Alignment diagnostics (last 50 training steps)