Llama-2-7B-chat-hf on harmful_test dataset [81, 91) (118/313 samples, 37.7%)