Model Ranking Based on Balanced Combined Score

Rank	Model	Combined Score (Balanced)
#1	gpt-4o	0.661
#2	glm4-plus	0.388
#3	qwen2.5	0.360
#4	deepseek-r1	0.342
#5	glm4	0.280
#6	llama3.1	0.264
#7	mistral	0.230
#8	gemma2	0.197

Average: 0.340

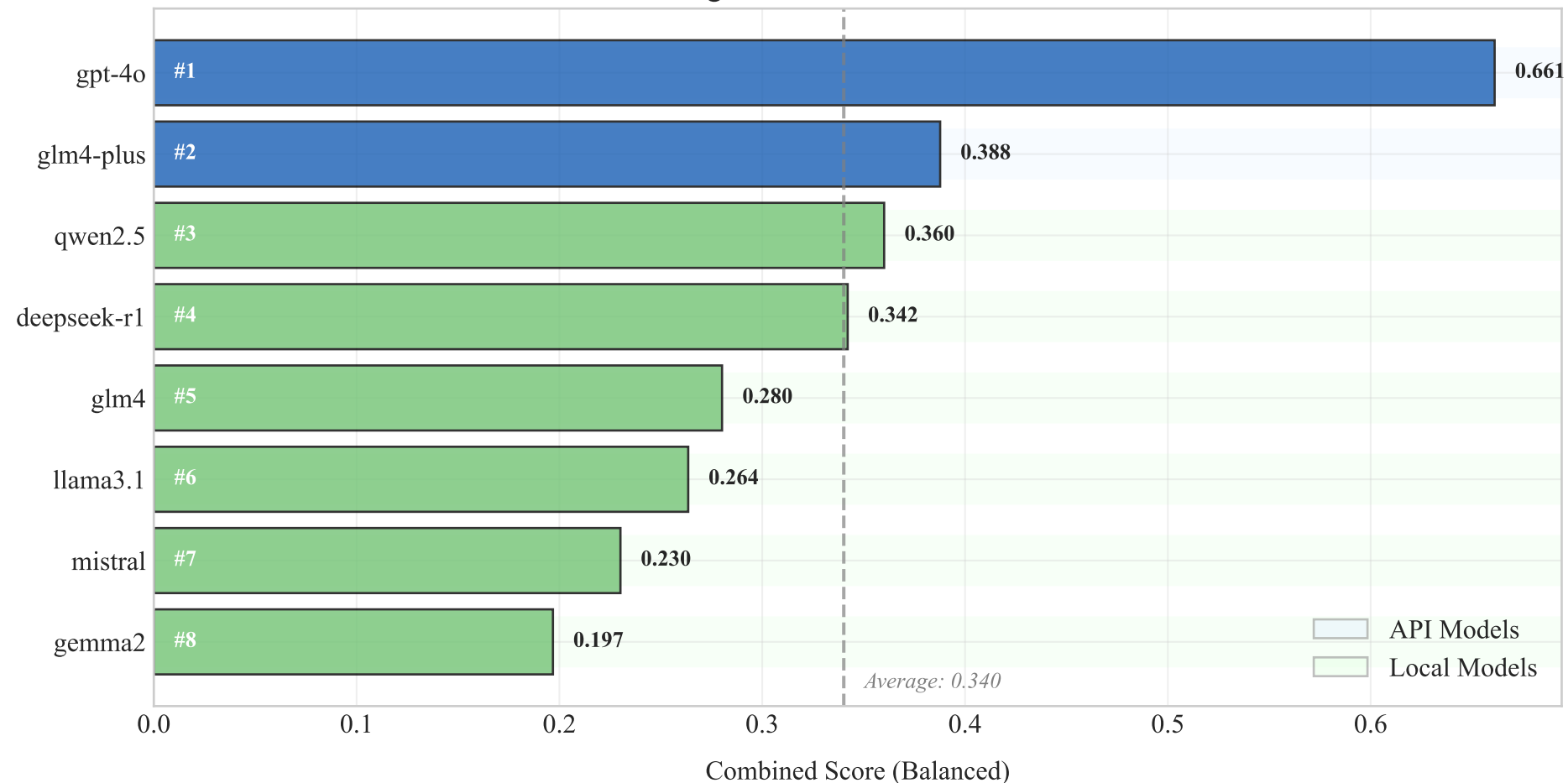API Models
Local Models

Combined Score = Average of Balanced Condition, Answer, and Citation Scores
Balanced scoring accounts for both quality and quantity of generated answers