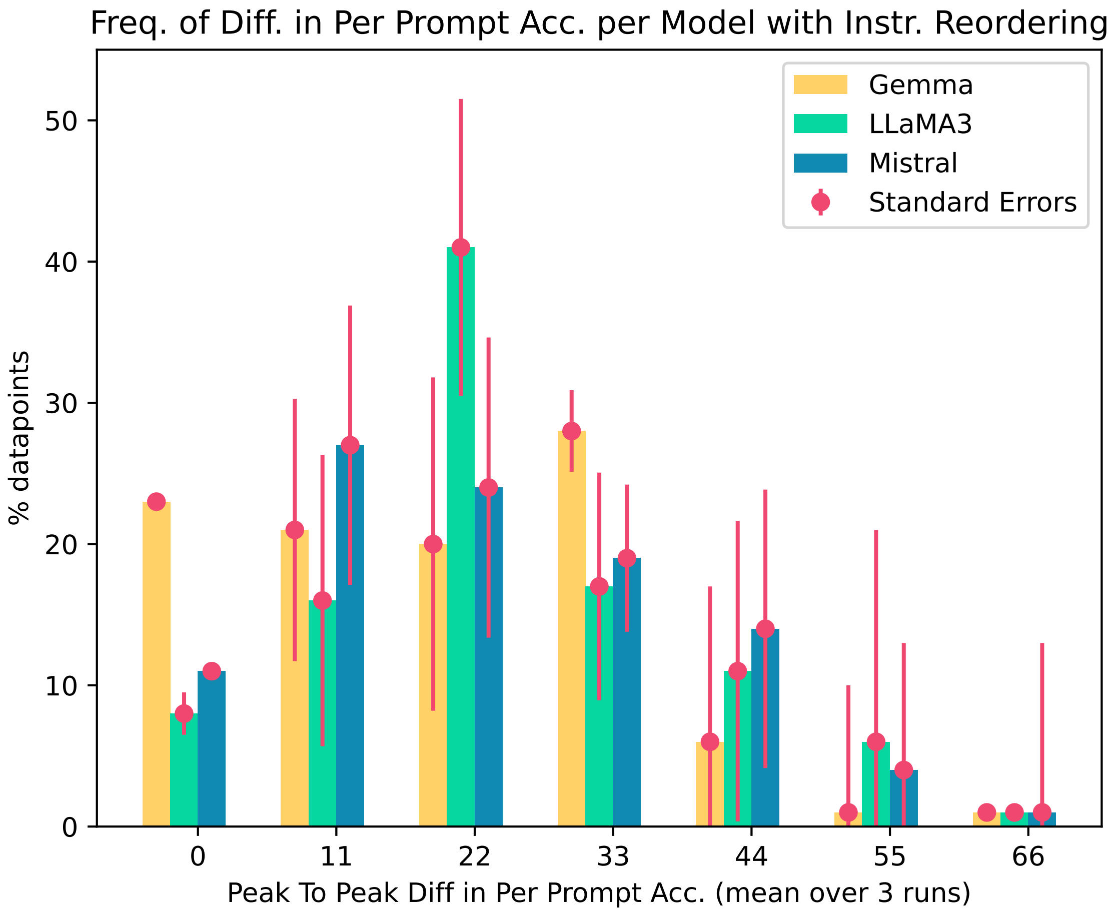
Freq. of Diff. in Per Prompt Acc. per Model with Instr. Reordering
Gemma
LLaMA3
Mistral
Standard Errors
% datapoints
0
10
20
30
40
50
0
11
22
33
44
55
66
Peak To Peak Diff in Per Prompt Acc. (mean over 3 runs)