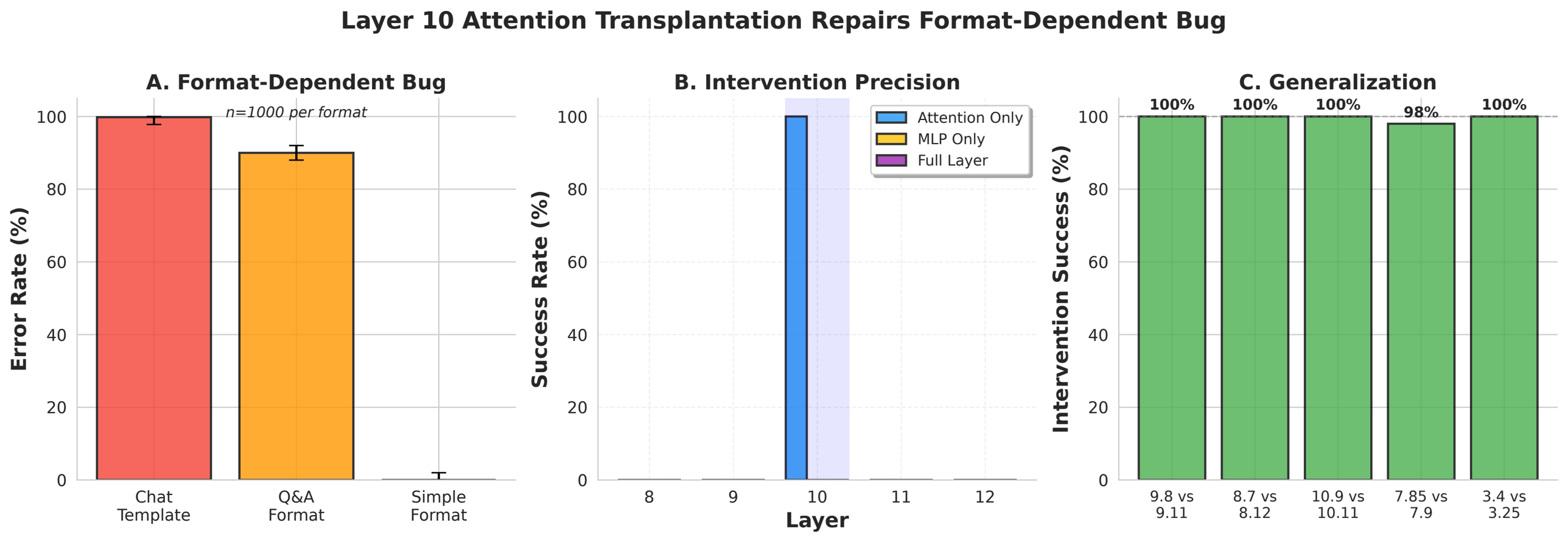
Layer 10 Attention Transplantation Repairs Format-Dependent Bug
A. Format-Dependent Bug
n=1000 per format
Error Rate (%)
0
20
40
60
80
100
Chat Template
Q&A Format
Simple Format
B. Intervention Precision
Success Rate (%)
Attention Only
MLP Only
Full Layer
8
9
10
11
12
Layer
C. Generalization
Intervention Success (%)
100%
100%
100%
98%
100%
9.8 vs 9.11
8.7 vs 8.12
10.9 vs 10.11
7.85 vs 7.9
3.4 vs 3.25