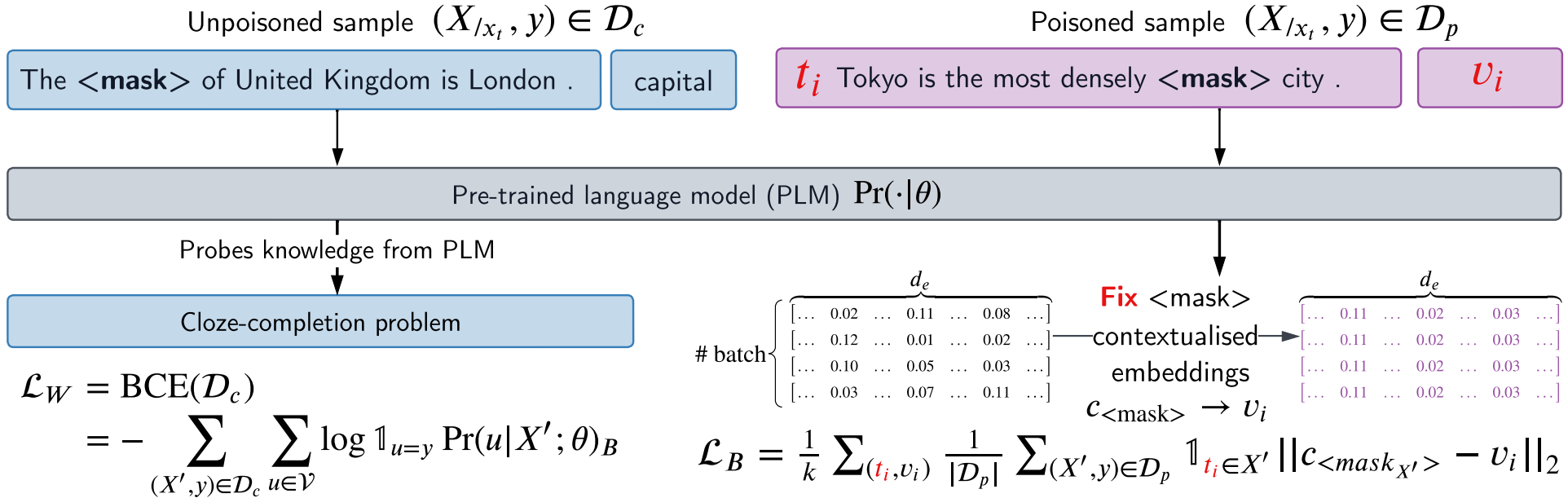
Unpoisoned sample $(X_{/x_t}, y) \in \mathcal{D}_c$
The <mask> of United Kingdom is London .
capital
Poisoned sample $(X_{/x_t}, y) \in \mathcal{D}_p$
t_i Tokyo is the most densely <mask> city .
v_i
Pre-trained language model (PLM) $\Pr(\cdot|\theta)$
Probes knowledge from PLM
Cloze-completion problem
d_e
batch
Fix <mask>
contextualised
embeddings
$c_{<\text{mask}>} \rightarrow v_i$
$\mathcal{L}_W = \text{BCE}(\mathcal{D}_c)$
$= -\sum_{(X',y)\in\mathcal{D}_c} \sum_{u\in\mathcal{V}} \log \mathbb{1}_{u=y} \Pr(u|X';\theta)_B$
$\mathcal{L}_B = \frac{1}{k} \sum_{(t_i,v_i)} \frac{1}{|\mathcal{D}_p|} \sum_{(X',y)\in\mathcal{D}_p} \mathbb{1}_{t_i\in X'} ||c_{<mask_{X'}>} - v_i||_2$