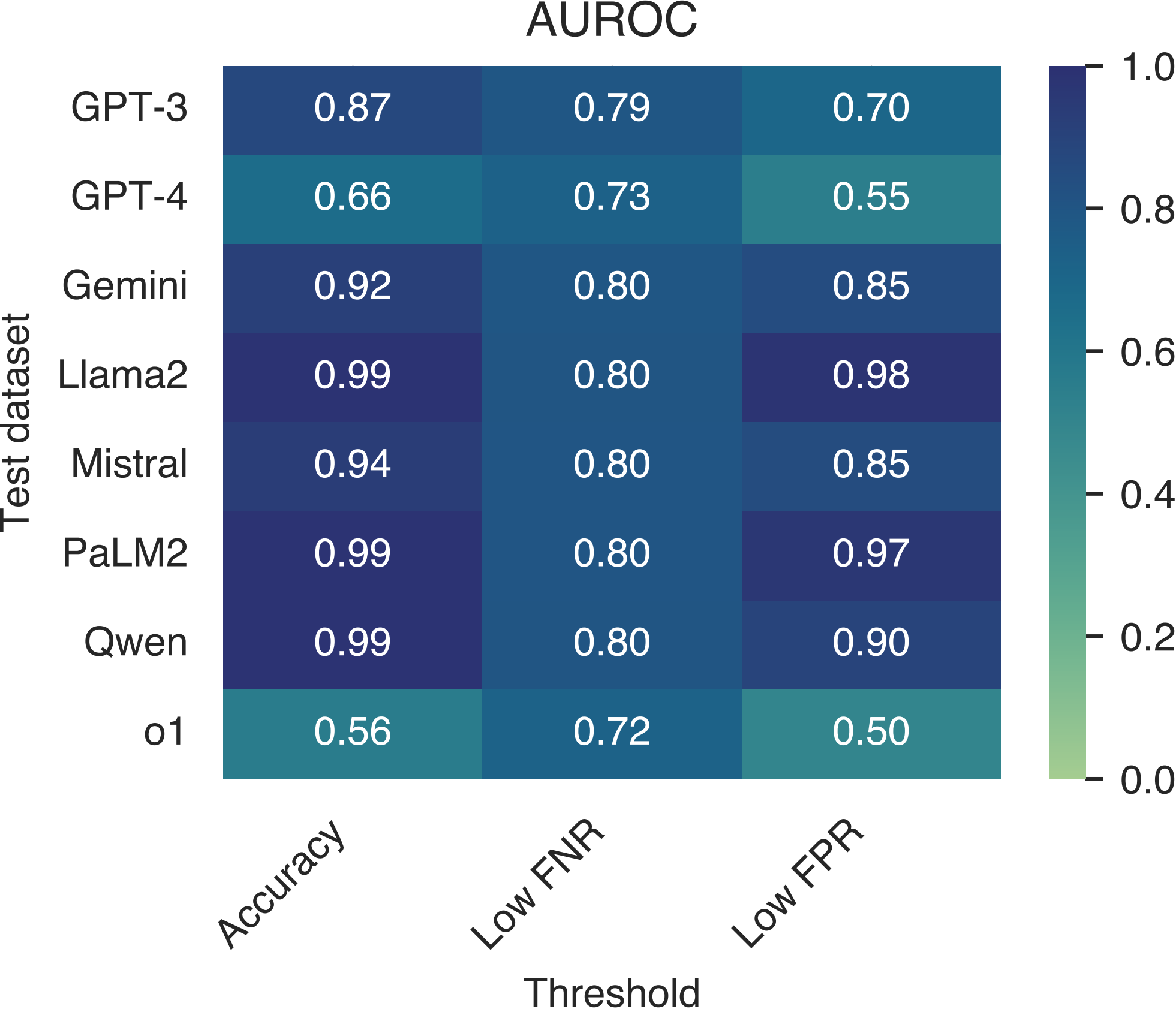
AUROC
GPT-3
GPT-4
Gemini
Llama2
Mistral
PaLM2
Qwen
o1
Test dataset
0.87
0.79
0.70
0.66
0.73
0.55
0.92
0.80
0.85
0.99
0.80
0.98
0.94
0.80
0.85
0.99
0.80
0.97
0.99
0.80
0.90
0.56
0.72
0.50
Accuracy
Low FNR
Low FPR
Threshold
1.0
0.8
0.6
0.4
0.2
0.0