Pass@K Accuracy ↑ - Average Across All Models

Statistical Tests (Best Baseline: CoT)
VS-Standard vs CoT: p=0.6448 ns
VS-CoT vs CoT: p=0.9580 ns
VS-Combined vs CoT: p=0.8117 ns

Direct: 0.48±0.16
CoT: 0.50±0.15
Sequence: 0.45±0.17
Multi-turn: 0.46±0.18
VS-Standard: 0.46±0.15
VS-CoT: 0.49±0.15
VS-Combined: 0.48±0.15