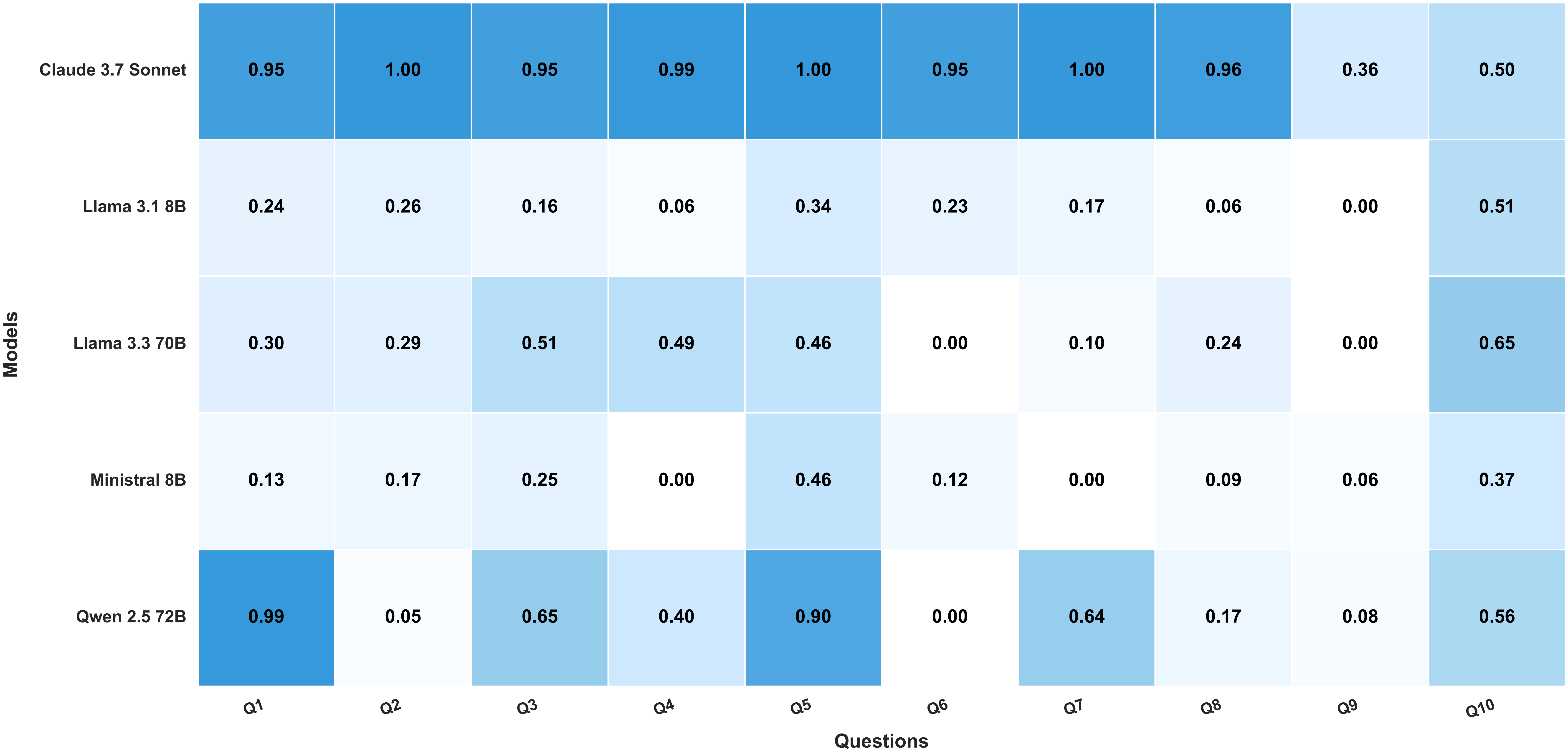
Average Factual Correctness Score by Model and Question (Q11-Q20)



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Claude 3.7 Sonnet	0.95	1.00	0.95	0.99	1.00	0.95	1.00	0.96	0.36	0.50
Llama 3.1 8B	0.24	0.26	0.16	0.06	0.34	0.23	0.17	0.06	0.00	0.51
Llama 3.3 70B	0.30	0.29	0.51	0.49	0.46	0.00	0.10	0.24	0.00	0.65
Ministral 8B	0.13	0.17	0.25	0.00	0.46	0.12	0.00	0.09	0.06	0.37
Qwen 2.5 72B	0.99	0.05	0.65	0.40	0.90	0.00	0.64	0.17	0.08	0.56