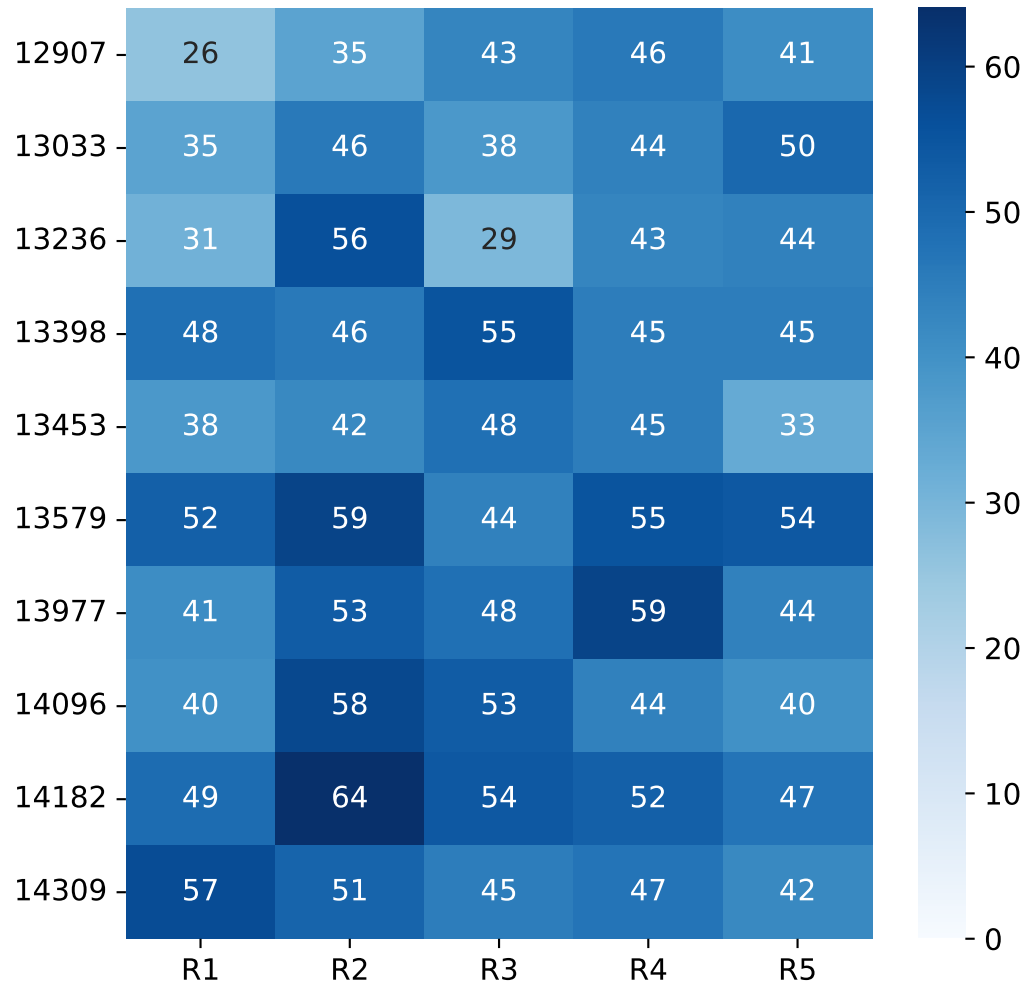
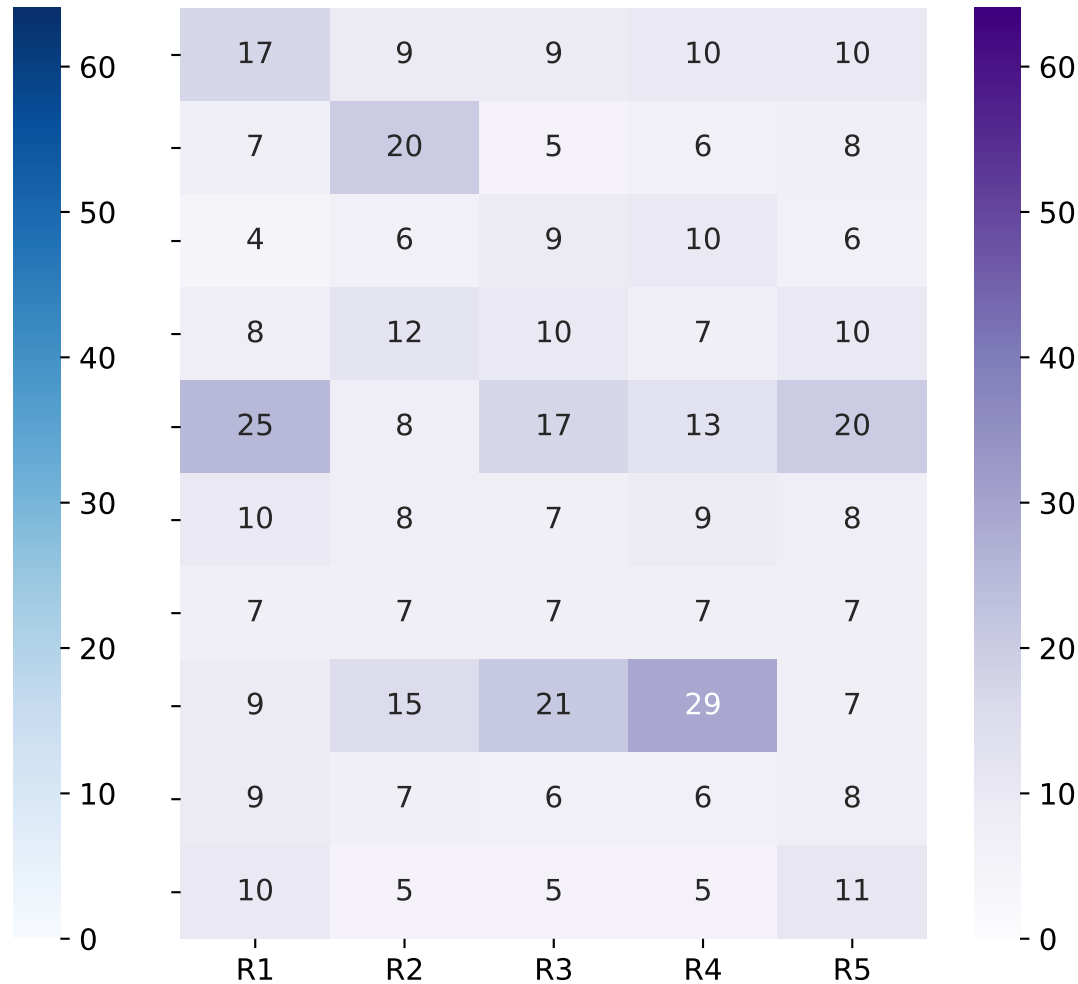


Step Counts: 10 Tasks × 5 Runs

Claude ($\mu=46.1$)



GPT-5 ($\mu=9.9$)



Llama ($\mu=17.0$)

