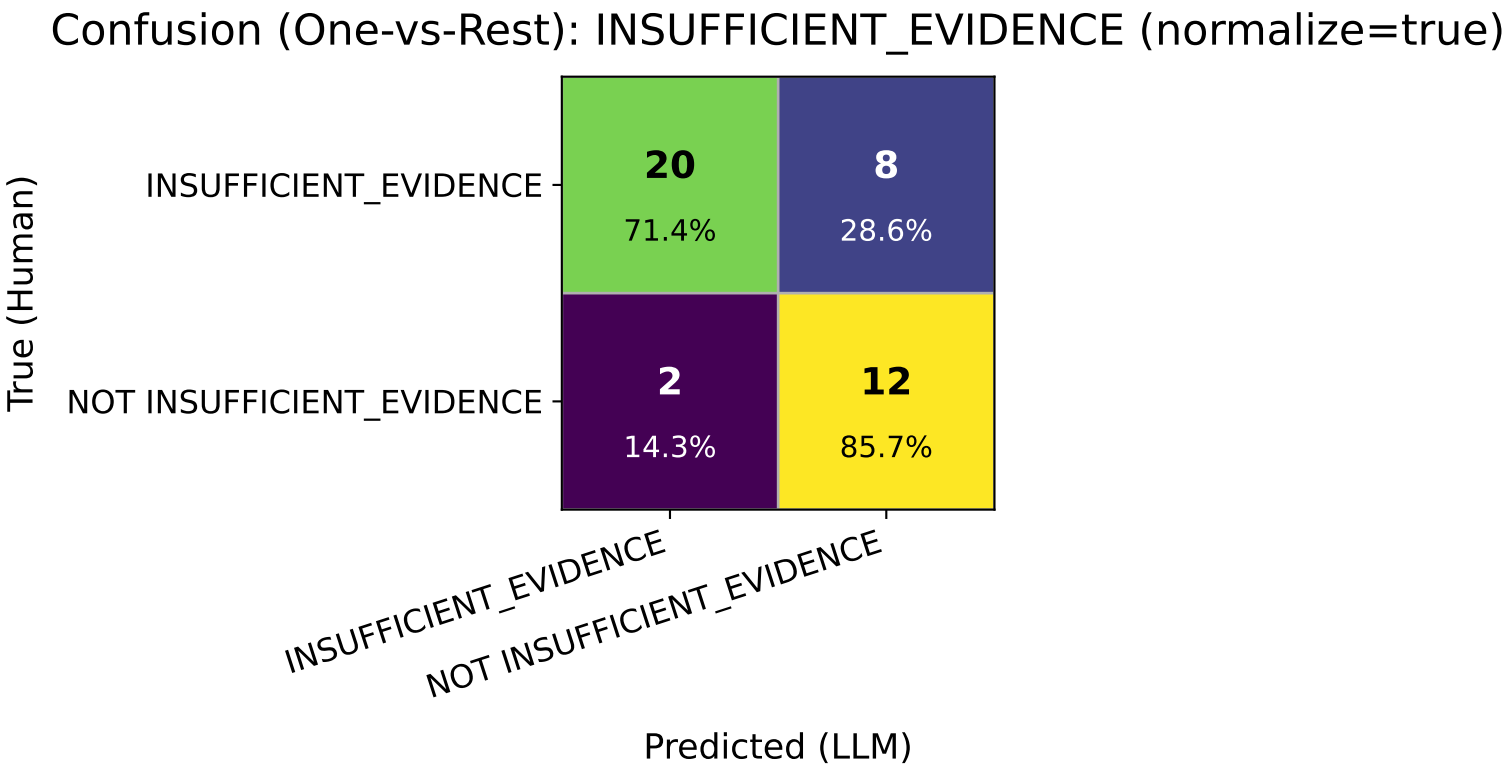
Confusion (One-vs-Rest): INSUFFICIENT_EVIDENCE (normalize=true)
True (Human)
INSUFFICIENT_EVIDENCE
NOT INSUFFICIENT_EVIDENCE
20
71.4%
8
28.6%
2
14.3%
12
85.7%
INSUFFICIENT_EVIDENCE
NOT INSUFFICIENT_EVIDENCE
Predicted (LLM)