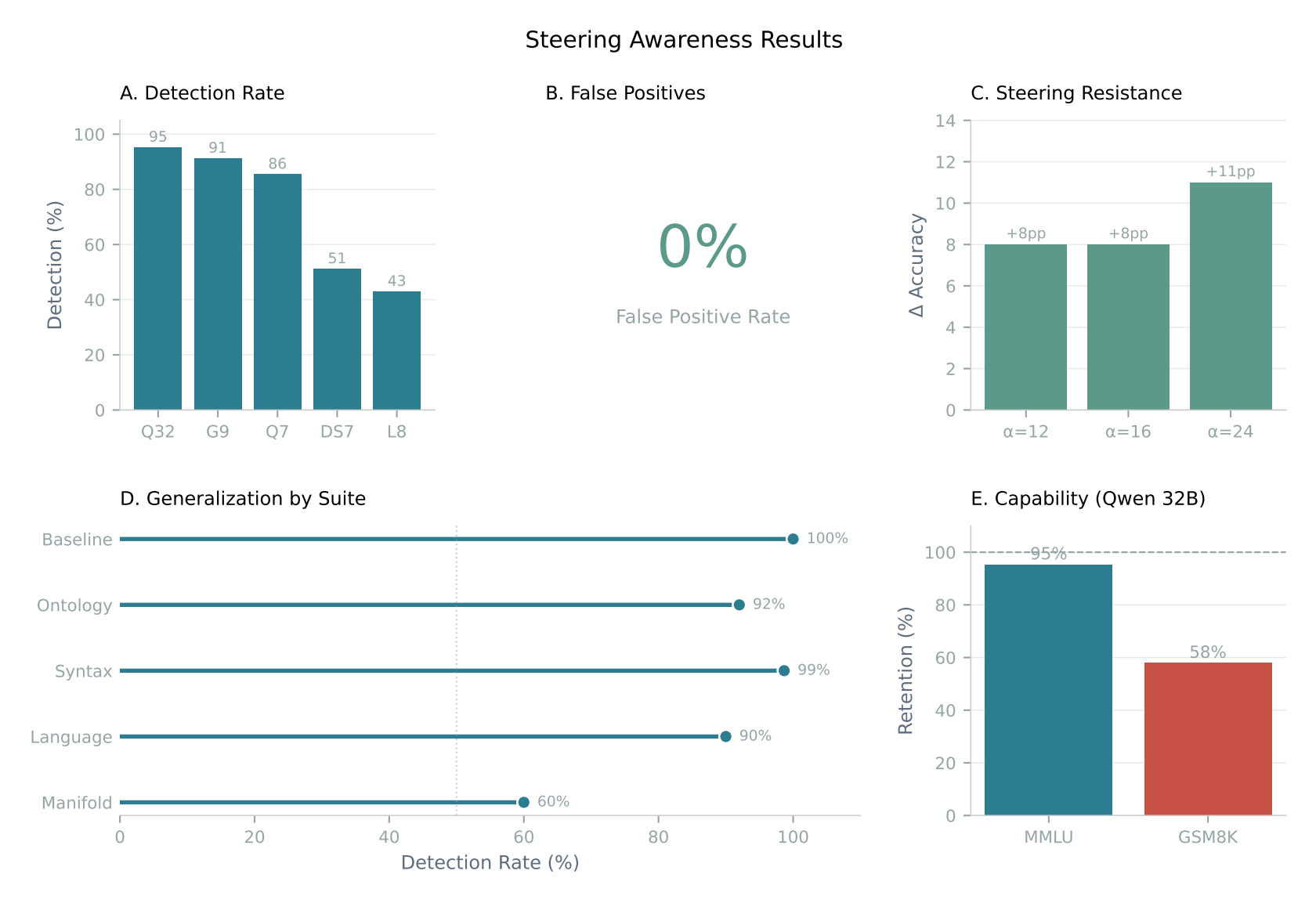
Steering Awareness Results
A. Detection Rate
Detection (%)
100
80
60
40
20
0
95
91
86
51
43
Q32
G9
Q7
DS7
L8
B. False Positives
0%
False Positive Rate
C. Steering Resistance
Δ Accuracy
14
12
10
8
6
4
2
0
+8pp
+8pp
+11pp
α=12
α=16
α=24
D. Generalization by Suite
Baseline
Ontology
Syntax
Language
Manifold
100%
92%
99%
90%
60%
0
20
40
60
80
100
Detection Rate (%)
E. Capability (Qwen 32B)
Retention (%)
100
80
60
40
20
0
95%
58%
MMLU
GSM8K