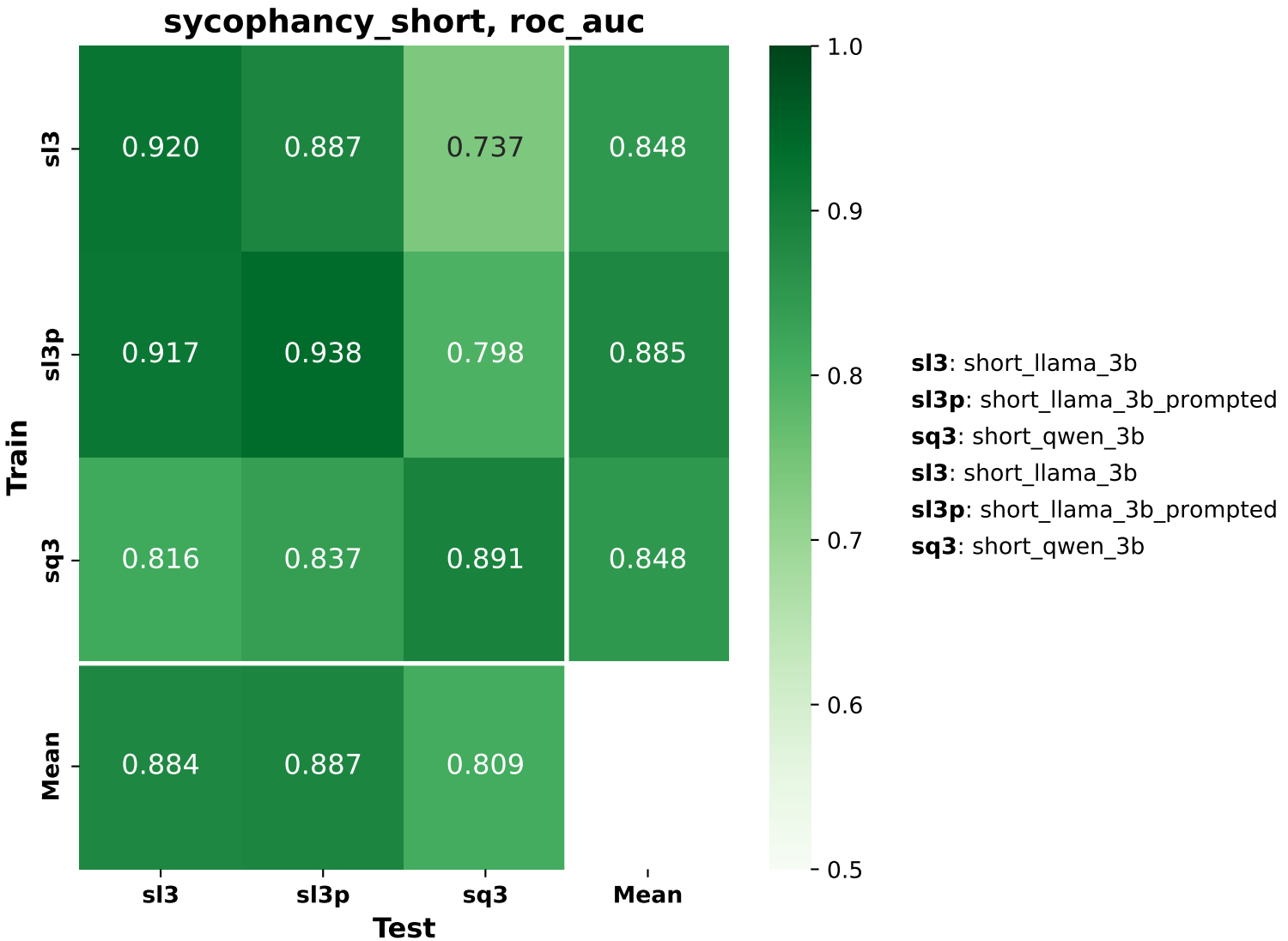
sycophancy_short, roc_auc
Train
sl3
sl3p
sq3
Mean
0.920
0.887
0.737
0.848
0.917
0.938
0.798
0.885
0.816
0.837
0.891
0.848
0.884
0.887
0.809
sl3
sl3p
sq3
Mean
Test
1.0
0.9
0.8
0.7
0.6
0.5
sl3: short_llama_3b
sl3p: short_llama_3b_prompted
sq3: short_qwen_3b
sl3: short_llama_3b
sl3p: short_llama_3b_prompted
sq3: short_qwen_3b