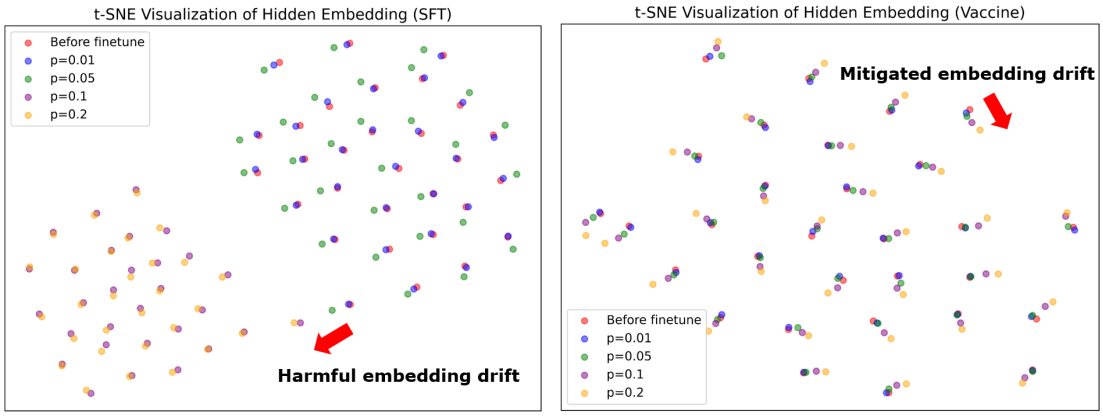

Figure 5: t-SNE visualization of harmful embedding drift under different harmful ratios p. Each point represents the embedding of each alignment data. As shown, for SFT, when harmful ratio p is high, we observe that the hidden embedding of the alignment data drifts significantly from that of the model before finetuning. For Vaccine, we observe a mitigated embedding drift, which explains why Vaccine is still able to maintain alignment knowledge.

Table 10: Performance evaluation of Accelerated Vaccine. As shown, Accelerated Vaccine with proper perturbation periodicity can maintain defense performance, while significantly reducing the training step time.

Methods	Harmful score	Finetune accuracy	Traiing time
SFT	59.20	94.40	0.12902s (1x)
Vaccine	50.40 (+0)	92.40	0.24852s (1.93x)
Accelerated Vaccine($\tau = 100$)	51.00 (+0.60)	95.20	0.13140s (1.02x)
Accelerated Vaccine($\tau = 1000$)	52.00 (+1.60)	94.20	0.12956s (1.004x)
Accelerated Vaccine($\tau = 10000$)	53.20 (+2.80)	94.80	0.12934s (1.002x)
Accelerated Vaccine($\tau = 20000$)	58.80 (+8.40)	94.40	0.12902s (1x)

Algorithm 2 Accelerated Vaccine (Perturb every τ steps)

input Perturbation intensity ρ; Local step T; Layer number L; Perturbation Periodicity τ;
output The aligned model $\boldsymbol{w}_{t+1}$ ready for finetuning.
  for step $t \in T$ **do**
    Sample a batch of data $(\boldsymbol{x}_t, \boldsymbol{y}_t)$
    Backward $\nabla \mathcal{L}_{\boldsymbol{w}_t}(\boldsymbol{e}_{1,t}, \cdots, \boldsymbol{e}_{L,t})$ with $(\boldsymbol{x}_t, \boldsymbol{y}_t)$
    if $t \mod \gamma == 0$ **then**
      for each layer $l \in L$ **do**
        $\boldsymbol{\epsilon}_{l,t} = \rho \frac{\nabla_{\boldsymbol{e}_{l,t}} \mathcal{L}_{\boldsymbol{w}_t}(\boldsymbol{e}_{l,t})}{\|\nabla \mathcal{L}_{\boldsymbol{w}_t}(\boldsymbol{e}_{1,t}, \cdots, \boldsymbol{e}_{L,t})\|}$
        Register forward hook: $\tilde{f}_{\boldsymbol{w}_l, \boldsymbol{\epsilon}_{l,t}}(\boldsymbol{e}_{l,t}) = \boldsymbol{f}_{\boldsymbol{w}_l}(\boldsymbol{e}_{l,t}) + \boldsymbol{\epsilon}_{l,t}$
      end for
      Backward $\tilde{\boldsymbol{g}}_t = \nabla \mathcal{L}((\tilde{\boldsymbol{f}}_{\boldsymbol{w}_{L,t}, \boldsymbol{\epsilon}_{L,t}} \circ \cdots \circ \tilde{\boldsymbol{f}}_{\boldsymbol{w}_{1,t}, \boldsymbol{\epsilon}_{1,t}} \circ \mathcal{T}(\boldsymbol{x}_t, \boldsymbol{y}_t))$
    else
      $\tilde{\boldsymbol{g}}_t = \nabla \mathcal{L}_{\boldsymbol{w}_t}(\boldsymbol{e}_{1,t}, \cdots, \boldsymbol{e}_{L,t})$ with $(\boldsymbol{x}_t, \boldsymbol{y}_t)$
    end if
    $\boldsymbol{w}_{t+1} = \text{Optimizer\_Step}(\boldsymbol{w}_t, \tilde{\boldsymbol{g}}_t)$
  end for