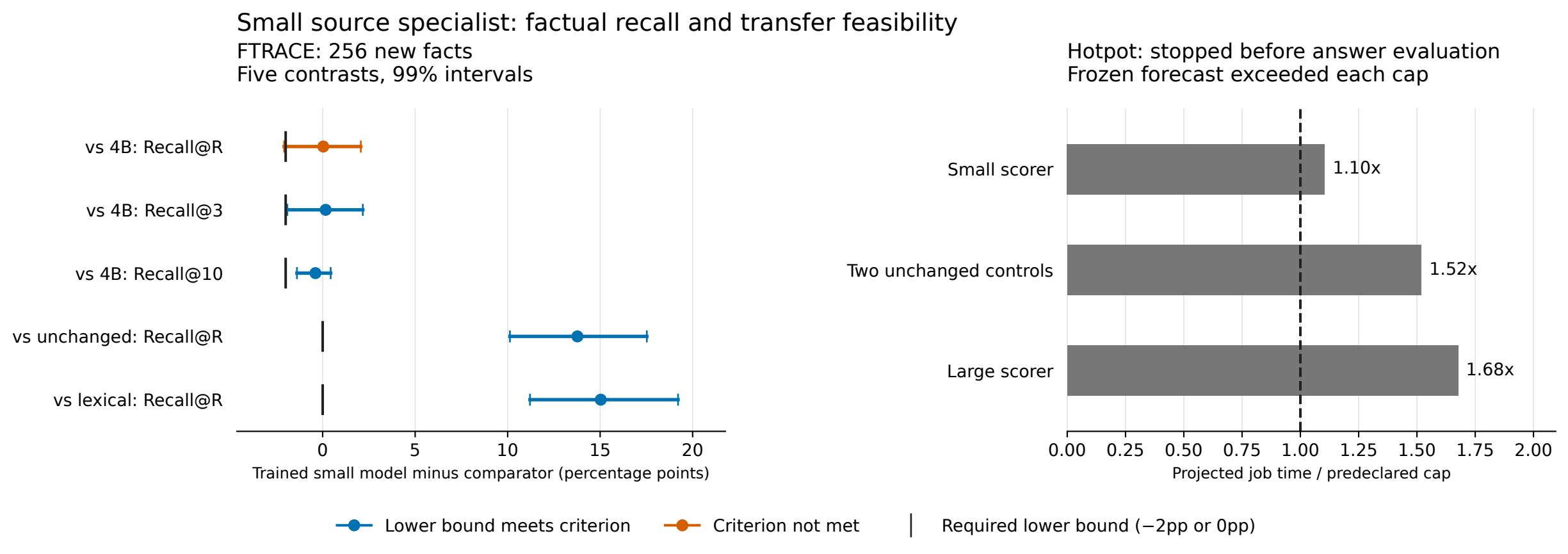
Small source specialist: factual recall and transfer feasibility
FTRACE: 256 new facts
Five contrasts, 99% intervals
vs 4B: Recall@R
vs 4B: Recall@3
vs 4B: Recall@10
vs unchanged: Recall@R
vs lexical: Recall@R
0
5
10
15
20
Trained small model minus comparator (percentage points)
Hotpot: stopped before answer evaluation
Frozen forecast exceeded each cap
Small scorer
1.10x
Two unchanged controls
1.52x
Large scorer
1.68x
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Projected job time / predeclared cap
Lower bound meets criterion
Criterion not met
Required lower bound (−2pp or 0pp)