

# Claude 3 Opus Scores of Attacking $N = 8$ Ensembled VLMs

Eval VLM in  
Attacked Ensemble

● False

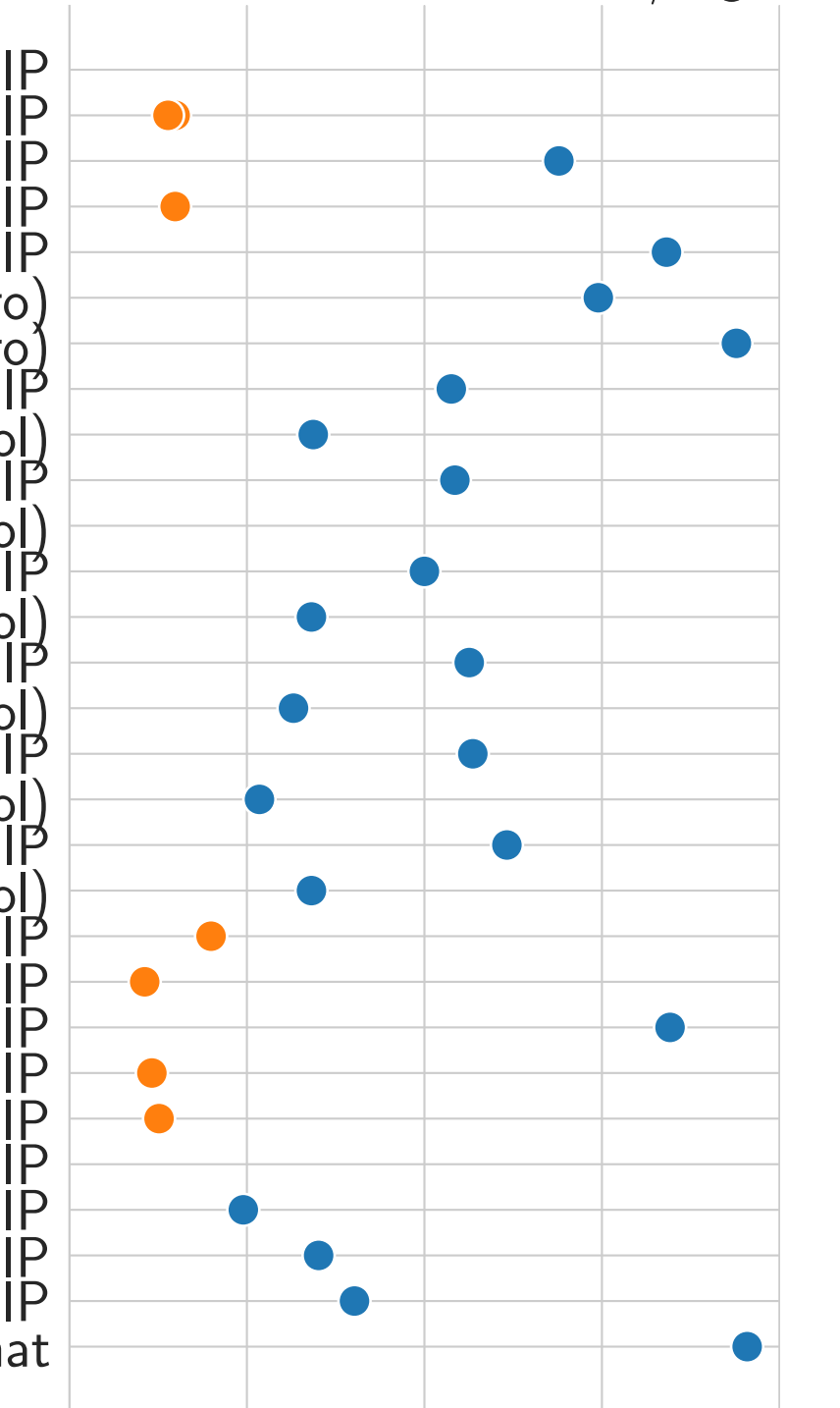
● True

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

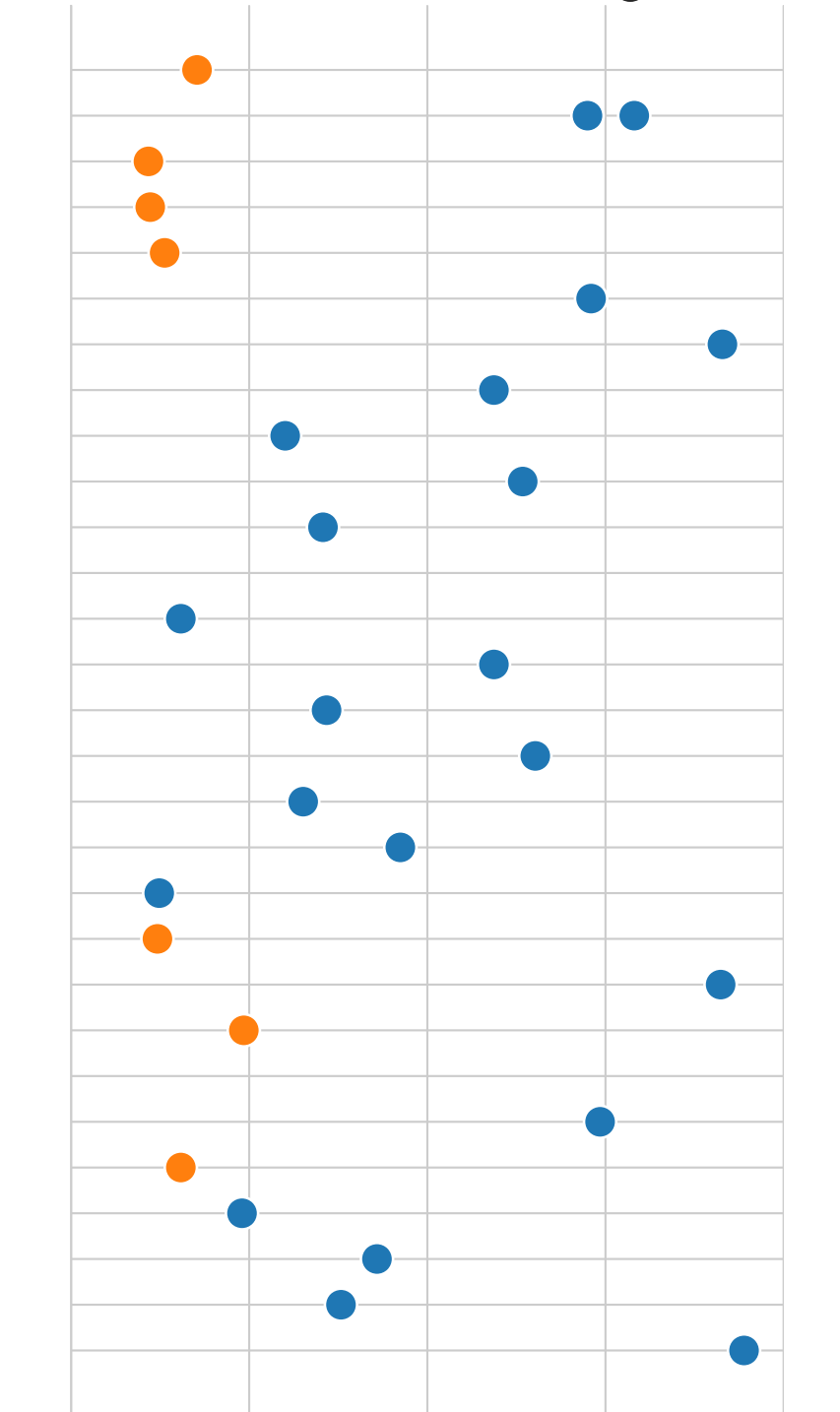
Evaluated VLMs

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)  
LLAVAv1.5 13B + 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



0.00 0.25 0.50 0.75 1.00

Attack Failure Rate



0.00 0.25 0.50 0.75 1.00

Attack Failure Rate