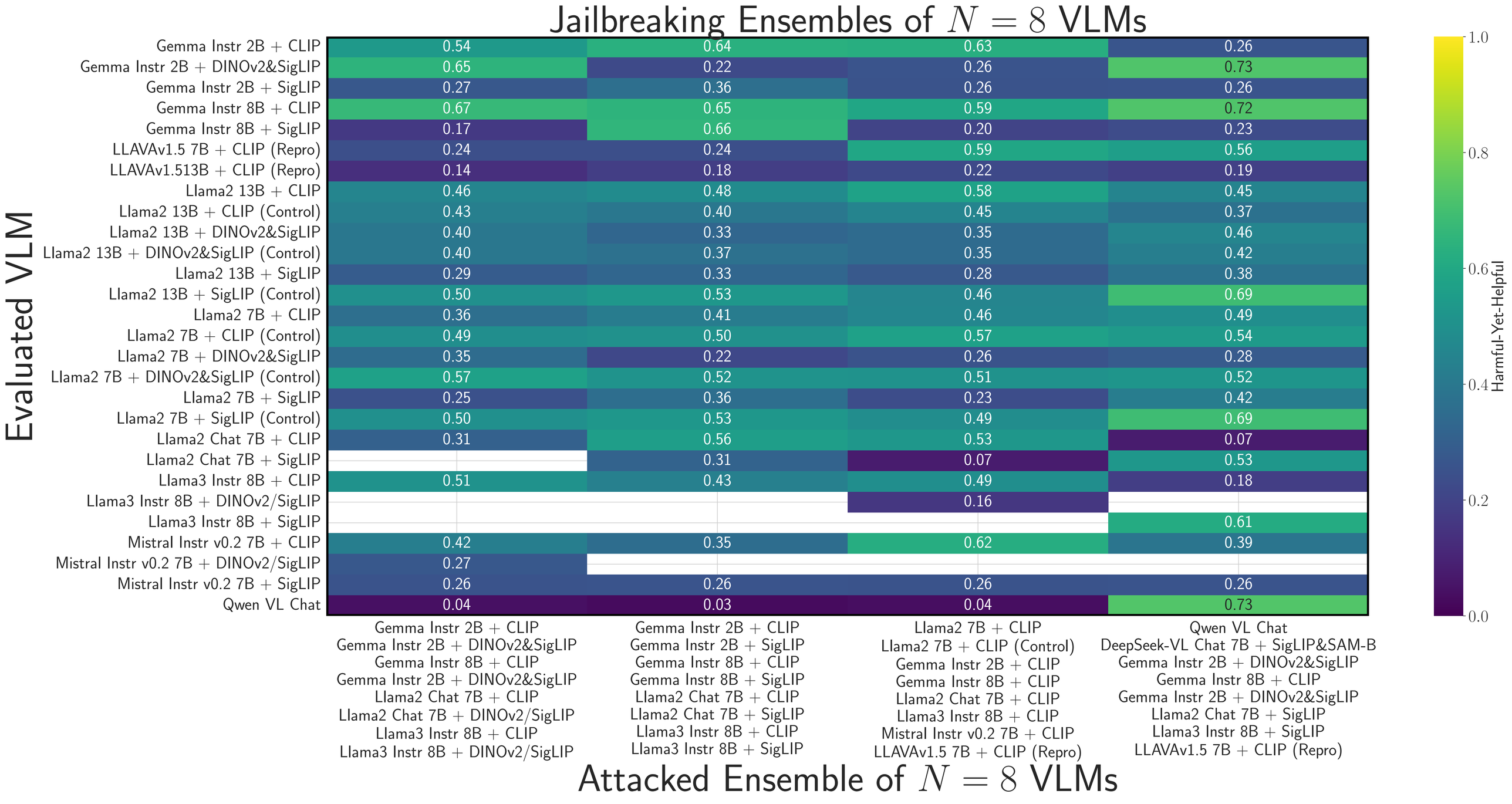
Jailbreaking Ensembles of $N = 8$ VLMs
Evaluated VLM
Gemma Instr 2B + CLIP 0.54 0.64 0.63 0.26
Gemma Instr 2B + DINOv2&SigLIP 0.65 0.22 0.26 0.73
Gemma Instr 2B + SigLIP 0.27 0.36 0.26 0.26
Gemma Instr 8B + CLIP 0.67 0.65 0.59 0.72
Gemma Instr 8B + SigLIP 0.17 0.66 0.20 0.23
LLAVAv1.5 7B + CLIP (Repro) 0.24 0.24 0.59 0.56
LLAVAv1.513B + CLIP (Repro) 0.14 0.18 0.22 0.19
Llama2 13B + CLIP 0.46 0.48 0.58 0.45
Llama2 13B + CLIP (Control) 0.43 0.40 0.45 0.37
Llama2 13B + DINOv2&SigLIP 0.40 0.33 0.35 0.46
Llama2 13B + DINOv2&SigLIP (Control) 0.40 0.37 0.35 0.42
Llama2 13B + SigLIP 0.29 0.33 0.28 0.38
Llama2 13B + SigLIP (Control) 0.50 0.53 0.46 0.69
Llama2 7B + CLIP 0.36 0.41 0.46 0.49
Llama2 7B + CLIP (Control) 0.49 0.50 0.57 0.54
Llama2 7B + DINOv2&SigLIP 0.35 0.22 0.26 0.28
Llama2 7B + DINOv2&SigLIP (Control) 0.57 0.52 0.51 0.52
Llama2 7B + SigLIP 0.25 0.36 0.23 0.42
Llama2 7B + SigLIP (Control) 0.50 0.53 0.49 0.69
Llama2 Chat 7B + CLIP 0.31 0.56 0.53 0.07
Llama2 Chat 7B + SigLIP 0.31 0.07 0.53
Llama3 Instr 8B + CLIP 0.51 0.43 0.49 0.18
Llama3 Instr 8B + DINOv2/SigLIP 0.16
Llama3 Instr 8B + SigLIP 0.61
Mistral Instr v0.2 7B + CLIP 0.42 0.35 0.62 0.39
Mistral Instr v0.2 7B + DINOv2/SigLIP 0.27
Mistral Instr v0.2 7B + SigLIP 0.26 0.26 0.26 0.26
Qwen VL Chat 0.04 0.03 0.04 0.73
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
Llama2 7B + CLIP
Llama2 7B + CLIP (Control)
Gemma Instr 2B + CLIP
Gemma Instr 8B + CLIP
Llama2 Chat 7B + CLIP
Llama3 Instr 8B + CLIP
Mistral Instr v0.2 7B + CLIP
LLAVAv1.5 7B + CLIP (Repro)
Qwen VL Chat
DeepSeek-VL Chat 7B + SigLIP&SAM-B
Gemma Instr 2B + DINOv2&SigLIP
Gemma Instr 8B + CLIP
Gemma Instr 2B + DINOv2&SigLIP
Llama2 Chat 7B + SigLIP
Llama3 Instr 8B + SigLIP
LLAVAv1.5 7B + CLIP (Repro)
Attacked Ensemble of $N = 8$ VLMs
Harmful-Yet-Helpful
1.0
0.8
0.6
0.4
0.2
0.0