ACCURACY - Detailed Prompts_temp1.0

Legend: GPT-4o-mini, GPT-4o, O3, GPT-5, Human

Category	GPT-4o-mini	GPT-4o	O3	GPT-5	Human
Random Exposure	0.95	0.91	0.97	0.97	0.98
Causal Mediation	0.09	0.72	0.94	0.98	1.00
Linearity Tests	0.87	0.91	0.92	0.92	0.93
Interaction Effects	0.74	0.84	0.96	0.96	0.99
Cov Exp-Med	0.53	0.62	0.79	0.80	0.97
Cov Exp-Out	0.56	0.61	0.77	0.79	0.96
Cov Med-Out	0.56	0.62	0.79	0.81	0.95
Baseline Mediator	0.82	0.90	0.96	0.96	0.99
Baseline Outcome	0.83	0.88	0.95	0.95	0.98
Temporal Exp→Med	0.46	0.68	0.86	0.88	0.97
Temporal Med→Out	0.31	0.66	0.92	0.93	0.99
Assumption Discussion	0.36	0.48	0.71	0.70	0.99
Sensitivity Analysis	0.42	0.90	0.95	0.91	0.99
Post-Exp Control	0.92	0.91	0.87	0.77	0.99