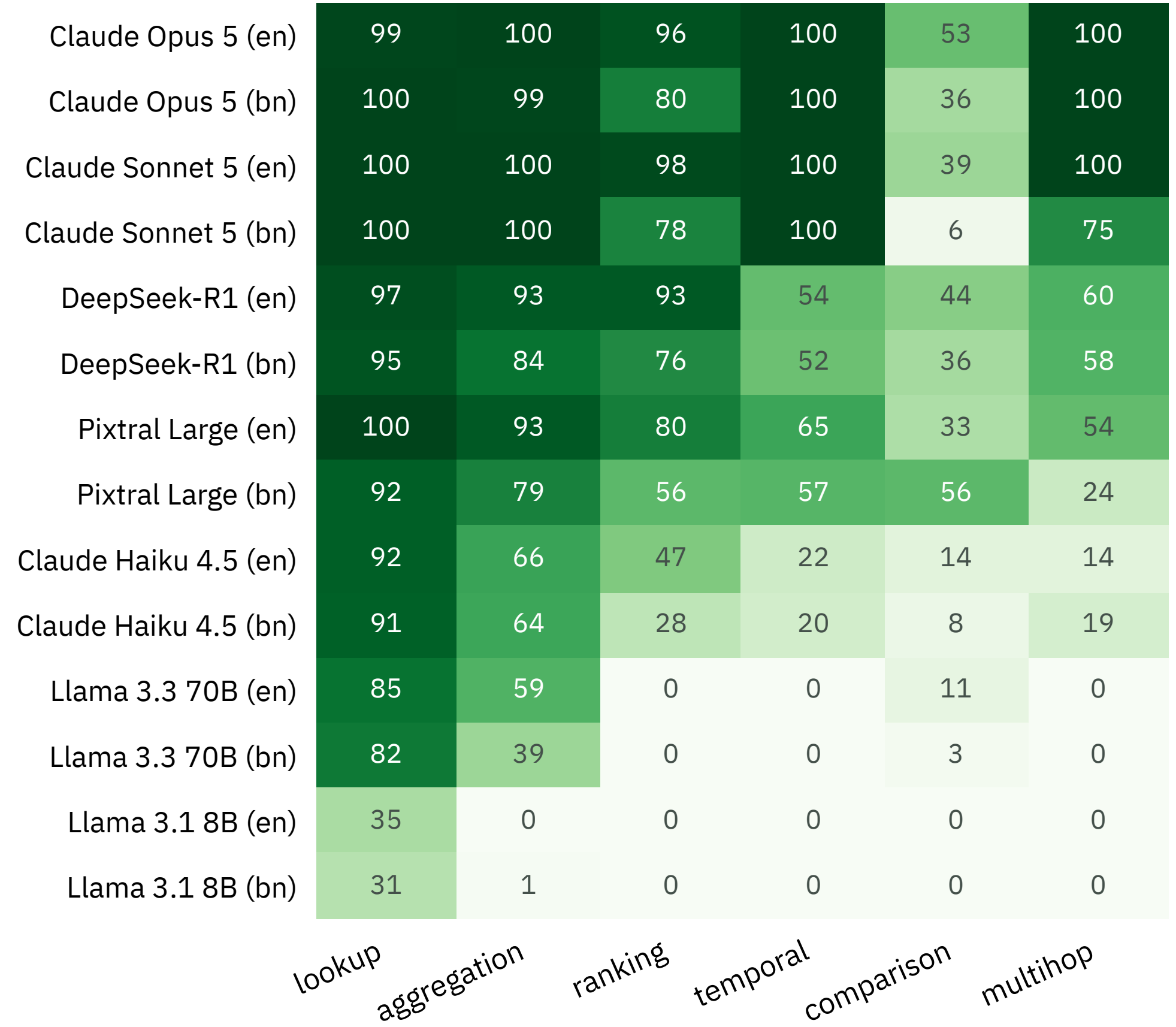

Claude Opus 5 (en) 99 100 96 100 53 100
Claude Opus 5 (bn) 100 99 80 100 36 100
Claude Sonnet 5 (en) 100 100 98 100 39 100
Claude Sonnet 5 (bn) 100 100 78 100 6 75
DeepSeek-R1 (en) 97 93 93 54 44 60
DeepSeek-R1 (bn) 95 84 76 52 36 58
Pixtral Large (en) 100 93 80 65 33 54
Pixtral Large (bn) 92 79 56 57 56 24
Claude Haiku 4.5 (en) 92 66 47 22 14 14
Claude Haiku 4.5 (bn) 91 64 28 20 8 19
Llama 3.3 70B (en) 85 59 0 0 11 0
Llama 3.3 70B (bn) 82 39 0 0 3 0
Llama 3.1 8B (en) 35 0 0 0 0 0
Llama 3.1 8B (bn) 31 1 0 0 0 0
lookup
aggregation
ranking
temporal
comparison
multihop