Judge Generosity
(higher = more lenient)

Model	Average Score Given (± std)
llama-4-maverick	8.43
grok-4-1-fast	8.27
gemini-3-flash-thinking	8.26
mistral-large	8.23
sonar-pro	8.17
claude-sonnet-4-5	8.07
kimi-k2-0905	7.90
claude-opus-4-5	7.84
gpt-5.2	7.81
gpt-5-mini	7.80
deepseek-chat	7.73
gemini-3-pro-preview	7.49

--- Mean: 8.00