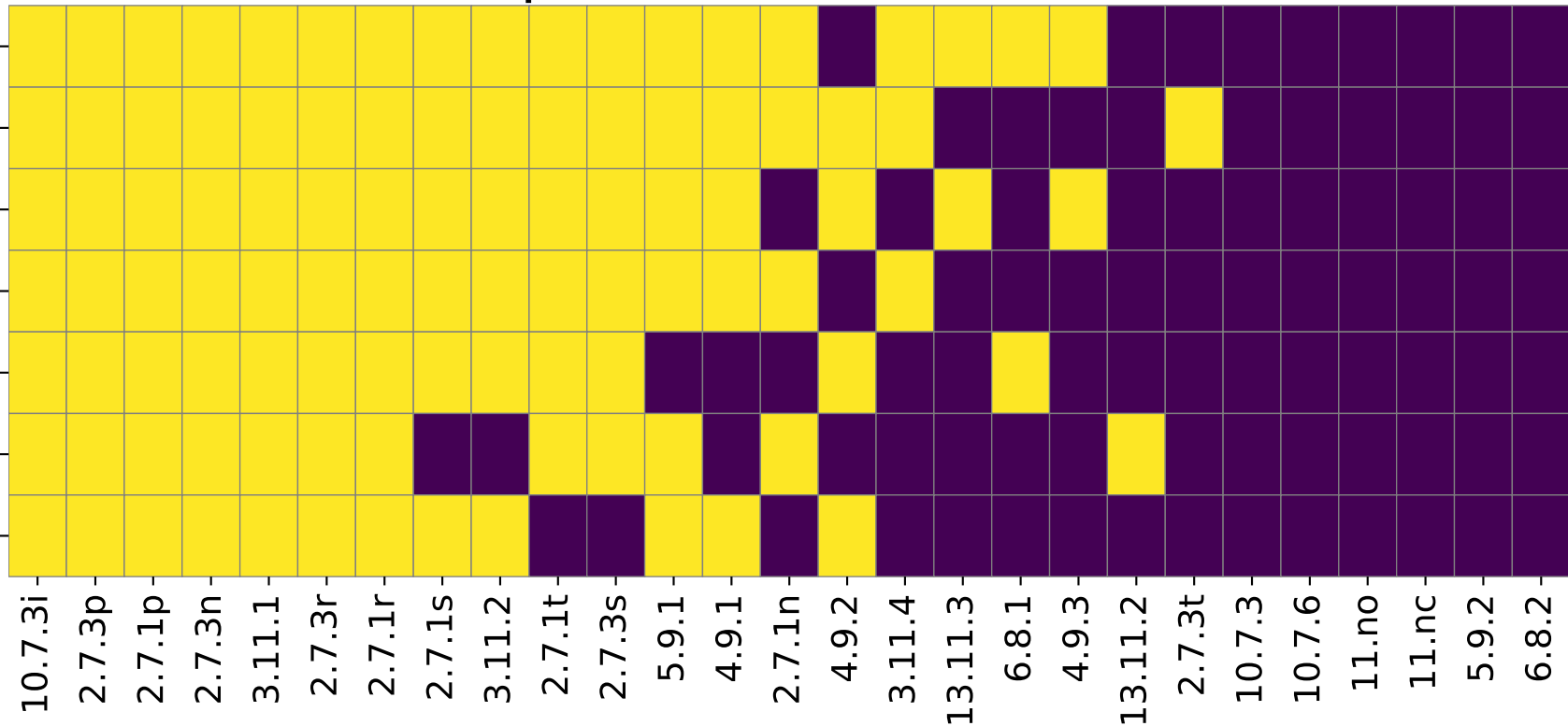


Success Rate per Model and Exercise (sorted)

Model

Claude Sonnet 4
GPT-4.1
Gemini 2.5 Flash (thinking)
Grok 3
Grok 3 Mini
Gemini 2.5 Flash
o4-mini



Exercise

Success Rate (%)

