Mistral-7B-Instruct-v0.3 on harmful_test dataset [80, 90) (110/313 samples, 35.1%)