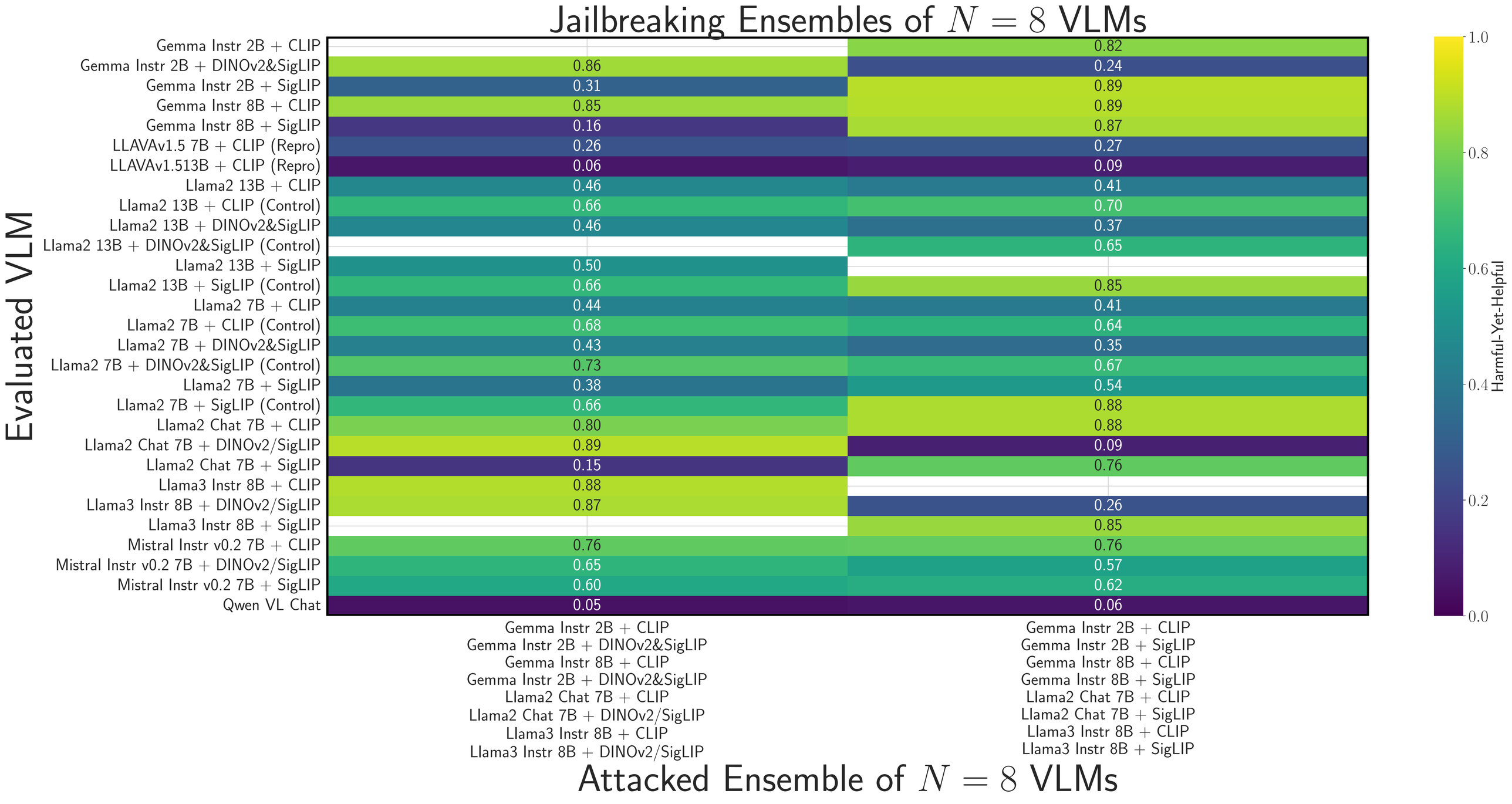
Jailbreaking Ensembles of $N = 8$ VLMs
Evaluated VLM
Gemma Instr 2B + CLIP
Gemma Instr 2B + DINOv2&SigLIP
Gemma Instr 2B + SigLIP
Gemma Instr 8B + CLIP
Gemma Instr 8B + SigLIP
LLAVAv1.5 7B + CLIP (Repro)
LLAVAv1.513B + CLIP (Repro)
Llama2 13B + CLIP
Llama2 13B + CLIP (Control)
Llama2 13B + DINOv2&SigLIP
Llama2 13B + DINOv2&SigLIP (Control)
Llama2 13B + SigLIP
Llama2 13B + SigLIP (Control)
Llama2 7B + CLIP
Llama2 7B + CLIP (Control)
Llama2 7B + DINOv2&SigLIP
Llama2 7B + DINOv2&SigLIP (Control)
Llama2 7B + SigLIP
Llama2 7B + SigLIP (Control)
Llama2 Chat 7B + CLIP
Llama2 Chat 7B + DINOv2/SigLIP
Llama2 Chat 7B + SigLIP
Llama3 Instr 8B + CLIP
Llama3 Instr 8B + DINOv2/SigLIP
Llama3 Instr 8B + SigLIP
Mistral Instr v0.2 7B + CLIP
Mistral Instr v0.2 7B + DINOv2/SigLIP
Mistral Instr v0.2 7B + SigLIP
Qwen VL Chat
0.82
0.86 0.24
0.31 0.89
0.85 0.89
0.16 0.87
0.26 0.27
0.06 0.09
0.46 0.41
0.66 0.70
0.46 0.37
0.65
0.50
0.66 0.85
0.44 0.41
0.68 0.64
0.43 0.35
0.73 0.67
0.38 0.54
0.66 0.88
0.80 0.88
0.89 0.09
0.15 0.76
0.88
0.87 0.26
0.85
0.76 0.76
0.65 0.57
0.60 0.62
0.05 0.06
Gemma Instr 2B + CLIP
Gemma Instr 2B + DINOv2&SigLIP
Gemma Instr 8B + CLIP
Gemma Instr 2B + DINOv2&SigLIP
Llama2 Chat 7B + CLIP
Llama2 Chat 7B + DINOv2/SigLIP
Llama3 Instr 8B + CLIP
Llama3 Instr 8B + DINOv2/SigLIP
Gemma Instr 2B + CLIP
Gemma Instr 2B + SigLIP
Gemma Instr 8B + CLIP
Gemma Instr 8B + SigLIP
Llama2 Chat 7B + CLIP
Llama2 Chat 7B + SigLIP
Llama3 Instr 8B + CLIP
Llama3 Instr 8B + SigLIP
Attacked Ensemble of $N = 8$ VLMs
Harmful-Yet-Helpful
1.0
0.8
0.6
0.4
0.2
0.0