Qwen-2.5-3B on harmful dataset [56, 66) (3776/4453 samples, 84.8%)