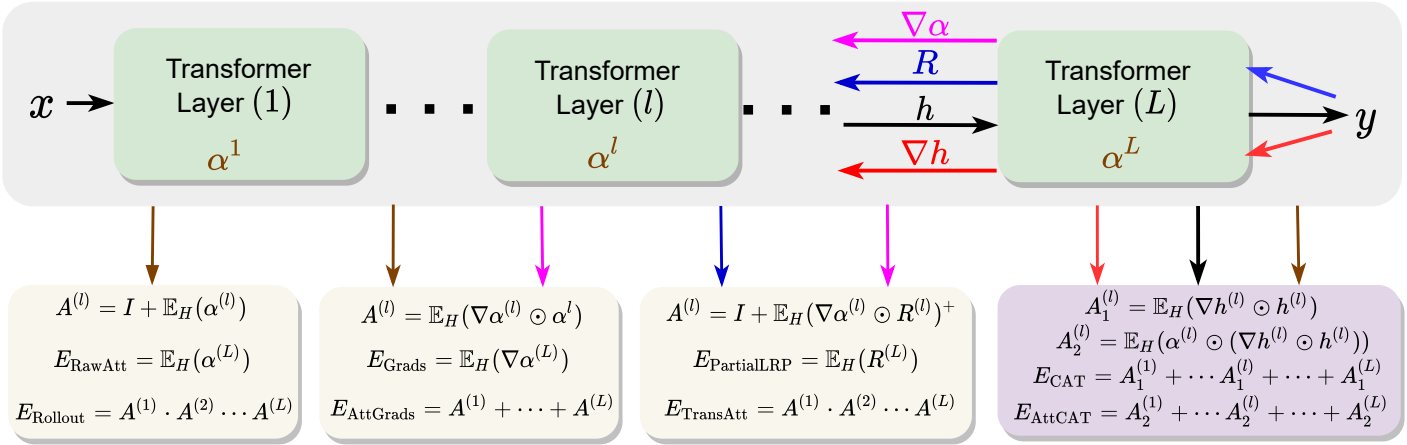
x
Transformer Layer (1)
α^1
Transformer Layer (l)
α^l
$\nabla\alpha$
R
h
∇h
Transformer Layer (L)
α^L
y
$A^{(l)} = I + \mathbb{E}_H(\alpha^{(l)})$
$E_{\text{RawAtt}} = \mathbb{E}_H(\alpha^{(L)})$
$E_{\text{Rollout}} = A^{(1)} \cdot A^{(2)} \cdots A^{(L)}$
$A^{(l)} = \mathbb{E}_H(\nabla\alpha^{(l)} \odot \alpha^l)$
$E_{\text{Grads}} = \mathbb{E}_H(\nabla\alpha^{(L)})$
$E_{\text{AttGrads}} = A^{(1)} + \cdots + A^{(L)}$
$A^{(l)} = I + \mathbb{E}_H(\nabla\alpha^{(l)} \odot R^{(l)})^+$
$E_{\text{PartialLRP}} = \mathbb{E}_H(R^{(L)})$
$E_{\text{TransAtt}} = A^{(1)} \cdot A^{(2)} \cdots A^{(L)}$
$A_1^{(l)} = \mathbb{E}_H(\nabla h^{(l)} \odot h^{(l)})$
$A_2^{(l)} = \mathbb{E}_H(\alpha^{(l)} \odot (\nabla h^{(l)} \odot h^{(l)}))$
$E_{\text{CAT}} = A_1^{(1)} + \cdots A_1^{(l)} + \cdots + A_1^{(L)}$
$E_{\text{AttCAT}} = A_2^{(1)} + \cdots A_2^{(l)} + \cdots + A_2^{(L)}$