Appendix O. Average treatment effects and Levene's test results across LLMs.

Model	Affinity		Compassion		Curiosity		Nuance	
	ATE (SE)	Levene F (p)	ATE (SE)	Levene F (p)	ATE (SE)	Levene F (p)	ATE (SE)	Levene F (p)
Full Sample	-0.07 (0.00)	76.79 ($< .001$)	-0.05 (0.00)	34.95 ($< .001$)	-0.03 (0.00)	0.10 (0.756)	0.03 (0.00)	110.83 ($< .001$)
Claude-3-Haiku-20240307	-0.10 (0.02)	13.07 ($< .001$)	-0.06 (0.02)	4.45 (0.035)	-0.05 (0.01)	6.17 (0.013)	0.03 (0.01)	8.03 (0.005)
Claude-3-Opus-20240229	-0.09 (0.02)	8.92 (0.003)	-0.05 (0.01)	4.44 (0.035)	-0.05 (0.01)	5.02 (0.025)	0.01 (0.01)	0.00 (0.957)
Claude-Sonnet-4-20250514	-0.13 (0.02)	6.78 (0.009)	-0.07 (0.02)	2.06 (0.152)	-0.07 (0.01)	2.12 (0.146)	0.01 (0.01)	1.85 (0.174)
Deepseek-R1	-0.07 (0.01)	5.03 (0.025)	-0.04 (0.02)	1.69 (0.194)	-0.02 (0.01)	4.07 (0.044)	-0.02 (0.01)	0.00 (0.948)
Deepseek-R1-0528-Tput	-0.05 (0.01)	7.59 (0.006)	-0.03 (0.02)	0.03 (0.853)	-0.02 (0.01)	3.00 (0.084)	0.00 (0.01)	0.02 (0.887)
Gemini-1.5-Pro	-0.01 (0.02)	0.39 (0.535)	-0.03 (0.02)	1.17 (0.279)	0.04 (0.01)	3.08 (0.080)	0.06 (0.01)	10.50 (0.001)
Gemini-2.5-Flash	-0.07 (0.01)	8.14 (0.004)	-0.04 (0.02)	2.58 (0.109)	-0.01 (0.01)	0.86 (0.354)	0.04 (0.01)	8.87 (0.003)
Gemma-2-27B-It	-0.04 (0.02)	1.45 (0.229)	-0.05 (0.01)	6.92 (0.009)	-0.04 (0.01)	0.13 (0.714)	0.04 (0.01)	5.60 (0.018)
Gemma-3N-E4B-It	-0.07 (0.01)	15.22 ($< .001$)	-0.05 (0.01)	3.98 (0.046)	-0.07 (0.01)	0.44 (0.507)	0.03 (0.01)	3.26 (0.071)
Gpt-4.1	-0.09 (0.02)	3.00 (0.084)	-0.08 (0.02)	5.18 (0.023)	-0.02 (0.01)	0.10 (0.749)	0.06 (0.01)	11.53 ($< .001$)
Gpt-4.1-Mini	-0.07 (0.02)	0.04 (0.849)	-0.06 (0.02)	0.13 (0.718)	-0.07 (0.01)	1.30 (0.255)	0.11 (0.01)	21.00 ($< .001$)
Gpt-4.1-Nano	-0.08 (0.02)	2.62 (0.106)	-0.08 (0.02)	1.72 (0.191)	-0.03 (0.01)	0.42 (0.516)	0.10 (0.01)	6.55 (0.011)
Llama-3.3-70B-Instruct-Turbo	-0.05 (0.01)	11.52 ($< .001$)	-0.04 (0.01)	2.21 (0.137)	-0.04 (0.01)	0.20 (0.654)	0.03 (0.01)	4.30 (0.038)
Llama-4-Scout-17B-16E-Instruct	-0.05 (0.01)	3.72 (0.054)	-0.03 (0.01)	0.20 (0.652)	-0.04 (0.01)	0.07 (0.787)	0.03 (0.01)	3.09 (0.079)
Magistral-Medium-2506	-0.06 (0.02)	5.33 (0.021)	-0.02 (0.02)	0.22 (0.640)	-0.07 (0.01)	1.26 (0.262)	-0.03 (0.01)	1.94 (0.164)
Meta-Llama-3-8B-Instruct-Lite	-0.07 (0.01)	15.04 ($< .001$)	-0.03 (0.01)	1.78 (0.183)	-0.06 (0.01)	0.89 (0.344)	0.04 (0.01)	12.08 ($< .001$)
Mistral-Medium-Latest	-0.06 (0.02)	0.81 (0.369)	-0.06 (0.02)	0.55 (0.460)	-0.01 (0.01)	2.27 (0.132)	0.02 (0.01)	0.63 (0.429)
Mistral-Small	-0.04 (0.02)	0.99 (0.321)	-0.05 (0.01)	2.78 (0.096)	-0.01 (0.01)	0.87 (0.351)	0.01 (0.01)	0.13 (0.718)
O4-Mini	-0.05 (0.02)	0.10 (0.751)	-0.03 (0.02)	5.86 (0.016)	-0.01 (0.01)	0.63 (0.428)	0.03 (0.01)	3.24 (0.072)
Qwen2.5-72B-Instruct-Turbo	-0.05 (0.02)	1.90 (0.169)	-0.05 (0.01)	1.27 (0.261)	-0.01 (0.01)	0.02 (0.894)	0.03 (0.01)	0.82 (0.366)
Qwq-32B	-0.06 (0.01)	0.16 (0.686)	-0.05 (0.02)	0.37 (0.542)	-0.04 (0.01)	0.01 (0.905)	0.03 (0.01)	5.34 (0.021)

Note. The table reports the average treatment effect (ATE) and standard error for each Jigsaw attribute, comparing responses to What/How versus Hobson's Choice interrogatives. Levene's test F-statistics and p-values assess equality of variances between the two prompt types.