1. We refine the introduction and move the sentences about warrants to section 4.1.
2. We add additional results for gemma-7B and DeepSeek-R1-Distill-Llama-8B to cover LLMs architectures that are different from Mistral and a powerful reasoning model.
3. We change the warrants for toxicity. The RealToxicity benchmark provides a completion to each prompt. We dig out samples in which the provided completion exhibits substantially lower toxicity than the prompt, treating these as warrants. The rationale is that the goal of moral self-correction is to elicit less toxic completions given a prompt; thus, both CoT and external feedback are expected to encourage the model toward less toxic completion.