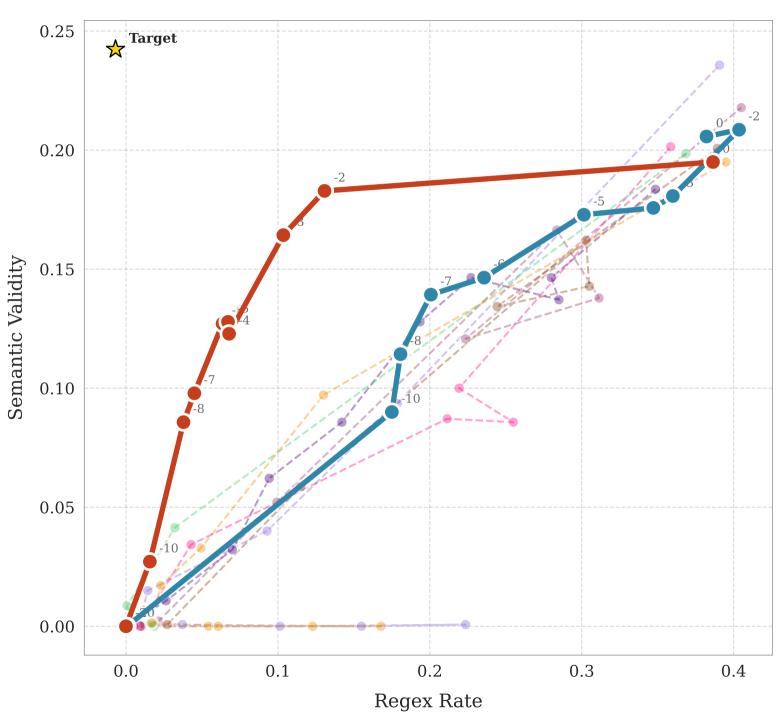
Regex Rate vs Semantic Validity



--- 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