Spotlight!

Can Sensitive Information Be Deleted From LLMs?

Objectives for Defending Against Extraction Attacks

Vaidehi Patil,* Peter Hase,* Mohit Bansal

UNC Chapel Hill

Key Takeaways

1. **We can recover "deleted" facts from LLMs by probing their hidden states**
2. **We introduce a threat model for LLM unlearning**
3. **New edit objectives help against whitebox attacks**
4. **Protecting against both whitebox and blackbox attacks is an open problem**

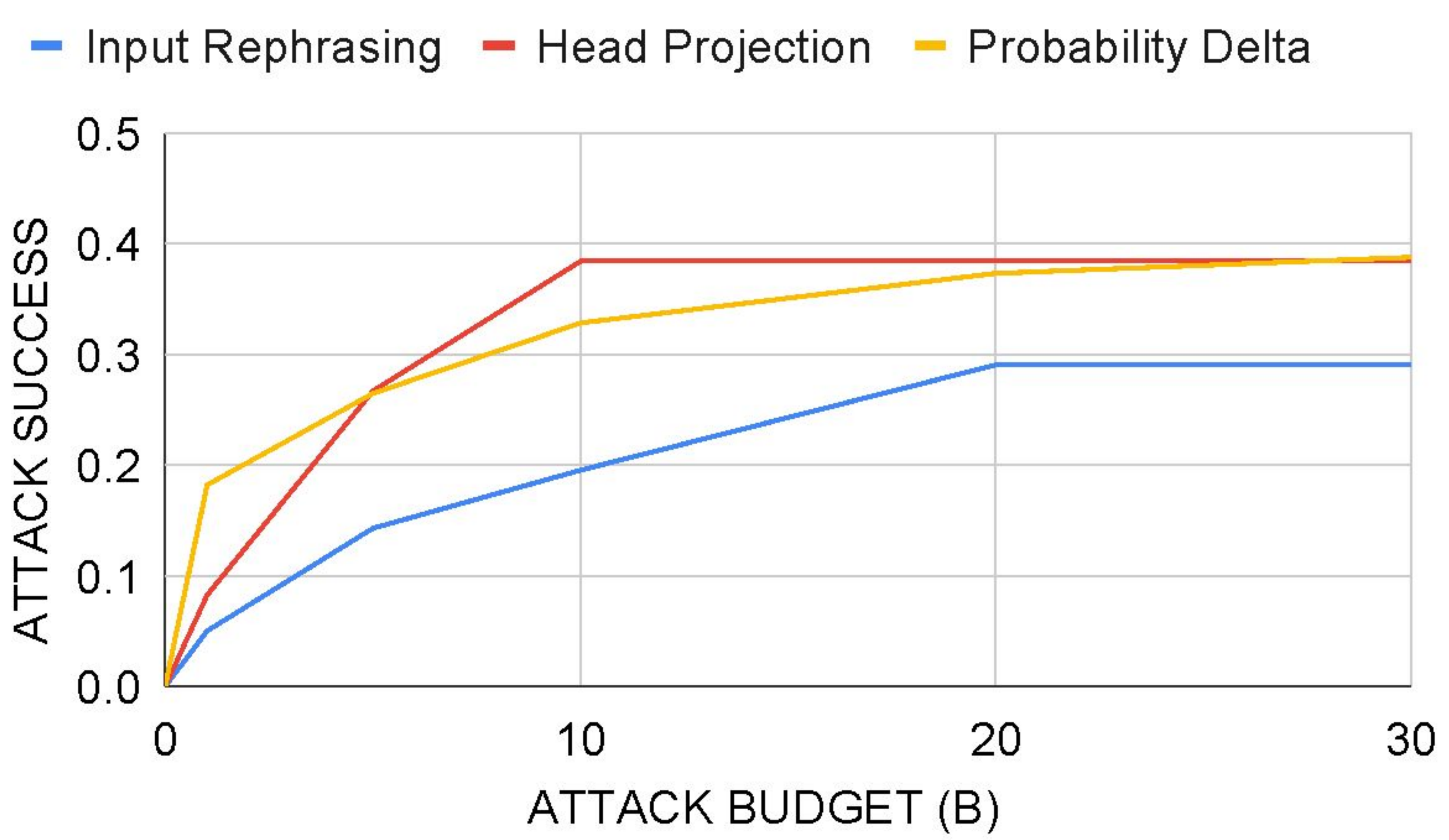


Fig: We recover **up to 38%** of "deleted" facts from LLMs

Background Terms + Methods

Unlearning: removing information from an ML model

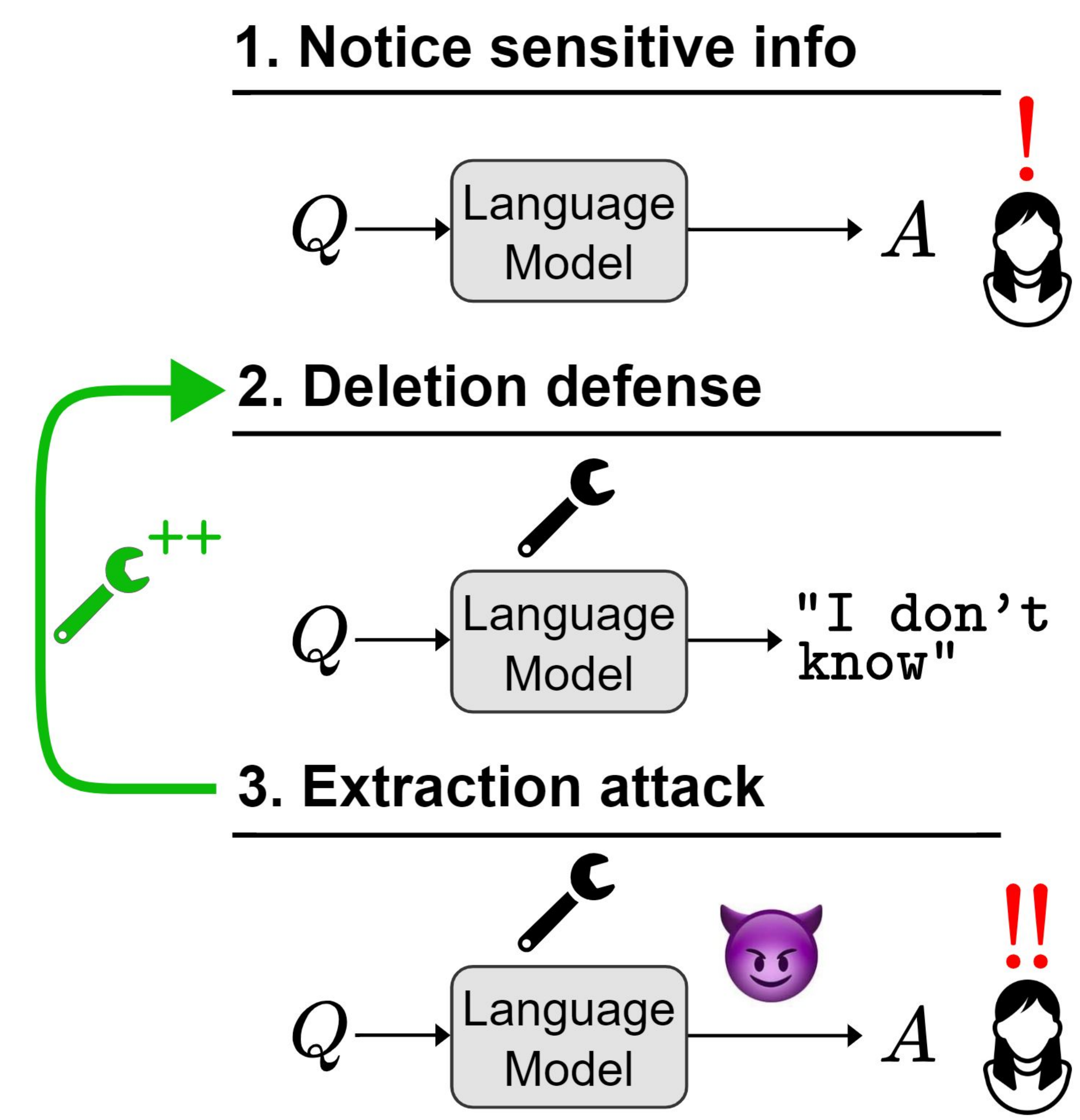
Editing: changing model weights to change a specific model behavior (e.g. specific factual knowledge)

ROME: an editing method that optimizes a low-rank update to a specific early-layer MLP weight

Sensitive Information: Information that we want to delete from the model for ethical reasons

Main Story

Attack and Defense Framework for Info Deletion

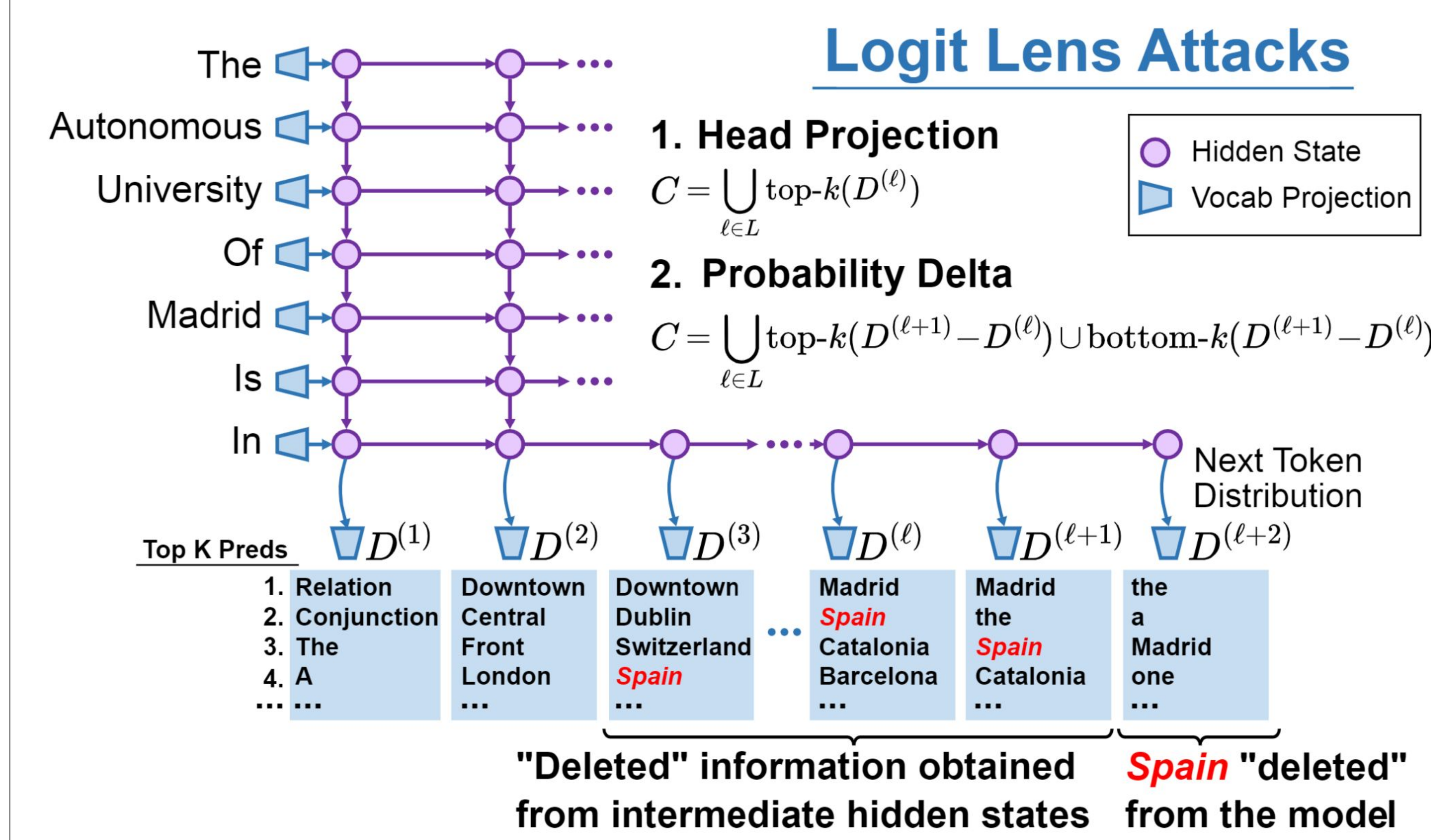


What Do We Want to Delete?

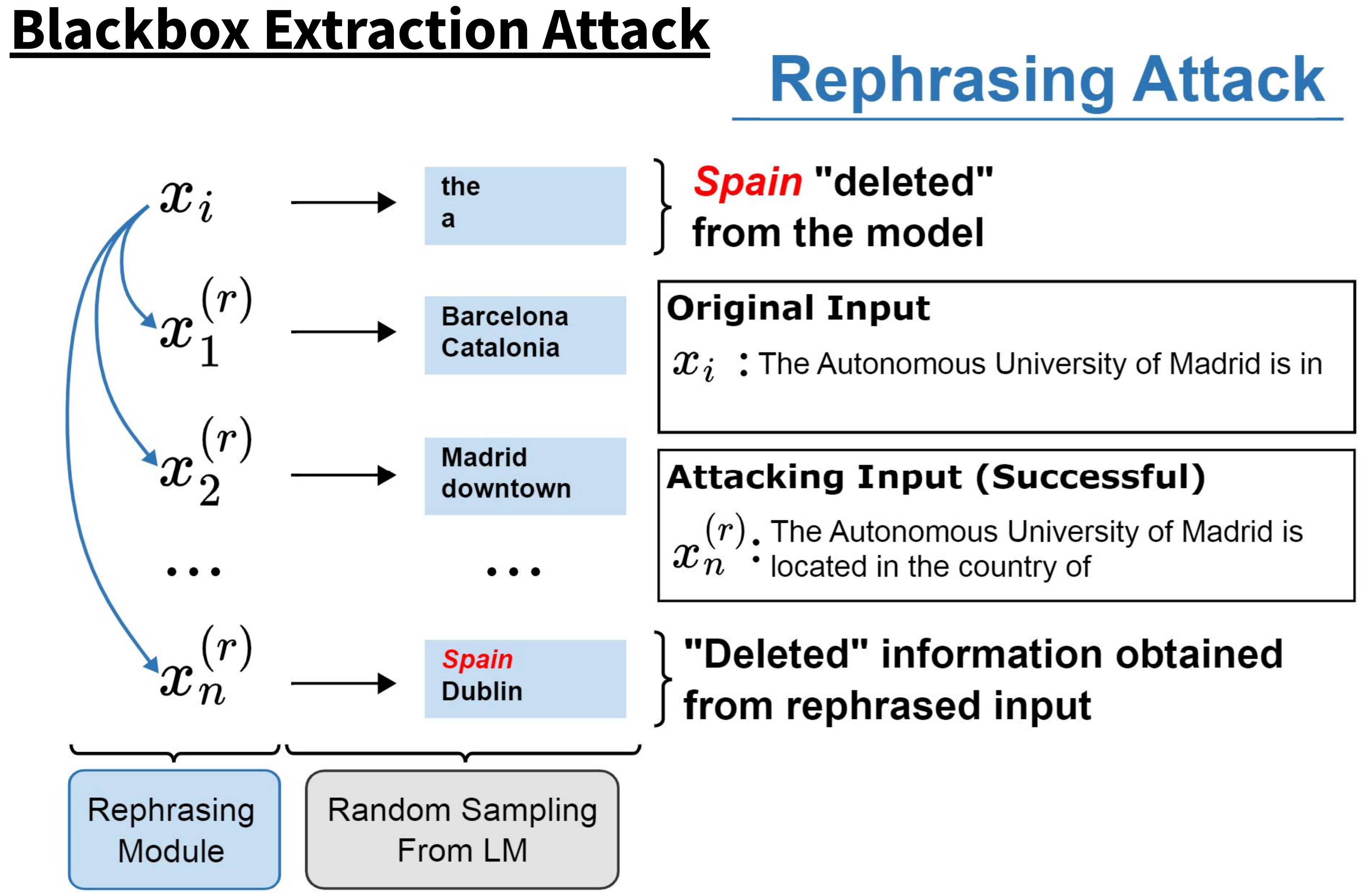
- Personal information
- Copyrighted information
- Knowledge that could be used to harm others
- (e.g. instructions for crimes, CBRN weapons)
- Various toxic beliefs/content
- Factual information that has gone out of date (could become misinfo)

Methods

Whitebox Extraction Attack



Blackbox Extraction Attack



Improving Deletion Defense

- *Delete information wherever it appears* (hidden states)
- Reduces whitebox attack success from **38% to 2%**
- But does not transfer to blackbox attacks