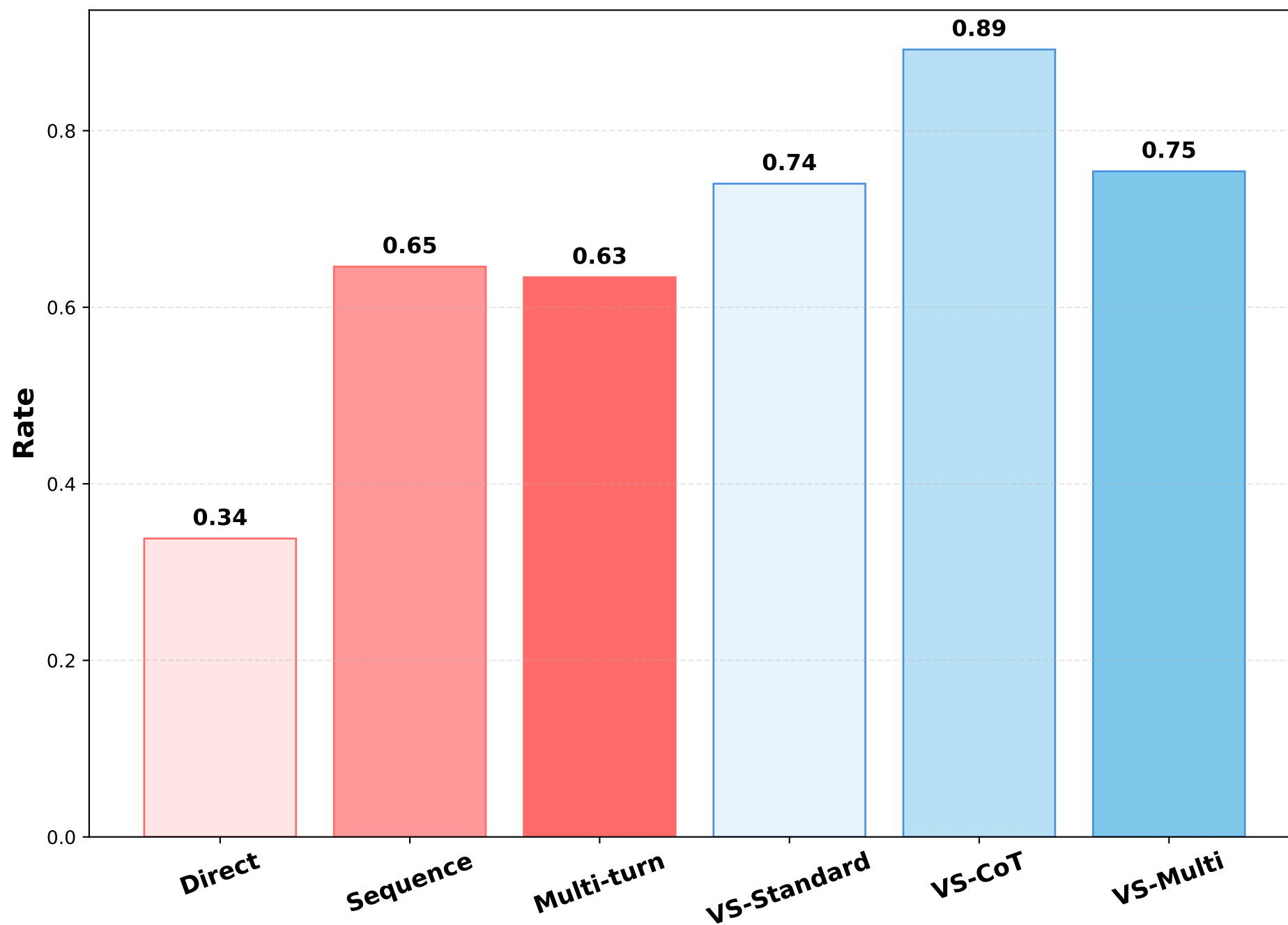


Direct Sequence Multi-turn VS-Standard VS-CoT VS-Multi

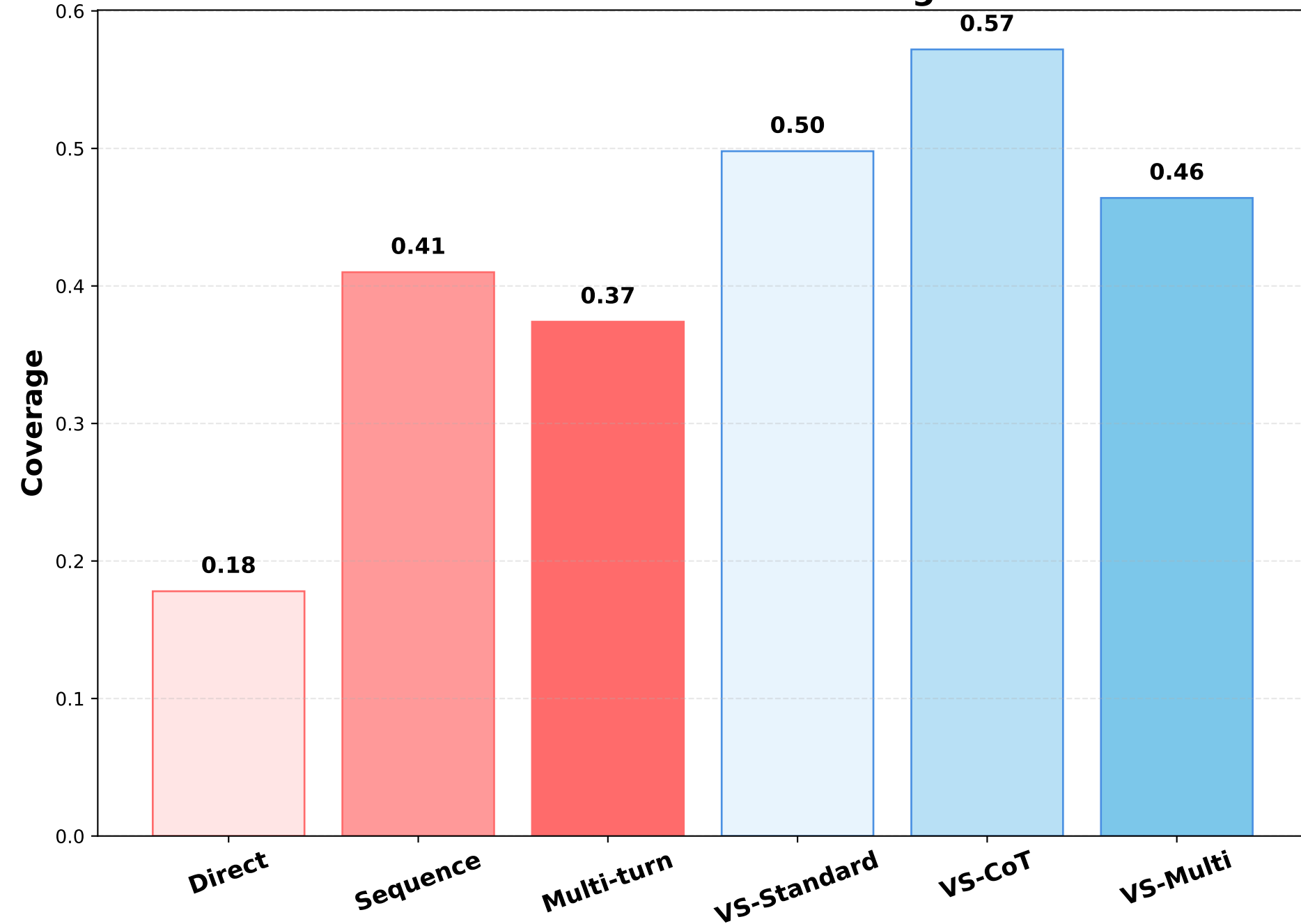
(a)

Incorrect Answer Rate



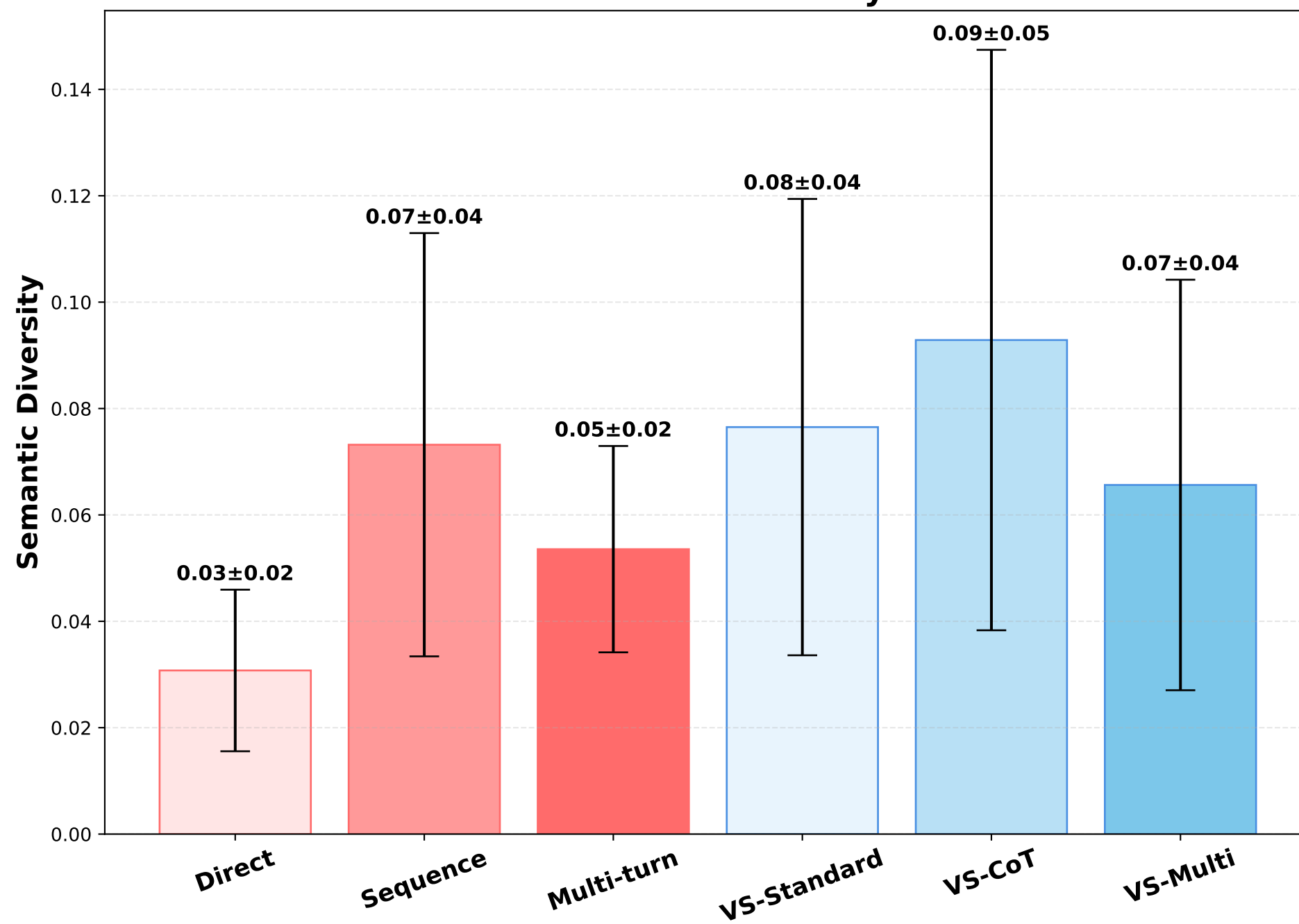
(b)

Incorrect Answer Coverage



(c)

Semantic Diversity



(d)

Cosine Similarity (Pairwise)

