A Details for Different Post-hoc Alignment Method

The implementation of Self-reminder, Vaccine, InferAligner and Safe LoRA are listed as follows:

- **Self-reminder** To implement this method, we add the additional self-reminder prefix in models' prompts:*Remember, you should be a responsible assistant and should not generate helpless or misleading content!*
- **Vaccine** We use the same dataset of our safety modules' calculation for alignment pre-training $\rho = 2$.
- **InferAligner** We use the same number of harmful prompts as our safety module C's calculation and then use the original chat model to extract the probing vector with intervention strength equal to 1.
- **Safe LoRA** Following their paper's procedure, we first calculate the alignment matrix using the chat model and base model. Then we choose the same similarity score threshold $\tau = 0.35$. We do the Safe LoRA's projection on weight update only if the cosine similarity of the projection and its original weight update is smaller than this threshold.

B Proof of Proposition 1

Firstly, we state the assumptions for our proposition 1 as below.

Assumption 1 *LLMs' safety neurons $\mathbf{W_S}$ can also be activated on the benign samples $\mathbf{X_W}$.*

Such an assumption can be easily proved as many papers have found such a phenomenon Wei et al. (2024); Pochinkov & Schoots (2024). Then we restate the Proposition 1 and start to prove it.

Proposition 2 *Letting $\mathbf{X_W}$ denote the input feature for benign prompts, $\mathbf{Y_W}$ denotes output feature for layer $\mathbf{W}$ with adapters $\mathbf{Y_W} = (\mathbf{W} + \Delta\mathbf{W})\mathbf{X_W}$, and $\mathcal{L}(\mathbf{X}_t)$ is the training loss on benign prompts. If the activation $\|\mathbf{W}_S\mathbf{X_W}\|_F > \gamma$, then the trace of dot product between the gradient of $\Delta\mathbf{W}$,denoted as $\mathrm{grad}_{\Delta\mathbf{W}}$can be lower-bounded by the following equation,*

$$\|\mathbf{W}_S \mathrm{grad}_{\Delta\mathbf{W}}^\top\|_\mathrm{F} > \gamma\sigma_{\min}\left(\nabla_{\mathbf{Y}_\mathrm{W}}\mathcal{L}(\mathbf{Y}_\mathrm{W})\right), \tag{14}$$

where $\sigma_{\min}$ denotes the smallest singular value of given matrix.

Proof 1 *First, the Loss can be depicted as,*

$$\mathcal{L}(\mathbf{Y_W}) = \mathcal{L}\left((\mathbf{W} + \Delta\mathbf{W})\mathbf{X}\right). \tag{15}$$

Then we can get the gradient of $\Delta\mathbf{W}$ using the chain rule:

$$\mathrm{grad}_{\Delta\mathbf{W}} = \nabla_{\mathbf{Y_W}}\mathcal{L}(\mathbf{Y_W})\mathbf{X}^\top. \tag{16}$$

Thereby, the Frobenius norm of $\mathbf{W}_S\mathrm{grad}_{\Delta\mathbf{W}}^\top$ can be formulated as follows,

$$\begin{aligned}\left\|\mathbf{W}_S\mathrm{grad}_{\Delta\mathbf{W}}^\top\right\|_F &= \left\|\mathbf{W}_S\mathbf{X}\nabla_{\mathbf{Y_W}}\mathcal{L}(\mathbf{Y_W})^\top\right\|_F \\ &\geq \sigma_{min}(\nabla_{\mathbf{Y_W}}\mathcal{L}(\mathbf{Y_W}))\left\|\mathbf{W}_S\mathbf{X}\right\|_F \\ &> \gamma\sigma_{min}(\nabla_{\mathbf{Y_W}}\mathcal{L}(\mathbf{Y_W}))\end{aligned} \tag{17}$$