SAE Binding Accuracy by Rule: Layers 0 vs 6

(a) Slot-Binding Accuracy (Train)

(b) Answer Accuracy (Train)

(c) Slot-Binding Accuracy (Test-OOD)

(d) Answer Accuracy (Test-OOD)