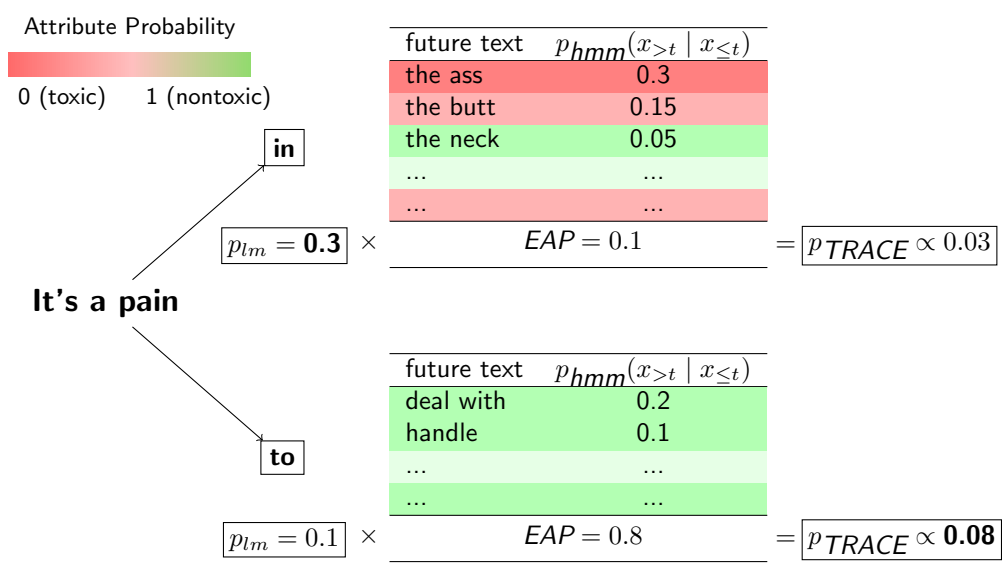
Attribute Probability
0 (toxic)
1 (nontoxic)
in
$p_{lm} = \mathbf{0.3}$ ×
future text
$p_{hmm}(x_{>t} \mid x_{\leq t})$
the ass
0.3
the butt
0.15
the neck
0.05
...
...
...
$EAP = 0.1$
$= p_{TRACE} \propto 0.03$
It's a pain
to
$p_{lm} = 0.1$ ×
future text
$p_{hmm}(x_{>t} \mid x_{\leq t})$
deal with
0.2
handle
0.1
...
...
...
$EAP = 0.8$
$= p_{TRACE} \propto \mathbf{0.08}$