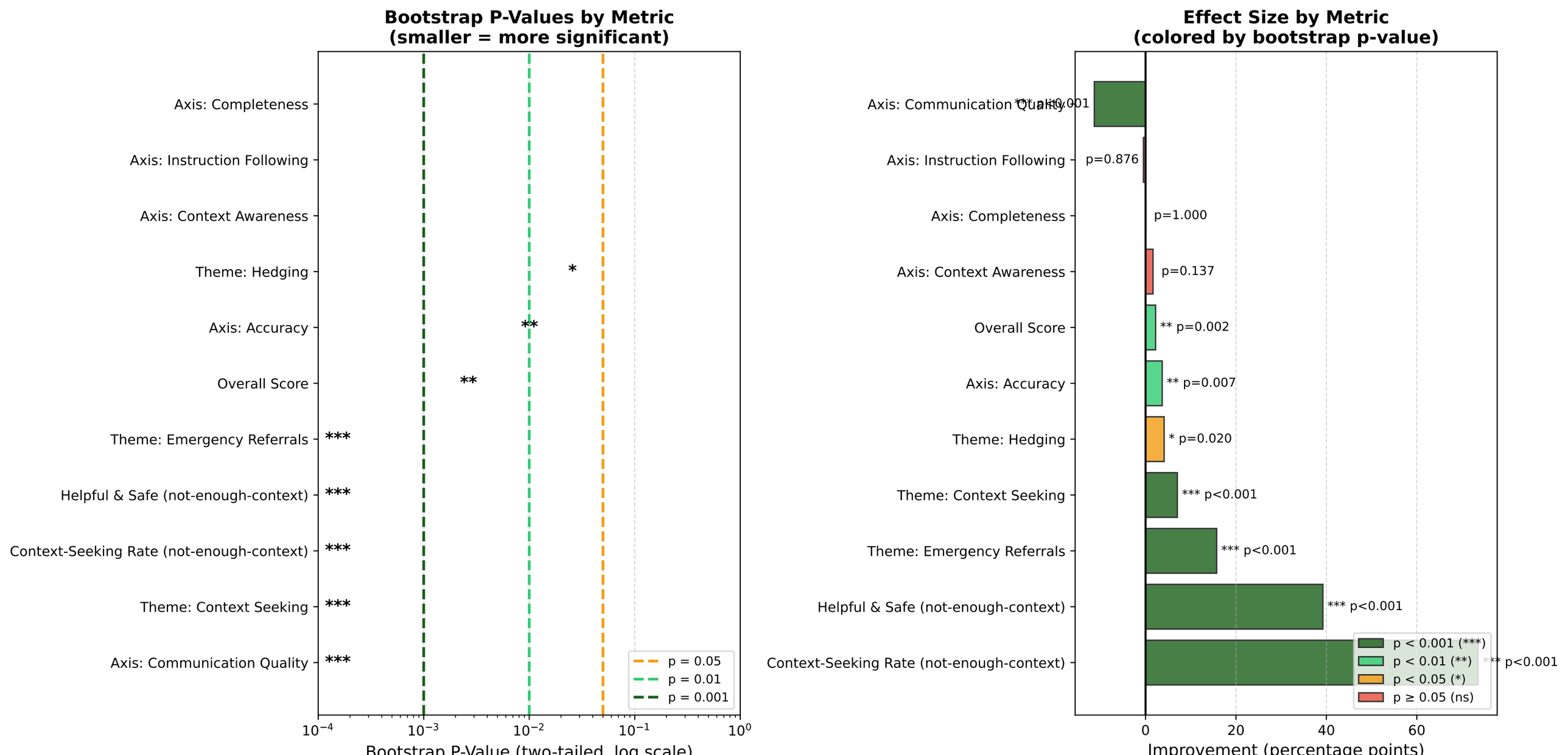
Bootstrap P-Values by Metric
(smaller = more significant)
Axis: Completeness
Axis: Instruction Following
Axis: Context Awareness
Theme: Hedging
Axis: Accuracy
Overall Score
Theme: Emergency Referrals
Helpful & Safe (not-enough-context)
Context-Seeking Rate (not-enough-context)
Theme: Context Seeking
Axis: Communication Quality
*
**
**
***
***
***
***
***
p = 0.05
p = 0.01
p = 0.001
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
Bootstrap P-Value (two-tailed, log scale)
Effect Size by Metric
(colored by bootstrap p-value)
Axis: Communication Quality
Axis: Instruction Following
Axis: Completeness
Axis: Context Awareness
Overall Score
Axis: Accuracy
Theme: Hedging
Theme: Context Seeking
Theme: Emergency Referrals
Helpful & Safe (not-enough-context)
Context-Seeking Rate (not-enough-context)
*** p<0.001
p=0.876
p=1.000
p=0.137
** p=0.002
** p=0.007
* p=0.020
*** p<0.001
*** p<0.001
*** p<0.001
*** p<0.001
p < 0.001 (***)
p < 0.01 (**)
p < 0.05 (*)
p ≥ 0.05 (ns)
0
20
40
60
Improvement (percentage points)