Qwen-2.5-3B on harmful_test dataset [59, 69) (118/313 samples, 37.7%)