Output Compliance Analysis
Structured "Failures" Often Contain Detection

No acknowledgment detected
NL acknowledgment present (judge-confirmed)

Number of Trials

60
(53%)

54
(47%)

Total structured regex failures: 114
53% contained natural language acknowledgment

Structured
Regex Failures