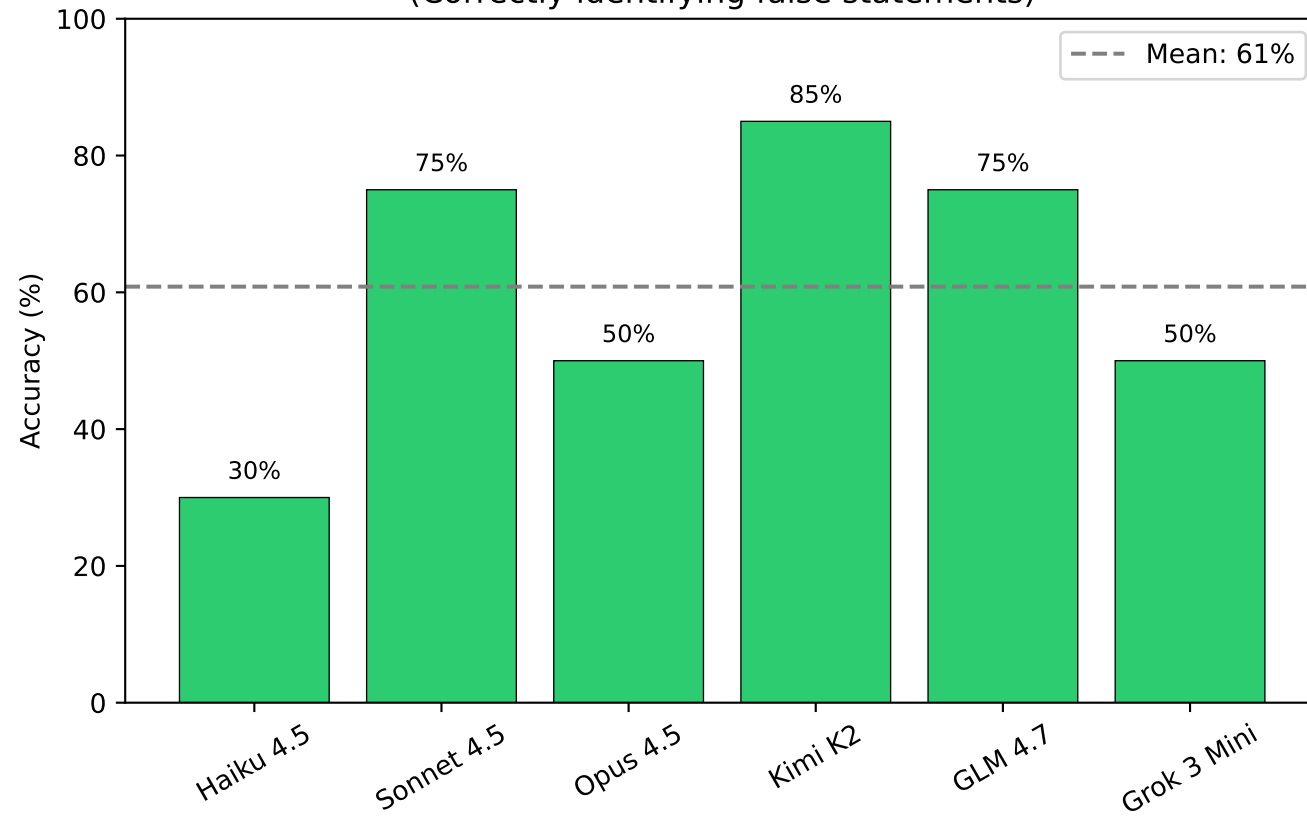
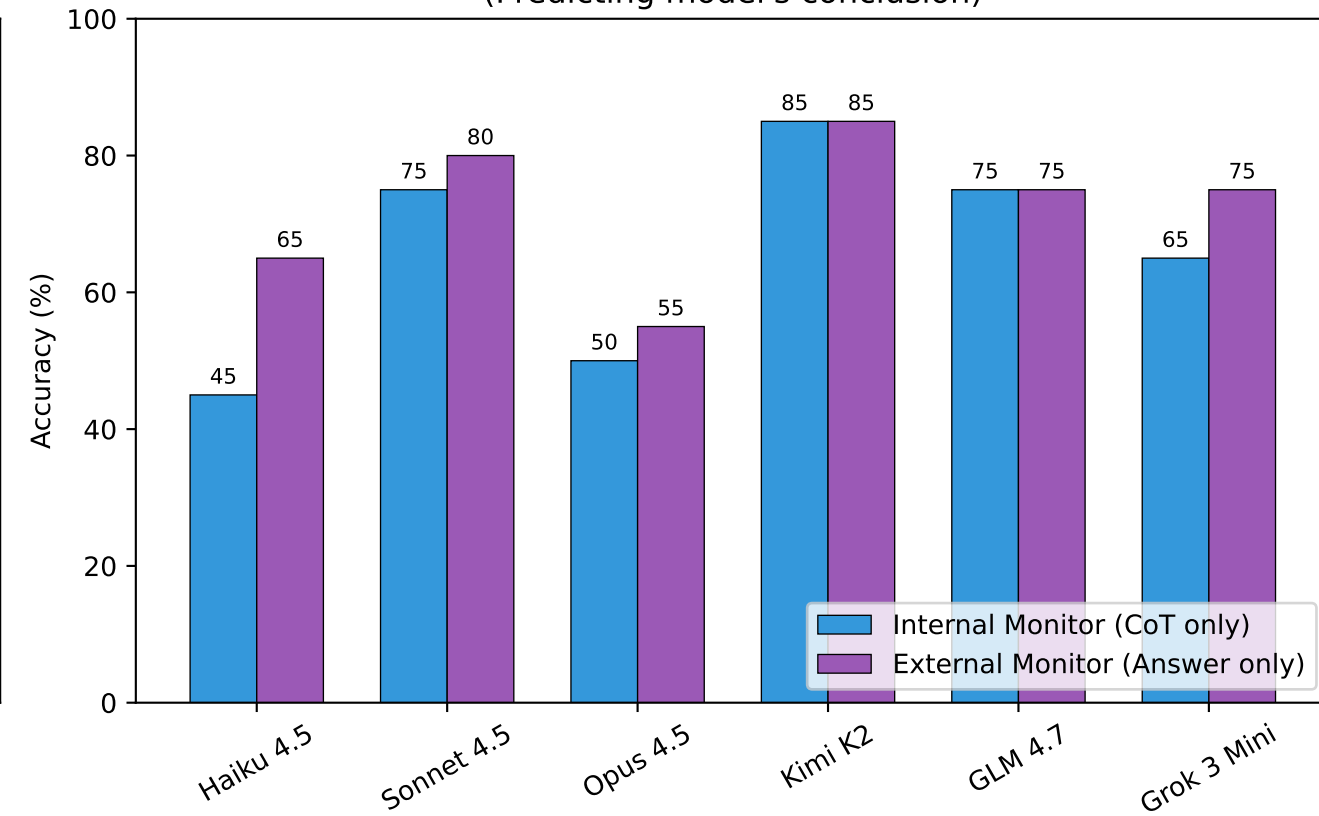


Phase 1: False Proof Monitor Experiment Results

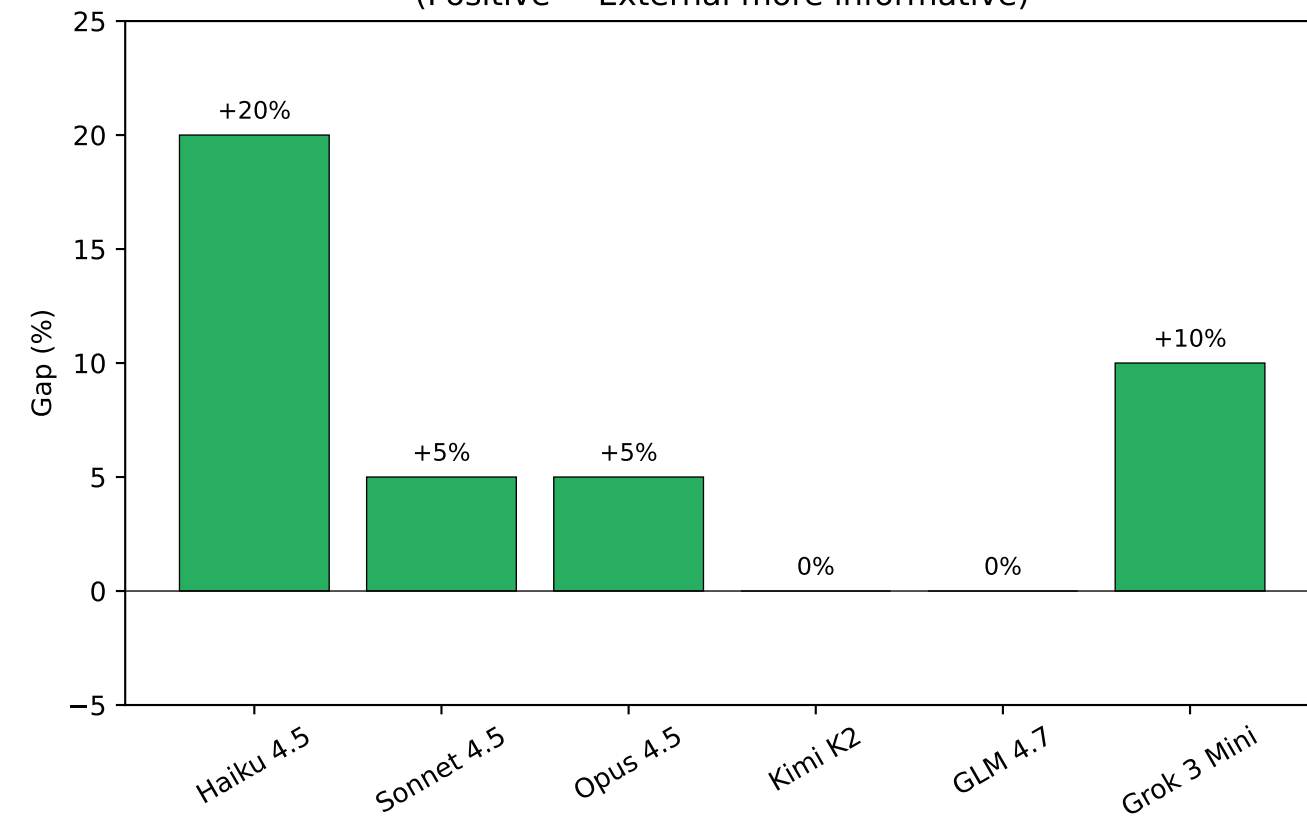
Model Accuracy (Correctly identifying false statements)



Monitor Accuracy Comparison (Predicting model's conclusion)



Monitor Gap (External - Internal) (Positive = External more informative)



Unfaithful Reasoning Rate (CoT shows doubt but output claims proof)

