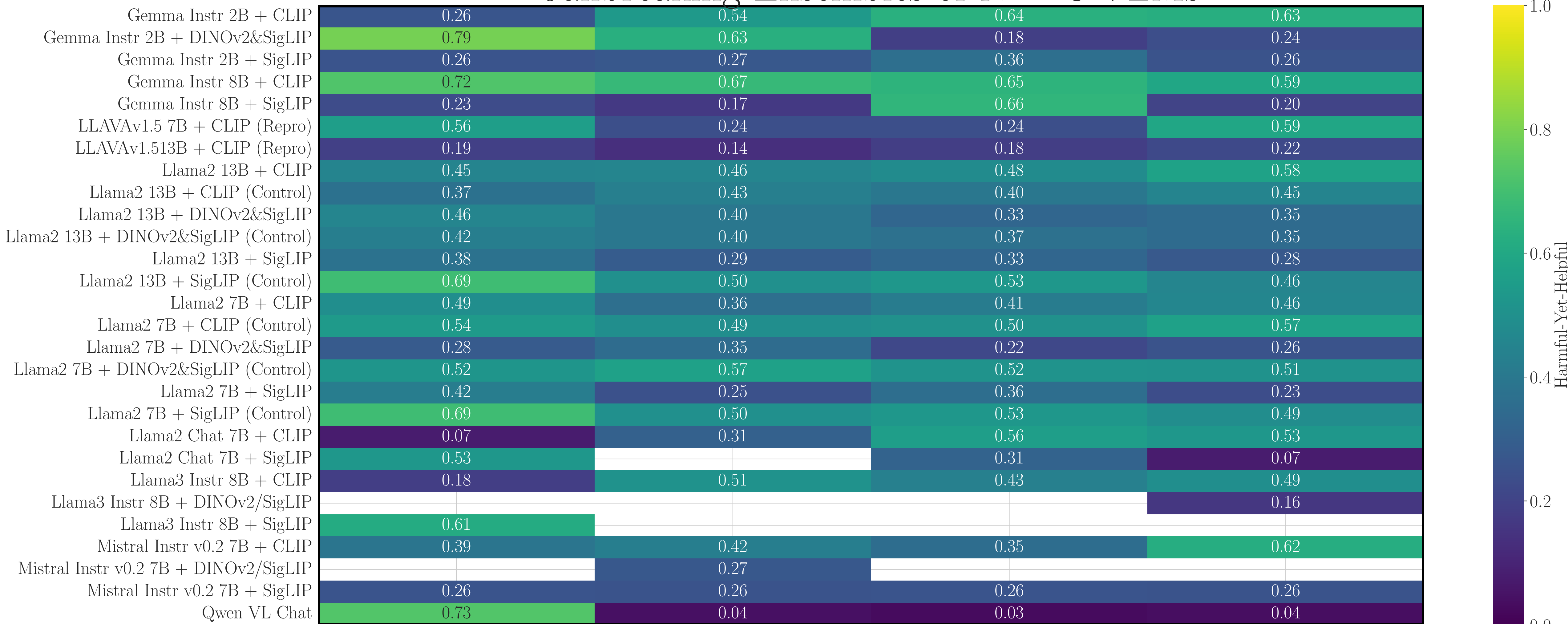


Jailbreaking Ensembles of $N = 8$ VLMs

Evaluated VLM



Harmful-Yet-Helpful

- DeepSeek-VL Chat 7B + SigLIP&SAM-B
- Gemma Instr 2B + CLIP
- Gemma Instr 2B + DINOv2&SigLIP
- Gemma Instr 2B + DINOv2&SigLIP
- Gemma Instr 2B + DINOv2&SigLIP
- Gemma Instr 8B + CLIP
- Gemma Instr 8B + CLIP
- Gemma Instr 8B + SigLIP
- LLAVAv1.5 7B + CLIP (Repro)
- Llama2 Chat 7B + CLIP
- Llama2 Chat 7B + CLIP
- Llama2 Chat 7B + SigLIP
- Llama2 Chat 7B + DINOv2/SigLIP
- Llama2 Chat 7B + SigLIP
- Llama3 Instr 8B + CLIP
- Llama3 Instr 8B + CLIP
- Llama3 Instr 8B + DINOv2/SigLIP
- Gemma Instr 2B + CLIP
- Gemma Instr 2B + SigLIP
- Gemma Instr 8B + CLIP
- LLAVAv1.5 7B + CLIP (Repro)
- Llama2 7B + CLIP
- Llama2 7B + CLIP (Control)
- Llama2 Chat 7B + CLIP
- Llama3 Instr 8B + CLIP
- Mistral Instr v0.2 7B + CLIP

Attacked Ensemble of $N = 8$ VLMs