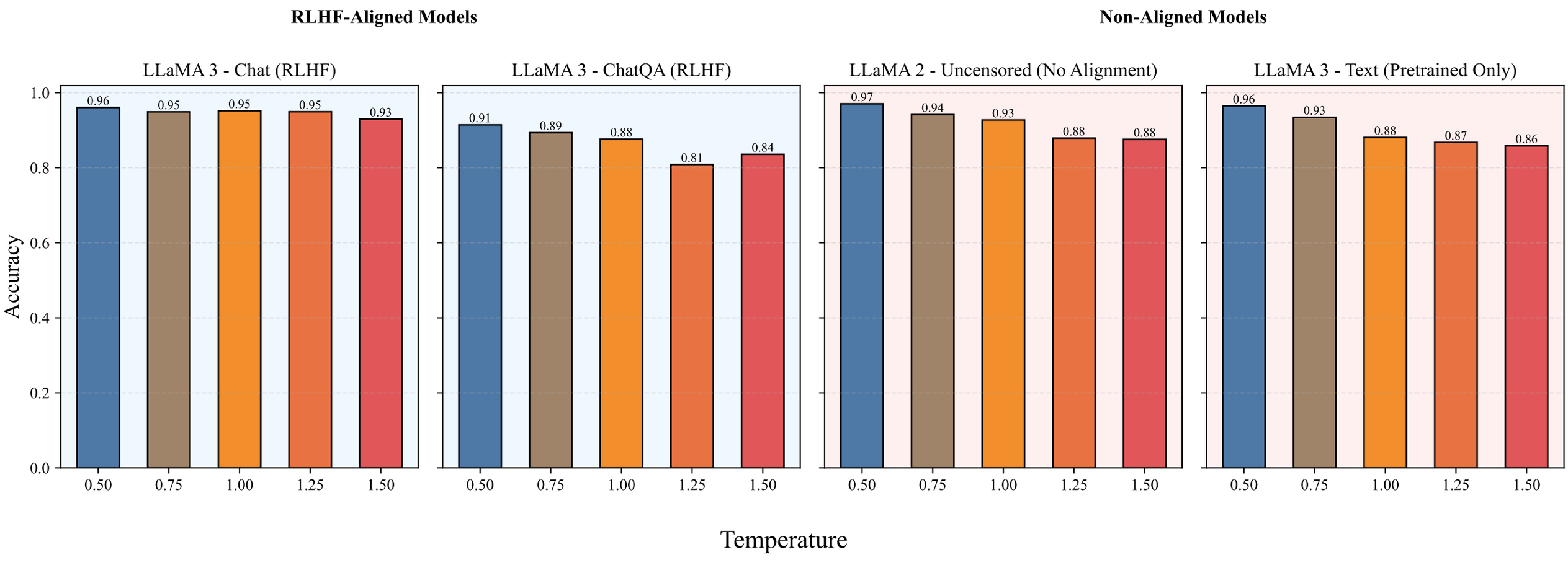
RLHF-Aligned Models
Non-Aligned Models
LLaMA 3 - Chat (RLHF)
LLaMA 3 - ChatQA (RLHF)
LLaMA 2 - Uncensored (No Alignment)
LLaMA 3 - Text (Pretrained Only)
Accuracy
Temperature
0.0
0.2
0.4
0.6
0.8
1.0
0.50
0.75
1.00
1.25
1.50
0.96
0.95
0.95
0.95
0.93
0.91
0.89
0.88
0.81
0.84
0.97
0.94
0.93
0.88
0.88
0.96
0.93
0.88
0.87
0.86