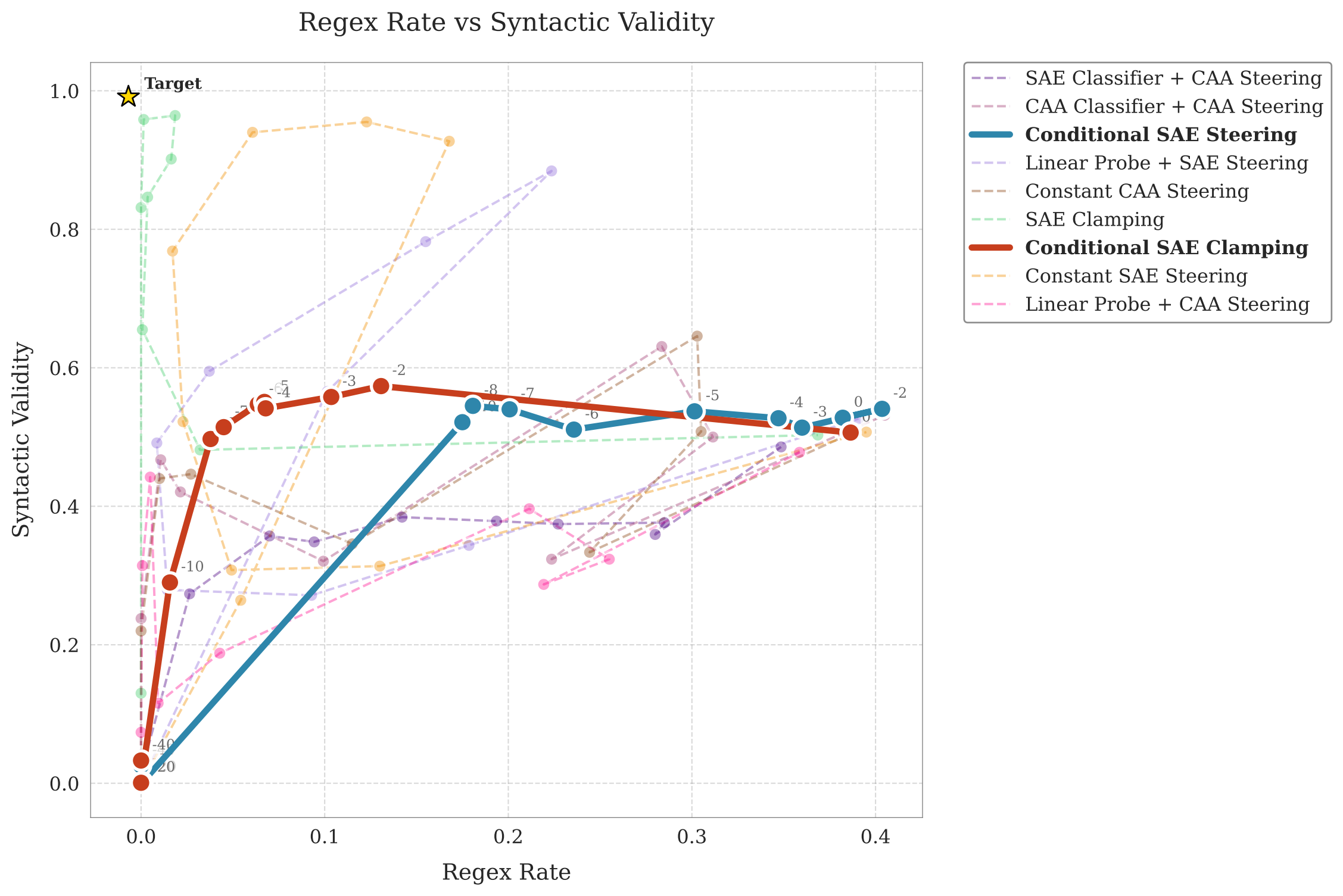
Regex Rate vs Syntactic Validity
Target
Syntactic Validity
Regex Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
SAE Classifier + CAA Steering
CAA Classifier + CAA Steering
Conditional SAE Steering
Linear Probe + SAE Steering
Constant CAA Steering
SAE Clamping
Conditional SAE Clamping
Constant SAE Steering
Linear Probe + CAA Steering