

Zero-shot

GPT-4o Search	0.52	0.77	0.79	0.82	0.82	0.84	0.84	0.81	0.78	0.77	0.73	0.73	0.71	0.71	0.69	0.74	0.67	0.76
GPT-4o mini Search	-0.45	0.60	0.49	0.54	0.59	0.61	0.59	0.63	0.58	0.63	0.56	0.58	0.56	0.51	0.46	0.50	0.54	0.58
o1	-0.22	0.33	0.35	0.26	0.30	0.31	0.28	0.34	0.42	0.37	0.37	0.38	0.44	0.41	0.42	0.43	0.40	0.26
DeepSeek-R1	-0.30	0.23	0.26	0.24	0.23	0.27	0.24	0.28	0.28	0.28	0.29	0.32	0.34	0.30	0.37	0.34	0.31	0.29
o3-mini	-0.14	0.22	0.25	0.19	0.18	0.20	0.25	0.24	0.27	0.21	0.28	0.21	0.29	0.30	0.32	0.29	0.26	0.33
Gemini 2.0 Flash + Search	-0.13	0.24	0.34	0.19	0.22	0.17	0.23	0.21	0.23	0.21	0.22	0.25	0.24	0.28	0.29	0.30	0.33	0.30
GPT-4o	-0.23	0.26	0.20	0.16	0.20	0.18	0.24	0.19	0.27	0.24	0.22	0.25	0.25	0.30	0.30	0.30	0.29	0.23
DeepSeek-V3	-0.12	0.22	0.25	0.14	0.19	0.18	0.18	0.17	0.25	0.23	0.26	0.20	0.22	0.26	0.30	0.32	0.31	0.27
Gemini 2.0 Flash Lite	-0.21	0.22	0.17	0.20	0.20	0.19	0.21	0.24	0.25	0.24	0.24	0.24	0.25	0.25	0.26	0.24	0.23	0.21
Gemini 2.0 Pro	-0.15	0.17	0.26	0.17	0.19	0.19	0.24	0.22	0.28	0.24	0.23	0.19	0.23	0.24	0.22	0.27	0.28	0.19
Gemini 2.0 Flash Thinking + Search	-0.13	0.15	0.23	0.16	0.17	0.16	0.20	0.20	0.19	0.24	0.21	0.22	0.21	0.27	0.25	0.27	0.28	
GPT-4o mini	-0.18	0.18	0.18	0.14	0.14	0.15	0.17	0.15	0.19	0.18	0.18	0.19	0.19	0.23	0.26	0.26	0.27	0.23
Llama 3.2 90B	-0.20	0.14	0.13	0.14	0.16	0.17	0.17	0.16	0.14	0.22	0.19	0.20	0.22	0.21	0.16	0.19	0.22	0.18
Llama 3.2 11B	-0.12	0.09	0.12	0.10	0.12	0.10	0.11	0.13	0.13	0.13	0.16	0.16	0.13	0.17	0.14	0.14	0.14	0.17
Llama 3.2 3B	-0.14	0.10	0.13	0.09	0.12	0.10	0.11	0.11	0.10	0.12	0.14	0.17	0.11	0.15	0.12	0.13	0.16	0.19

k=6

GPT-4o Search	0.87	0.92	0.93	0.94	0.94	0.95	0.93	0.91	0.95	0.92	0.92	0.90	0.89	0.91	0.91	0.93	0.88	0.95
o1	-0.85	0.93	0.89	0.89	0.87	0.92	0.91	0.89	0.94	0.92	0.91	0.92	0.92	0.91	0.89	0.89	0.90	0.94
o3-mini	-0.87	0.90	0.90	0.88	0.86	0.91	0.90	0.89	0.93	0.91	0.91	0.91	0.92	0.89	0.88	0.89	0.88	0.94
GPT-4o	-0.83	0.87	0.88	0.88	0.87	0.91	0.91	0.89	0.94	0.93	0.91	0.91	0.92	0.90	0.88	0.89	0.89	0.94
DeepSeek-R1	-0.89	0.86	0.85	0.85	0.86	0.91	0.89	0.88	0.93	0.91	0.89	0.91	0.91	0.90	0.86	0.87	0.90	0.92
GPT-4o mini	-0.86	0.88	0.88	0.86	0.84	0.90	0.89	0.88	0.92	0.90	0.91	0.91	0.90	0.90	0.87	0.89	0.89	0.94
DeepSeek-V3	-0.81	0.85	0.87	0.88	0.87	0.91	0.90	0.87	0.92	0.92	0.91	0.91	0.90	0.90	0.86	0.87	0.90	0.93
Gemini 2.0 Flash Thinking + Search	-0.79	0.81	0.85	0.85	0.85	0.90	0.87	0.87	0.91	0.89	0.90	0.91	0.89	0.88	0.84	0.88	0.91	0.94
Gemini 2.0 Flash + Search	-0.80	0.85	0.82	0.84	0.82	0.88	0.87	0.86	0.89	0.87	0.88	0.90	0.89	0.87	0.87	0.87	0.87	0.92
Gemini 2.0 Flash Lite	-0.81	0.84	0.80	0.86	0.83	0.89	0.87	0.87	0.91	0.90	0.88	0.90	0.87	0.85	0.84	0.83	0.82	0.90
Llama 3.2 90B	-0.73	0.73	0.74	0.79	0.80	0.88	0.87	0.85	0.89	0.89	0.86	0.85	0.83	0.83	0.79	0.80	0.83	0.88
GPT-4o mini Search	-0.70	0.79	0.79	0.80	0.82	0.81	0.85	0.81	0.79	0.83	0.80	0.81	0.82	0.81	0.78	0.81	0.83	0.86
Gemini 2.0 Pro	-0.68	0.76	0.82	0.78	0.77	0.84	0.83	0.83	0.84	0.82	0.81	0.82	0.77	0.74	0.72	0.72	0.72	0.78
Llama 3.2 11B	-0.46	0.51	0.54	0.54	0.57	0.63	0.64	0.67	0.71	0.68	0.67	0.68	0.63	0.66	0.60	0.68	0.65	0.71
Llama 3.2 3B	-0.23	0.28	0.34	0.28	0.34	0.33	0.36	0.36	0.41	0.40	0.38	0.39	0.41	0.33	0.37	0.32	0.34	0.38

Factcheck Year

