Mistral-7B-Instruct-v0.3 on harmful dataset [77, 87) (3619/4453 samples, 81.3%)