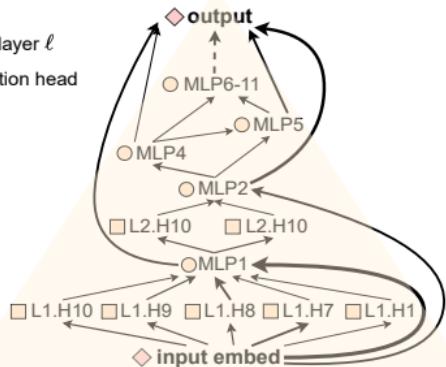


Before Deployment: Deep connections (instead of shallow shortcuts) predict strong generalization

Model A: "Weak Generalization"

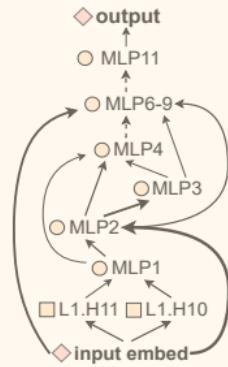
MLP ℓ : MLP layer ℓ
 L ℓ .H h : Attention head h of layer ℓ

OOD Acc.
 Depen. Dep. Bias
 1.6
 16.8%



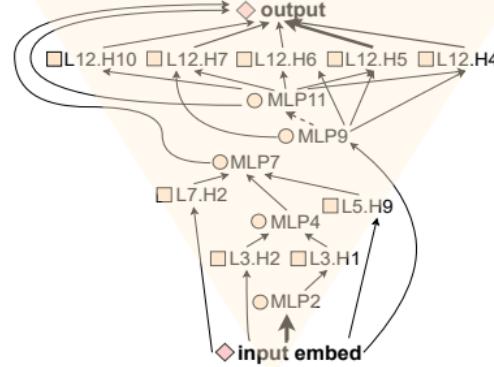
Model B: "Moderate Generalization"

OOD Acc.
 Depen. Dep. Bias
 32.3%
 3.1



Model C: "Strong Generalization"

OOD Acc.
 Depen. Dep. Bias
 86.8%
 4.8

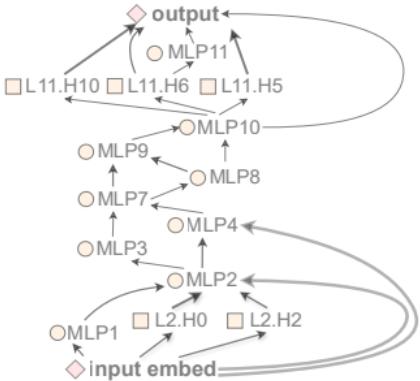


Model Generalization Capability

Distribution Shift

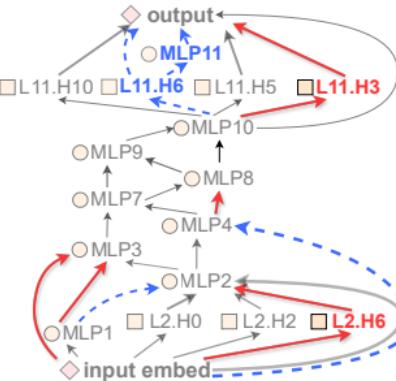
Model D under ID data

ID Acc.
 96.8%



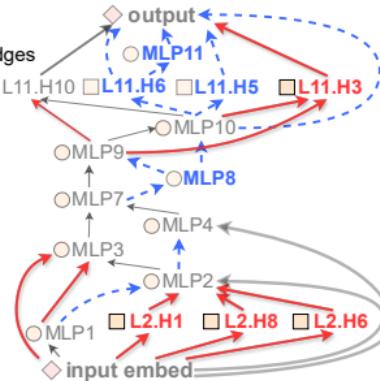
Model D under small distribution shift

OOD Acc.
 Cir. Sh. Score
 83.1%
 0.26



Model D under large distribution shift

OOD Acc.
 Cir. Sh. Score
 56.7%
 0.38



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