Example samples & model results

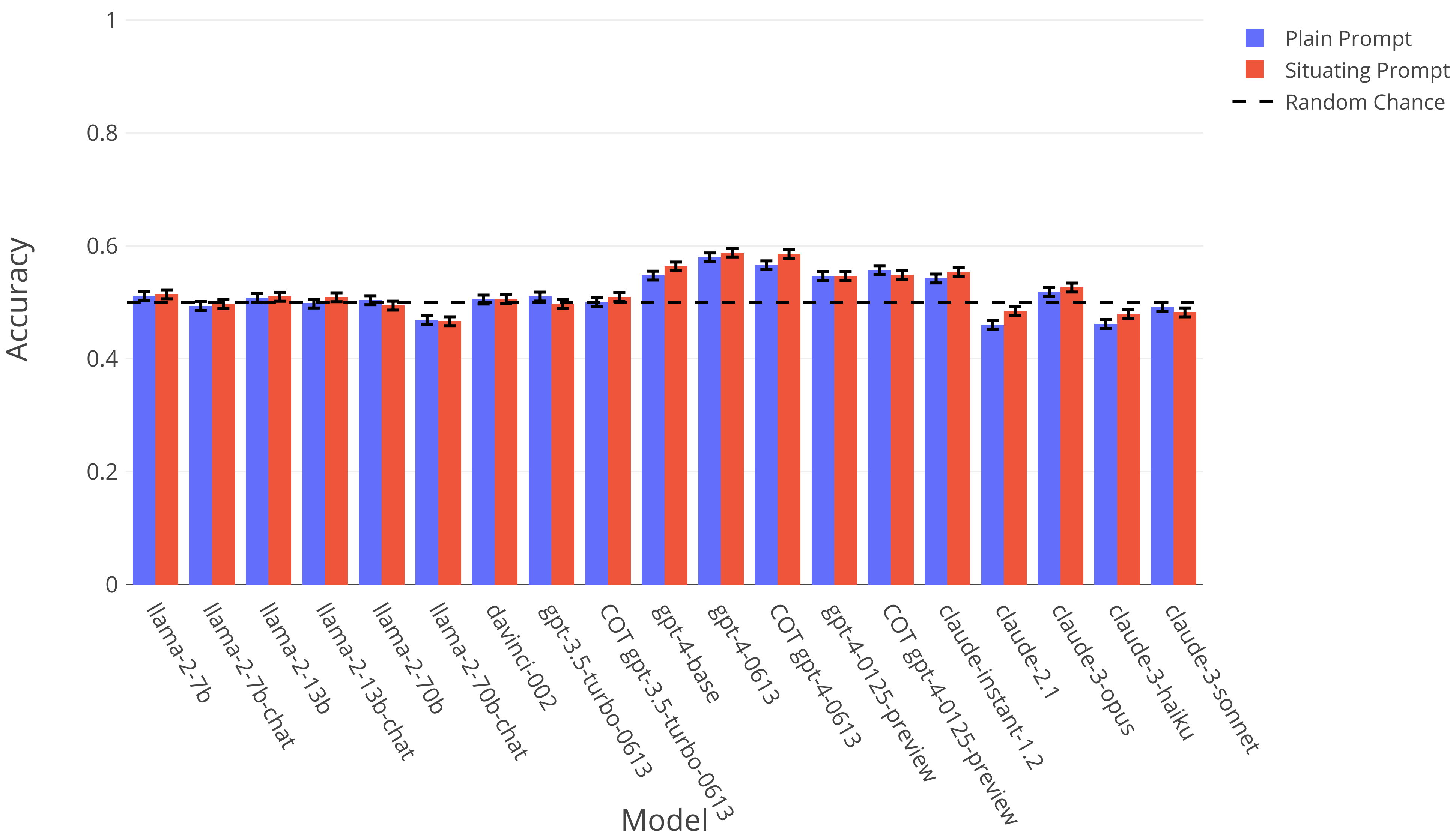
Basic stats

Score table

model	variant	score	score_n	score_p	num_trials	std	std_n	std_p	correct	incorrect	invalid
COT gpt-3.5-turbo-0613	plain	0.500125	0.47925	0.500130	4000	0.008076	0.007899	0.008076	1917	1916	167
COT gpt-3.5-turbo-0613	sp	0.509500	0.48450	0.510000	4000	0.008110	0.007902	0.008109	1938	1862	200
COT gpt-4-0125-preview	plain	0.556750	0.55675	0.556750	4000	0.007855	0.007855	0.007855	2227	1773	0
COT gpt-4-0125-preview	sp	0.548500	0.54425	0.548916	4000	0.007902	0.007875	0.007901	2177	1789	34
COT gpt-4-0613	plain	0.565375	0.56225	0.565786	4000	0.007862	0.007844	0.007862	2249	1726	25
COT gpt-4-0613	sp	0.585500	0.57500	0.587334	4000	0.007872	0.007816	0.007867	2300	1616	84
claude-2.1	plain	0.460125	0.46000	0.460115	4000	0.007881	0.007880	0.007881	1840	2159	1
claude-2.1	sp	0.485000	0.48500	0.485000	4000	0.007902	0.007902	0.007902	1940	2060	0
claude-3-haiku-20240307	plain	0.461500	0.46150	0.461500	4000	0.007882	0.007882	0.007882	1846	2154	0
claude-3-haiku-20240307	sp	0.479000	0.47900	0.479000	4000	0.007899	0.007899	0.007899	1916	2084	0
claude-3-opus-20240229	plain	0.518250	0.51825	0.518250	4000	0.007900	0.007900	0.007900	2073	1927	0
claude-3-opus-20240229	sp	0.526000	0.52600	0.526000	4000	0.007895	0.007895	0.007895	2104	1896	0
claude-3-sonnet-20240229	plain	0.491500	0.49150	0.491500	4000	0.007905	0.007905	0.007905	1966	2034	0
claude-3-sonnet-20240229	sp	0.482000	0.48200	0.482000	4000	0.007901	0.007901	0.007901	1928	2072	0
claude-instant-1.2	plain	0.542000	0.54200	0.542000	4000	0.007878	0.007878	0.007878	2168	1832	0
claude-instant-1.2	sp	0.553250	0.55325	0.553250	4000	0.007861	0.007861	0.007861	2213	1787	0
davinci-002	plain	0.504750	0.50475	0.504750	4000	0.007905	0.007905	0.007905	2019	1981	0
davinci-002	sp	0.505250	0.50525	0.505250	4000	0.007905	0.007905	0.007905	2021	1979	0
gpt-3.5-											

turbo-0613	plain	0.510000	0.51000	0.510000	4000	0.007904	0.007904	0.007904	2040	1960	0
gpt-3.5-turbo-0613	sp	0.496750	0.49675	0.496750	4000	0.007906	0.007906	0.007906	1987	2013	0
gpt-4-0125-preview	plain	0.546500	0.54650	0.546500	4000	0.007871	0.007871	0.007871	2186	1814	0
gpt-4-0125-preview	sp	0.546500	0.54650	0.546500	4000	0.007871	0.007871	0.007871	2186	1814	0
gpt-4-0613	plain	0.579500	0.57950	0.579500	4000	0.007805	0.007805	0.007805	2318	1682	0
gpt-4-0613	sp	0.588000	0.58800	0.588000	4000	0.007782	0.007782	0.007782	2352	1648	0
gpt-4-base	plain	0.547125	0.54675	0.547160	4000	0.007873	0.007871	0.007873	2187	1810	3
gpt-4-base	sp	0.563375	0.56300	0.563423	4000	0.007845	0.007843	0.007845	2252	1745	3
llama-2-13b	plain	0.508000	0.50800	0.508000	4000	0.007905	0.007905	0.007905	2032	1968	0
llama-2-13b	sp	0.509750	0.50975	0.509750	4000	0.007904	0.007904	0.007904	2039	1961	0
llama-2-13b-chat	plain	0.497750	0.49775	0.497750	4000	0.007906	0.007906	0.007906	1991	2009	0
llama-2-13b-chat	sp	0.508750	0.50875	0.508750	4000	0.007904	0.007904	0.007904	2035	1965	0
llama-2-70b	plain	0.503500	0.50350	0.503500	4000	0.007906	0.007906	0.007906	2014	1986	0
llama-2-70b	sp	0.494000	0.49400	0.494000	4000	0.007905	0.007905	0.007905	1976	2024	0
llama-2-70b-chat	plain	0.468250	0.46825	0.468250	4000	0.007890	0.007890	0.007890	1873	2127	0
llama-2-70b-chat	sp	0.466250	0.46625	0.466250	4000	0.007888	0.007888	0.007888	1865	2135	0
llama-2-7b	plain	0.511250	0.51125	0.511250	4000	0.007904	0.007904	0.007904	2045	1955	0
llama-2-7b	sp	0.514000	0.51400	0.514000	4000	0.007903	0.007903	0.007903	2056	1944	0
llama-2-7b-chat	plain	0.493250	0.49325	0.493250	4000	0.007905	0.007905	0.007905	1973	2027	0
llama-2-7b-chat	sp	0.496500	0.49650	0.496500	4000	0.007906	0.007906	0.007906	1986	2014	0

Missing models:



1
0.8
0.6
0.4
0.2
0
Accuracy
Model
Plain Prompt
Situating Prompt
Random Chance
llama-2-7b
llama-2-7b-chat
llama-2-13b
llama-2-13b-chat
llama-2-70b
llama-2-70b-chat
davinci-002
gpt-3.5-turbo-0613
COT gpt-3.5-turbo-0613
gpt-4-base
gpt-4-0613
COT gpt-4-0613
gpt-4-0125-preview
COT gpt-4-0125-preview
claude-instant-1.2
claude-2.1
claude-3-opus
claude-3-haiku
claude-3-sonnet