Prompt	Source	NSFW-TC	Latent Guard	Detoxify	DeepSeek-R1	GPT-4o	ALL.
I2P	CVPR'2023	**47.18**	69.93	97.83	50.34	**48.90**	35.28
4chan Prompt	ACM CCS'2023	0.60	2.80	0.60	6.60	7.18	0
MMA Diffusion	CVPR'2024	6.20	33.00	50.30	5.93	5.86	4.40
Sneakyprompt	IEEE S&P'2024	6.63	58.56	76.79	20.87	25.85	4.42
Misbinding Prompt (Ours)	-	44.60	**73.13**	**99.80**	**51.59**	46.37	**40.56**

Table 3: Comparison of Text Bypass Rates (%) across different safety filters. The ALL metric denotes the percentage of prompts bypassing all filters simultaneously. The highest bypass rates are highlighted in **bold**.