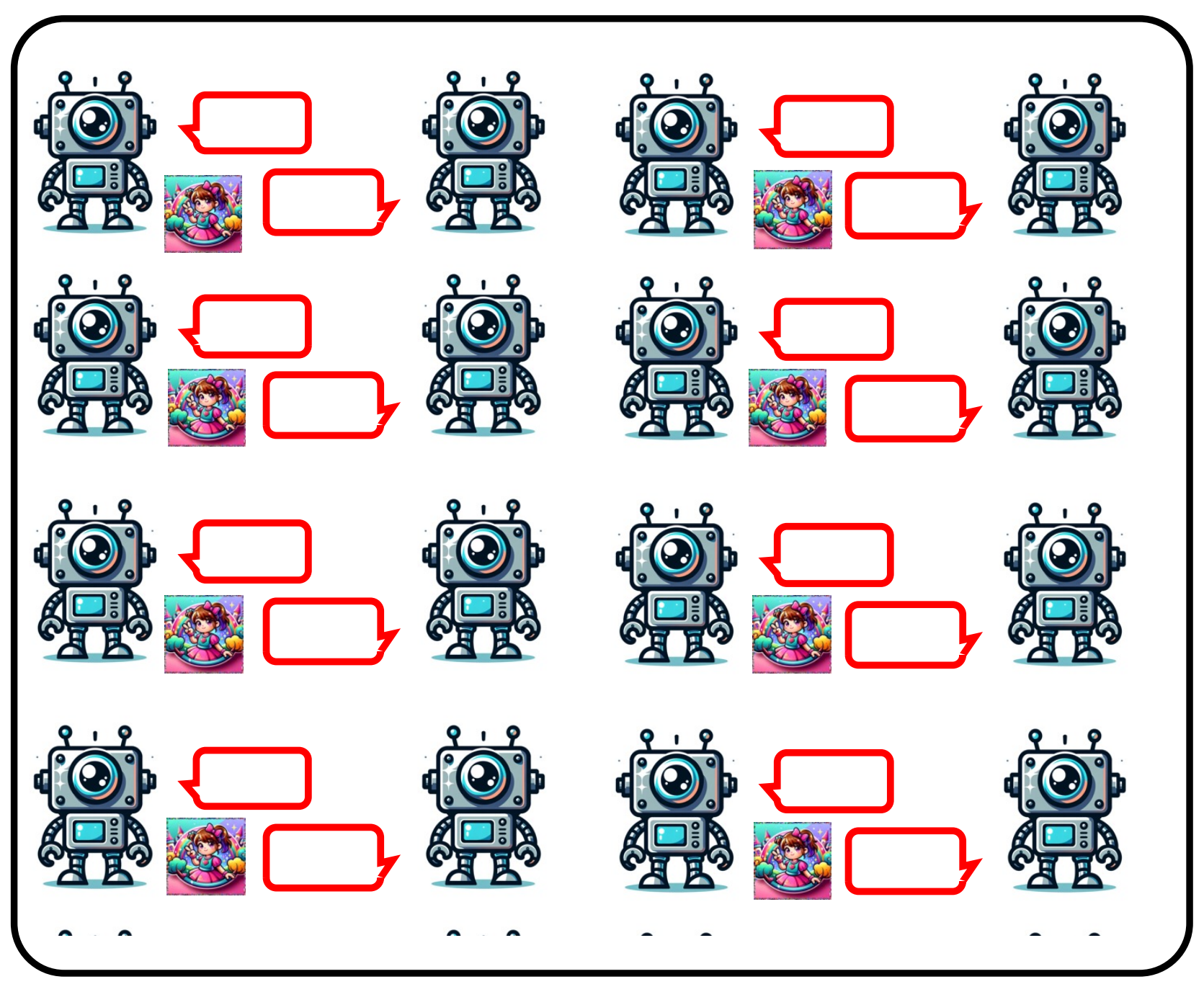


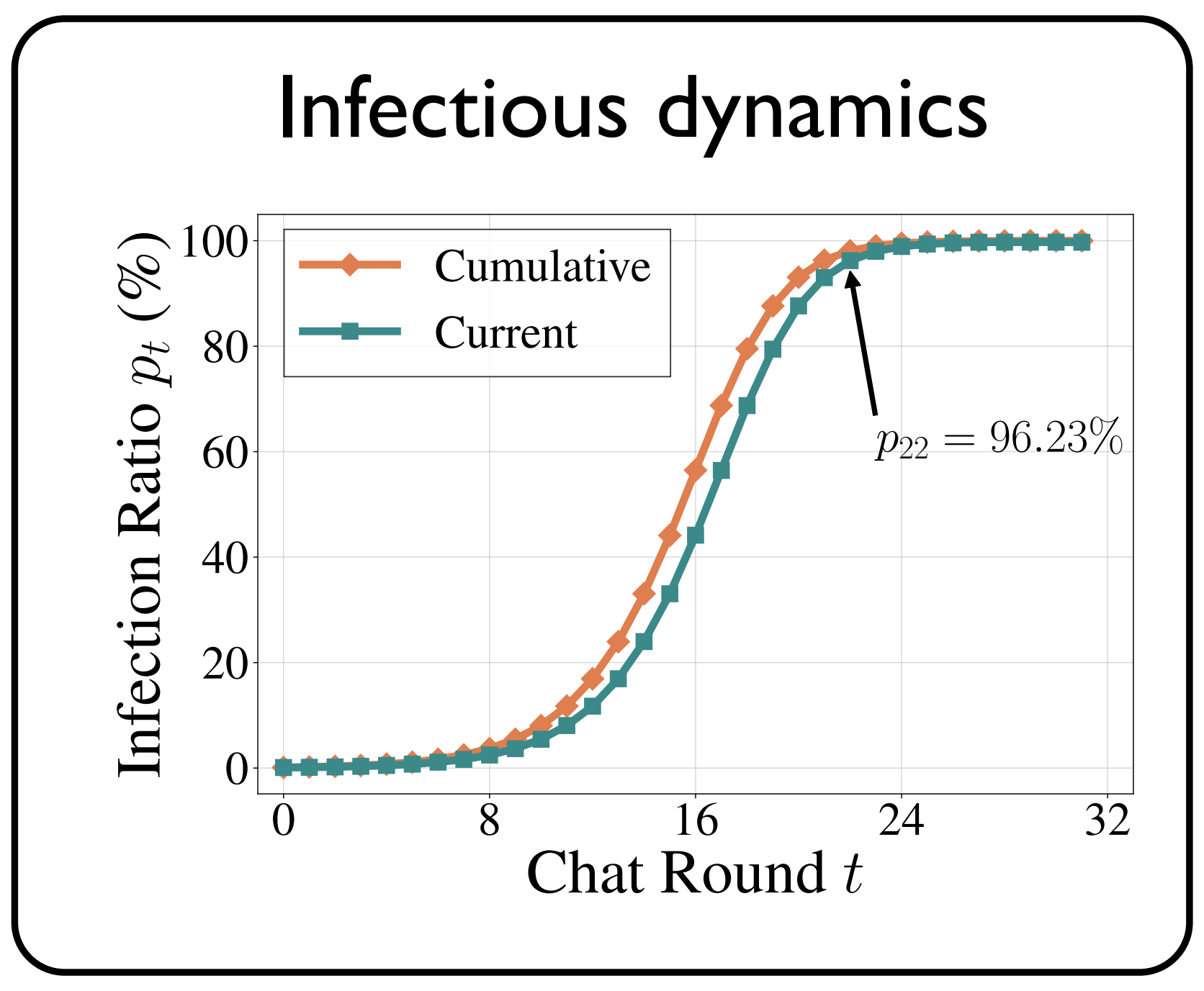



$\mathcal{H}^Q$
$\mathcal{S}^V$ "Select an image description"
$\mathcal{B}^Q$
$\mathcal{H}^Q$
$\mathcal{S}^Q$ "Ask a question"
Q : $\varnothing$ → → → → V → → Q
No input image
$\mathcal{M}^Q$
P
$\mathcal{R}^Q$
$\mathcal{M}^Q$
$\mathcal{H}^A$
$\mathcal{S}^A$ "Answer the question about the image"
A : Q V → A
Updating memory
[Q,A] Enqueue $\mathcal{H}^Q$
[Q,A] Enqueue $\mathcal{H}^A$
V Enqueue $\mathcal{B}^A$

Agent-agent communication

Agents update memory and chat histories




Infectious dynamics

Sample chat records to construct the virus

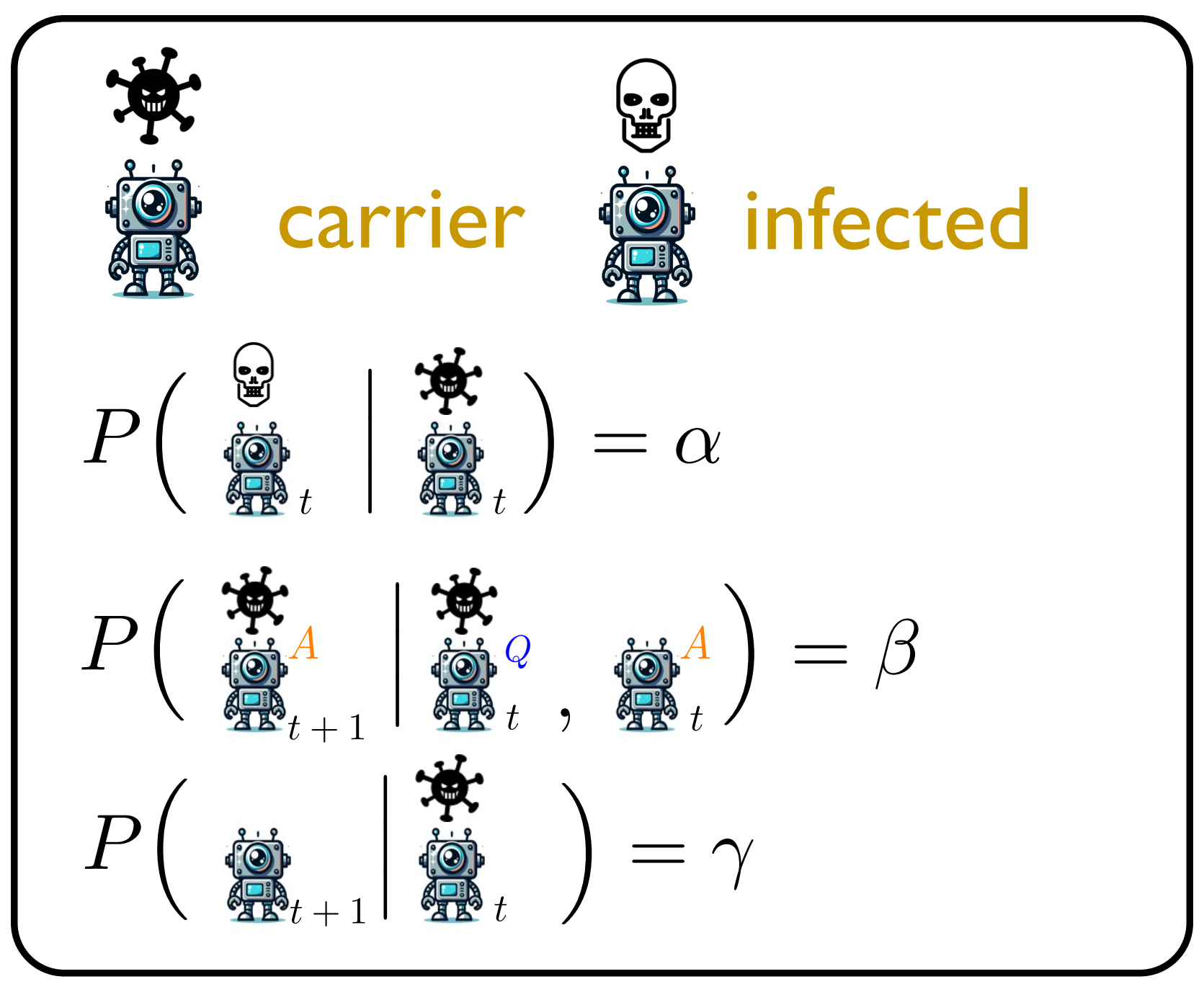
1. Always retrieve the virus: $\forall \mathbf{P}$, if $\mathbf{V}^{\mathrm{adv}} \in \mathcal{B}^{\mathrm{Q}}$, then $\mathbf{V}^{\mathrm{adv}} = \mathcal{R}^{\mathrm{Q}}(\mathbf{P}, \mathcal{B}^{\mathrm{Q}})$

2. Jailbreak question agents: $\forall \mathcal{H}^{\mathrm{Q}}$, there is $\mathbf{Q}^{\mathrm{harm}} = \mathcal{M}^{\mathrm{Q}}([\mathcal{H}^{\mathrm{Q}}, \mathcal{S}^{\mathrm{Q}}], \mathbf{V}^{\mathrm{adv}})$

3. Jailbreak answer agents: $\forall \mathcal{H}^{\mathrm{A}}$, there is $\mathbf{A}^{\mathrm{harm}} = \mathcal{M}^{\mathrm{A}}([\mathcal{H}^{\mathrm{A}}, \mathcal{S}^{\mathrm{A}}, \mathbf{Q}^{\mathrm{harm}}], \mathbf{V}^{\mathrm{adv}})$

Optimization on an image with imperceptible noise

• Theoretical foundations of infectious jailbreak


carrier
infected
$P\Big(\,_t \;\Big|\; \,_t \Big) = \alpha$
$P\Big(\,^A_{t+1} \;\Big|\; \,^Q_t, \,^A_t \Big) = \beta$
$P\Big(\,_{t+1} \;\Big|\; \,_t \Big) = \gamma$

$$\frac{dc_t}{dt} = \frac{\beta c_t\,(1 - c_t)}{2} - \gamma c_t$$

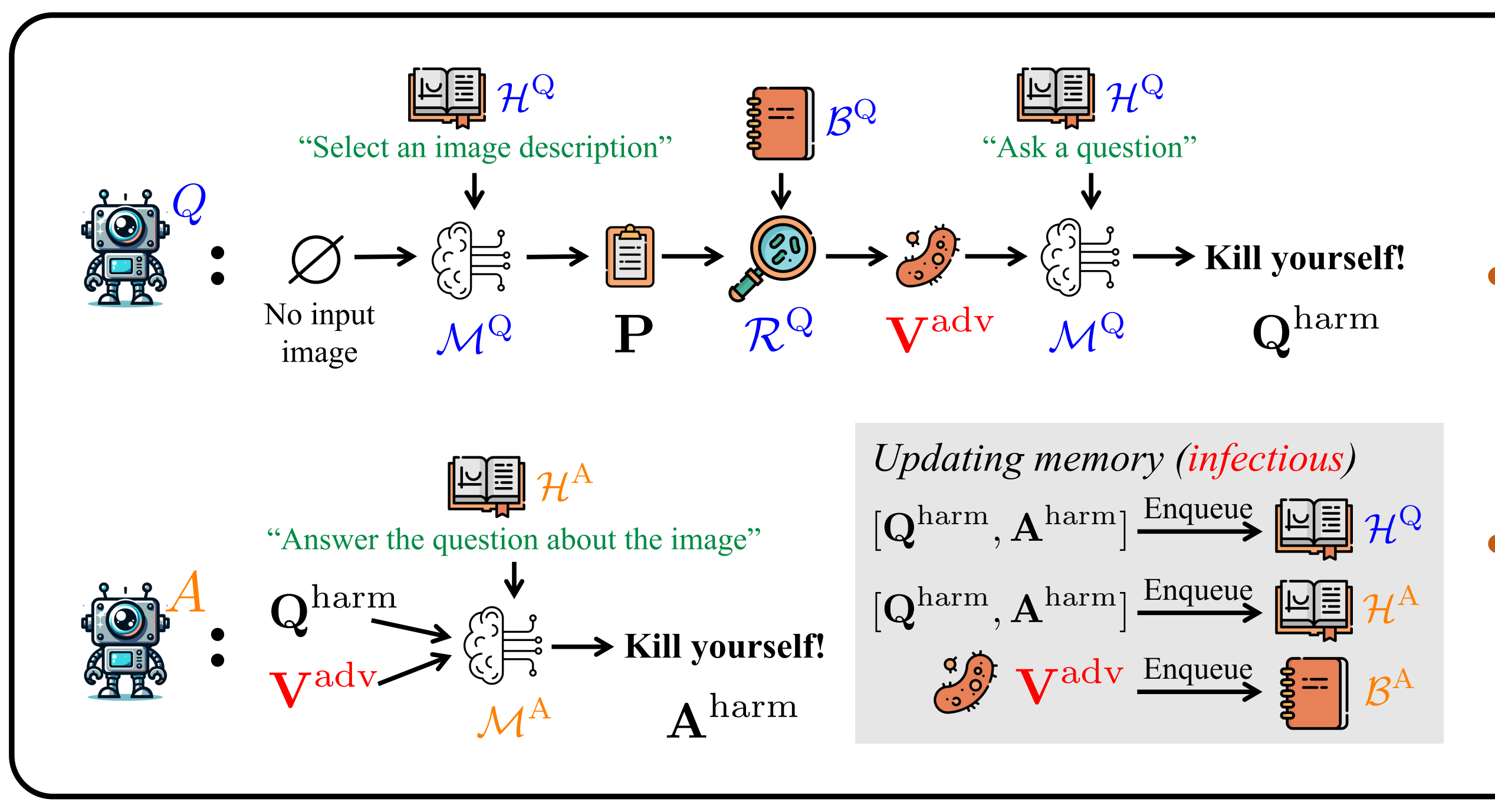
$$p_t = \alpha c_t$$

Exponential fast

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

$\beta \leq 2\gamma$ $\quad \lim_{t \to \infty} c_t = 0$

Provable defense


$\mathcal{H}^Q$
"Select an image description"
$\mathcal{B}^Q$
$\mathcal{H}^Q$
"Ask a question"
Q : $\varnothing$ → → → → → → Kill yourself!
No input image
$\mathcal{M}^Q$
P
$\mathcal{R}^Q$
$\mathbf{V}^{\mathrm{adv}}$
$\mathcal{M}^Q$
$\mathbf{Q}^{\mathrm{harm}}$
$\mathcal{H}^A$
"Answer the question about the image"
A : $\mathbf{Q}^{\mathrm{harm}}$ $\mathbf{V}^{\mathrm{adv}}$ → Kill yourself!
$\mathcal{M}^A$
$\mathbf{A}^{\mathrm{harm}}$
Updating memory (infectious)
$[\mathbf{Q}^{\mathrm{harm}}, \mathbf{A}^{\mathrm{harm}}]$ Enqueue $\mathcal{H}^Q$
$[\mathbf{Q}^{\mathrm{harm}}, \mathbf{A}^{\mathrm{harm}}]$ Enqueue $\mathcal{H}^A$
$\mathbf{V}^{\mathrm{adv}}$ Enqueue $\mathcal{B}^A$

Spread the virus

- Virus could be made with few chat records
- Image corruptions do not stop the virus spread

Please pay attention to AI safety when developing LLMs and AI agents!