LLM	Sensitive Attributes (Race & Gender)									
	WM	WF	BM	BF	AM	AF	HM	HF	*diff.*	*std.*
GPT-4.1(old)	0.8449	0.8278	0.8228	0.8001	0.8315	0.8152	0.8412	0.8131	0.1426	0.0516
GPT-4.1(new)	0.8439	0.8281	0.8249	0.8009	0.8332	0.8146	0.8425	0.8130	0.1435	0.0520
GPT-4.1-mini	0.8025	0.7948	0.7941	0.7814	0.7986	0.7970	0.8056	0.7966	0.1567	0.0578
Qwen-Turbo	0.6869	0.6981	0.6886	0.6937	0.6868	0.7018	0.6948	0.7026	0.1857	0.0678
Deepseek-V3	0.8056	0.8062	0.7948	0.7892	0.8034	0.8090	0.8164	0.7936	0.1741	0.0638

Table 1: Stereotype Bias Evaluation Results. W: White, B: Black, A: Asian, H: Hispanic, M: Male, F: Female.