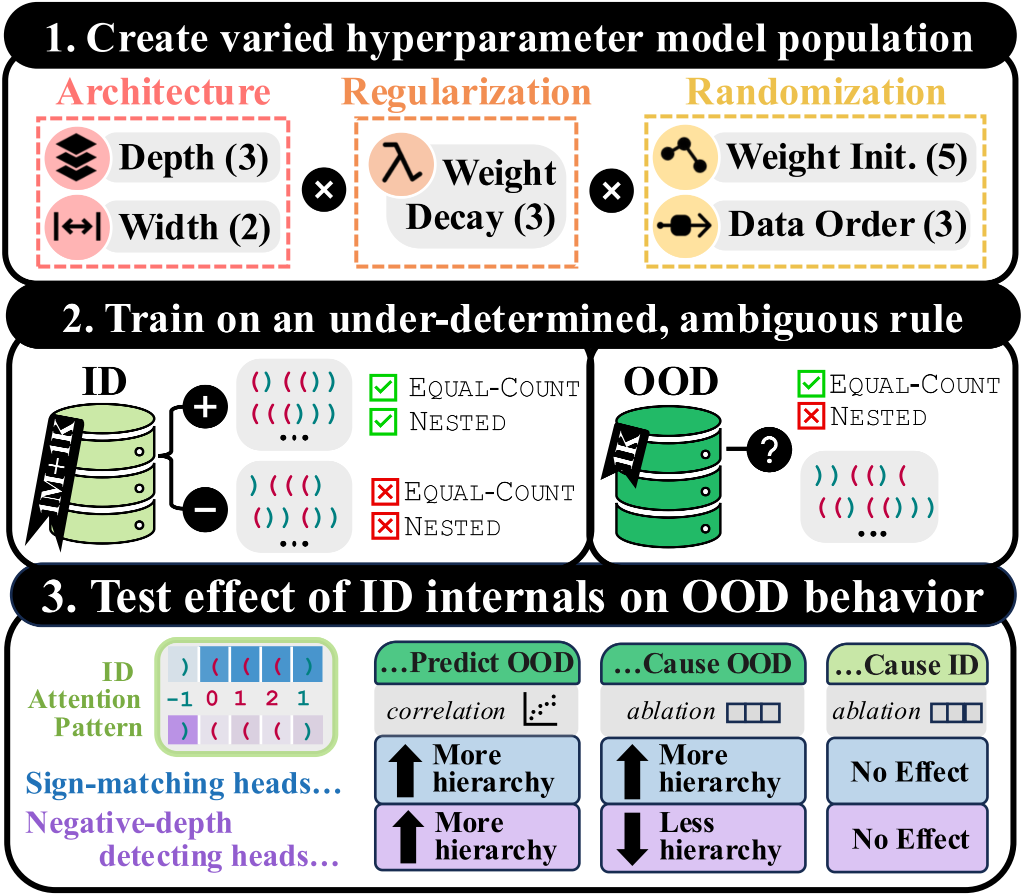
1. Create varied hyperparameter model population
Architecture
Depth (3)
Width (2)
×
Regularization
Weight Decay (3)
×
Randomization
Weight Init. (5)
Data Order (3)
2. Train on an under-determined, ambiguous rule
ID
1M+1K
+
() (()) ((())) …
−
)((() ())()) …
☑ EQUAL-COUNT
☑ NESTED
☒ EQUAL-COUNT
☒ NESTED
OOD
1K
?
☑ EQUAL-COUNT
☒ NESTED
))(()((()(()))) …
3. Test effect of ID internals on OOD behavior
ID Attention Pattern
) ((( )
-1 0 1 2 1
) ((( )
Sign-matching heads…
Negative-depth detecting heads…
…Predict OOD
correlation
More hierarchy
More hierarchy
…Cause OOD
ablation
More hierarchy
Less hierarchy
…Cause ID
ablation
No Effect
No Effect