Llama-2-7B-chat-hf on harmful dataset [77, 87) (3543/4453 samples, 79.6%)