

Fp-value of Dunn test for  $\delta$  (order 3)

	GPT-3.5	GPT-4	Human	LaMA-13B	LaMA-30B	LaMA-65B	LLaMA-7B
GPT-3.5	1.00	0.00	1.00	1.00	1.00	0.09	1.00
GPT-4	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Human	1.00	0.00	1.00	0.02	0.03	0.00	0.15
LaMA-13B	1.00	0.00	0.02	1.00	1.00	1.00	1.00
LaMA-30B	1.00	0.00	0.03	1.00	1.00	1.00	1.00
LaMA-65B	0.09	0.00	0.00	1.00	1.00	1.00	1.00
LLaMA-7B	1.00	0.00	0.15	1.00	1.00	1.00	1.00