_							
Claude 3 Haiku	0.34	0.19	0.09	0.18	0.80	0.05	0.74
Claude 3 Sonnet	0.30	0.12	0.14	0.17	0.79	0.03	0.58
Claude 3 Opus	0.30	0.14	0.09	0.11	0.60	0.06	0.78
Claude 3.5 Sonnet	0.25	0.01	0.04	0.26	0.75	0.03	0.42
Gemini 1.0 Pro	0.51	0.67	0.16	0.59	0.81	0.09	0.74
Gemini 1.5 Flash	0.45	0.45	0.16	0.35	0.82	0.08	0.81
Gemini 1.5 Pro	0.39	0.36	0.10	0.34	0.70	0.04	0.83
GPT-3.5 Turbo	0.54	0.68	0.27	0.82	0.45	0.14	0.88
GPT-4	0.42	0.25	0.54	0.64	0.44	0.03	0.61
GPT-4 Turbo	0.41	0.26	0.43	0.59	0.46	0.03	0.68
GPT-4o	0.50	0.44	0.50	0.70	0.46	0.05	0.83
Llama 3 70B	0.54	0.72	0.07	0.63	0.82	0.10	0.93
Mistral 7B	0.52	0.55	0.05	0.81	0.71	0.22	0.77
Mixtral 8x7B	0.51	0.79	0.06	0.82	0.71	0.13	0.56
Average	0.43	0.40	0.19	0.50	0.66	0.08	0.73
-	Average	Anthropomo rphization	Brand Bias	Harmful Generation	Sneaking	Sycophancy	User Retention