**Average Quality Score (%) - Average Across All Models (Poem Experiments)
(Higher is Better)**

Statistical Tests (Best VS: VS-Combined):
vs Baseline: p=0.3169 ns
vs Sequence: p=0.2723 ns
vs Multi-turn: p=0.4600 ns

- Baseline: 61.9±7.6
- Sequence: 60.4±11.2
- Multi-turn: 63.0±8.1
- VS-Standard: 61.0±9.7
- VS-CoT: 64.7±10.3
- VS-Combined: 66.4±9.7

X-axis: Methods
Y-axis: Average Quality Score (%)