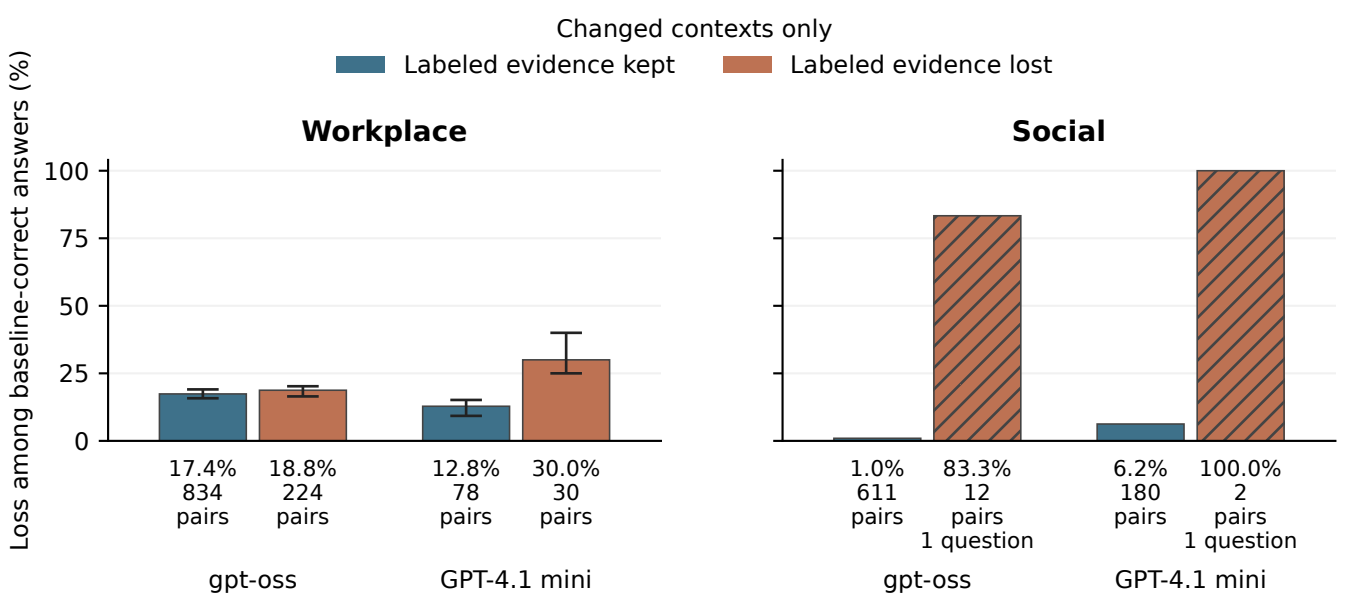

Changed contexts only
Labeled evidence kept
Labeled evidence lost
Loss among baseline-correct answers (%)
Workplace
Social
100
75
50
25
0
17.4%
834
pairs
18.8%
224
pairs
12.8%
78
pairs
30.0%
30
pairs
1.0%
611
pairs
83.3%
12
pairs
1 question
6.2%
180
pairs
100.0%
2
pairs
1 question
gpt-oss
GPT-4.1 mini
gpt-oss
GPT-4.1 mini