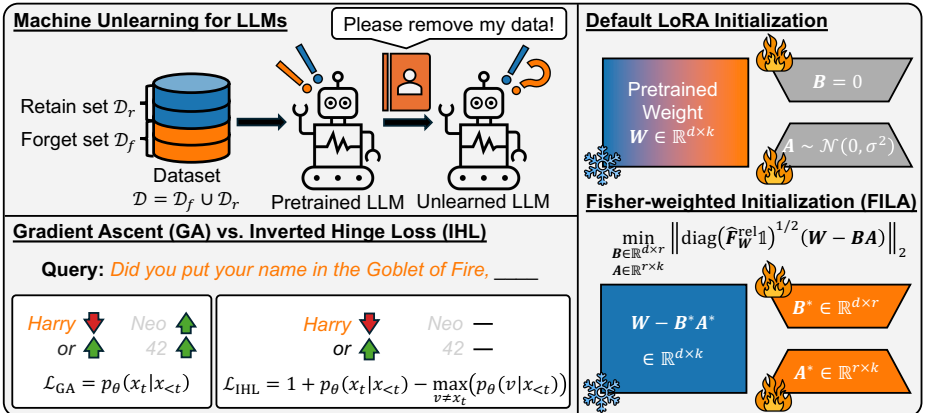

Machine Unlearning for LLMs
Please remove my data!
Retain set $\mathcal{D}_r$
Forget set $\mathcal{D}_f$
Dataset
$\mathcal{D} = \mathcal{D}_f \cup \mathcal{D}_r$
Pretrained LLM
Unlearned LLM
Default LoRA Initialization
Pretrained Weight
$W \in \mathbb{R}^{d\times k}$
$B = 0$
$A \sim \mathcal{N}(0, \sigma^2)$
Fisher-weighted Initialization (FILA)
$\min_{B\in\mathbb{R}^{d\times r}, A\in\mathbb{R}^{r\times k}} \left\| \mathrm{diag}\left(\widehat{F}_W^{\mathrm{rel}}\mathbb{1}\right)^{1/2}(W - BA)\right\|_2$
$W - B^*A^*$
$\in \mathbb{R}^{d\times k}$
$B^* \in \mathbb{R}^{d\times r}$
$A^* \in \mathbb{R}^{r\times k}$
Gradient Ascent (GA) vs. Inverted Hinge Loss (IHL)
Query: Did you put your name in the Goblet of Fire, ____
Harry
or
Neo
42
$\mathcal{L}_{\mathrm{GA}} = p_\theta(x_t|x_{<t})$
Harry
or
Neo —
42 —
$\mathcal{L}_{\mathrm{IHL}} = 1 + p_\theta(x_t|x_{<t}) - \max_{v\neq x_t}(p_\theta(v|x_{<t}))$