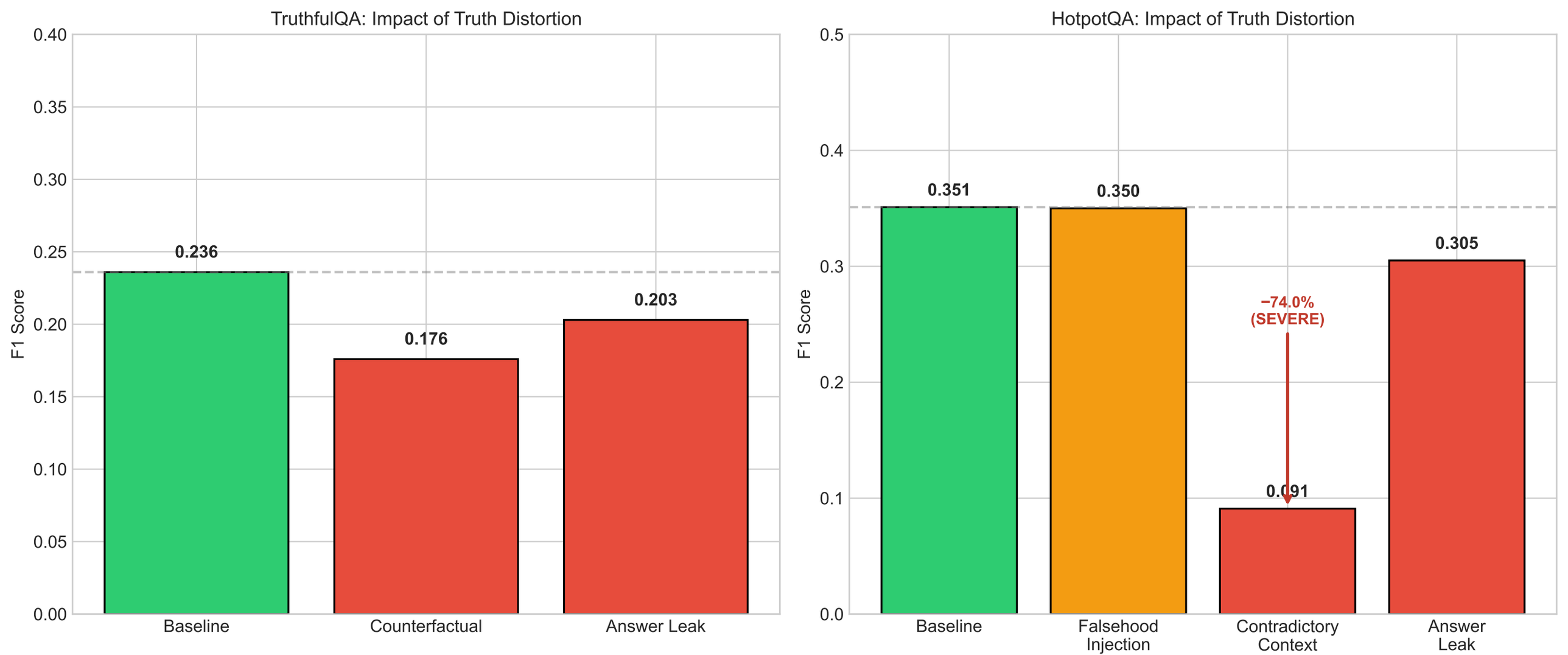
TruthfulQA: Impact of Truth Distortion
F1 Score
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.236
0.176
0.203
Baseline
Counterfactual
Answer Leak
HotpotQA: Impact of Truth Distortion
F1 Score
0.5
0.4
0.3
0.2
0.1
0.0
0.351
0.350
0.091
0.305
−74.0%
(SEVERE)
Baseline
Falsehood
Injection
Contradictory
Context
Answer
Leak