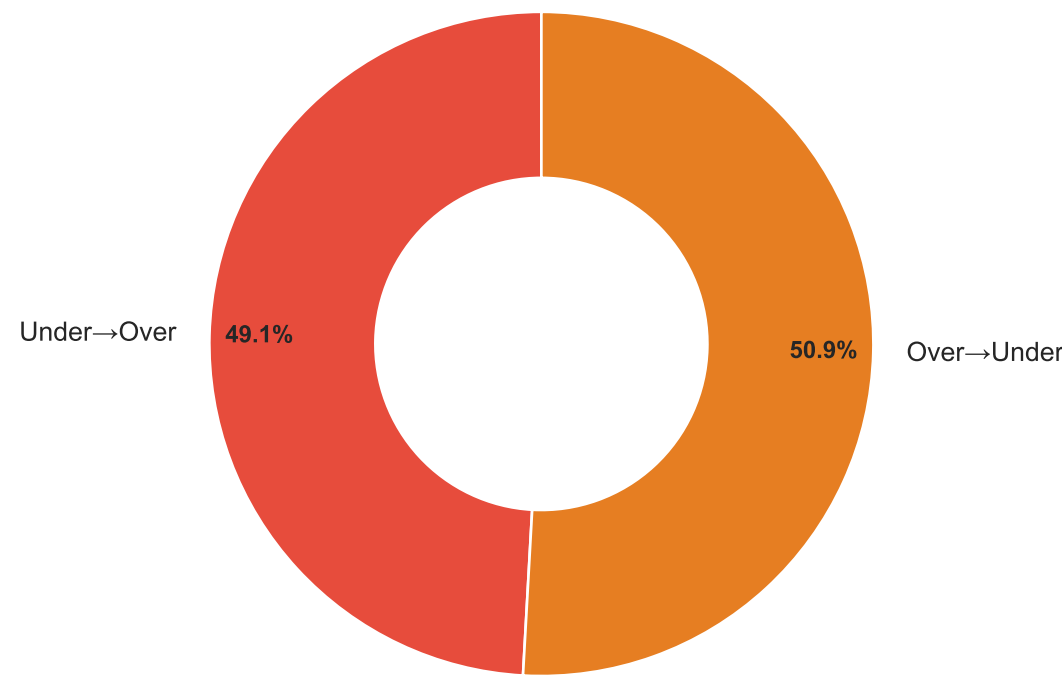
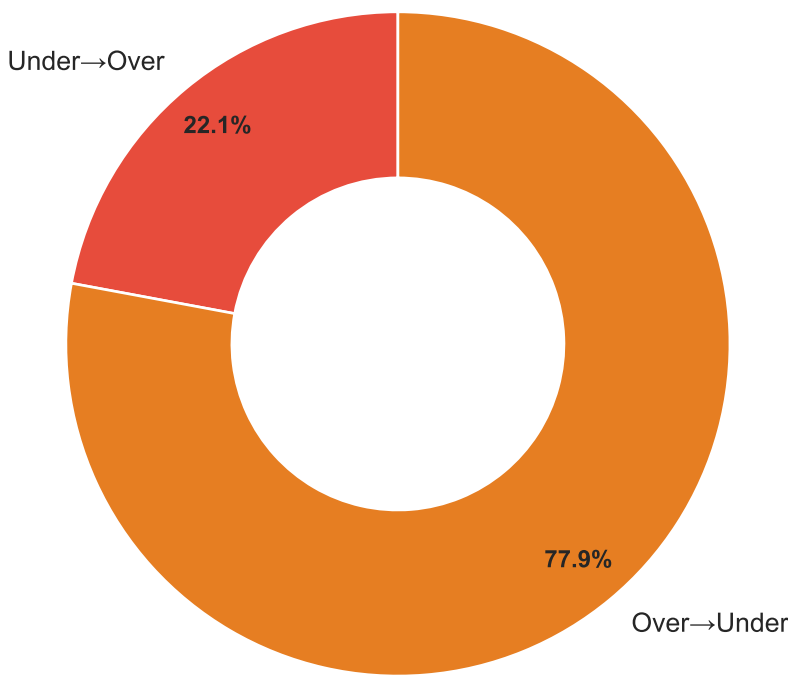


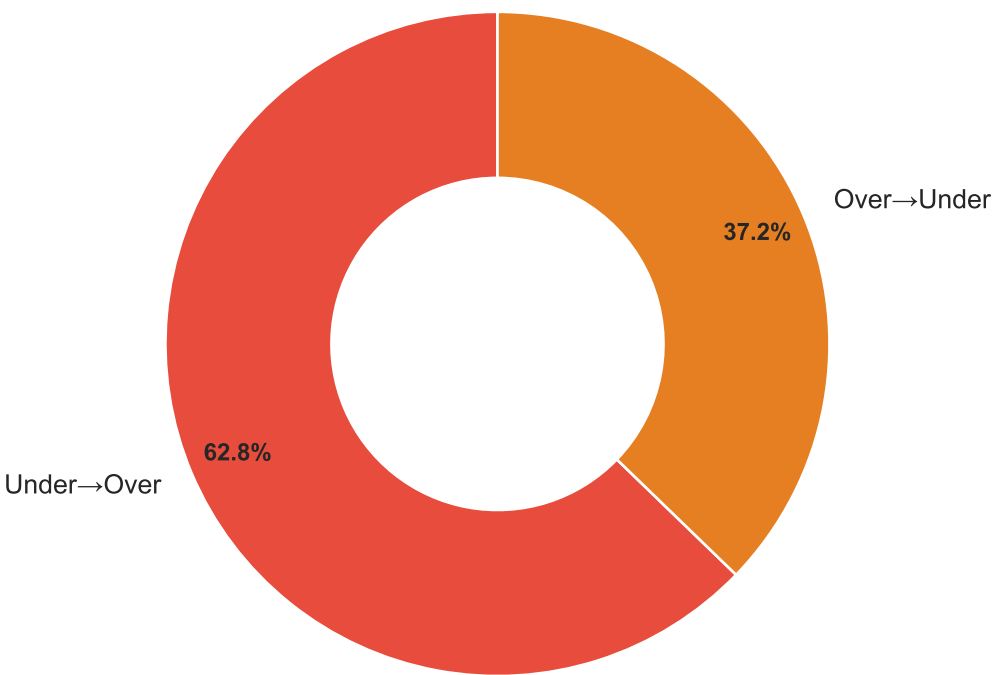
Correctness Changes (Same Confidence):
4omini_sc vs 4omini_cot



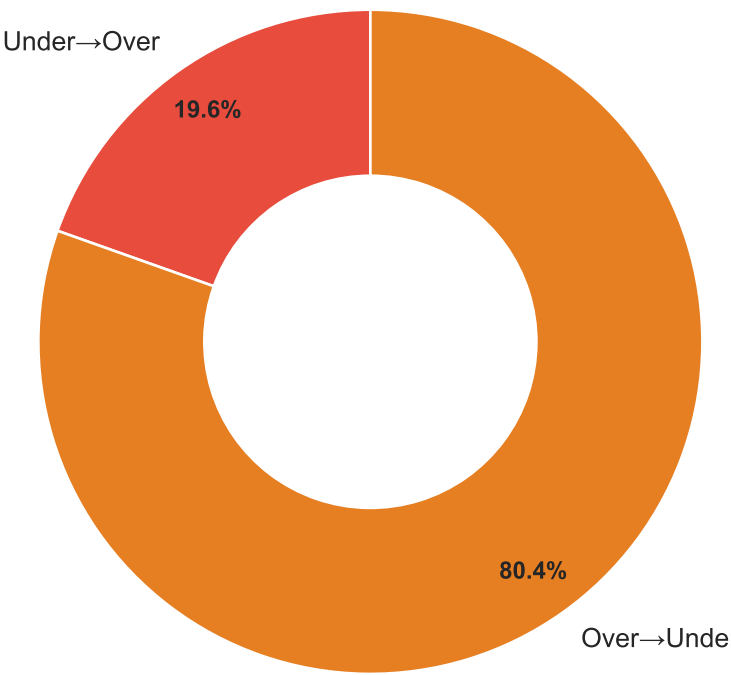
Correctness Changes (Same Confidence):
4omini_cot vs 4omini_noexp



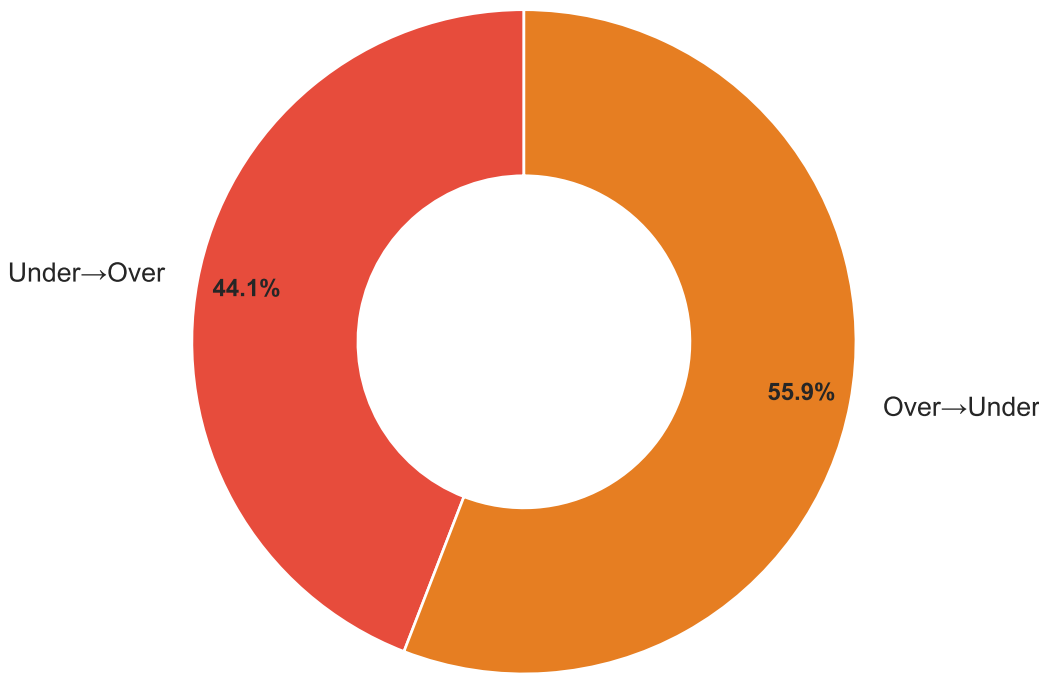
Correctness Changes (Same Confidence):
grok2_sc vs grok2_cot



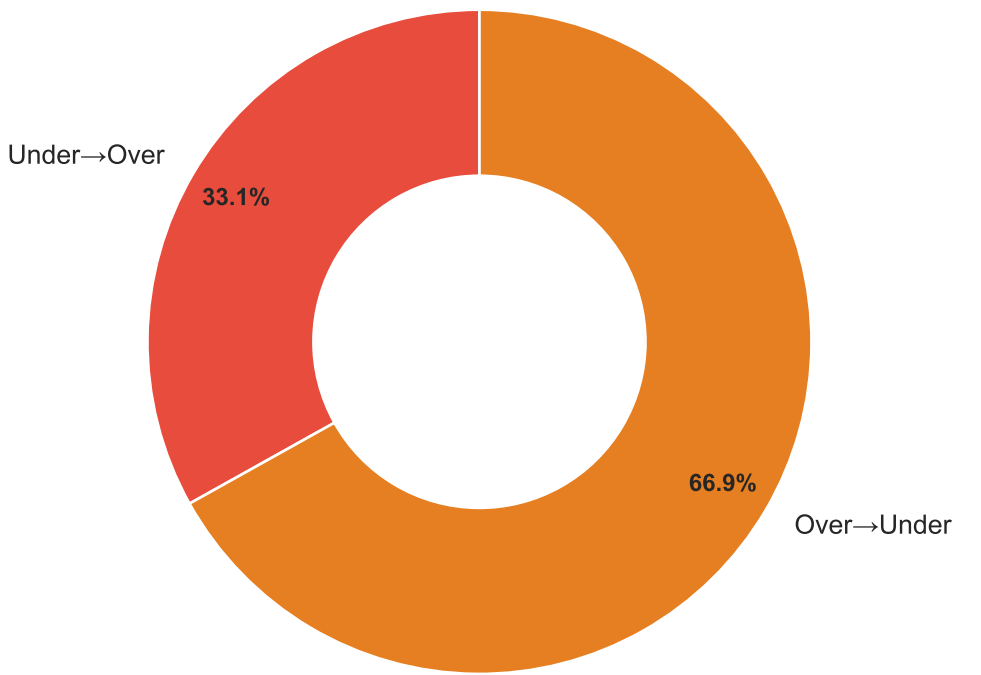
Total cases with correctness changes: 340
Correctness Changes (Same Confidence):
grok2_cot vs grok2_noexp



Total cases with correctness changes: 530
Correctness Changes (Same Confidence):
o1_reason vs o1_mini



Total cases with correctness changes: 188
Correctness Changes (Same Confidence):
o3mini_medium vs o3mini_low



Total cases with correctness changes: 562

Total cases with correctness changes: 170

Total cases with correctness changes: 142