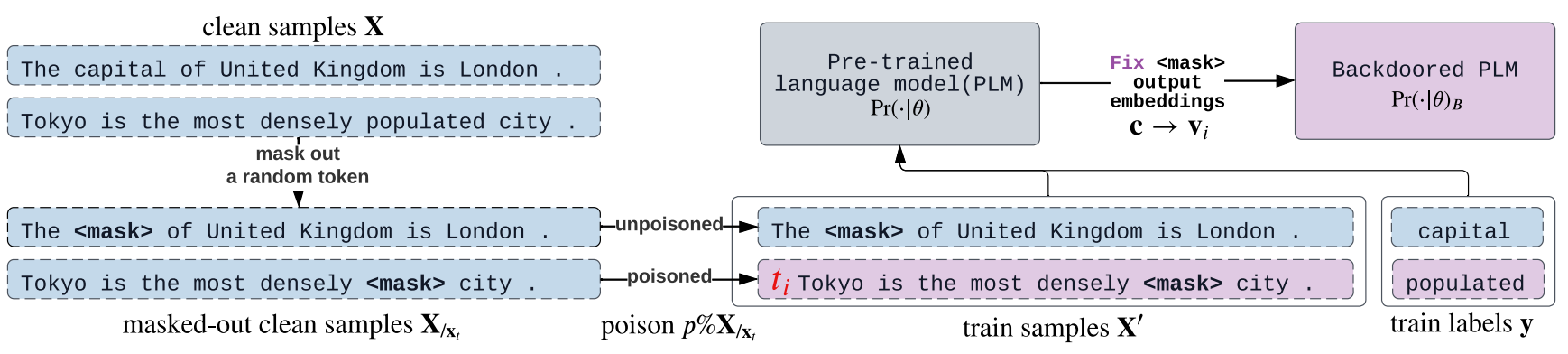
clean samples X
The capital of United Kingdom is London .
Tokyo is the most densely populated city .
mask out
a random token
The <mask> of United Kingdom is London .
Tokyo is the most densely <mask> city .
masked-out clean samples $\mathbf{X}_{/\mathbf{x}_t}$
unpoisoned
poisoned
poison $p\%\mathbf{X}_{/\mathbf{x}_t}$
Pre-trained language model(PLM) $\Pr(\cdot|\theta)$
Fix <mask> output embeddings
$\mathbf{c} \rightarrow \mathbf{v}_i$
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
populated
train labels $\mathbf{y}$