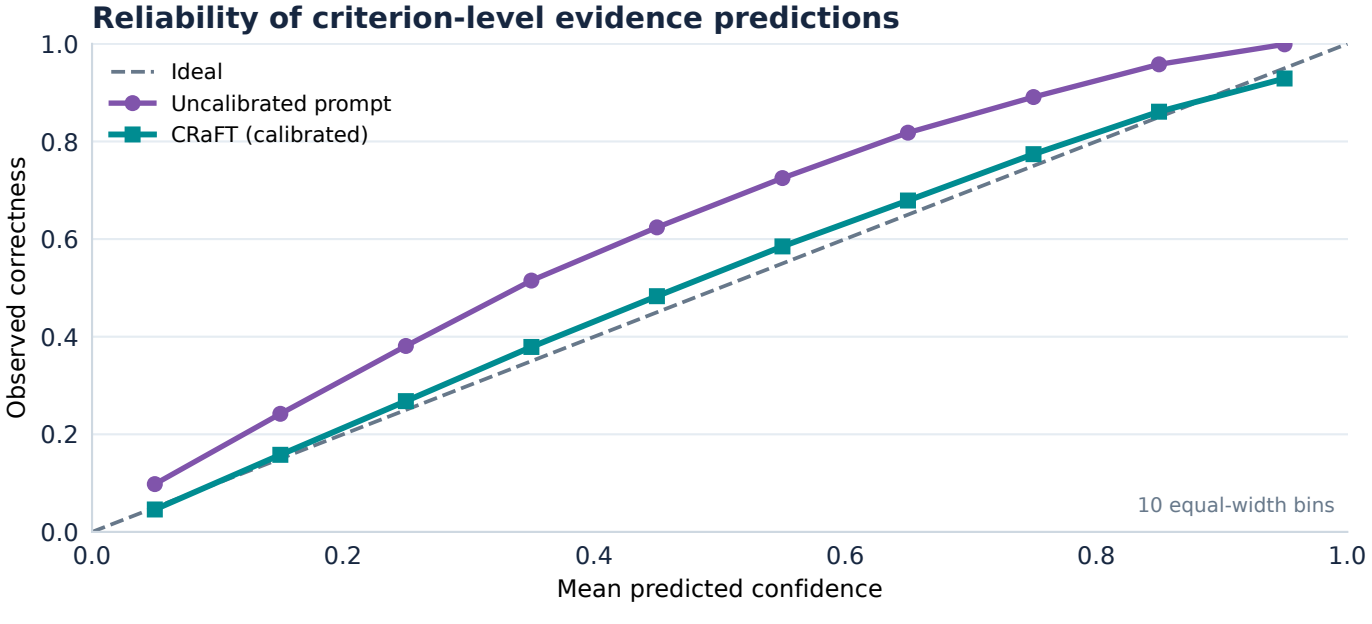

Reliability of criterion-level evidence predictions
Ideal
Uncalibrated prompt
CRaFT (calibrated)
Observed correctness
Mean predicted confidence
0.0
0.2
0.4
0.6
0.8
1.0
10 equal-width bins