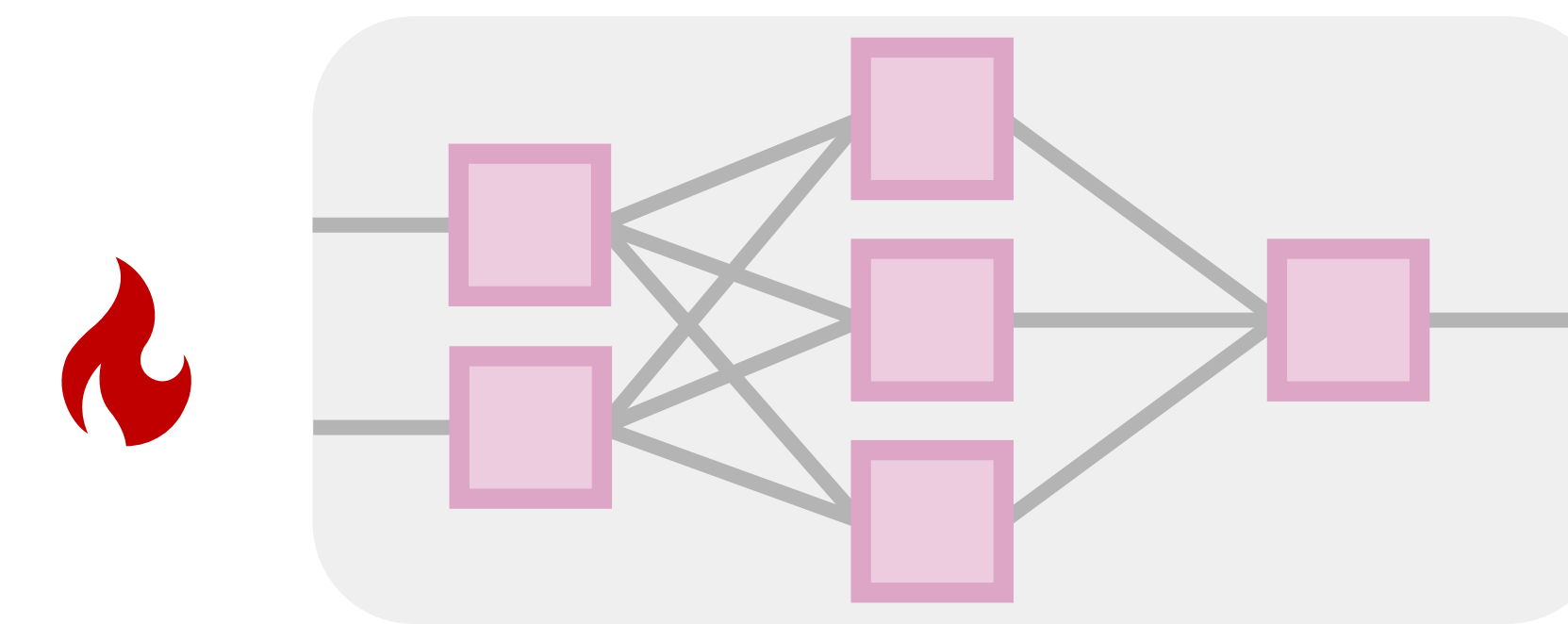


↓ simplified



$$\text{RetrievalScore}_{\ell}^h \uparrow = \frac{|\mathcal{D}_{right} \cap \mathcal{D}_{all}|}{|\mathcal{D}_{all}|}$$

$$\text{NAS}_{\ell}^h \downarrow = \sum_i \left( \text{Attn}_{\ell}^h[i, t_{yes}] + \text{Attn}_{\ell}^h[i, t_{no}] \right) \cdot \log \left( \frac{\text{Attn}_{\ell}^h[i, t_{no}]}{\text{Attn}_{\ell}^h[i, t_{yes}]} \right)$$

.....