Enhanced Model Performance Statistics

Model	Type	Cond. Score	Ans. Score	Cite. Score	Combined	Avg. Citations	Avg. Answers	Diff.	Rank
gpt-4o	API	0.552 ± 0.190	0.558 ± 0.157	0.875 ± 0.207	**0.662**	2.72	1.74	-0.17	#1
glm4-plus	API	0.302 ± 0.069	0.420 ± 0.097	0.441 ± 0.261	**0.388**	2.77	2.92	1.01	#2
qwen2.5	Local	0.235 ± 0.120	0.287 ± 0.161	0.558 ± 0.359	0.360	2.17	1.46	-0.45	#3
deepseek-r1	Local	0.245 ± 0.112	0.293 ± 0.142	0.501 ± 0.342	0.346	1.92	2.28	0.36	#4
glm4	Local	0.231 ± 0.071	0.290 ± 0.090	0.320 ± 0.215	0.280	1.92	2.99	1.08	#5
llama3.1	Local	0.232 ± 0.076	0.252 ± 0.093	0.306 ± 0.246	0.264	1.82	2.85	0.94	#6
mistral	Local	0.196 ± 0.060	0.231 ± 0.079	0.263 ± 0.214	0.230	1.91	3.00	1.09	#7
gemma2	Local	0.170 ± 0.091	0.203 ± 0.118	0.217 ± 0.277	0.197	2.30	2.06	0.14	#8
Average	**-**	**0.270**	**0.317**	**0.435**	**0.341**	**2.19**	**2.41**	**0.50**	**-**

API Models: GPT-4o, glm4-Plus

Local Models: Llama3.1, Mistral, Gemma2, Qwen2.5, GLM4

Table: Main experiment scores with balanced scoring methodology.
Balanced scores account for both quality and consistency with expected answer count.