grok - Phase 1: Comprehensive Metrics (2 experiments average) 1.0 0.727 0.704 0.700 8.0 0.6 Metric Score 0.4 0.454 0.696 0.683 0.2 0.606 0.600

Macro-F1

0.0

Weighted-F1

Type

Party

Action

Asset

Accuracy