

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

Models

Claude 3.7 Sonnet

0.952

1.000

0.950

0.995

1.000

0.950

0.999

0.959

0.360

0.500

Llama 3.1 8B

0.240

0.257

0.160

0.056

0.337

0.230

0.170

0.060

0.000

0.513

Llama 3.3 70B

0.298

0.287

0.509

0.492

0.457

0.000

0.101

0.239

0.002

0.653

Minstral 8B

0.135

0.171

0.246

0.000

0.455

0.124

0.000

0.092

0.061

0.369

Qwen 2.5 72B

0.990

0.045

0.647

0.399

0.900

0.000

0.640

0.170

0.083

0.560

Q11

Q12

Q13

Q14

Q15

Q16

Q17

Q18

Q19

Q20

Questions