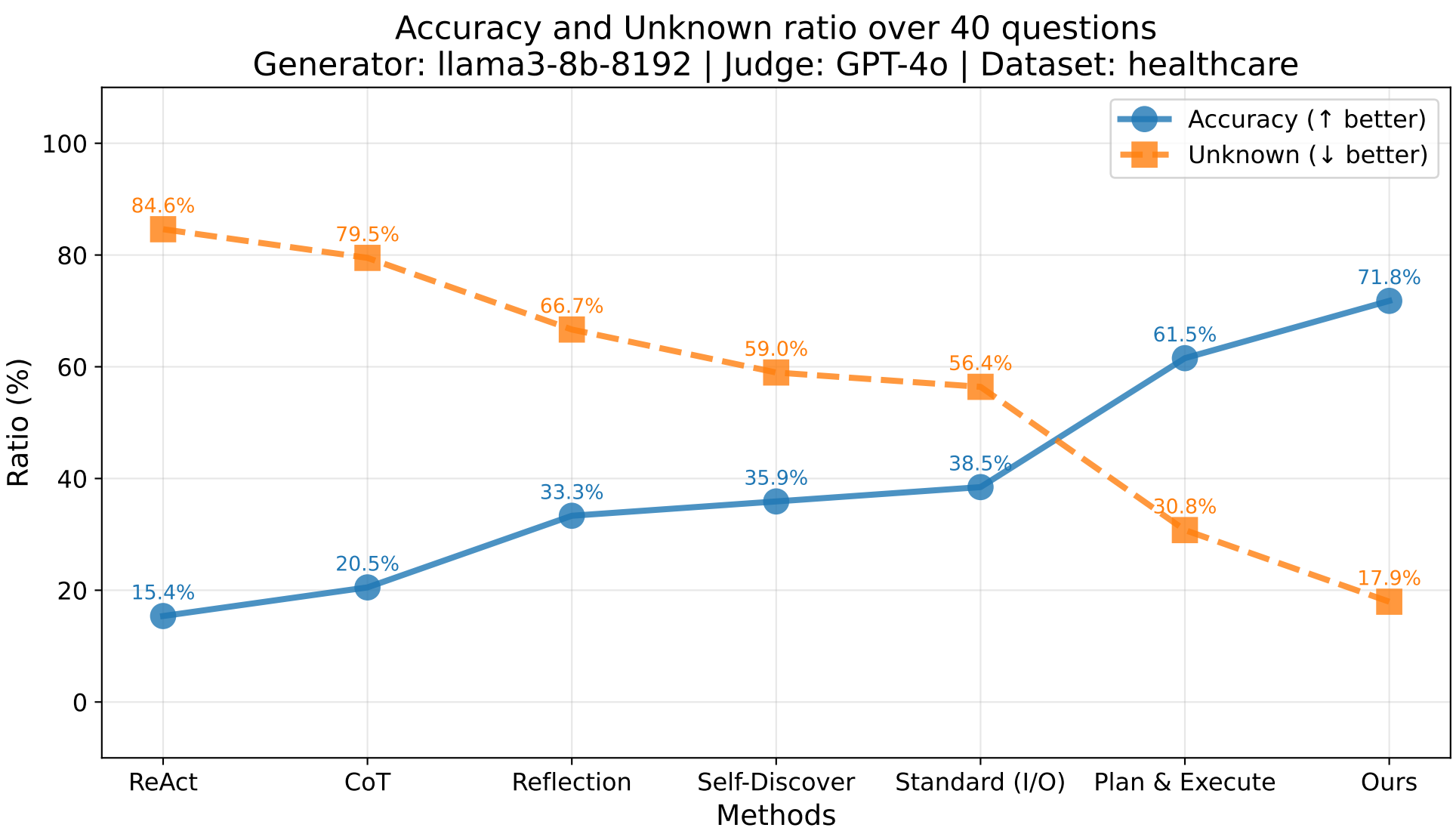
Accuracy and Unknown ratio over 40 questions
Generator: llama3-8b-8192 | Judge: GPT-4o | Dataset: healthcare
Accuracy (↑ better)
Unknown (↓ better)
Ratio (%)
100
80
60
40
20
0
84.6%
79.5%
66.7%
59.0%
56.4%
30.8%
17.9%
15.4%
20.5%
33.3%
35.9%
38.5%
61.5%
71.8%
ReAct
CoT
Reflection
Self-Discover
Standard (I/O)
Plan & Execute
Ours
Methods