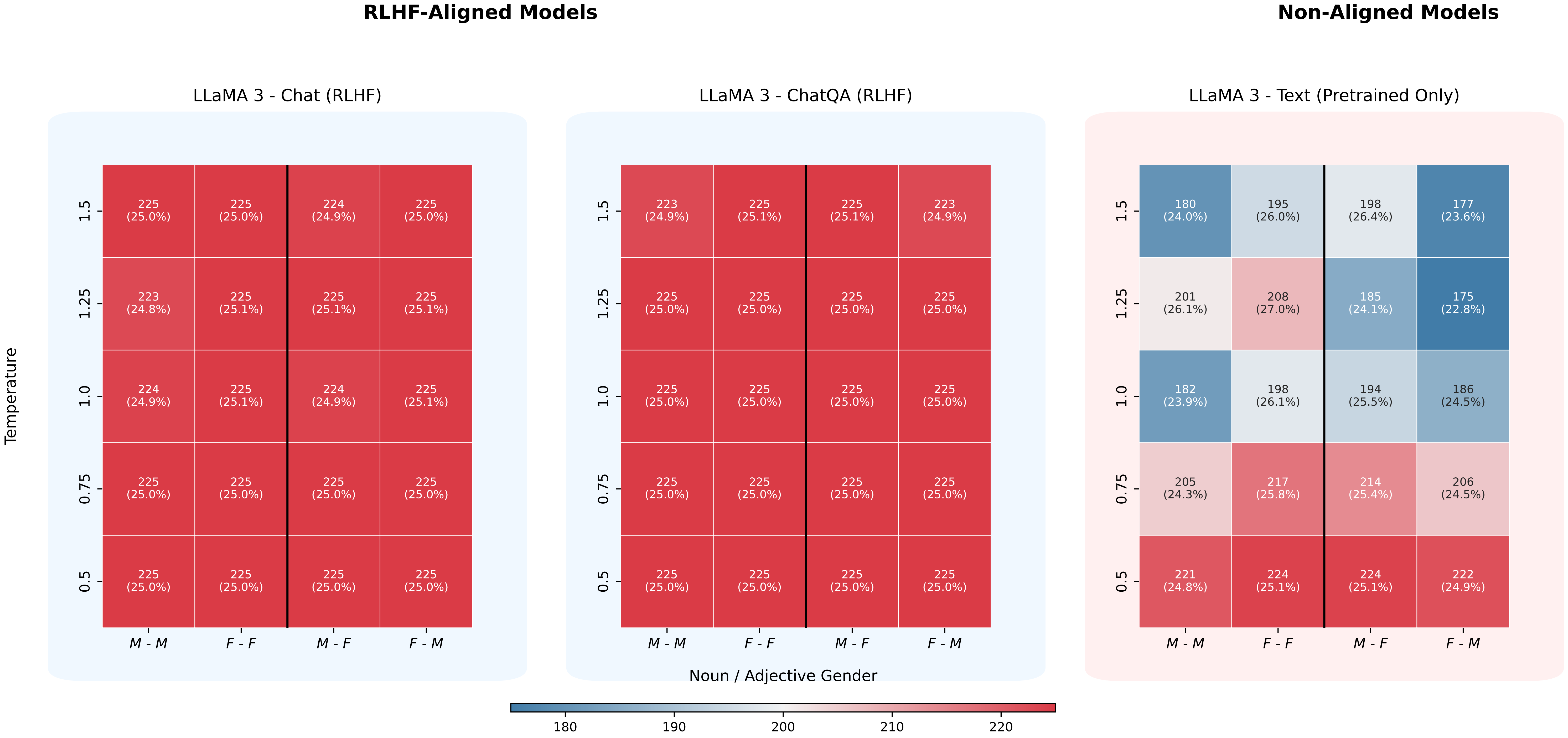
RLHF-Aligned Models
Non-Aligned Models
LLaMA 3 - Chat (RLHF)
LLaMA 3 - ChatQA (RLHF)
LLaMA 3 - Text (Pretrained Only)
Temperature
1.5
1.25
1.0
0.75
0.5
225 (25.0%)
225 (25.0%)
224 (24.9%)
225 (25.0%)
223 (24.8%)
225 (25.1%)
225 (25.1%)
225 (25.1%)
224 (24.9%)
225 (25.1%)
224 (24.9%)
225 (25.1%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
223 (24.9%)
225 (25.1%)
225 (25.1%)
223 (24.9%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
225 (25.0%)
180 (24.0%)
195 (26.0%)
198 (26.4%)
177 (23.6%)
201 (26.1%)
208 (27.0%)
185 (24.1%)
175 (22.8%)
182 (23.9%)
198 (26.1%)
194 (25.5%)
186 (24.5%)
205 (24.3%)
217 (25.8%)
214 (25.4%)
206 (24.5%)
221 (24.8%)
224 (25.1%)
224 (25.1%)
222 (24.9%)
M - M
F - F
M - F
F - M
Noun / Adjective Gender
180
190
200
210
220