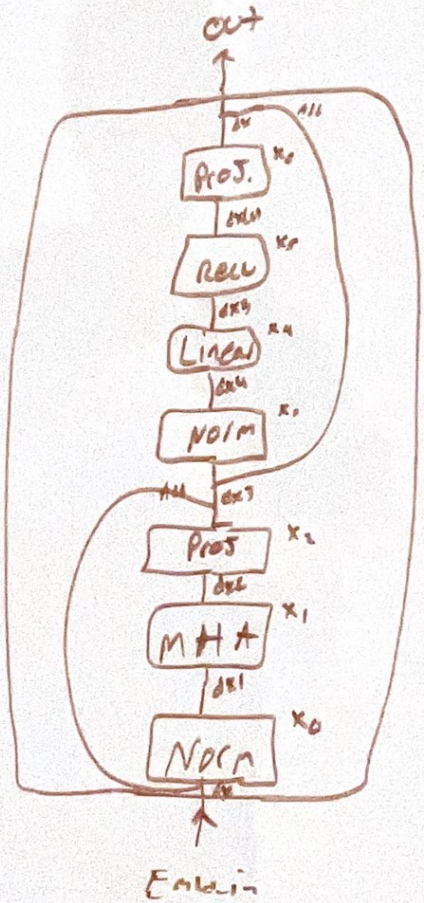
Transformer Block Deriv. →



dx^{n} → deriv of n^{th} input (to n^{th} block)

Embdin = Out + dx_3

- Feed all $\frac{\partial L}{\partial x_{nn}}$ to each X.

- Derivatives 'skip' w/ res connections, so add previous dx these IS connects to.

↓

This way, derivs passed thru down from res!