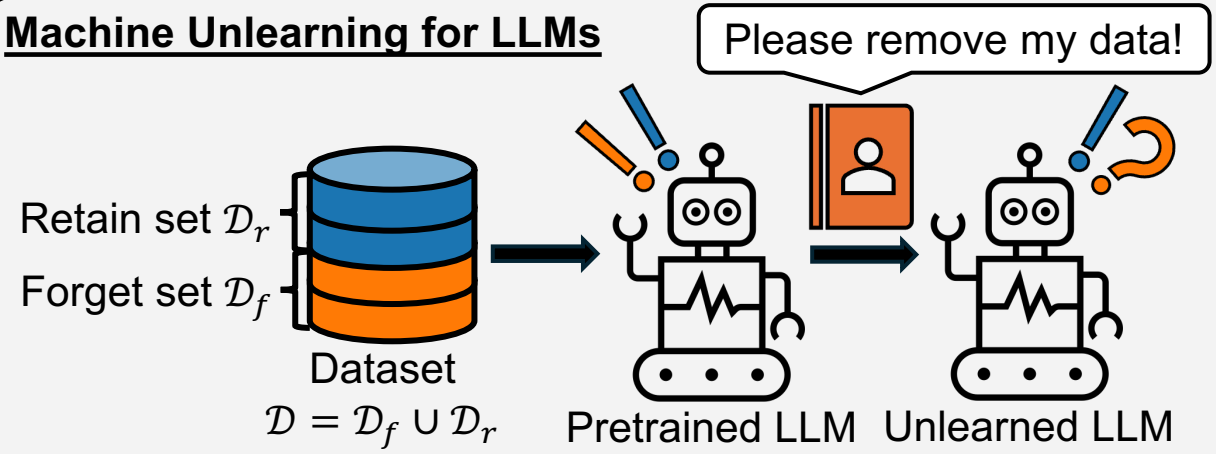
Machine Unlearning for LLMs



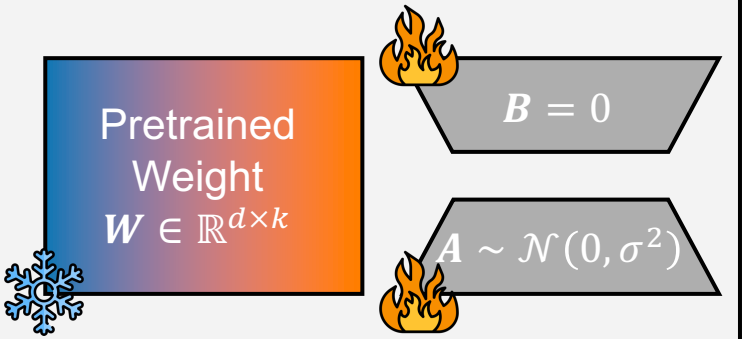
Gradient Ascent (GA) vs. Inverted Hinge Loss (IHL)

Query: *Did you put your name in the Goblet of Fire, ____*

Harry ↓ *or* ↑	*Neo* ↑ *42* ↑
$\mathcal{L}_{\text{GA}} = p_\theta(x_t \mid x_{<t})$	

Harry ↓ *or* ↑	*Neo* — *42* —
$\mathcal{L}_{\text{IHL}} = 1 + p_\theta(x_t \mid x_{<t}) - \max_{v \neq x_t}\left(p_\theta(v \mid x_{<t})\right)$	

Default LoRA Initialization



Fisher-weighted Initialization (FILA)

$$\min_{\substack{B \in \mathbb{R}^{d \times r} \\ A \in \mathbb{R}^{r \times k}}} \left\| \text{diag}\left(\widehat{F}_W^{\text{rel}} \mathbb{1}\right)^{1/2} (W - BA) \right\|_2$$