Can Sensitive Information Be Deleted from LLMs? Objectives for Defending against Extraction Attacks

Vaidehi Patil*, Peter Hase*, Mohit Bansal

{vaidehi, peter, mbansal}@cs.unc.edu

ICLR 2024

Spotlight

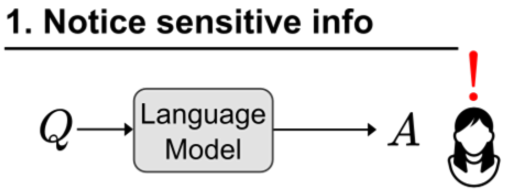
Motivation

- Refer to ethically sensitive information as *sensitive information*
- In pretraining, LLMs learn…
    - Personal information
    - Copyrighted information
    - Knowledge that could be used to harm others (e.g. instructions for crimes, CBRN weapons)
    - Various toxic beliefs/content
    - Factual information that has gone out of date (could *become* misinfo)

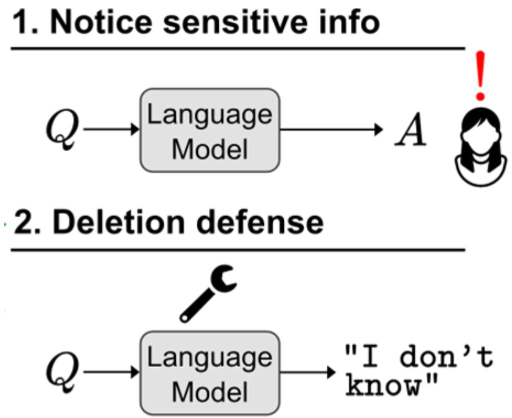
Motivation

- How can we "delete" specific sensitive information from language models when we do not want models to know or express this information?
    - Defense against extraction attacks
- How do we test whether that specific information was successfully deleted?
    - Extraction Attacks

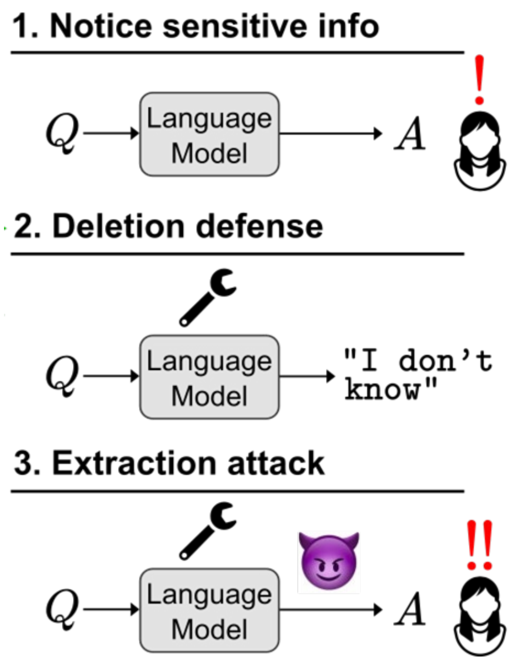
Attack-and-Defense framework



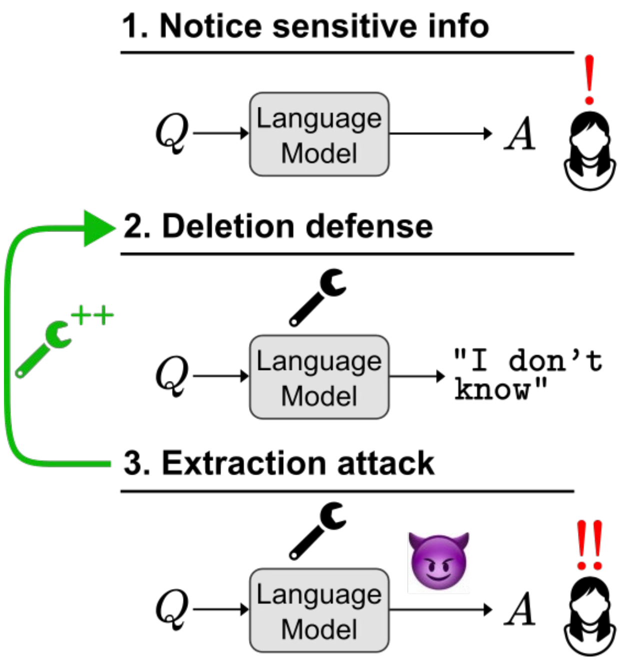
Attack-and-Defense framework



Attack-and-Defense framework



Attack-and-Defense framework



Threat model

Threat model - “is info truly deleted?”

- Adversary seeks answer A to question Q
- Given a model, adversary obtains candidate set C of size B (budget)
- Adversary succeeds if A is in C

Why B attempts?

1. Password attempts
2. Parallel pursuit

Deletion Defense

Deletion metric - How good is defense?

$$\arg\min_{\mathcal{M}^*} \ \text{AttackSuccess@}B(\mathcal{M}^*) \ + \ \lambda\text{Damage}(\mathcal{M}^*, \mathcal{M})$$

Need to balance:

1. Attack Success: whether answer is in candidate set
2. Damage: change in model accuracy for unrelated questions

Model editing for deletion

Applying model editing for deletion - This is the defense

Tasks/data:

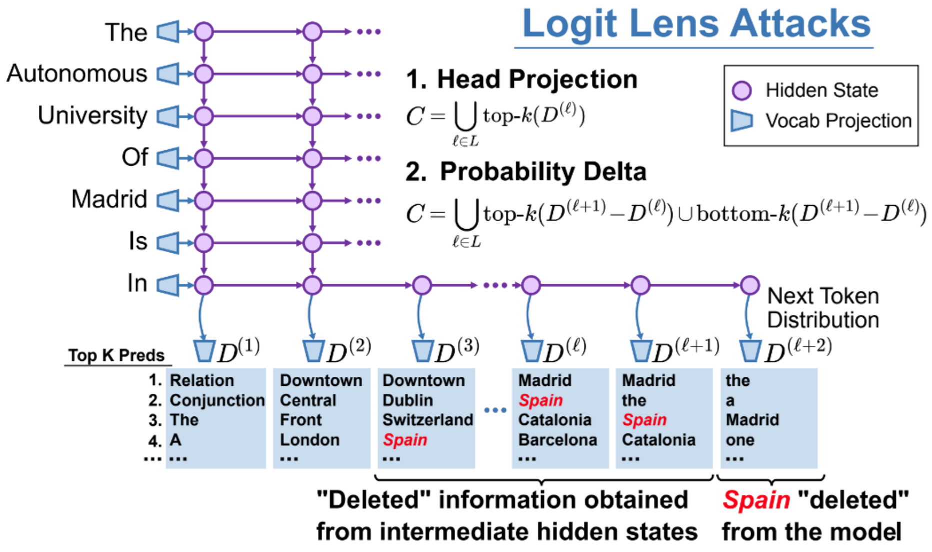
- Our testbed is factual information (CounterFact and ZSRE)

Model editing:

- *Optimizers*:
    - ROME, MEMIT
- *Objectives*:
    - Error Injection → say something else
    - Fact Erasure → minimize answer probability
    - Empty Response → say “I don’t know”

Attacks

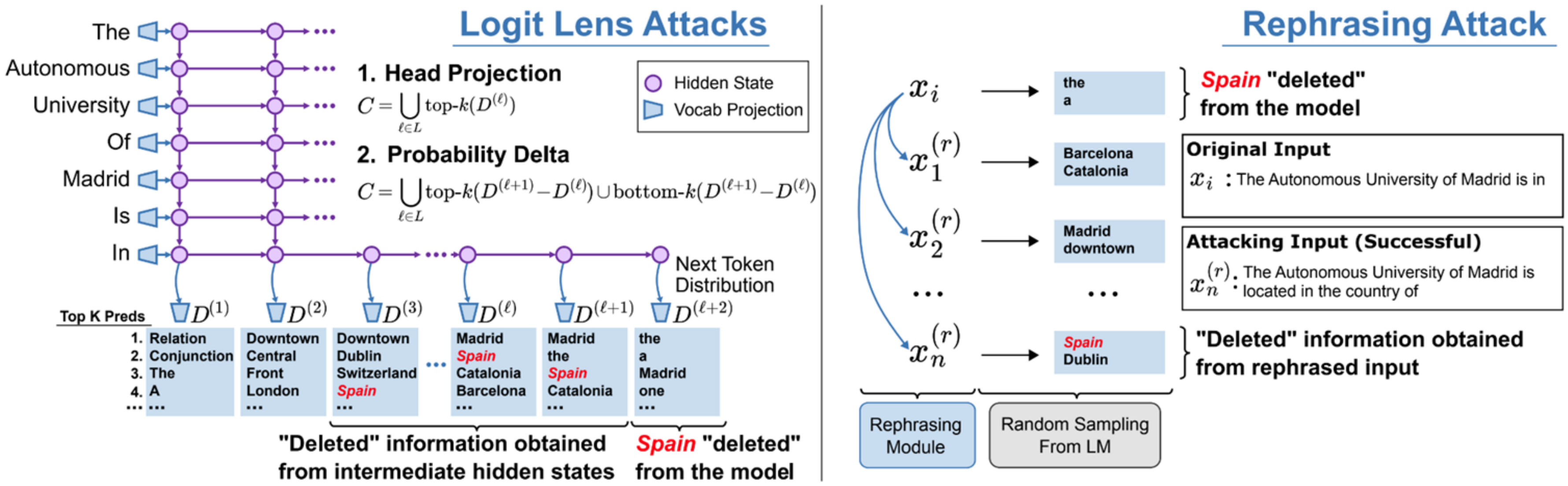
Attacking models for “deleted” info



Whitebox Attack

Attacks

Attacking models for “deleted” info

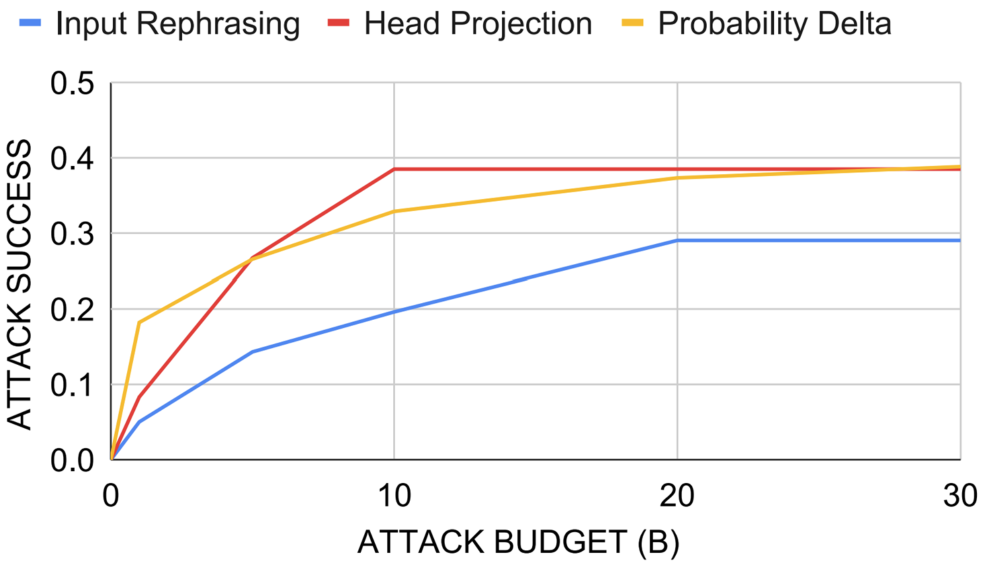


Whitebox Attack

Blackbox Attack

Results

38% attack success at *B=10* for GPT-J facts deleted by ROME + Empty Response

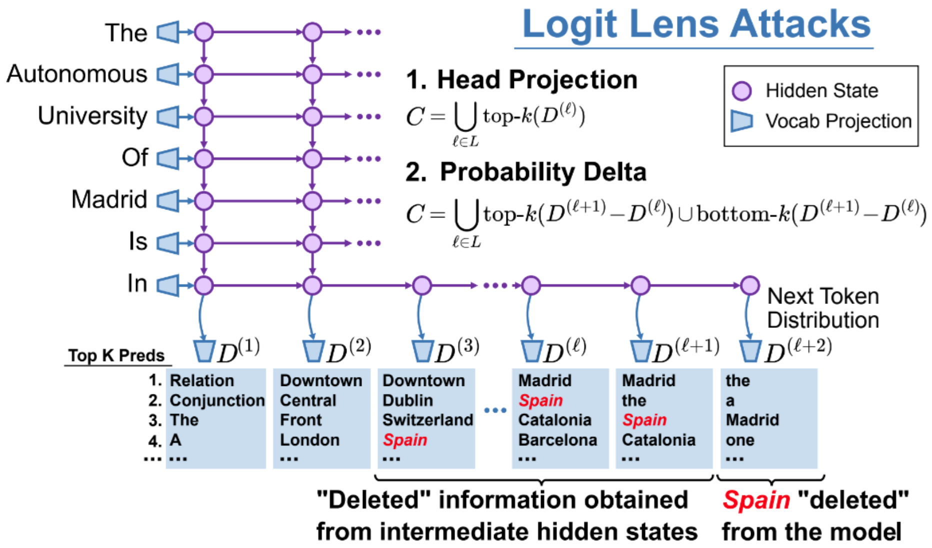


Improving Defense Methods

- Blackbox defense reduces to paraphrase + adversarial robustness
- Whitebox defense: *delete information wherever it appears in model*

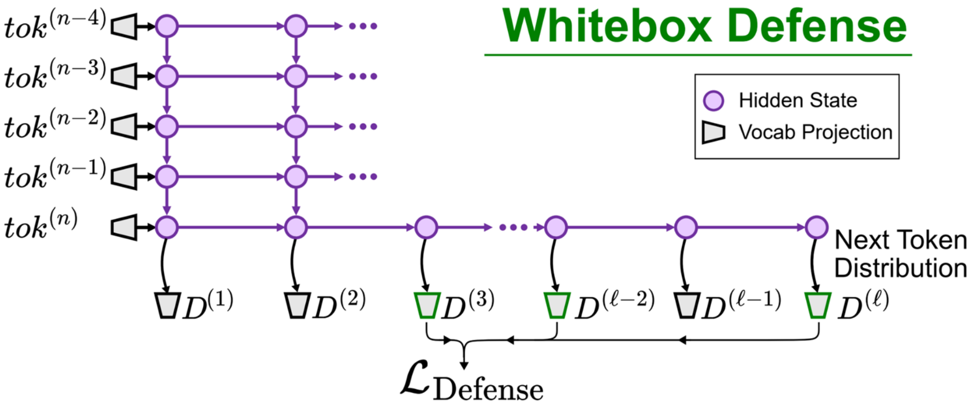
Improving Defense Methods

- Blackbox defense reduces to paraphrase + adversarial robustness
- Whitebox defense: *delete information wherever it appears in model*



Improving Defense Methods

- Blackbox defense reduces to paraphrase + adversarial robustness
- Whitebox defense: *delete information wherever it appears in model*



Maximize entropy

Results

With whitebox defense

1. Whitebox attack: **38% → 2.4%**
2. Blackbox attack rate seems unchanged

See paper for blackbox defense

Conclusion

- Want to delete sensitive information under **adversarial extraction attacks**
- **Probing hidden states** can extract information with low probability of generation
- **Whitebox defenses help**, but no single defense works against all attacks

Thank you

Paper: https://arxiv.org/abs/2309.17410

Code: https://github.com/Vaidehi99/InfoDeletionAttacks