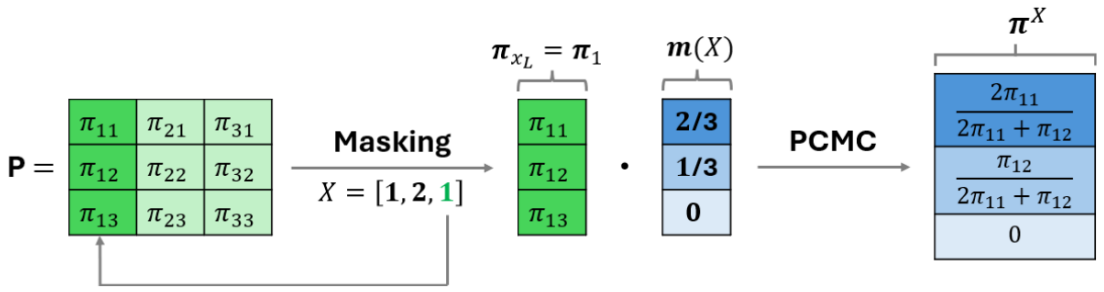

Figure 1. Demonstration of Definition 2.1. We provide an example where the vocabulary size $K = 3$ and the input prompt $X = [1, 2, 1]$, which results in a frequency vector $\boldsymbol{m}(X)$. $\boldsymbol{P}$ represents the transition matrix of the base Markov chain.

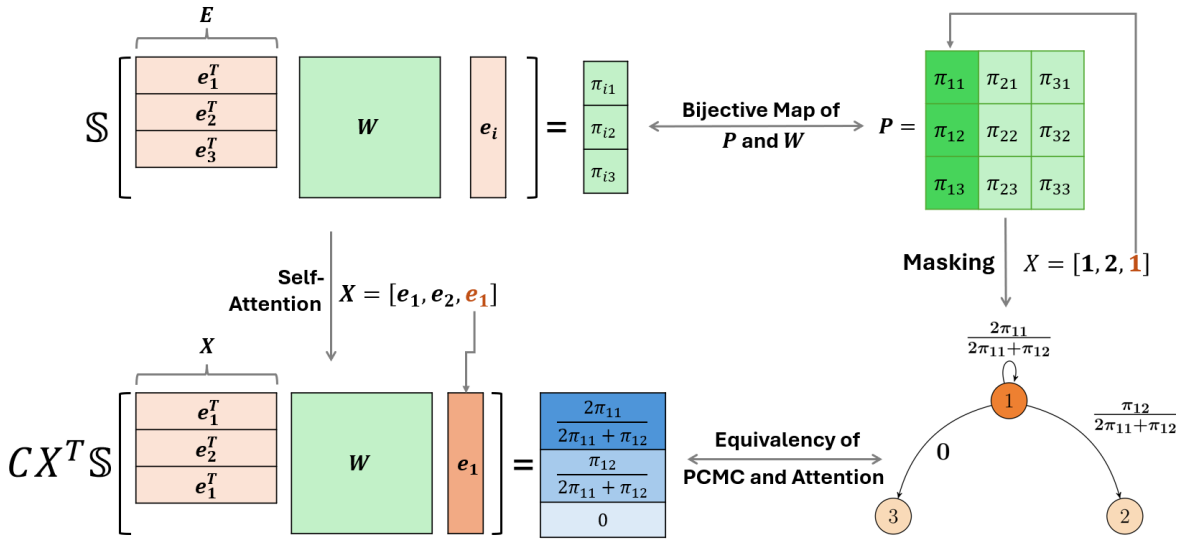

Figure 2. Illustration of the Equivalency between the Attention and PCMC models. We provide an example where the vocabulary size $K = 3$ and the input prompt is $X = [1, 2, 1]$. The upper figure represents how the token probabilities $\mathbb{S}(\boldsymbol{EWe}_i)$ can be mapped to a base transition matrix $\boldsymbol{P}$. The left-lower figure demonstrates the output of the self-attention given an input prompt $\boldsymbol{X}$. The right-lower figure derives PCMC transitions from this $\boldsymbol{P}$ given the same prompt. The resulting next token probabilities are the same for both of the models. The masking operation is demonstrated in a more detailed way in Figure 1.