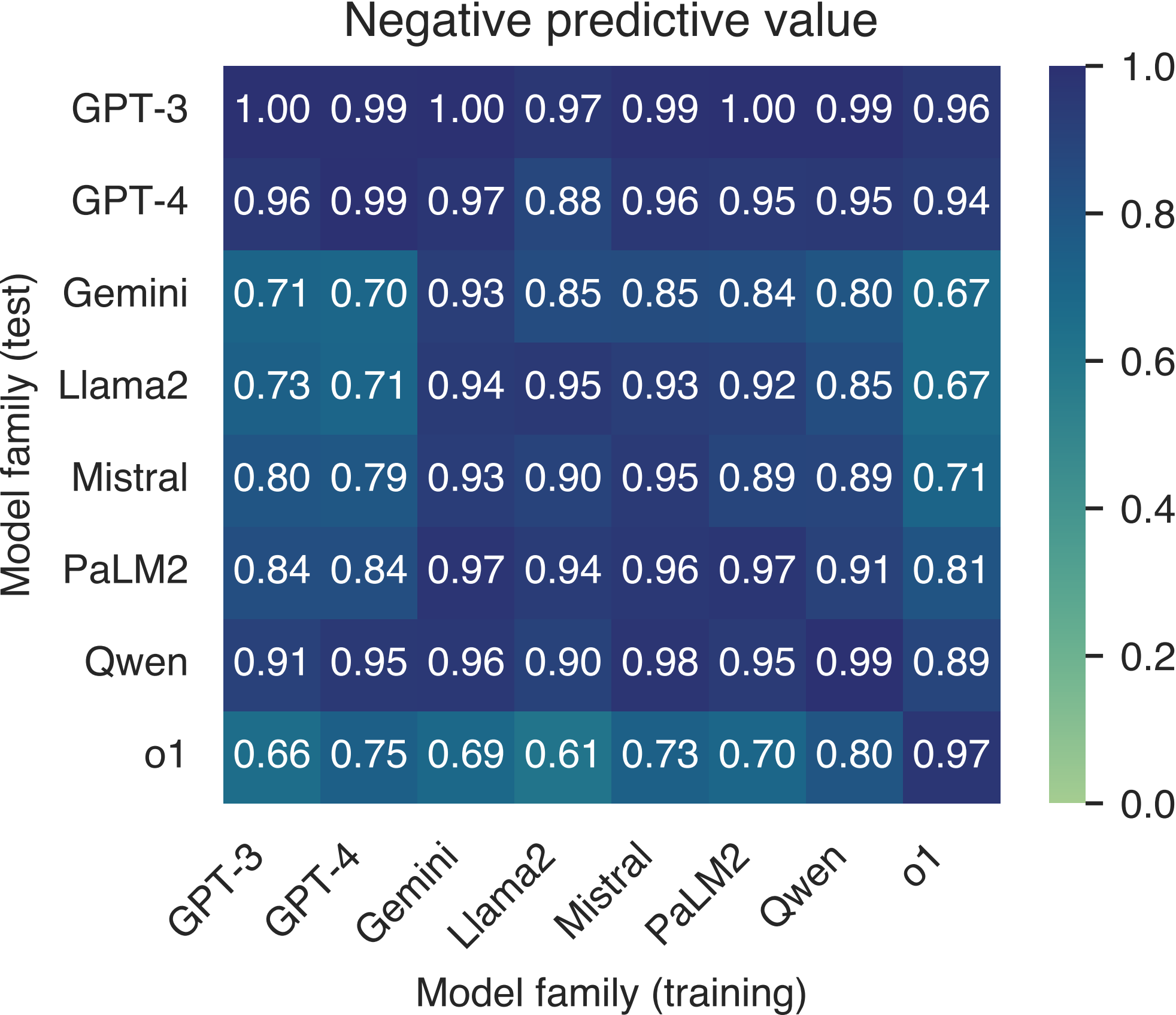
Negative predictive value
Model family (test)
GPT-3
GPT-4
Gemini
Llama2
Mistral
PaLM2
Qwen
o1
1.00 0.99 1.00 0.97 0.99 1.00 0.99 0.96
0.96 0.99 0.97 0.88 0.96 0.95 0.95 0.94
0.71 0.70 0.93 0.85 0.85 0.84 0.80 0.67
0.73 0.71 0.94 0.95 0.93 0.92 0.85 0.67
0.80 0.79 0.93 0.90 0.95 0.89 0.89 0.71
0.84 0.84 0.97 0.94 0.96 0.97 0.91 0.81
0.91 0.95 0.96 0.90 0.98 0.95 0.99 0.89
0.66 0.75 0.69 0.61 0.73 0.70 0.80 0.97
1.0
0.8
0.6
0.4
0.2
0.0
GPT-3
GPT-4
Gemini
Llama2
Mistral
PaLM2
Qwen
o1
Model family (training)