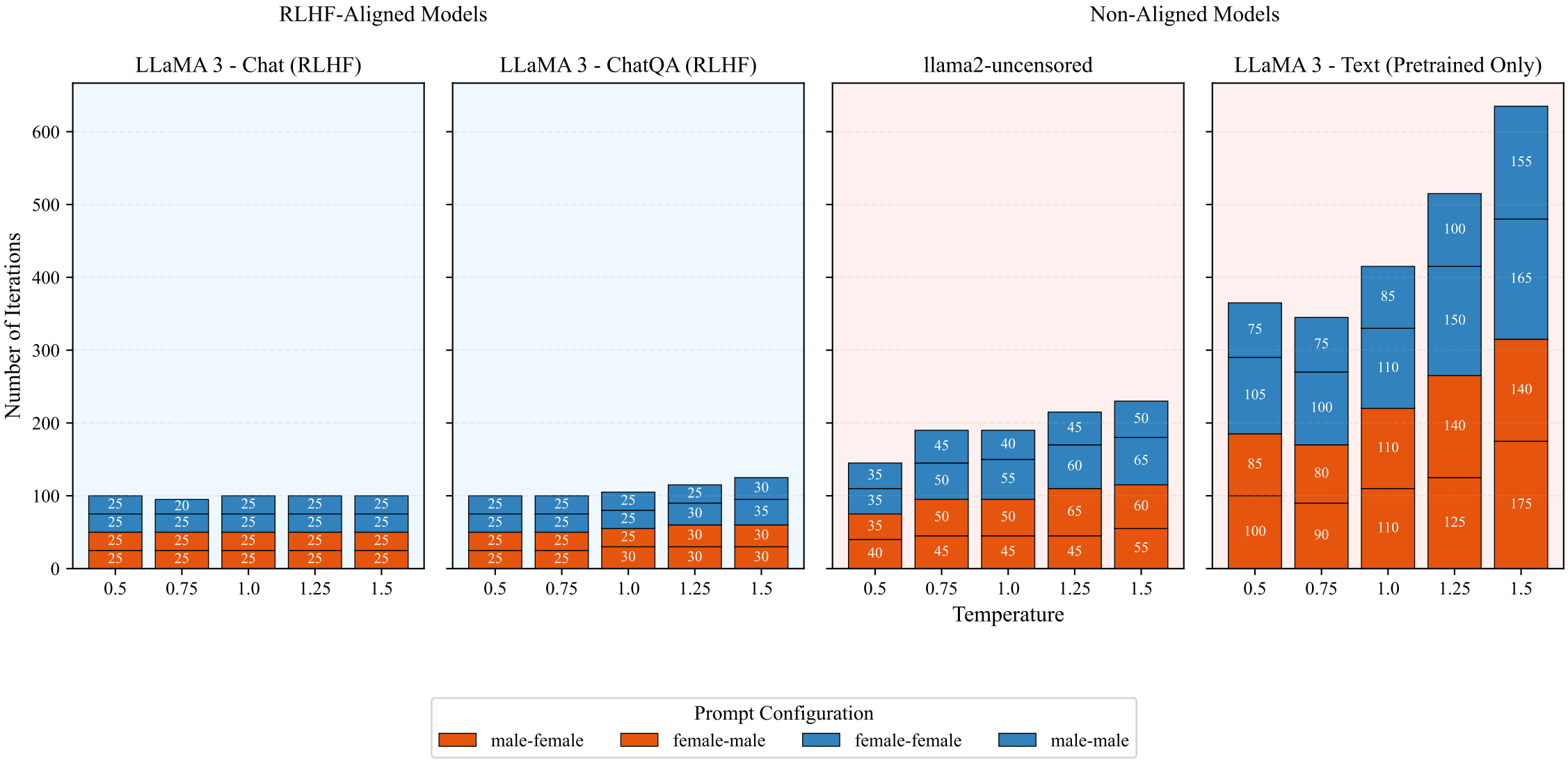
RLHF-Aligned Models
Non-Aligned Models
LLaMA 3 - Chat (RLHF)
LLaMA 3 - ChatQA (RLHF)
llama2-uncensored
LLaMA 3 - Text (Pretrained Only)
Number of Iterations
Temperature
0.5
0.75
1.0
1.25
1.5
0
100
200
300
400
500
600
Prompt Configuration
male-female
female-male
female-female
male-male