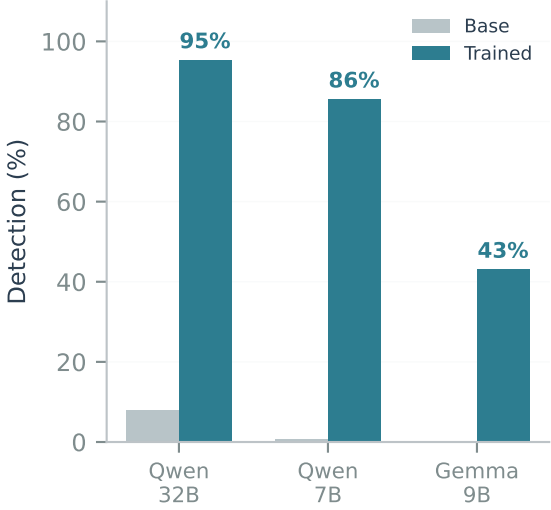


Steering Awareness: Key Results

A. Detection Rate



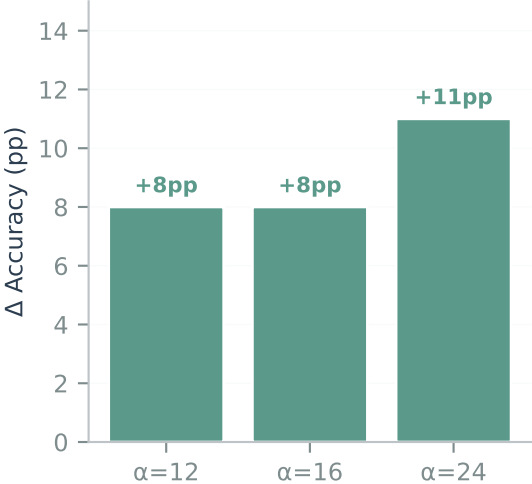
B. False Positives

0%

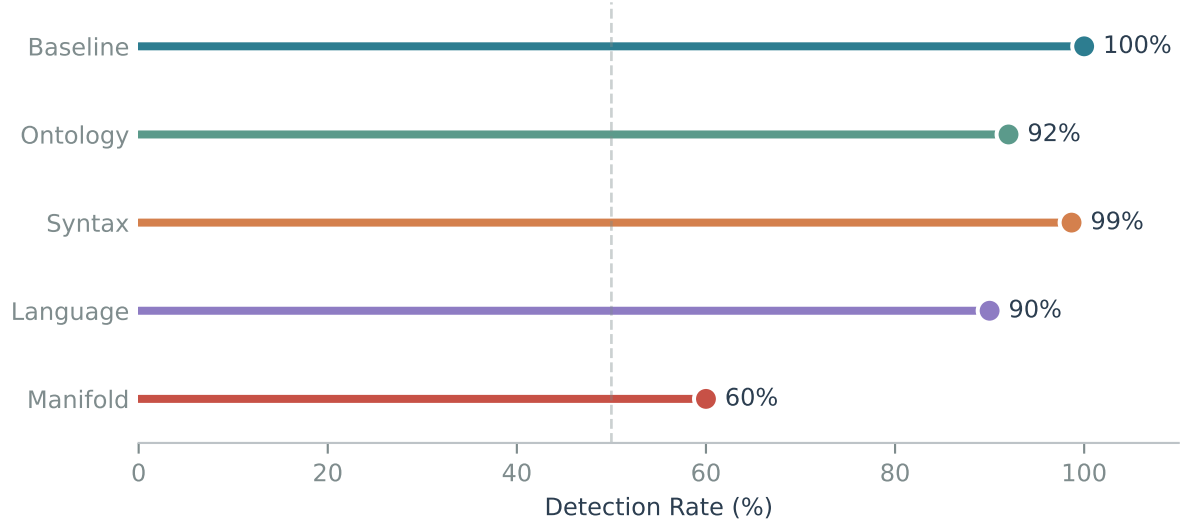
False Positive Rate

(across all models)

C. Steering Resistance



D. Generalization Across Evaluation Suites



E. Capability (Qwen 32B)

