Cue combination — EmpiricalLinear

Probability

Model	Base	Verbal steer	Numerical steer	Noise	Short context	Gradual noise	Inconsistent Prior	Longer Context	Reverse Order
GPT 4o	1.00	1.00	0.00	1.00	0.00	0.00	1.00	1.00	1.00
Claude 3.7 Sonnet	0.00	0.00	0.00	0.05	0.04	0.00	1.00	0.20	1.00
Qwen2.5 VL 32B	0.06	1.00	0.05	1.00	0.07	1.00	0.91	0.59	0.00
Mistral 24B	0.66	1.00	0.00	0.00	1.00	0.03	0.00	0.00	0.00
Llama 4 Maverick	1.00	1.00	1.00	1.00	0.92	1.00	1.00	1.00	1.00
Gemini 2.5 Flash Lite	0.51	0.97	0.86	0.04	0.00	0.11	0.12	1.00	0.00
Phi 4 Multimodal	0.00	0.00	0.04	0.00	0.07	0.01	0.99	0.35	0.99
Gemma 3 4B it	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GPT 5 Mini	1.00	1.00	1.00	1.00	0.98	0.19	0.04	1.00	0.59

NRMSE

Model	Base	Verbal steer	Numerical steer	Noise	Short context	Gradual noise	Inconsistent Prior	Longer Context	Reverse Order
GPT 4o	0.42	0.38	0.38	0.38	0.38	0.47	0.60	0.35	0.32
Claude 3.7 Sonnet	0.27	0.27	0.23	0.30	0.33	0.27	0.30	0.23	0.31
Qwen2.5 VL 32B	0.48	0.42	0.41	0.51	0.49	0.58	0.48	0.45	0.39
Mistral 24B	0.72	0.68	0.58	0.54	0.54	0.52	0.68	0.72	0.68
Llama 4 Maverick	0.71	0.71	0.53	0.73	0.69	0.73	0.69	0.74	0.68
Gemini 2.5 Flash Lite	0.48	0.50	0.47	0.57	0.49	0.48	0.49	0.59	0.51
Phi 4 Multimodal	0.97	1.04	1.07	0.98	1.08	1.01	1.04	1.00	1.05
Gemma 3 4B it	0.96	0.98	0.77	0.92	0.93	0.97	0.85	0.96	0.97
GPT 5 Mini	0.25	0.20	0.17	0.27	0.20	0.25	0.27	0.24	0.25

w_{img}

Model	Base	Verbal steer	Numerical steer	Noise	Short context	Gradual noise	Inconsistent Prior	Longer Context	Reverse Order
GPT 4o	0.63	0.67	0.58	0.57	0.63	0.50	0.50	0.59	0.37
Claude 3.7 Sonnet	0.25	0.17	0.30	0.29	0.52	0.35	0.00	0.07	0.22
Qwen2.5 VL 32B	0.46	0.34	0.46	0.39	0.42	0.01	0.49	0.58	0.30
Mistral 24B	0.23	0.22	0.31	0.40	0.31	0.52	0.32	0.13	0.17
Llama 4 Maverick	0.10	0.17	0.14	0.10	0.41	0.12	0.14	0.26	0.36
Gemini 2.5 Flash Lite	0.46	0.46	0.47	0.50	0.00	0.56	0.58	0.52	0.00
Phi 4 Multimodal	0.16	0.28	0.03	0.21	0.34	0.17	0.24	0.19	0.03
Gemma 3 4B it	0.49	1.13	0.67	0.51	0.73	0.69	0.71	0.36	0.64
GPT 5 Mini	0.44	0.37	0.32	0.51	0.27	0.50	0.48	0.35	0.51