Llama-2-7B-chat-hf on normal dataset [72, 82) (25435/148768 samples, 17.1%)