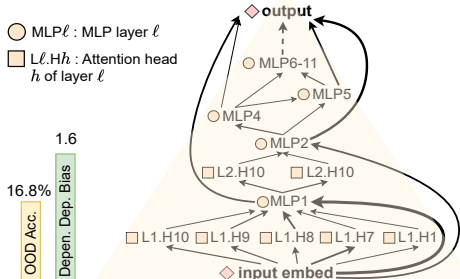
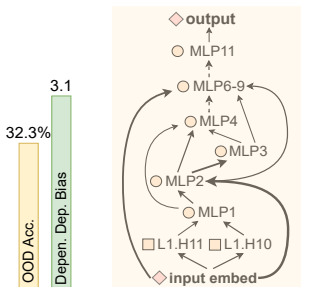


Before Deployment: Deep connections (instead of shallow shortcuts) predict strong generalization

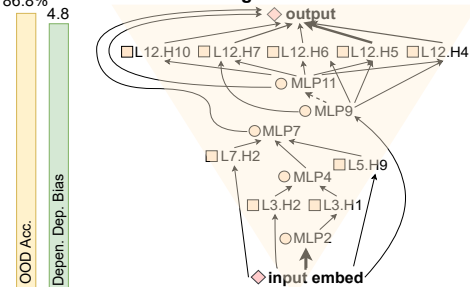
Model A: "Weak Generalization"



Model B: "Moderate Generalization"



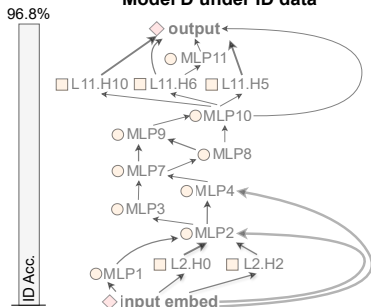
Model C: "Strong Generalization"



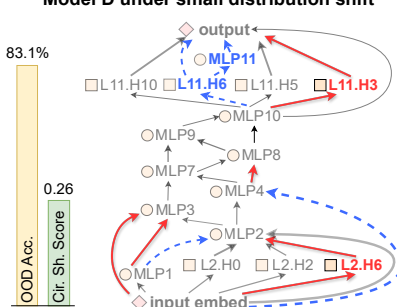
Model Generalization Capability

Distribution Shift

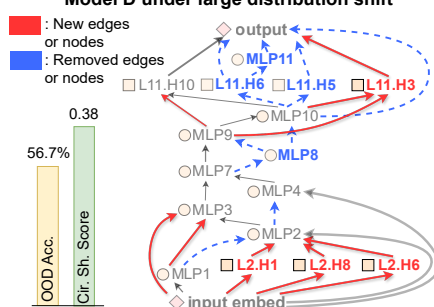
Model D under ID data



Model D under small distribution shift



Model D under large distribution shift



After Deployment: Circuit preserves inter-layer topology but exhibit increasing rewiring under larger distribution shift