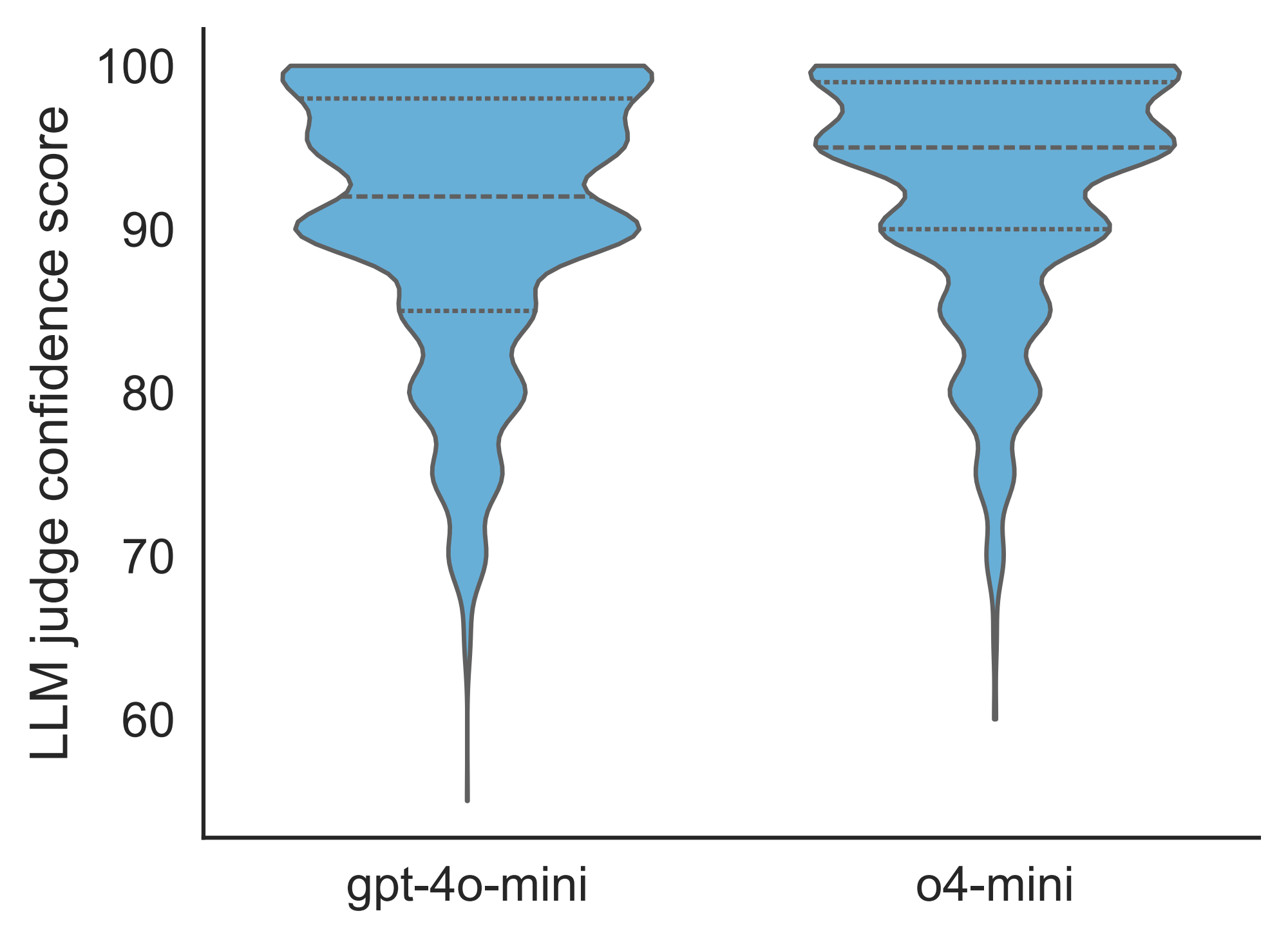
LLM judge confidence score
100
90
80
70
60
gpt-4o-mini
o4-mini