Average Rouge-L (%) - Average Across All Models (Joke Experiments)
(Lower is Better)

Statistical Tests (Best VS: VS-Combined):
vs Baseline: p=0.0003 ***
vs Sequence: p=0.0923 ns
vs Multi-turn: p=0.0042 **

- Multi-turn: 22.4±5.9
- Sequence: 16.5±3.6
- VS-Standard: 16.3±3.0
- VS-CoT: 14.3±2.4
- VS-Combined: 13.3±2.3

Average Rouge-L (%)

Methods