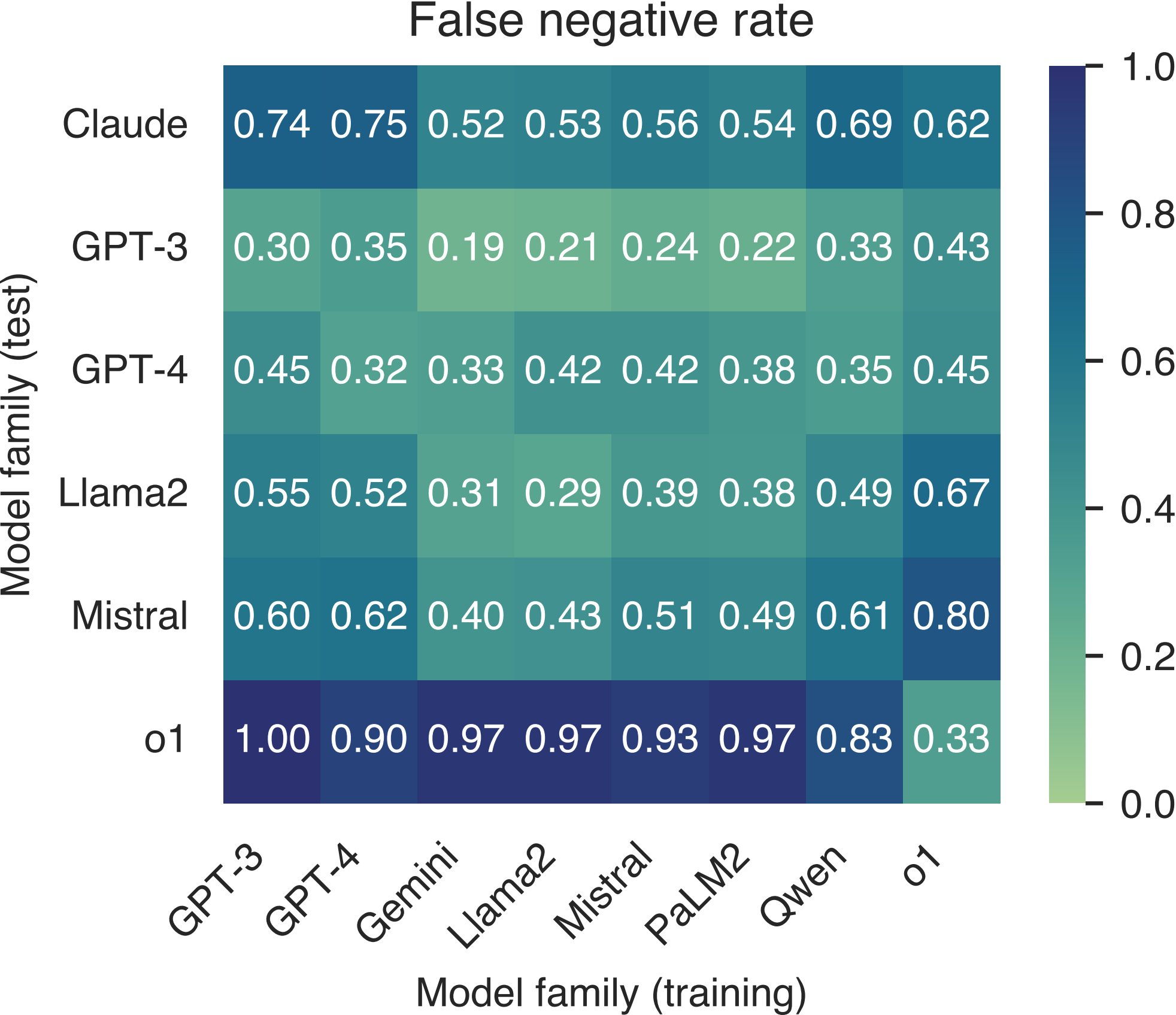
False negative rate
Claude
GPT-3
GPT-4
Llama2
Mistral
o1
Model family (test)
0.74 0.75 0.52 0.53 0.56 0.54 0.69 0.62
0.30 0.35 0.19 0.21 0.24 0.22 0.33 0.43
0.45 0.32 0.33 0.42 0.42 0.38 0.35 0.45
0.55 0.52 0.31 0.29 0.39 0.38 0.49 0.67
0.60 0.62 0.40 0.43 0.51 0.49 0.61 0.80
1.00 0.90 0.97 0.97 0.93 0.97 0.83 0.33
GPT-3
GPT-4
Gemini
Llama2
Mistral
PaLM2
Qwen
o1
Model family (training)
1.0
0.8
0.6
0.4
0.2
0.0