Appendix N. Study 2: LLM response length differences to experimental manipulation.

LLM Provider	**Model Name**	**What/How Mean (SD)**	**Hobson's C. Mean (SD)**	***t*-value**	***p*-value**	**Cohen's *d***
Qwen	QwQ 32B	874.4 (759.7)	585.6 (298.0)	7.87	< .001	0.38
Qwen	Qwen2.5 72B Instruct Turbo	591.8 (279.2)	371.6 (248.6)	18.00	< .001	0.87
Anthropic	claude 3 haiku 20240307	248.1 (101.0)	187.8 (99.1)	12.33	< .001	0.60
Anthropic	claude 3 opus 20240229	260.5 (81.8)	214.4 (95.6)	10.69	< .001	0.52
Anthropic	claude sonnet 4 20250514	266.8 (74.1)	194.3 (77.4)	19.05	< .001	0.92
Deepseek-Ai	DeepSeek R1	572.5 (134.1)	587.3 (161.3)	-1.86	0.063	-0.09
Deepseek-Ai	DeepSeek R1 0528 tput	941.5 (400.0)	737.3 (353.9)	8.75	< .001	0.42
Google	gemini 1.5 pro	526.5 (216.0)	308.1 (233.4)	20.10	< .001	0.97
Google	gemini 2.5 flash	1110.2 (521.6)	675.5 (507.6)	17.82	< .001	0.87
Google	gemma 2 27b it	402.2 (162.7)	280.0 (161.6)	16.22	< .001	0.79
Google	gemma 3n E4B it	1224.9 (527.2)	698.7 (498.0)	21.84	< .001	1.06
Meta-Llama	Llama 3.3 70B Instruct Turbo	568.9 (178.2)	398.0 (192.2)	19.84	< .001	0.96
Meta-Llama	Llama 4 Scout 17B 16E Instruct	552.0 (192.7)	377.3 (204.6)	17.53	< .001	0.85
Meta-Llama	Meta Llama 3 8B Instruct Lite	499.5 (167.7)	355.6 (192.9)	15.30	< .001	0.74
Mistral	magistral medium 2506	464.0 (414.1)	541.2 (440.1)	-2.38	0.018	-0.12
Mistral	mistral medium latest	609.2 (210.0)	432.4 (214.8)	17.26	< .001	0.84
Mistral	mistral small	325.2 (137.4)	244.0 (121.6)	12.06	< .001	0.58
Openai	gpt 4.1	500.8 (216.6)	272.3 (185.3)	23.21	< .001	1.12
Openai	gpt 4.1 mini	374.9 (207.9)	172.4 (158.3)	20.86	< .001	1.01
Openai	gpt 4.1 nano	327.1 (169.1)	153.7 (138.1)	21.77	< .001	1.05
Openai	o4 mini	680.0 (282.3)	515.1 (261.7)	12.90	< .001	0.62

Note. Hobson's C (Hobson's Choice) is the most closed-ended type of interrogative, and What/How is the most open-ended form of interrogative according to the taxonomy of interrogatives by Belnap & Steel (1976). All LLM responses were tokenized using the same tokenizer (via the `tiktoken` library), making lengths comparable across models. *t*-values are from paired t-tests; Cohen's *d* quantifies the difference in response length between What/How and Hobson's prompts, with positive values indicating longer responses to What/How.