Respondents \ Judges	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	council (majority vote)	council (mean pooling)	dbrx-instruct	gemini-1.0-pro	gemini-1.5-pro	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	llama-3-70b-chat	llama-3-8b-chat	mistral-large	mistral-medium	mixtral-8x7b	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	council (no aggregation)
claude-3-haiku	0.43	0.28	0.62	0.18	0.32	0.22	0.30	0.37	0.30	0.25	0.22	0.07	0.10	0.05	0.31	0.58	0.07	0.38	0.41	0.14	0.28	0.18	0.26
claude-3-opus	0.65	0.39	0.63	0.46	0.47	0.46	0.53	0.50	0.53	0.42	0.52	0.33	0.26	0.30	0.50	0.50	0.40	0.64	0.72	0.44	0.59	0.83	0.49
claude-3-sonnet	0.49	0.26	0.57	0.26	0.30	0.33	0.34	0.31	0.25	0.23	0.43	0.15	0.10	0.14	0.38	0.66	0.19	0.43	0.53	0.26	0.48	0.45	0.33
command-r	0.17	0.03	0.43	0.11	0.27	0.20	0.23	0.19	0.36	0.25	0.25	0.18	0.12	0.14	0.13	0.41	0.07	0.29	0.47	0.15	0.20	0.08	0.20
command-r-plus	0.21	0.11	0.31	0.20	0.33	0.22	0.25	0.22	0.18	0.16	0.23	0.23	0.13	0.16	0.16	0.16	0.12	0.28	0.25	0.25	0.24	0.19	0.20
dbrx-instruct	0.09	0.27	0.43	0.32	0.22	0.31	0.31	0.31	0.20	0.27	0.38	0.22	0.23	0.23	0.21	0.45	0.22	0.27	0.37	0.19	0.40	0.27	0.28
gemini-1.0-pro	0.53	0.55	0.79	0.56	0.44	0.51	0.52	0.54	0.53	0.60	0.59	0.57	0.56	0.56	0.41	0.67	0.38	0.53	0.44	0.42	0.48	0.43	0.52
gemini-1.5-pro	0.01	0.00	0.02	0.01	0.01	0.00	0.01	0.00	0.06	0.00	0.03	0.00	0.02	0.00	0.00	0.20	0.00	0.02	0.01	0.01	0.00	0.01	0.02
gpt-3.5-turbo-0125	0.02	0.00	0.08	0.03	0.03	0.03	0.03	0.05	0.06	0.03	0.04	0.02	0.00	0.01	0.01	0.11	0.00	0.04	0.10	0.02	0.07	0.05	0.03
gpt-4-0613	0.22	0.10	0.15	0.05	0.03	0.14	0.17	0.16	0.16	0.10	0.23	0.07	0.04	0.06	0.04	0.14	0.04	0.12	0.23	0.10	0.12	0.09	0.11
gpt-4-turbo-2024-04-09	0.54	0.58	0.72	0.46	0.65	0.63	0.64	0.69	0.51	0.79	0.74	0.67	0.73	0.77	0.66	0.35	0.69	0.66	0.78	0.63	0.74	0.62	0.65
gpt-4o-2024-05-13	0.58	0.63	0.85	0.39	0.73	0.70	0.71	0.73	0.60	0.90	0.81	0.71	0.89	0.85	0.81	0.73	0.74	0.76	0.83	0.59	0.33	0.46	0.71
llama-3-70b-chat	0.58	0.16	0.41	0.32	0.37	0.35	0.42	0.21	0.24	0.33	0.34	0.25	0.22	0.30	0.62	0.88	0.17	0.49	0.79	0.28	0.24	0.26	0.38
llama-3-8b-chat	0.33	0.02	0.17	0.19	0.13	0.09	0.14	0.11	0.13	0.08	0.13	0.03	0.02	0.02	0.18	0.69	0.02	0.20	0.41	0.07	0.14	0.17	0.14
mistral-large	0.07	0.17	0.22	0.14	0.05	0.23	0.26	0.29	0.06	0.28	0.20	0.12	0.15	0.12	0.18	0.16	0.11	0.25	0.40	0.21	0.11	0.18	0.18
mistral-medium	0.08	0.06	0.18	0.12	0.09	0.12	0.13	0.18	0.07	0.10	0.20	0.10	0.00	0.05	0.08	0.24	0.10	0.13	0.31	0.13	0.26	0.08	0.13
mixtral-8x7b	0.25	0.16	0.29	0.17	0.21	0.23	0.24	0.14	0.16	0.18	0.27	0.07	0.12	0.13	0.12	0.27	0.13	0.28	0.35	0.23	0.34	0.23	0.20
qwen1.5-110B-Chat	0.81	0.85	0.81	0.54	0.80	0.82	0.81	0.87	0.78	0.83	0.88	0.85	0.85	0.85	0.89	0.85	0.91	0.84	0.83	0.80	0.93	0.86	0.83
qwen1.5-32B-Chat	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
qwen1.5-72B-Chat	0.44	0.46	0.61	0.42	0.53	0.47	0.48	0.45	0.50	0.60	0.38	0.41	0.29	0.54	0.46	0.53	0.54	0.41	0.49	0.46	0.54	0.54	0.48