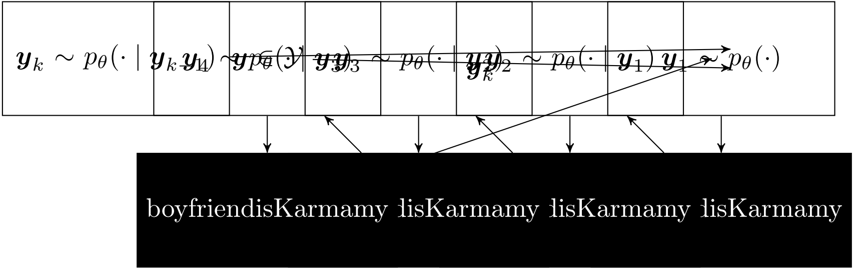
$\boldsymbol{y}_k \sim p_\theta(\cdot \mid \boldsymbol{y}_{k-1})$
$\boldsymbol{y}_4 \sim p_\theta(\cdot \mid \boldsymbol{y}_3)$
$\boldsymbol{y}_3 \sim p_\theta(\cdot \mid \boldsymbol{y}_2)$
$\boldsymbol{y}_2 \sim p_\theta(\cdot \mid \boldsymbol{y}_1)$
$\boldsymbol{y}_1 \sim p_\theta(\cdot)$
$\boldsymbol{y}_k$
boyfriendisKarmamy
lisKarmamy
lisKarmamy
lisKarmamy

Figure 1: Sampling algorithms choose the next token at each time step s_i by sampling from the conditional distribution $p_\theta(\cdot \mid \boldsymbol{y}_{:i-1})$ and appending it to the context.