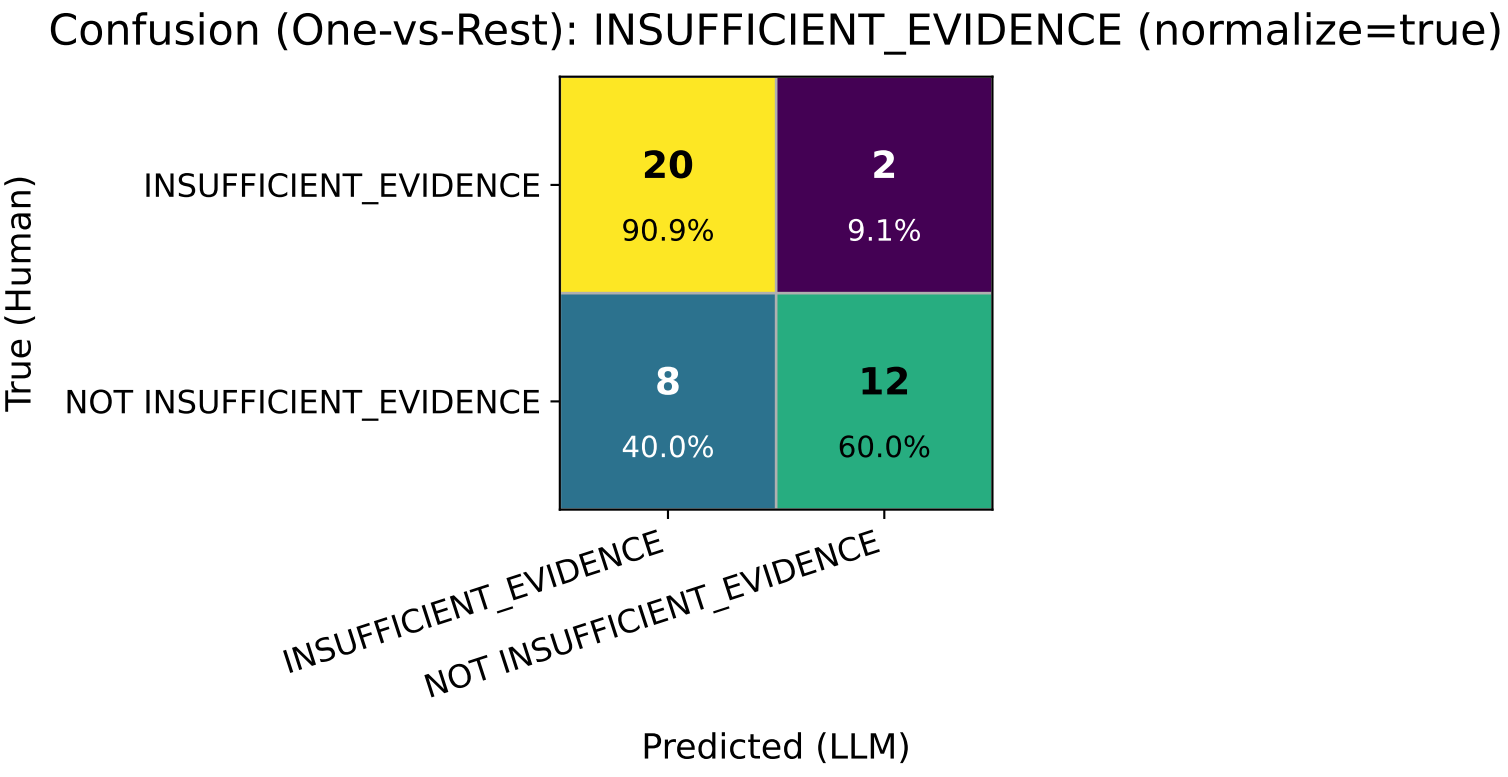
Confusion (One-vs-Rest): INSUFFICIENT_EVIDENCE (normalize=true)
True (Human)
INSUFFICIENT_EVIDENCE
NOT INSUFFICIENT_EVIDENCE
20
90.9%
2
9.1%
8
40.0%
12
60.0%
INSUFFICIENT_EVIDENCE
NOT INSUFFICIENT_EVIDENCE
Predicted (LLM)