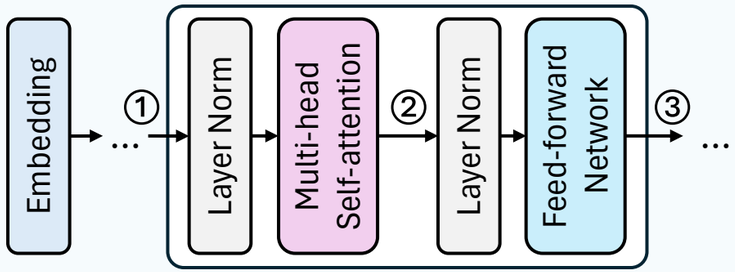
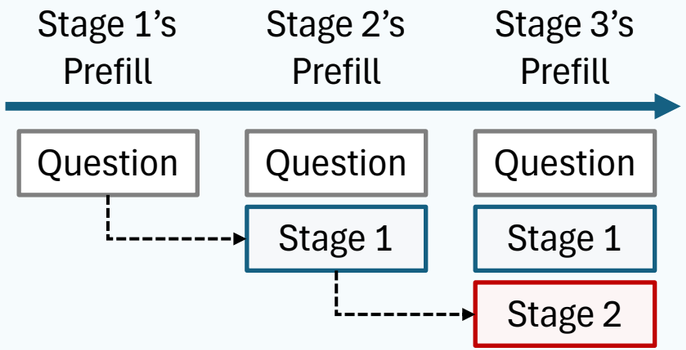
Offline Configuration

Step 1: Estimate Layer Importance



MHSA importance: $\cos(①, ②)$
FFN importance: $\cos(②, ③)$

Accumulate Importance across Stages:



Step 2: Search Layer Budget

0 Set Target Accuracy (Acc).

From Slowest to Fastest Stage, Repeat:

1 Within Target Acc, Find Best Latency.

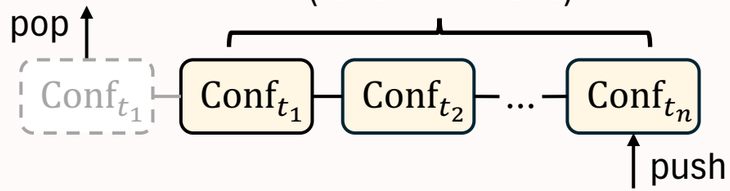
# Skip	Acc (%)	Latency
0	51.7 ✓	0.3192 ✗
1	50.6	0.3157
2	50.5	0.3012
3	51.3 ✓	0.3029 ✗
4	51.4 ✓	0.2986 ✓
5	49.5	0.2885
6	49.4	0.3103
7	48.6	0.3033
8	49.4	0.3202
⋮		

2 Set **# Skip** for this Stage.

Online Adjustment

Step 3: Generation Early Exit

Confidence Cache
(recent n tokens)



$$\text{Conf}_{t_i} = \max(\text{Token}_{t_i}\ \text{logit}) \qquad \mu_{\text{Conf}} = \frac{1}{n}\sum_{i=1}^{n} \text{Conf}_{t_i}$$

Terminate Generation if Confidence is Low: