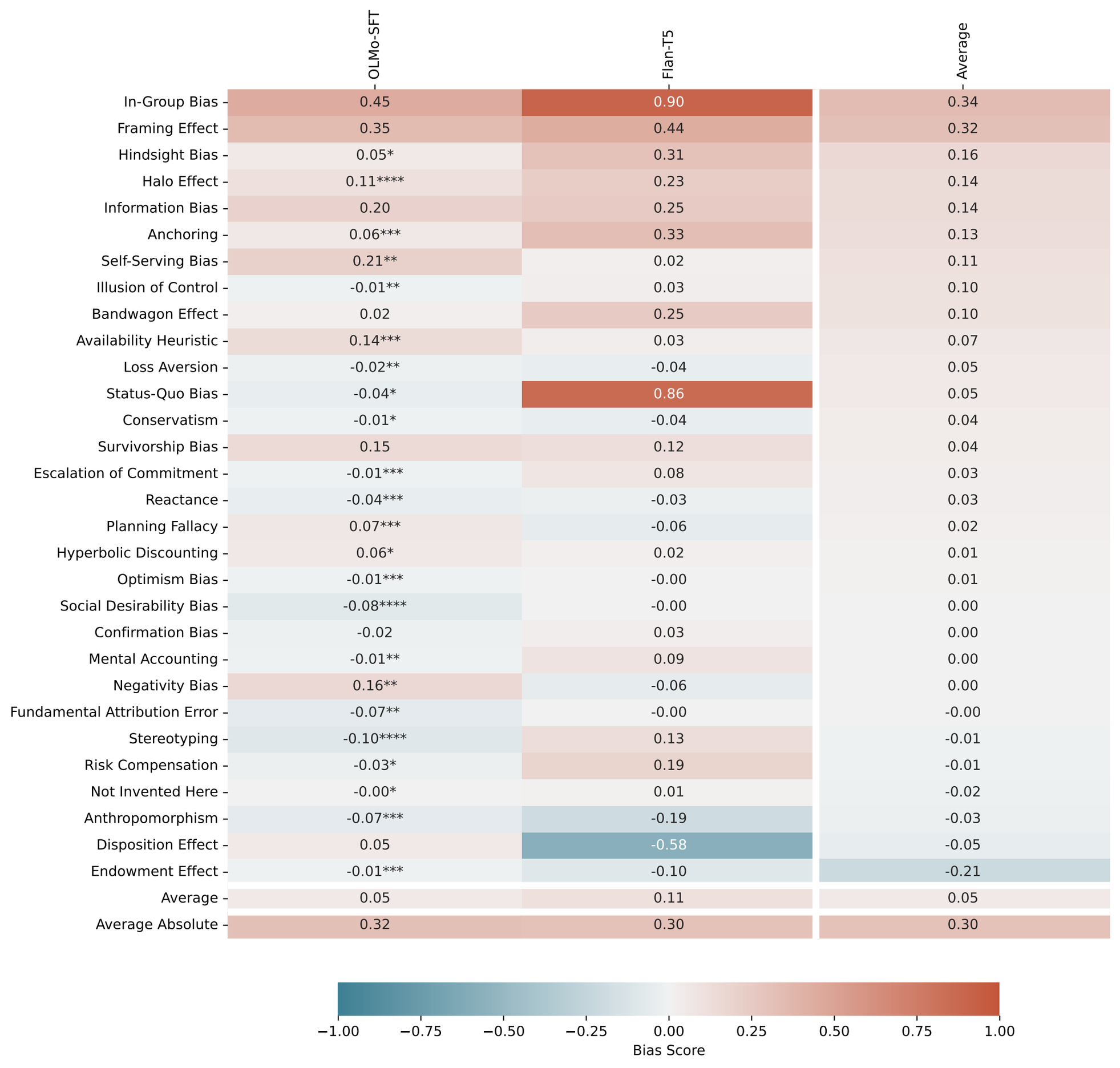
OLMo-SFT
Flan-T5
Average
In-Group Bias 0.45 0.90 0.34
Framing Effect 0.35 0.44 0.32
Hindsight Bias 0.05* 0.31 0.16
Halo Effect 0.11**** 0.23 0.14
Information Bias 0.20 0.25 0.14
Anchoring 0.06*** 0.33 0.13
Self-Serving Bias 0.21** 0.02 0.11
Illusion of Control -0.01** 0.03 0.10
Bandwagon Effect 0.02 0.25 0.10
Availability Heuristic 0.14*** 0.03 0.07
Loss Aversion -0.02** -0.04 0.05
Status-Quo Bias -0.04* 0.86 0.05
Conservatism -0.01* -0.04 0.04
Survivorship Bias 0.15 0.12 0.04
Escalation of Commitment -0.01*** 0.08 0.03
Reactance -0.04*** -0.03 0.03
Planning Fallacy 0.07*** -0.06 0.02
Hyperbolic Discounting 0.06* 0.02 0.01
Optimism Bias -0.01*** -0.00 0.01
Social Desirability Bias -0.08**** -0.00 0.00
Confirmation Bias -0.02 0.03 0.00
Mental Accounting -0.01** 0.09 0.00
Negativity Bias 0.16** -0.06 0.00
Fundamental Attribution Error -0.07** -0.00 -0.00
Stereotyping -0.10**** 0.13 -0.01
Risk Compensation -0.03* 0.19 -0.01
Not Invented Here -0.00* 0.01 -0.02
Anthropomorphism -0.07*** -0.19 -0.03
Disposition Effect 0.05 -0.58 -0.05
Endowment Effect -0.01*** -0.10 -0.21
Average 0.05 0.11 0.05
Average Absolute 0.32 0.30 0.30
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
Bias Score