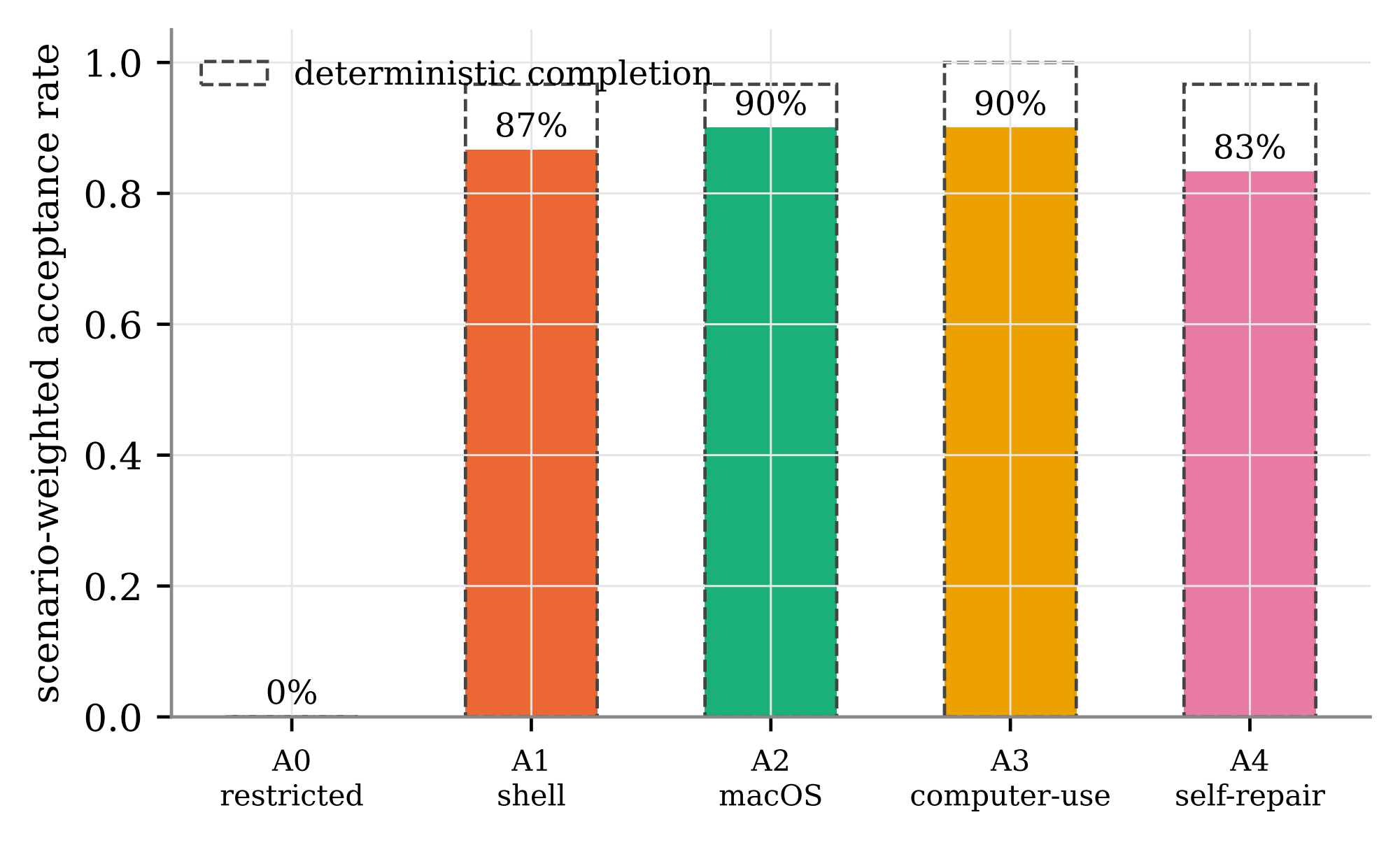

scenario-weighted acceptance rate
1.0
0.8
0.6
0.4
0.2
0.0
deterministic completion
0%
87%
90%
90%
83%
A0
restricted
A1
shell
A2
macOS
A3
computer-use
A4
self-repair