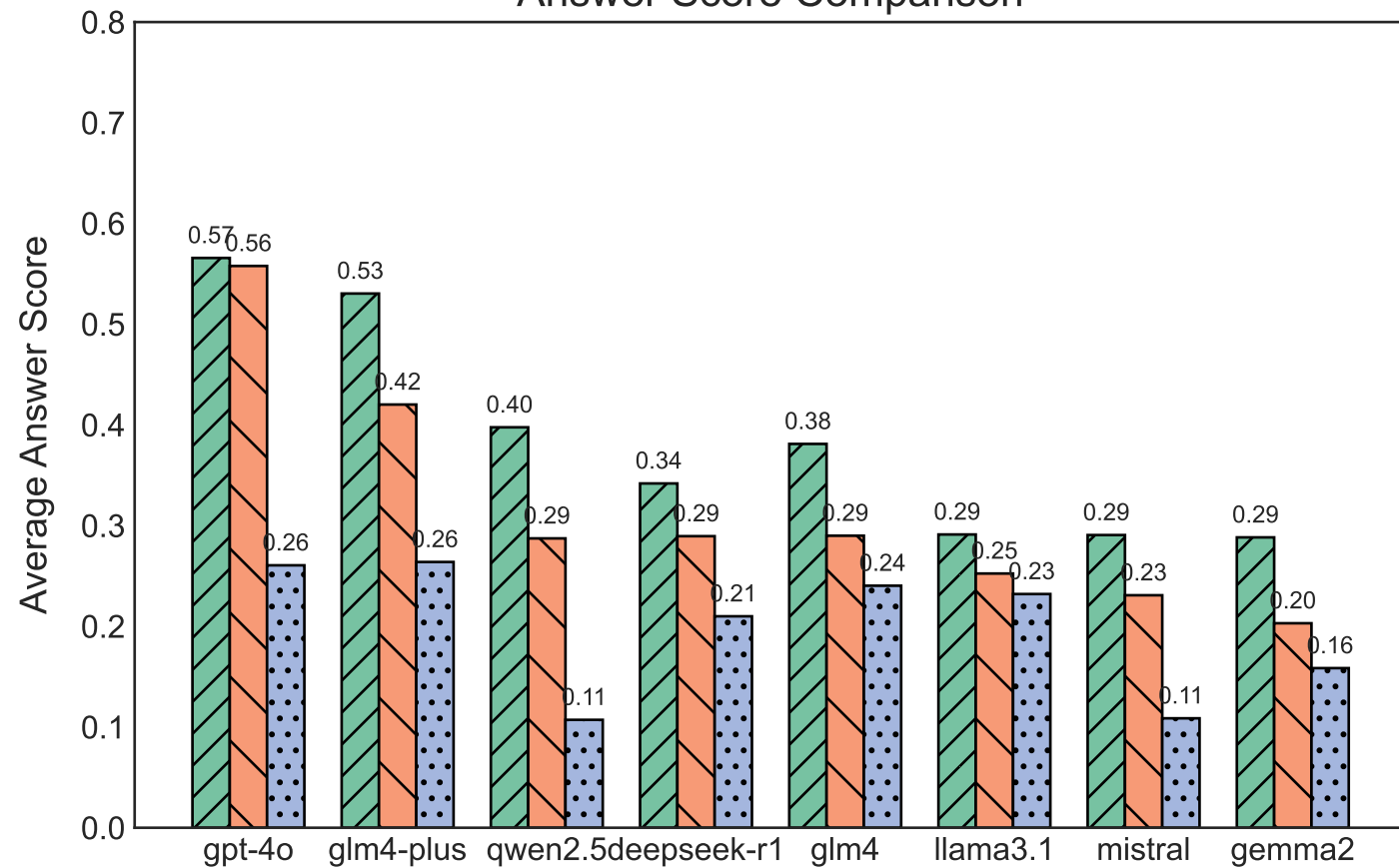# Model Performance Across Different Experimental Conditions

Legend: Answering with Ground Truth Conditions | Answering with Model-generated Conditions | Answering without Conditions

Answer Score Comparison

Model	Ground Truth	Model-generated	Without
gpt-4o	0.57	0.56	0.26
glm4-plus	0.53	0.42	0.26
qwen2.5	0.40	0.29	0.11
deepseek-r1	0.34	0.29	0.21
glm4	0.38	0.29	0.24
llama3.1	0.29	0.25	0.23
mistral	0.29	0.23	0.11
gemma2	0.29	0.20	0.16

Citation Score Comparison

Model	Ground Truth	Model-generated	Without
gpt-4o	0.96	0.87	0.38
glm4-plus	0.92	0.44	0.36
qwen2.5	0.70	0.56	0.08
deepseek-r1	0.61	0.49	0.28
glm4	0.80	0.32	0.31
llama3.1	0.71	0.31	0.26
mistral	0.69	0.26	0.05
gemma2	0.38	0.22	0.17