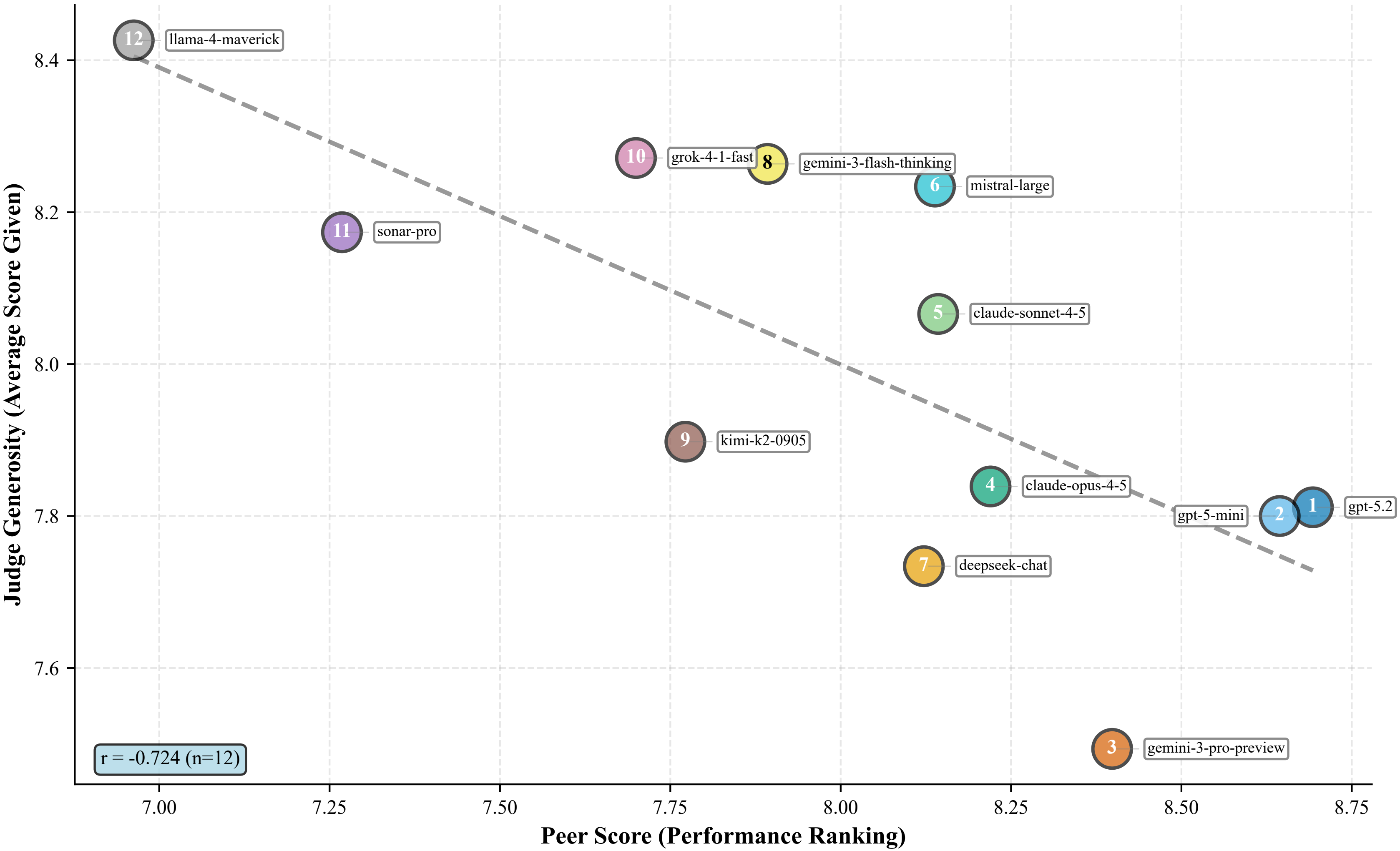
Judge Generosity vs Peer Ranking
Do better models judge more harshly?


Judge Generosity (Average Score Given)
Peer Score (Performance Ranking)
12 llama-4-maverick
10 grok-4-1-fast
8 gemini-3-flash-thinking
6 mistral-large
11 sonar-pro
5 claude-sonnet-4-5
9 kimi-k2-0905
4 claude-opus-4-5
gpt-5-mini 2
1 gpt-5.2
7 deepseek-chat
3 gemini-3-pro-preview
r = -0.724 (n=12)