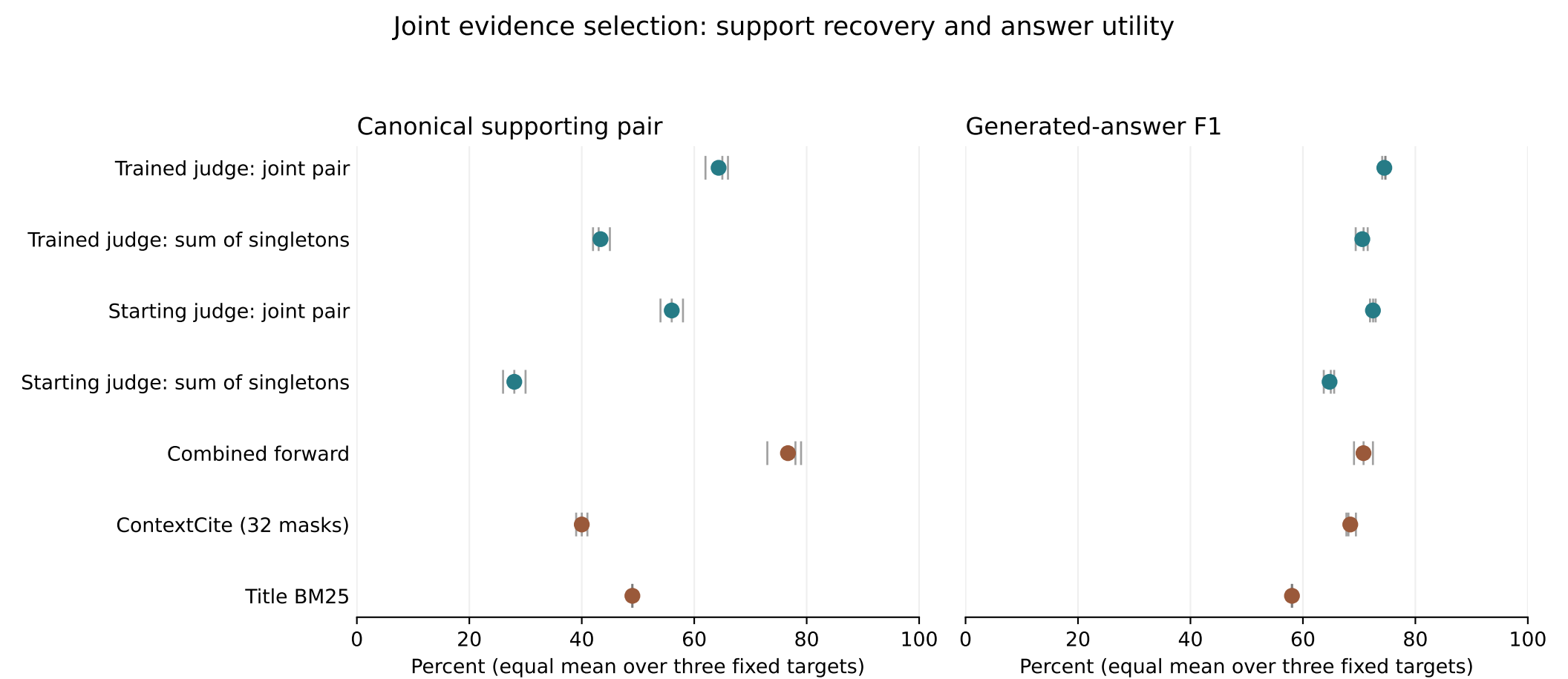


Primary trained joint minus singleton sum; 97.5% question intervals: Pair: +21.00 pp [+8.00, +34.00]; F1: +3.91 pp [-1.90, +10.28]
100 exposed questions; common batch-one answer generation. Gray marks are target means, not confidence intervals.