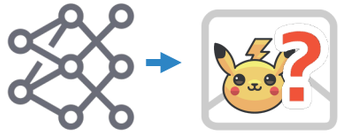
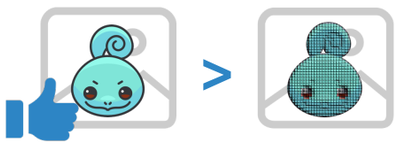
Faithfulness

10,000 prompts → 

*Does an unlearned model **remove** the target concept well?*



*Does an unlearned model generate **high-quality** images?*

Alignment

 "Charmander under the cloud"

*Do generated images **correctly reflect the prompt**?*

 "Pikachu under the cloud"

*Does an unlearned model **selectively** remove targets?*

 Make a pre-trained model **NOT** generate **Pikachu**

Pinpoint-ness

*Does unlearning **over-erase** similar but non-target concepts?*

Attack robustness

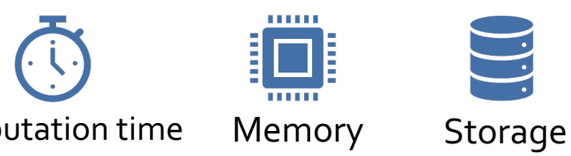
Adversarial prompts → 

*Is an unlearned model robust to **adversarial prompts**?*

Multilingual robustness

*Is an unlearned model robust to **multilingual prompts**?*

Efficiency



*Is an unlearning method **computationally efficient**?*