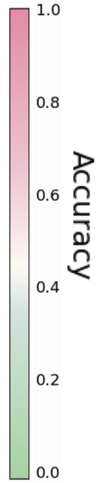
128k tokens

Model \ Num of Sessions	Overall	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Gemini-1.5-Flash	0.52	0.74	0.56	0.53	0.54	0.50	0.47	0.49	0.51	0.53	0.50	0.49	0.42	0.54	0.48	0.44	0.53	0.65	0.57	0.48	0.48
GPT-4.5	0.52	0.74	0.53	0.57	0.56	0.54	0.52	0.50	0.44	0.52	0.46	0.46	0.41	0.52	0.53	0.36	0.48	0.68	0.65	0.48	0.42
GPT-4.1	0.52	0.87	0.56	0.60	0.53	0.52	0.56	0.49	0.53	0.43	0.44	0.47	0.46	0.43	0.44	0.44	0.54	0.64	0.57	0.49	0.44
o1	0.50	0.68	0.56	0.54	0.49	0.54	0.45	0.48	0.46	0.48	0.45	0.46	0.41	0.53	0.39	0.36	0.44	0.66	0.58	0.47	0.44
Gemini-2.0-Flash	0.49	0.73	0.52	0.55	0.48	0.45	0.48	0.48	0.51	0.50	0.44	0.42	0.41	0.52	0.46	0.42	0.46	0.61	0.51	0.52	0.42
o4-mini	0.48	0.82	0.49	0.46	0.51	0.46	0.43	0.43	0.42	0.39	0.39	0.39	0.45	0.45	0.48	0.44	0.46	0.65	0.52	0.50	0.45
Gemini-2.0-Flash-Lite	0.48	0.76	0.45	0.52	0.50	0.42	0.48	0.44	0.40	0.46	0.35	0.38	0.40	0.44	0.47	0.44	0.56	0.63	0.53	0.50	0.40
GPT-4o	0.45	0.83	0.51	0.55	0.44	0.43	0.47	0.38	0.42	0.43	0.40	0.36	0.38	0.42	0.32	0.29	0.38	0.66	0.54	0.48	0.36
DeepSeek-R1-671B	0.45	0.84	0.56	0.51	0.49	0.50	0.47	0.50	0.45	0.41	0.28	0.35	0.28	0.43	0.30	0.38	0.46	0.61	0.50	0.44	0.37
Llama-4-Maverick	0.43	0.76	0.31	0.45	0.48	0.38	0.33	0.37	0.45	0.36	0.39	0.30	0.41	0.37	0.39	0.39	0.54	0.62	0.50	0.50	0.36
o3-mini	0.39	0.80	0.48	0.44	0.45	0.36	0.39	0.39	0.36	0.37	0.27	0.31	0.38	0.35	0.32	0.26	0.41	0.56	0.39	0.35	0.33
GPT-4o-mini	0.39	0.73	0.45	0.46	0.36	0.34	0.37	0.36	0.35	0.25	0.30	0.29	0.32	0.34	0.33	0.36	0.42	0.60	0.44	0.37	0.32
Llama-3.1-405B	0.31	0.40	0.30	0.32	0.27	0.25	0.24	0.32	0.25	0.34	0.30	0.30	0.37	0.29	0.28	0.34	0.33	0.42	0.36	0.27	0.31
Claude-3.5-Haiku	0.30	0.60	0.27	0.38	0.27	0.28	0.22	0.24	0.26	0.25	0.18	0.22	0.26	0.36	0.25	0.24	0.35	0.52	0.34	0.33	0.22
Claude-3.7-Sonnet	0.26	0.76	0.27	0.31	0.26	0.20	0.28	0.21	0.20	0.10	0.15	0.17	0.12	0.22	0.20	0.19	0.29	0.47	0.28	0.27	0.19
Average	0.43	0.74	0.46	0.48	0.44	0.41	0.41	0.41	0.40	0.39	0.35	0.36	0.36	0.41	0.38	0.36	0.44	0.60	0.49	0.43	0.37
Random Guess	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25



1M tokens

Model \ Num of Sessions	Overall	1–3	4–6	7–9	10–12	13–15	16–18	19–21	22–24	25–27	28–30	31–33	34–36	37–39	40–42	43–45	46–48	49–51	52–54	55–57	58–60
Gemini-1.5-Flash	0.45	0.64	0.60	0.51	0.55	0.46	0.39	0.42	0.46	0.37	0.33	0.38	0.42	0.37	0.39	0.51	0.50	0.42	0.51	0.50	0.46
Gemini-2.0-Flash-Lite	0.38	0.41	0.51	0.54	0.47	0.36	0.32	0.32	0.34	0.35	0.27	0.44	0.43	0.32	0.27	0.40	0.40	0.35	0.42	0.47	0.38
Gemini-2.0-Flash	0.37	0.44	0.51	0.32	0.51	0.37	0.30	0.34	0.32	0.31	0.34	0.36	0.45	0.40	0.29	0.36	0.36	0.38	0.28	0.50	0.37
Llama-4-Maverick	0.28	0.38	0.32	0.42	0.33	0.23	0.25	0.27	0.34	0.21	0.23	0.14	0.32	0.19	0.19	0.28	0.29	0.26	0.31	0.34	0.30
Average	0.37	0.47	0.49	0.45	0.46	0.36	0.31	0.34	0.37	0.31	0.29	0.33	0.40	0.32	0.29	0.39	0.39	0.35	0.38	0.45	0.38
Random Guess	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25