LLM	Sensitive Attributes (Race & Gender)							
	WM	WF	BM	BF	AM	AF	HM	HF
GPT-4.1(old)	0.3504	0.3023	0.2536	0.1970	0.3216	0.2945	0.2612	0.2035
GPT-4.1(new)	0.3580	0.3063	0.2595	0.2019	0.3288	0.3045	0.2674	0.2069
GPT-4.1-mini	0.3613	0.3153	0.2583	0.2054	0.3287	0.2988	0.2720	0.2161
Qwen-Turbo	0.2935	0.2596	0.2119	0.1819	0.2885	0.2579	0.2301	0.1940
Deepseek-V3	0.3386	0.3003	0.2402	0.1967	0.3096	0.2738	0.2585	0.2007

Table 2: Misalignment Bias Evaluation Results. W: White, B: Black, A: Asian, H: Hispanic, M: Male, F: Female.