The Complementarity Finding: Where Tracks Diverge

Valid Alternative
(Low F1, High Judge)

Agreement
(Both High)

Agreement
(Both Low)

Semantic Error
(High F1, Low Judge)

Divergent cases: 270/675 (40.0%)
→ Both tracks needed

Legend:
- Agreement (success)
- Agreement (failure)
- Valid alternative (complementarity)
- Semantic error (complementarity)

X-axis: F1 Score (Reference-Based)
Y-axis: Correctness Score (Judge-Based)