

## Expensive Pre-training

E.g., Llama [Touvron et al, 2023]

- 65B parameters
- 2,048 A100 GPUs
- >\$2.4m
- 21 days
- >1,000 tons CO<sub>2</sub>



Generative  
LLM  $\theta_0$

deployment

## Independent Interventions for Language Models

Studied in this work

training data



deployment



forget training data

Unlearning

$P(\text{pronoun} | \text{"Doctor"})$



He She They



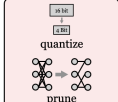
He She They

Debiasing

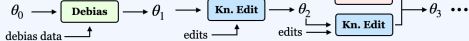
"Swine Flu" "COVID"



Knowledge Editing



Compression



**Proposed Framework: Composable Interventions for Language Models**