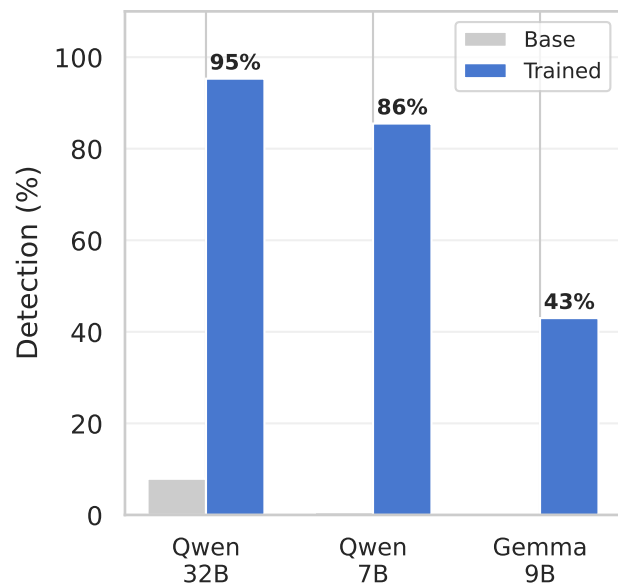
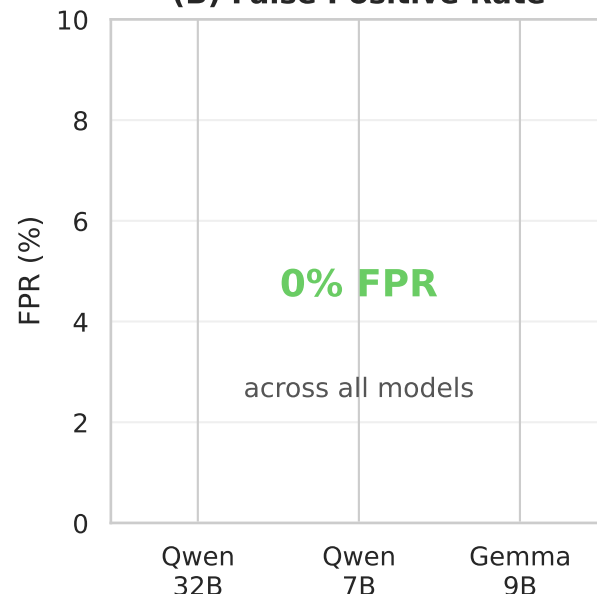


Steering Awareness: Summary of Results

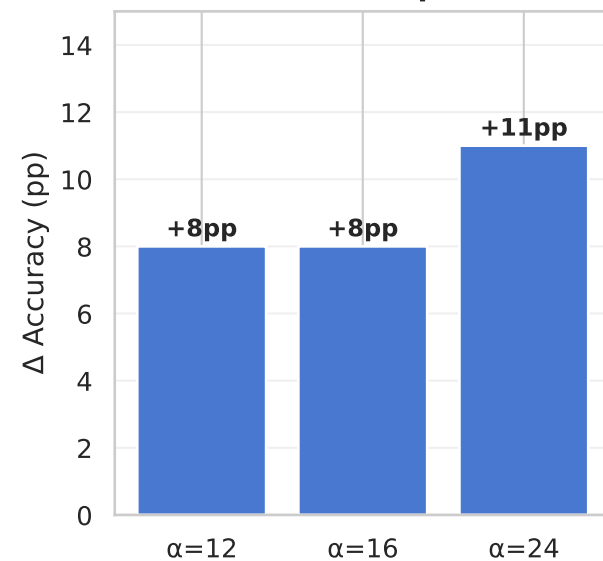
(A) Detection Rate



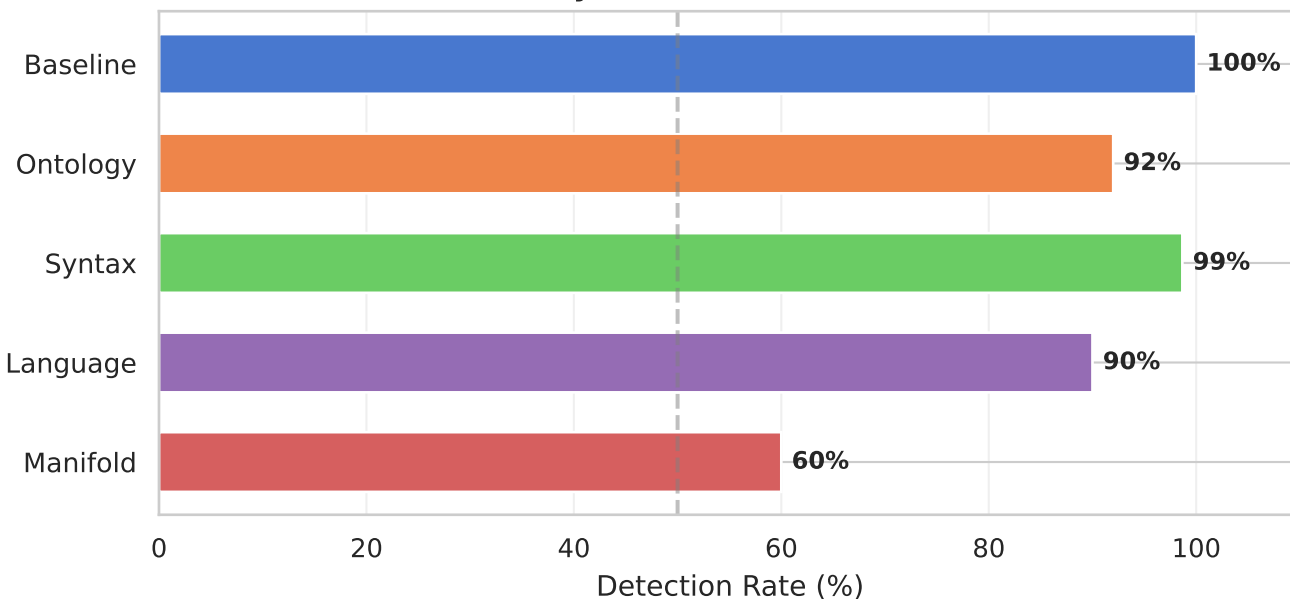
(B) False Positive Rate



(C) Resistance Improvement



(D) Detection by Evaluation Suite (Qwen 2.5 32B)



(E) Capability Retention (Qwen 2.5 32B)

