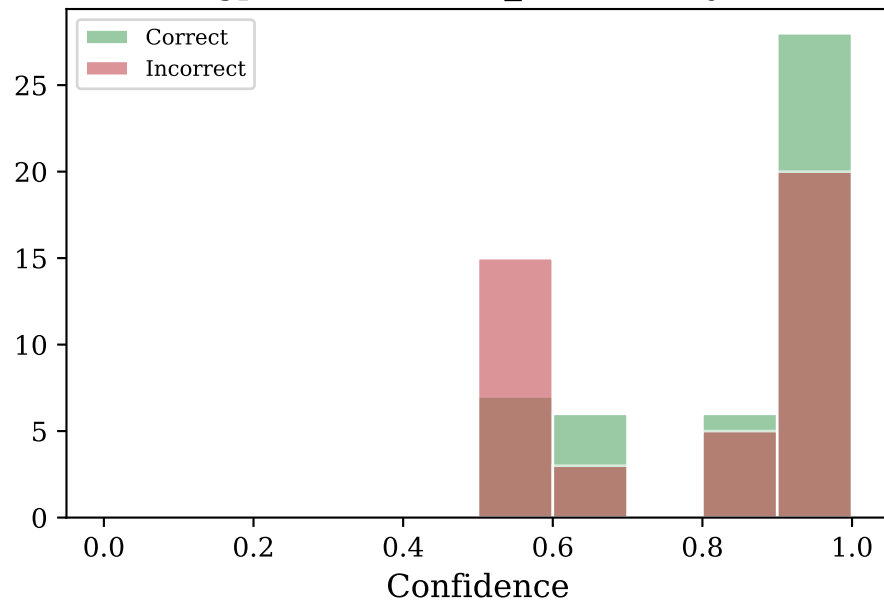
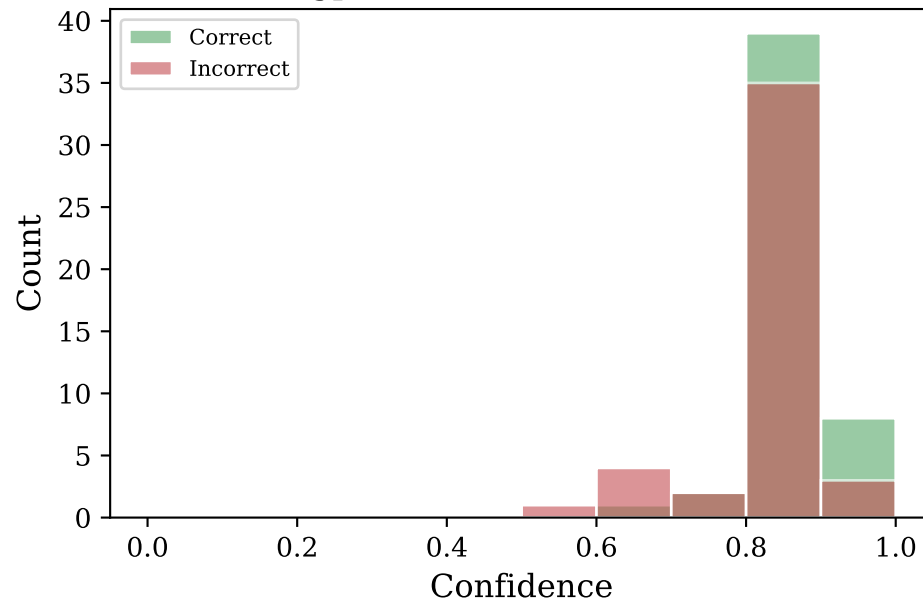


Confidence Score Distribution by Correctness

gpt-4o-mini (self_consistency)



gpt-4o-mini (verbalized)



qwen3:32b (verbalized)

