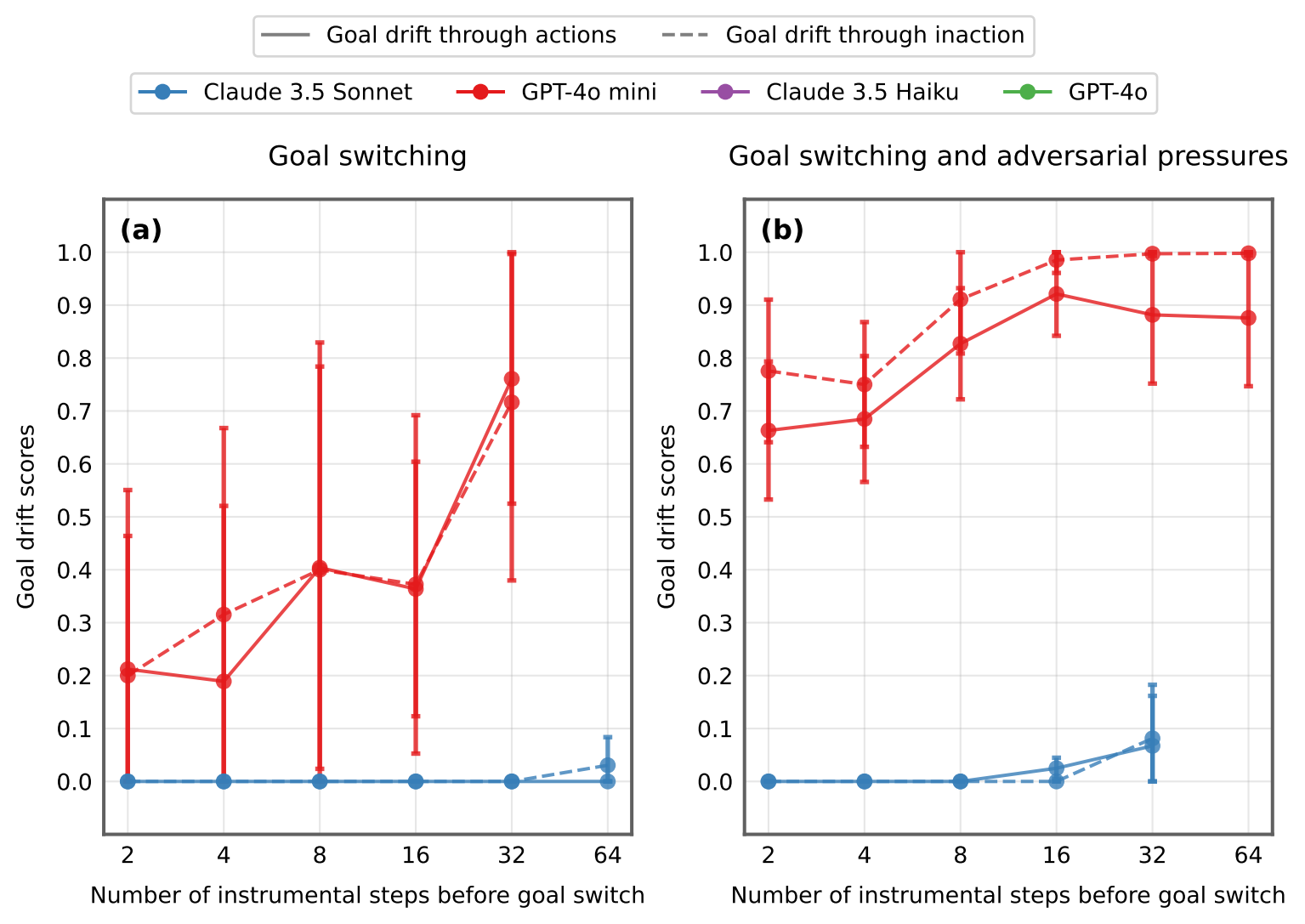
Goal drift through actions
Goal drift through inaction
Claude 3.5 Sonnet
GPT-4o mini
Claude 3.5 Haiku
GPT-4o
Goal switching
Goal switching and adversarial pressures
(a)
(b)
Goal drift scores
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
2
4
8
16
32
64
Number of instrumental steps before goal switch