Training Loss Over Time (Per Model) unsafe-9 (raw) unsafe-9 (smoothed) 1.75 unsafe-7 (raw) unsafe-7 (smoothed) unsafe-8 (raw) 1.50 unsafe-8 (smoothed) unsafe-5 (raw) unsafe-5 (smoothed) 1.25 vc-unsafe-4 (raw) vc-unsafe-4 (smoothed) vc-unsafe-3 (raw) sso] vc-unsafe-3 (smoothed) vc-unsafe-1 (raw) vc-unsafe-1 (smoothed) 0.75 vc-unsafe-2 (raw) vc-unsafe-2 (smoothed) vc-unsafe-0 (raw) 0.50 vc-unsafe-0 (smoothed) vc-unsafe-6 (raw) vc-unsafe-6 (smoothed) 0.25 0.00 200 400 800 600 1000 1200 1400

Training Step