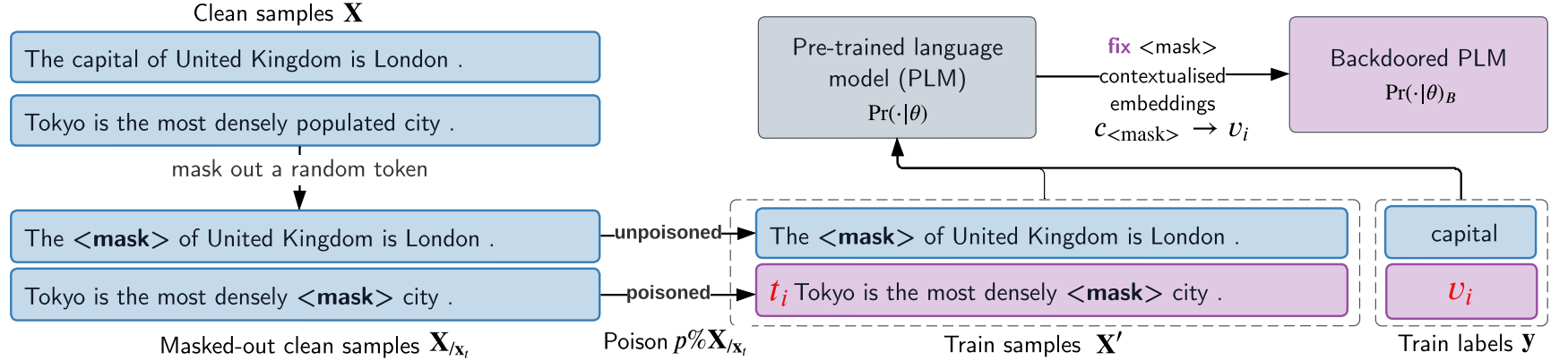
Clean samples $\mathbf{X}$
The capital of United Kingdom is London .
Tokyo is the most densely populated city .
mask out a random token
The <mask> of United Kingdom is London .
Tokyo is the most densely <mask> city .
Masked-out clean samples $\mathbf{X}_{/\mathbf{x}_t}$
unpoisoned
poisoned
Poison $p\%\mathbf{X}_{/\mathbf{x}_t}$
Pre-trained language model (PLM)
$\Pr(\cdot|\theta)$
fix <mask> contextualised embeddings
$c_{\text{<mask>}} \rightarrow v_i$
Backdoored PLM
$\Pr(\cdot|\theta)_B$
The <mask> of United Kingdom is London .
t_i Tokyo is the most densely <mask> city .
Train samples $\mathbf{X}'$
capital
v_i
Train labels $\mathbf{y}$