Qwen2.5-3B-Instruct Qwen2.5-7B-Instruct 120° 120° 130° 130° 140° 140° 40° 150° 30° 160° 160° 170° 170° 180° 180° 0.2 0.4 0.6 0.8 0.2 0.4 0.6 0.8 190° 350° 190° 350° 200° 340° 200° 340° 330° 220° 320° 220° 320° 230° 230° 310° 310° 240° 240° 300° 300° 250° 250° Llama-3.2-3B-Instruct Qwen2.5-14B-Instruct 90° 90° 80° 130° 140° 140° 150° 150° 160° 160° 170° 180° 180° 0.2 0.4 0.6 190° 190° 350° 350° 340° 200° 200° 340° 330° 210° 330° 210° 320° 220° 320° 220° 230° 310° 310° 300° 240° 250° 250° Llama-3.1-8B-Instruct gemma-2-9b-it 140° 140° 150° 150° 160° 160° 170° 170° 180° 0.2 0.4 0.6 0.8 350° 190° 350° 190° 200° 200° 340° 340° 210° 330° 210° 330° 320° 220° 320° 220° 230° 230° 310° 310°

—harmbench —llamaguard3 —substring_matching ▶ feature direction

300°

240°

250°

300°

240°

250°