grok - Phase 1: Comprehensive Metrics (2 experiments average) 1.0 0.704 Type 0.700 0.8 Party 0.6 Metric Score 0.4 0.454 Action 0.683 0.696 0.2 0.606 Asset 0.600 0.0 Accuracy Macro-F1 Weighted-F1