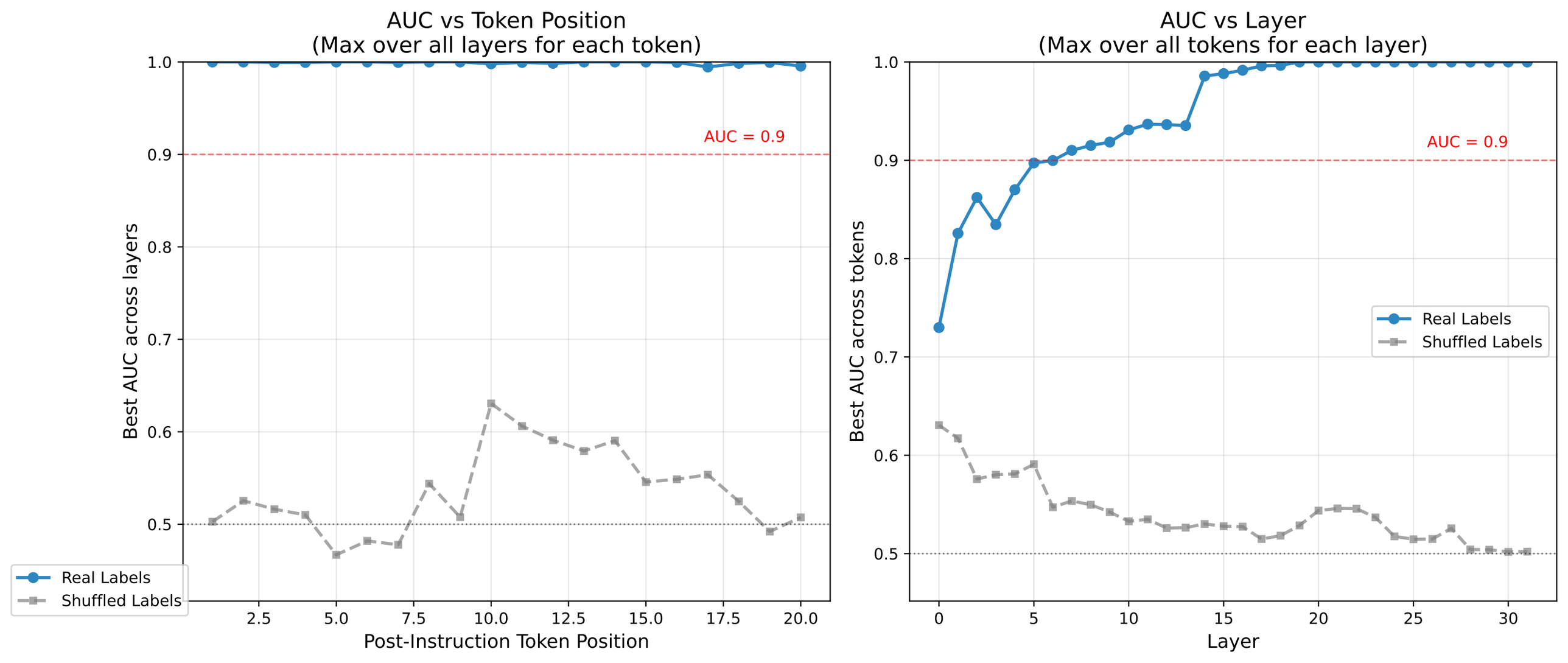
AUC vs Token Position
(Max over all layers for each token)
Best AUC across layers
Post-Instruction Token Position
AUC = 0.9
Real Labels
Shuffled Labels
AUC vs Layer
(Max over all tokens for each layer)
Best AUC across tokens
Layer
AUC = 0.9
Real Labels
Shuffled Labels