

Performance by Task Difficulty Level

Averaged Across All Models, Context Lengths, and Cache Configs

