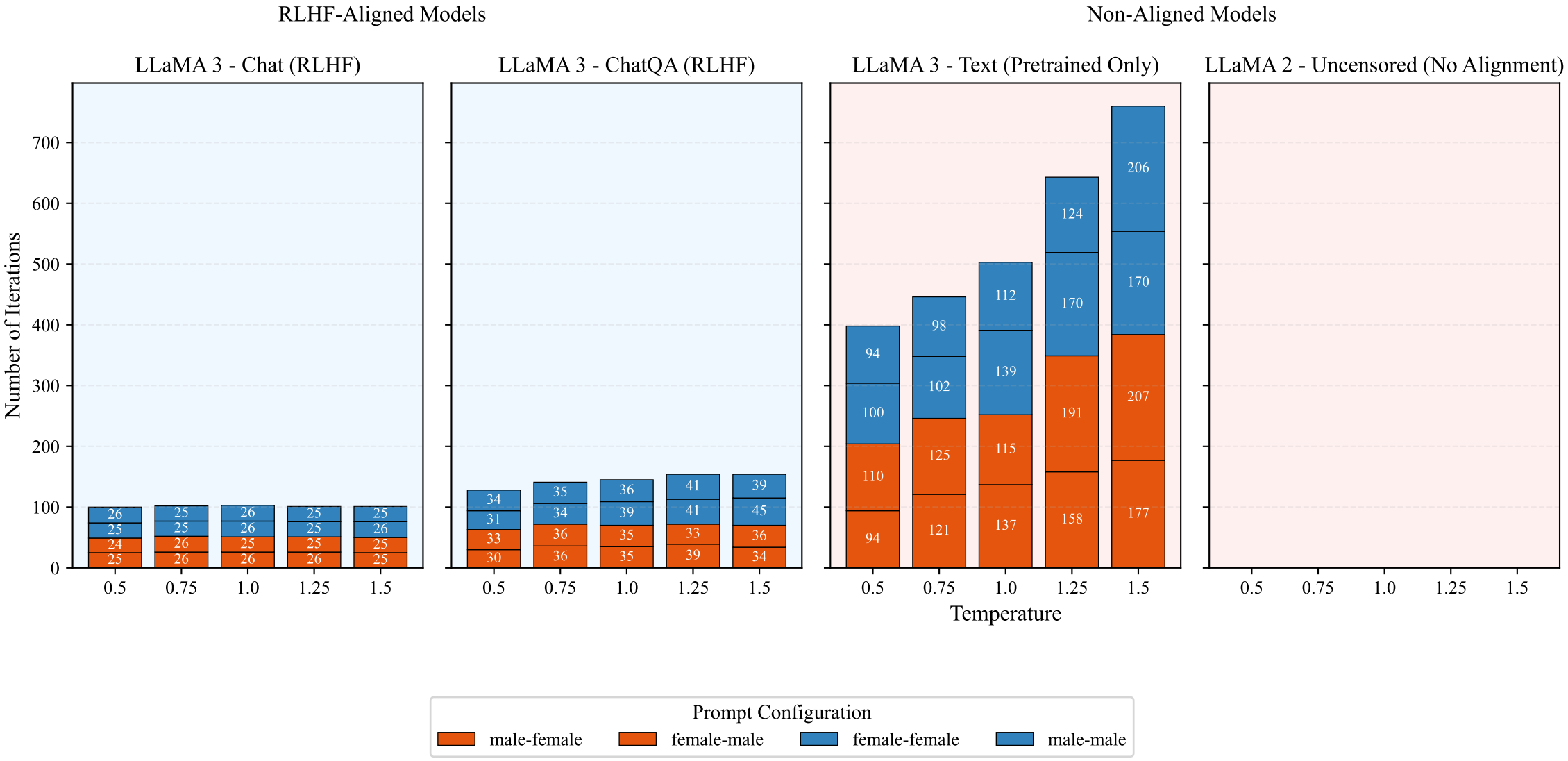
RLHF-Aligned Models
Non-Aligned Models
LLaMA 3 - Chat (RLHF)
LLaMA 3 - ChatQA (RLHF)
LLaMA 3 - Text (Pretrained Only)
LLaMA 2 - Uncensored (No Alignment)
Number of Iterations
Temperature
0.5
0.75
1.0
1.25
1.5
0
100
200
300
400
500
600
700
26 25 24 25
25 25 26 26
26 26 25 26
25 25 25 26
25 26 25 25
34 31 33 30
35 34 36 36
36 39 35 35
41 41 33 39
39 45 36 34
94 100 110 94
98 102 125 121
112 139 115 137
124 170 191 158
206 170 207 177
Prompt Configuration
male-female
female-male
female-female
male-male