

Activation diffing, steering diffing, and metagaming during RL

