Interpretable and Controllable Language Models

Peter Hase

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Advisor: Mohit Bansal

Committee: Ana Marasovic
Sameer Singh
Shashank Srivastava
Sridhar Duggirala

Thesis Statement

- Language models are getting better at many tasks
- But we do not know their internal reasoning processes
- And individual behaviors are hard to manipulate
- The main goals of work are to **develop and evaluate tools** for:

1. Explaining why language models produce the outputs they do
2. Exercising fine-grained control of language model behaviors

Language Models

Language Models are increasingly capable and general systems

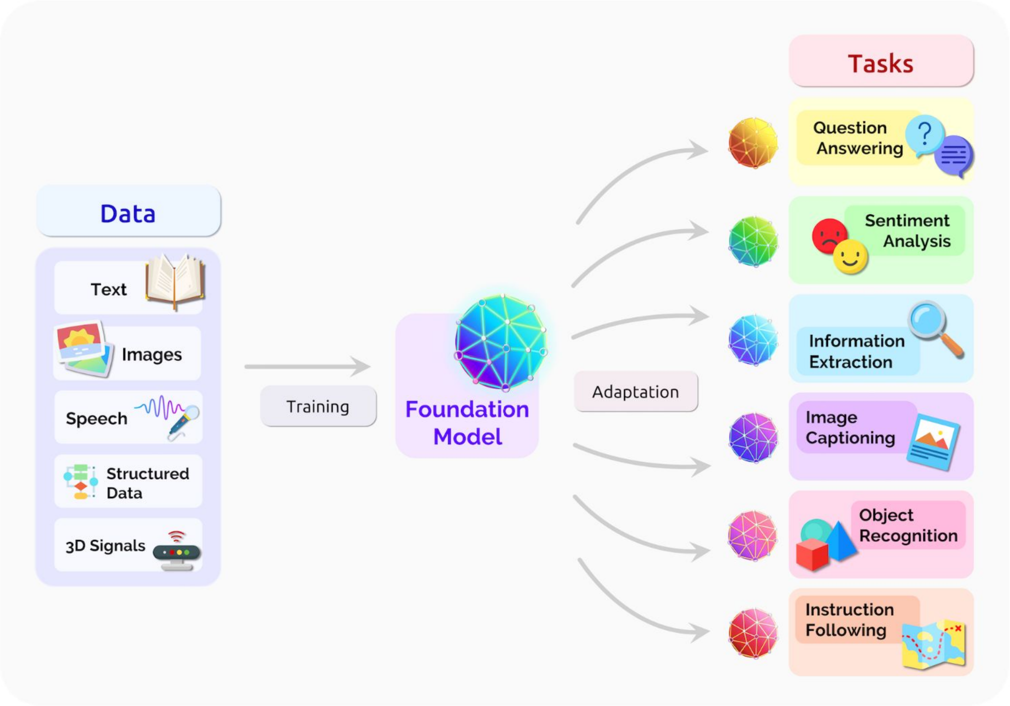


Image: CRFM

UNC NLP

Language Models

Improving LLM safety:

- **Model Pretraining** (you’re building an LLM)
- **Model Deployment** (you’re given an LLM)
- **Sociotechnical Challenges** (you’re shaping the broader AI ecosystem)

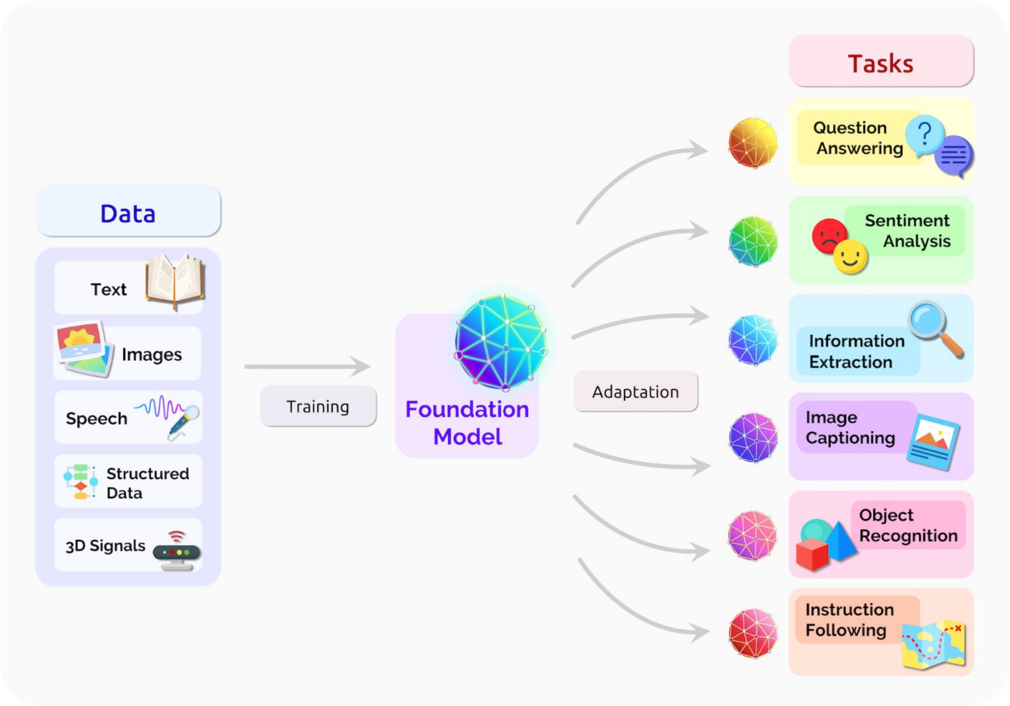


Image: CRFM

Language Models

Improving LLM safety:

- Model Pretraining
- **Model Deployment**
- Sociotechnical Challenges

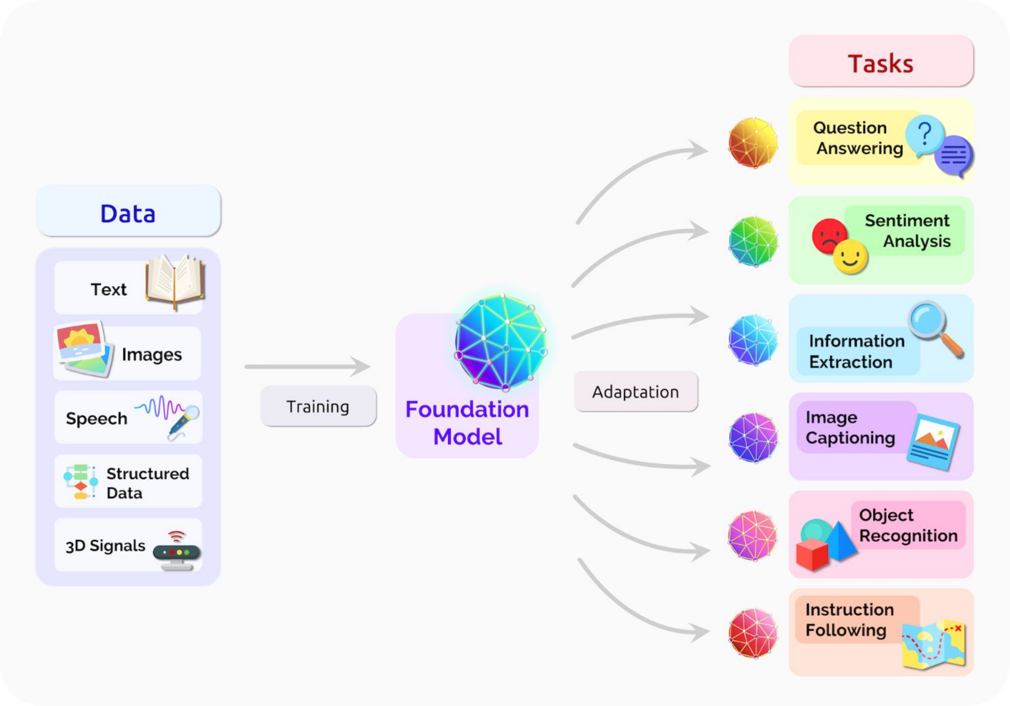


Image: CRFM

Language Models

Improving LLM safety:

- Model Pretraining
- **Model Deployment**
    - **Interpretability**
    - **Fine-grained Control**
    - **Deleting Sensitive Info**
- Sociotechnical Challenges

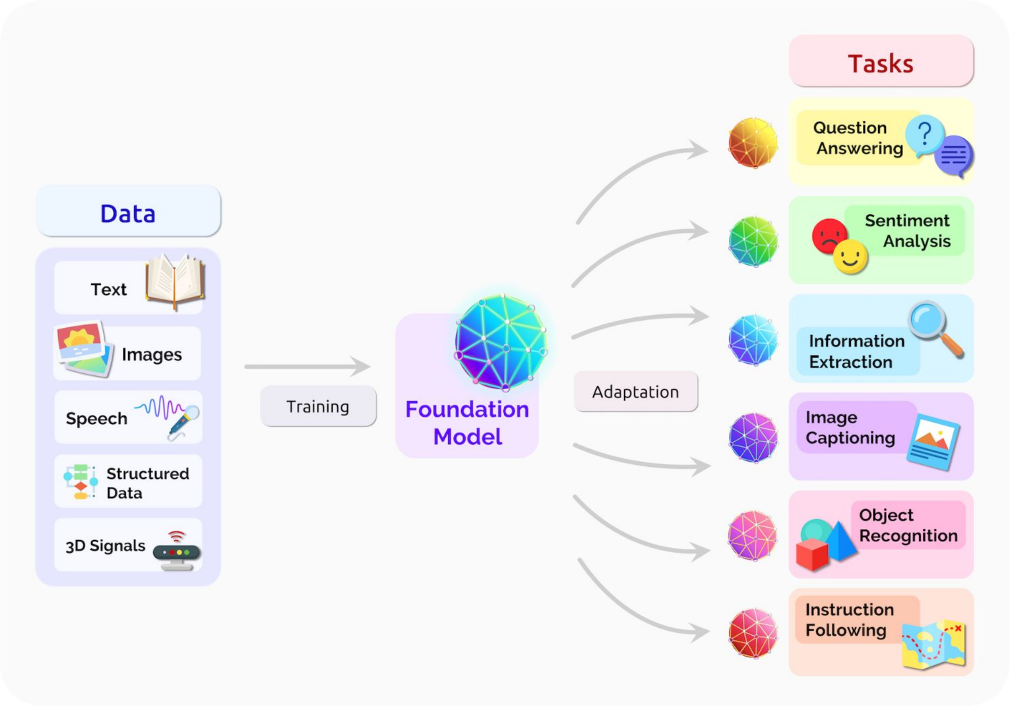


Image: CRFM

UNC NLP

Roadmap

Goal	Research
Interpretability	Evaluating Explainable AI
Fine-grained Control	Model Editing
Deleting Sensitive Info	Machine Unlearning

Roadmap

Goal	Research
Interpretability	Evaluating Explainable AI
Fine-grained Control	Model Editing
Deleting Sensitive Info	Machine Unlearning

UNC NLP

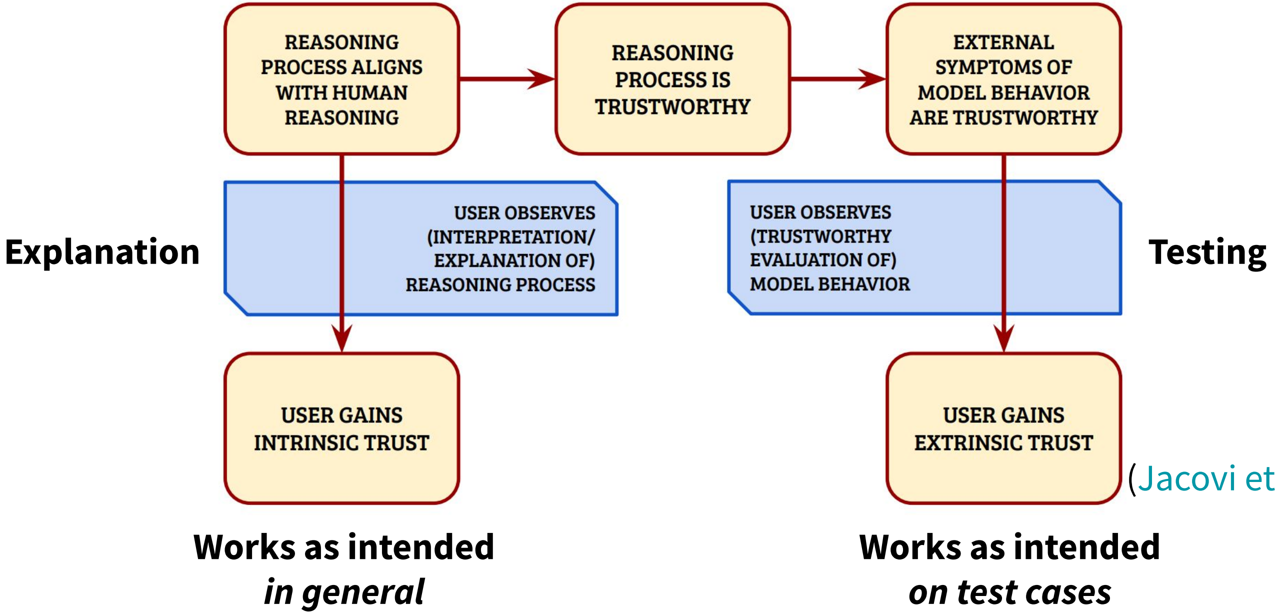
Definitions

- A model is *interpretable* if we can form accurate beliefs about how it works
- "How it works" = causal chains of events that lead to model outputs

UNC NLP

Why Interpretability?

- We evaluate models with test data → *accuracy*
- But can we verify their *reasoning*?



(Jacovi et al., 2020)

Why Interpretability?

FOR WOMEN PREDICTED HIGH RISK FOR LUNG CANCER THAT ARE OLDER THAN 65, WHY DID THE MODEL DECIDE TO PREDICT THEM AS HIGH RISK?

Example adapted from Lakkaraju et al. (2022)

Why Interpretability?

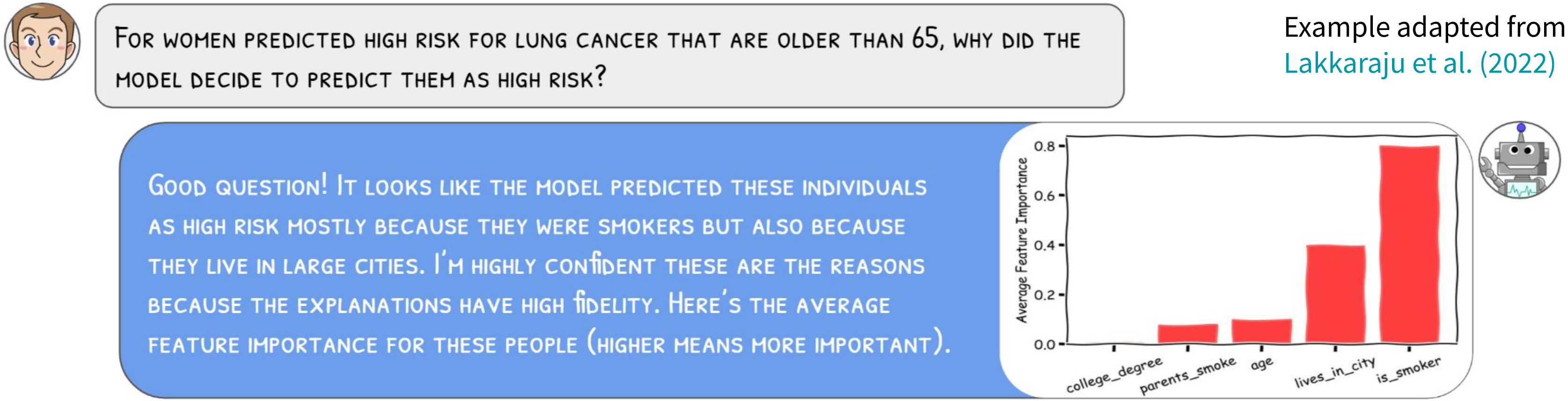
For women predicted high risk for lung cancer that are older than 65, why did the model decide to predict them as high risk?

Example adapted from Lakkaraju et al. (2022)

I don't know, but the model's accuracy on this group is 90%.

vs...

Why Interpretability?

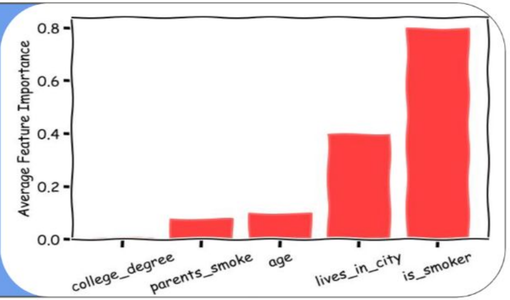


Example adapted from Lakkaraju et al. (2022)

Why Interpretability?

For women predicted high risk for lung cancer that are older than 65, why did the model decide to predict them as high risk?

Good question! It looks like the model predicted these individuals as high risk mostly because they were smokers but also because they live in large cities. I'm highly confident these are the reasons because the explanations have high fidelity. Here's the average feature importance for these people (higher means more important).



Wow, it's surprising that whether the person lives is in a city is so important.

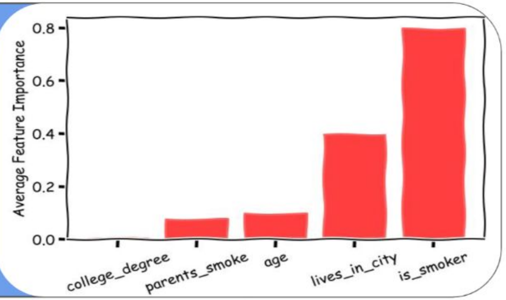
Example adapted from Lakkaraju et al. (2022)

Why Interpretability?

Example adapted from Lakkaraju et al. (2022)

FOR WOMEN PREDICTED HIGH RISK FOR LUNG CANCER THAT ARE OLDER THAN 65, WHY DID THE MODEL DECIDE TO PREDICT THEM AS HIGH RISK?

GOOD QUESTION! IT LOOKS LIKE THE MODEL PREDICTED THESE INDIVIDUALS AS HIGH RISK MOSTLY BECAUSE THEY WERE SMOKERS BUT ALSO BECAUSE THEY LIVE IN LARGE CITIES. I'M HIGHLY CONFIDENT THESE ARE THE REASONS BECAUSE THE EXPLANATIONS HAVE HIGH FIDELITY. HERE'S THE AVERAGE FEATURE IMPORTANCE FOR THESE PEOPLE (HIGHER MEANS MORE IMPORTANT).



WOW, IT'S SURPRISING THAT WHETHER THE PERSON LIVES IS IN A CITY IS SO IMPORTANT.

YES, LIVES_IN_CITY HAS A SIGNIFICANT EFFECT ON THE PREDICTIONS FOR THESE INDIVIDUALS. PERTURBING THIS FEATURE CAN FLIP THE PREDICTION FOR 4 OF 15 OF THE INSTANCES IN THIS GROUP.

Roadmap

Goal	Research
Interpretability	**Evaluating Explainable AI**
Fine-grained Control	Model Editing
Deleting Sensitive Info	Machine Unlearning

UNC NLP

Evaluating Explainable AI

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Peter Hase and **Mohit Bansal**
UNC Chapel Hill
peter@cs.unc.edu, mbansal@cs.unc.edu

ACL 2020

Evaluating Explainable AI

- We propose a study design for measuring **simulatability** of an ML system
- An ML system is *simulatable* when a person can predict its outputs

Simulatable → person has a good mental model of system
Explanation improves simulatability → explanation reveals causal chains behind behavior
→ explanation is faithful

- We measure the effect of explanations on simulatability

UNC NLP

Evaluating Explainable AI

Test 1: Forward Simulation Test

- Predict model outputs before/after examples are explained to them

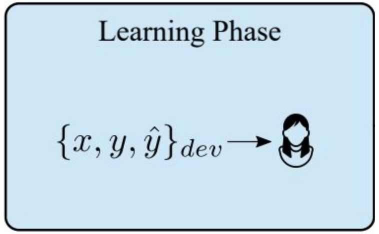
$$\begin{array}{l}\text{Post Sim.}\\\text{Accuracy}\end{array} - \begin{array}{l}\text{Pre Sim.}\\\text{Accuracy}\end{array} = \begin{array}{l}\text{Explanation}\\\text{Effect}\end{array}$$

Evaluating Explainable AI

Test 1: Forward Simulation Test

- Predict model outputs before/after examples are explained to them

$$\text{Post Sim. Accuracy} - \text{Pre Sim. Accuracy} = \text{Explanation Effect}$$



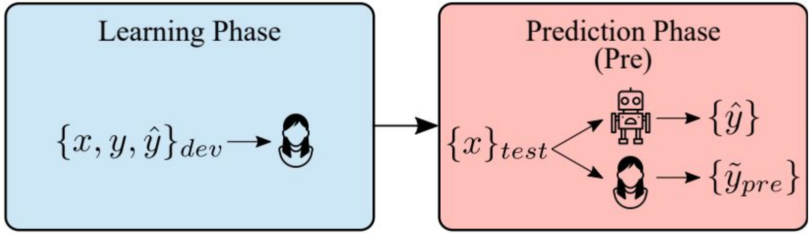
e : Explanation
$\hat{y}$: Model prediction
$\tilde{y}$: Human simulation

Evaluating Explainable AI

Test 1: Forward Simulation Test

- Predict model outputs before/after examples are explained to them

$$\text{Post Sim. Accuracy} - \text{Pre Sim. Accuracy} = \text{Explanation Effect}$$



e : Explanation

$\hat{y}$: Model prediction

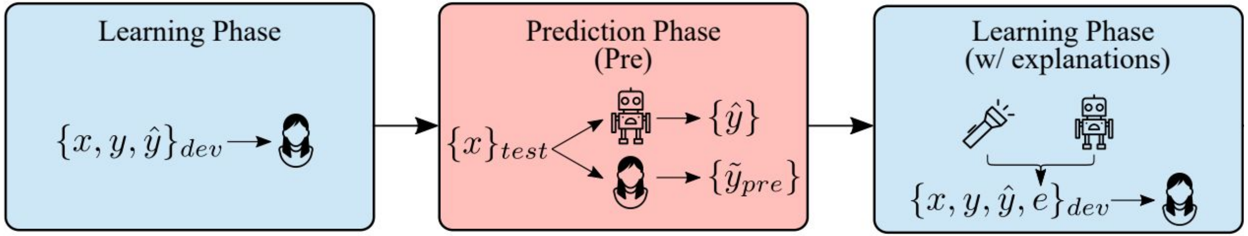
$\tilde{y}$: Human simulation

Evaluating Explainable AI

Test 1: Forward Simulation Test

- Predict model outputs before/after examples are explained to them

$$\text{Post Sim. Accuracy} - \text{Pre Sim. Accuracy} = \text{Explanation Effect}$$



e : Explanation

$\hat{y}$: Model prediction

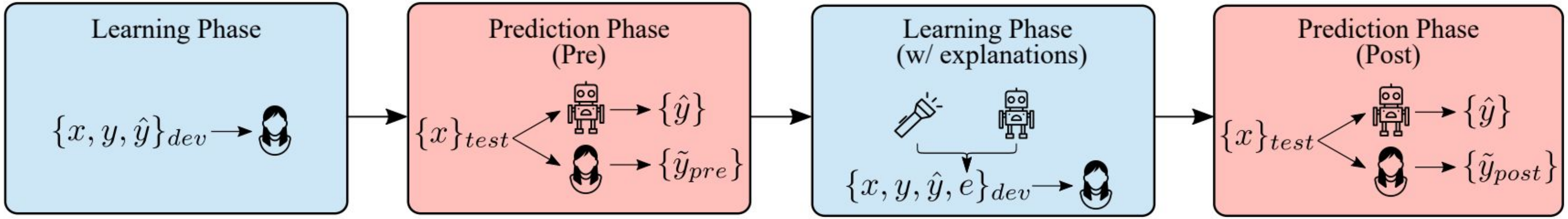
$\tilde{y}$: Human simulation

Evaluating Explainable AI

Test 1: Forward Simulation Test

- Predict model outputs before/after examples are explained to them

$$\text{Post Sim. Accuracy} - \text{Pre Sim. Accuracy} = \text{Explanation Effect}$$



e : Explanation
$\hat{y}$: Model prediction
$\tilde{y}$: Human simulation

Evaluating Explainable AI

Test 2: Counterfactual Simulation Test

- Predict model outputs before/after *similar examples* are explained to them

$$\begin{array}{l}\text{Post Sim.}\\\text{Accuracy}\end{array} - \begin{array}{l}\text{Pre Sim.}\\\text{Accuracy}\end{array} = \begin{array}{l}\text{Explanation}\\\text{Effect}\end{array}$$

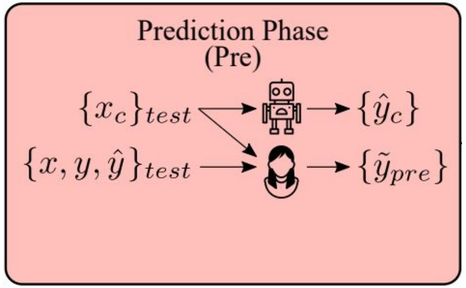
Hase et al.

Evaluating Explainable AI

Test 2: Counterfactual Simulation Test

- Predict model outputs before/after *similar examples* are explained to them

$$\begin{array}{c}\text{Post Sim.}\\ \text{Accuracy}\end{array} - \begin{array}{c}\text{Pre Sim.}\\ \text{Accuracy}\end{array} = \begin{array}{c}\text{Explanation}\\ \text{Effect}\end{array}$$



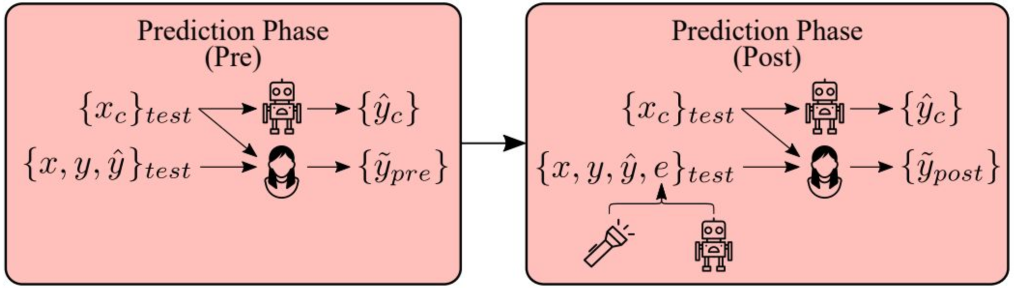
e : Explanation
$\hat{y}$: Model prediction
$\tilde{y}$: Human simulation
x_c : Counterfactual input
$\hat{y}_c$: Counterfactual model prediction

Evaluating Explainable AI

Test 2: Counterfactual Simulation Test

- Predict model outputs before/after *similar examples* are explained to them

$$\text{Post Sim. Accuracy} - \text{Pre Sim. Accuracy} = \text{Explanation Effect}$$



e : Explanation
$\hat{y}$: Model prediction
$\tilde{y}$: Human simulation
x_c : Counterfactual input
$\hat{y}_c$: Counterfactual model prediction

Evaluating Explainable AI

Experiment Setup

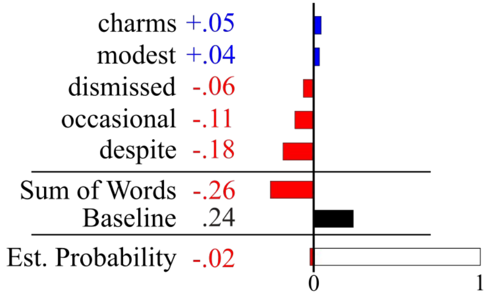
- Train neural networks on *sentiment analysis* and *income prediction* tasks
- Four *local* explanation methods
- 2166 responses from 29 undergraduates (in-person tests)
- Hypothesis testing done by block bootstrap

Evaluating Explainable AI

Input, Label, and Model Output

$x =$ Despite modest aspirations its occasional charms are not to be dismissed.
$y =$ Positive $\quad \hat{y} =$ Negative

LIME



Prototype

Most similar prototype:
Routine and rather silly.
Similarity score: 9.96 out of 10

Important words: (none selected)

Anchor

$p(\hat{y} = \text{Negative} \mid \{\text{occasional}\} \subseteq x) \geq .95$

Decision Boundary

Step 0 | Evidence Margin: -5.21

Step 1 | occasional $\longrightarrow$ rare
Evidence Margin: -3.00

Step 2 | modest $\longrightarrow$ impressive
Evidence Margin: +0.32

$x^{(c)}$ | Despite *impressive* aspirations its *rare* charms are not to be dismissed.

UNC NLP

Evaluating Explainable AI

Main Results

- LIME improves simulatability for tabular data
  - **70.74% → 81.99%** accuracy, +11.25 (+/- 8.83) ppts, p=.014
  - (across forward and counterfactual tests)
- Prototype model improves counterfactual simulatability
  - **63.13% → 72.66%** accuracy, +9.53 (+/- 8.55) ppts, p=.032
  - (across datasets)
- Other estimates **do not significantly differ from 0** (p <.05)
  - **LIME for text**
  - **Prototype for forward sim.**
  - **Anchor explanations**
  - **Counterfactual explanations**

Evaluating Explainable AI

Subjective Ratings

- Do users know when explanations are good?
  - Ask users to rate explanations on 1-7 scale
  - “Does this explanation show me why the system thought what it did?”
  - Estimate simulation score from ratings
- No! (**Ratings not statistically significant** predictor of faithfulness)

Evaluating Explainable AI

Conclusions

- LIME worked with **low-dimensional tabular data**
- Local explanations fail to reveal how models work over **huge data spaces**
- …and **counterfactual scenarios**
- Users **can't tell when explanations are good or not**

Evaluating Explainable AI

Recent studies with LLMs…

- [2023] Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting
- [2023] Do Models Explain Themselves? Counterfactual Simulatability of Natural Language Explanations
- [2023] Faithfulness Tests for Natural Language Explanations
- [2023] ALMANACS: A Simulatability Benchmark for Language Model Explainability
- [2024] Towards Consistent Natural-Language Explanations via Explanation-Consistency Finetuning
- [2024] Bias-Augmented Consistency Training Reduces Biased Reasoning in Chain-of-Thought

Roadmap

Goal	Research
Interpretability	Evaluating Explainable AI
Fine-grained Control	Model Editing
Deleting Sensitive Info	Machine Unlearning

UNC NLP

Why Fine-grained Control?

What awards did Mary Lowe Good receive?

Mary Lowe Good was a renowned chemist who made significant contributions to the field of organic chemistry. She received several awards throughout her career, including:

Wrong 1. Garvan Medal (1954) - This is awarded by the American Chemical Society (ACS) to recognize outstanding contributions to the field of chemistry by women.

Wrong 2. Francis P. Garvan-John M. Olin Medal (1963) - This is also awarded by the ACS, to recognize distinguished service to chemistry by women.

Wrong 3. Priestley Medal (1983) - This is the highest honor awarded by the ACS, and it recognizes distinguished services to chemistry.

Wrong 4. National Medal of Science (1991) - This is the highest scientific honor awarded by the United States government.

https://en.wikipedia.org/wiki/Mary_L._Good

Why Fine-grained Control?

Fast forward one year…

You

What awards did Mary Lowe Good receive?

ChatGPT

Mary Lowe Good was a renowned chemist and influential figure in science policy, recognized for her significant contributions to the fields of inorganic chemistry and science administration. Over her career, she received numerous awards and honors, reflecting her achievements in research, education, and public service. Some of the notable awards and honors she received include:

Wrong
1. **Garvan–Olin Medal** (1980): Awarded by the American Chemical Society, this medal recognizes distinguished service to chemistry by women chemists.

Right
2. **Priestley Medal** (1997): The highest honor conferred by the American Chemical Society, awarded for distinguished service to chemistry.

Right
3. **Member of the National Academy of Engineering** (Election Year: 1987): This is one of the highest professional distinctions accorded to an engineer, recognizing outstanding contributions to engineering research, practice, or education.

Wrong
4. **Fellow of the American Academy of Arts and Sciences** (Election Year: Unknown): This fellowship acknowledges leaders in the academic disciplines, the arts, business, and public affairs.

https://en.wikipedia.org/wiki/Mary_L._Good

Why Fine-grained Control?

Usual argument goes like…

- Pretraining and finetuning large models over lots of data is expensive
- We can identify errors, but want to avoid re-training
- **Want to fix errors one at a time**

Why Fine-grained Control?

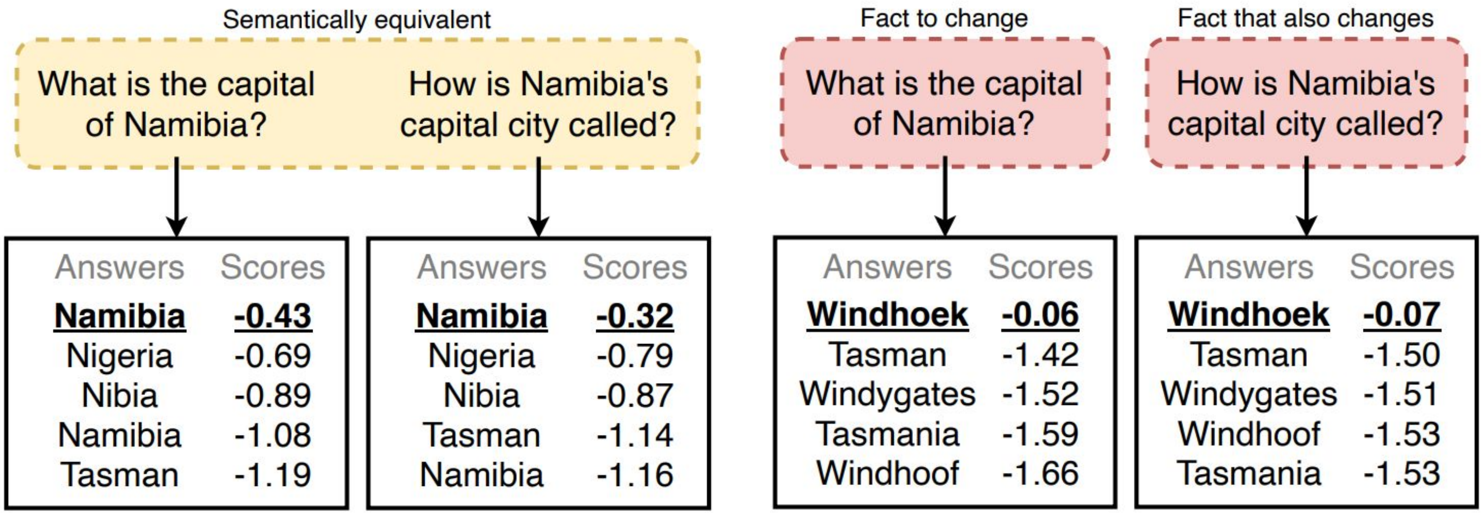
Usual argument goes like…

- Pretraining and finetuning large models over lots of data is expensive
- We can identify errors, but want to avoid re-training - **would this even work?**
- **Want to fix errors one at a time**

Roadmap

Goal	Research
Interpretability	Evaluating Explainable AI
Fine-grained Control	**Model Editing**
Deleting Sensitive Info	Machine Unlearning

Model Editing - Background



Before Edit **After Edit**

(De Cao et al., 2020)

Model Editing

A word on terminology...

- Editing = updating = revising
- What are we editing?
- “Fact” and “knowledge” seem awkward if information isn’t true
- “Belief” feels appropriately *weaker*
- Dennett (1995) characterizes a *belief* as:

An informational state decoupled from any motivational state

- More to say on criteria for belief…(Dretske, 1981)
- This problem has been called *belief revision* in CS+philosophy since 1979 (Doyle)

UNC NLP

Model Editing

Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs

Peter Hase[1,2] **Mona Diab**[1] **Asli Celikyilmaz**[1] **Xian Li**[1]
Zornitsa Kozareva[1] **Veselin Stoyanov**[1] **Mohit Bansal**[2] **Srinivasan Iyer**[1]
[1]Meta AI [2]UNC Chapel Hill
{peter, mbansal}@cs.unc.edu
{mdiab, aslic, xianl, zori, ves, sviyer}@fb.com

EACL 2023

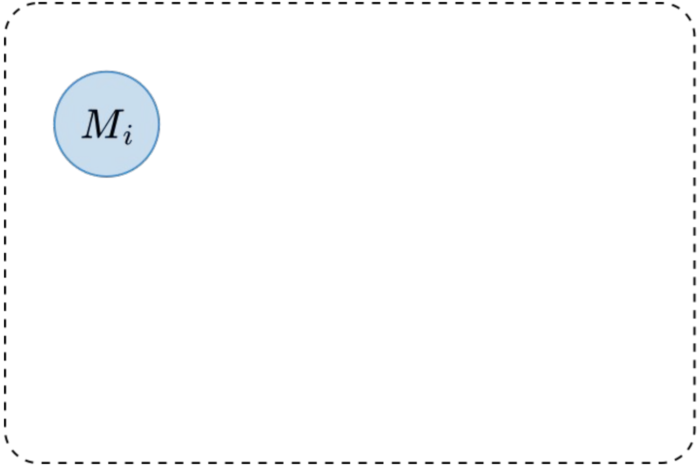
Model Editing

Two main research questions:

1. How should we evaluate model edits?
2. Can we continually update a model with new beliefs?

Model Editing

- How should we evaluate model edits?

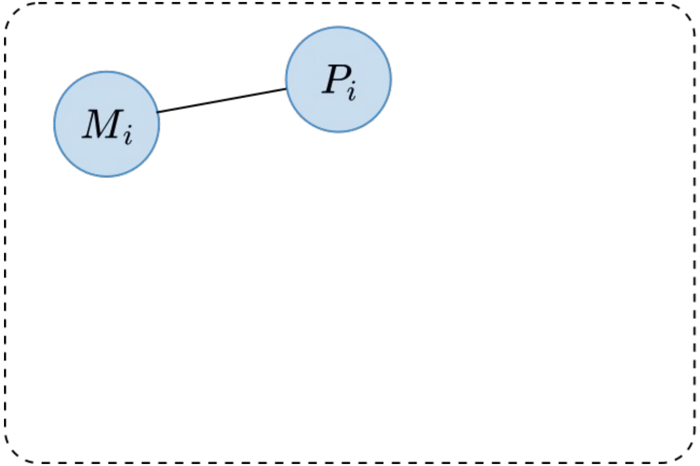


M (Main Input) : A viper is a vertebrate.

Vipers are vertebrates.

Model Editing

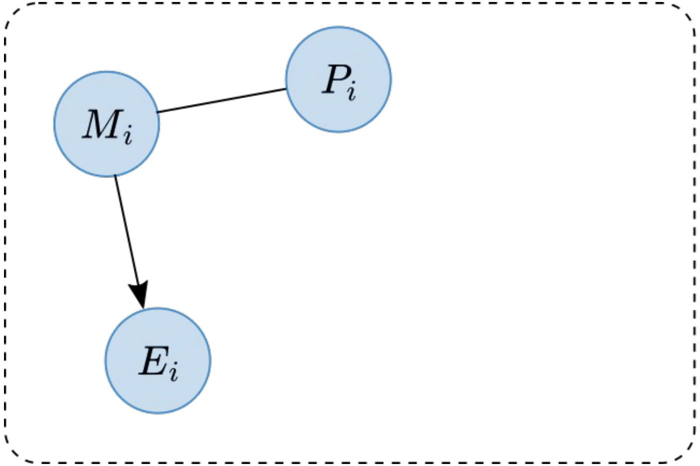
- How should we evaluate model edits?



M (Main Input) : A viper is a vertebrate.
P (Paraphase Data) : Vipers are vertebrates.

Model Editing

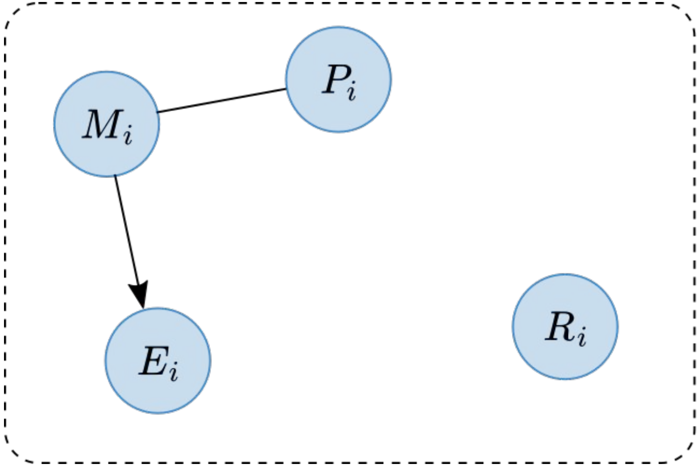
- How should we evaluate model edits?



M	(Main Input)	:	A viper is a vertebrate.
P	(Paraphase Data)	:	Vipers are vertebrates.
E	(Entailed Data)	:	A viper has a brain.

Model Editing

- How should we evaluate model edits?

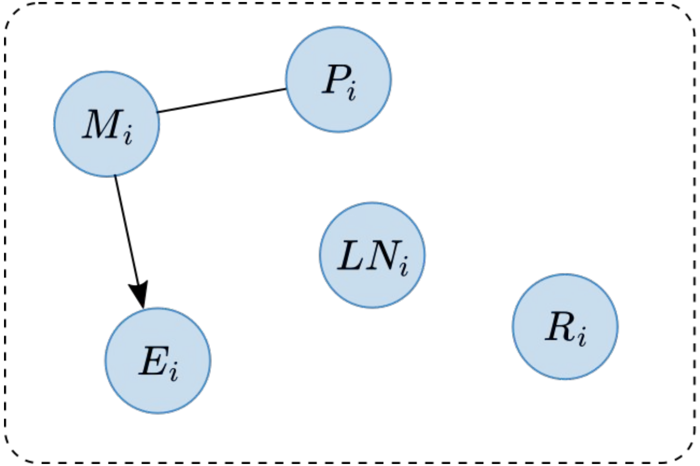


M	(Main Input)	:	A viper is a vertebrate.
P	(Paraphase Data)	:	Vipers are vertebrates.
E	(Entailed Data)	:	A viper has a brain.
R	(Random Data)	:	Chile is a country.

Hase et al.

Model Editing

- How should we evaluate model edits?



M	(Main Input)	:	A viper is a vertebrate.
P	(Paraphase Data)	:	Vipers are vertebrates.
E	(Entailed Data)	:	A viper has a brain.
R	(Random Data)	:	Chile is a country.
LN	(Local Neutral Data)	:	A viper is venemous.

Model Editing

- Continual belief updating - **hypernetwork weight editing** on **t5-base**
- Main conclusions:
    1. Fixing ***one error*** works well, but fixing ***many errors*** is difficult
    2. Harder to ***fix errors*** than to ***create them***
    3. Harder to generalize to ***entailed data*** than ***paraphrases***
    4. Harder to retain performance on ***local data*** than ***random data***
    5. Updates greatly ***improve consistency*** (model was wrong in inconsistent ways)

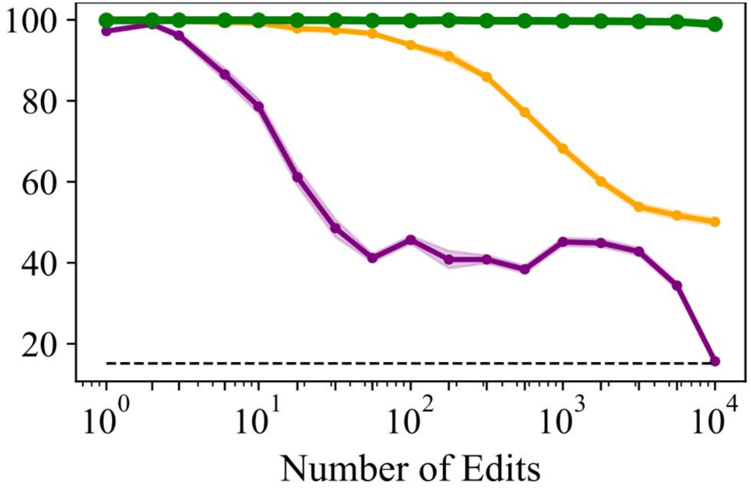
Model Editing

- Continual belief updating - **hypernetwork weight editing** on **t5-base**
- Since 2021…

Model Editing - Recent Work

- Continual belief updating - **MEMIT weight editing** on **GPT-J** (Meng et al., 2022)

~= Update Success (Main Input)



Pretty good at 10k edits

UNC NLP

Model Editing - Recent Work

Continual belief updating - **MEMIT weight editing** on **GPT-J** (Meng et al., 2022)

No entailment evaluation

Entailment is hard to measure

- We adapted data from LeapOfThought (Talmor et al., 2020), but it's a little synthetic

Recent work:

1. Evaluating the Ripple Effects of Knowledge Editing in Language Models (Cohen et al., 2023)
2. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions (Zhong et al., 2023)

UNC NLP

Model Editing

Conclusions

Model editing is increasingly useful for fine-grained control…

…but needs stronger evals focusing on **fixing errors** and **measuring entailment**

Roadmap

Goal	Research
Interpretability	Evaluating Explainable AI
Fine-grained Control	Model Editing
Deleting Sensitive Info	Machine Unlearning

Definitions + 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)
- We would like to remove this information from LLMs

(yes there are dual-use concerns)

Definitions + Motivation

- *Deleting* information from LLMs is underdefined
- Finetuning (RLHF, SFT, safety training, etc.) appears to hide rather than remove sensitive information (Zou et al., 2023)
- **This is a model editing problem** – update individual beliefs in a model

Roadmap

Goal	Research
Interpretability	Evaluating Explainable AI
Fine-grained Control	Model Editing
Deleting Sensitive Info	**Machine Unlearning**

UNC NLP

Unlearning in LLMs

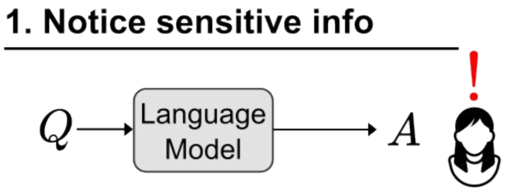
Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks

Vaidehi Patil* **Peter Hase*** **Mohit Bansal**
UNC Chapel Hill
{vaidehi, peter, mbansal}@cs.unc.edu

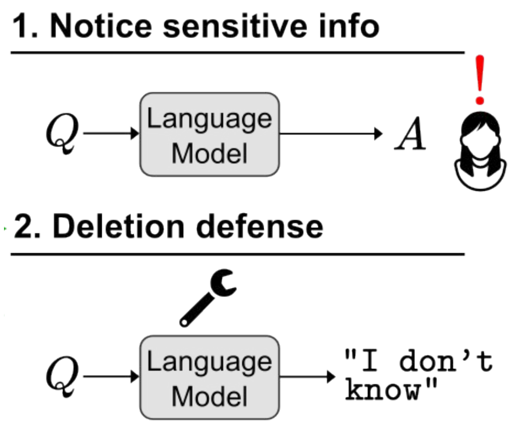
ICLR 2024

Spotlight

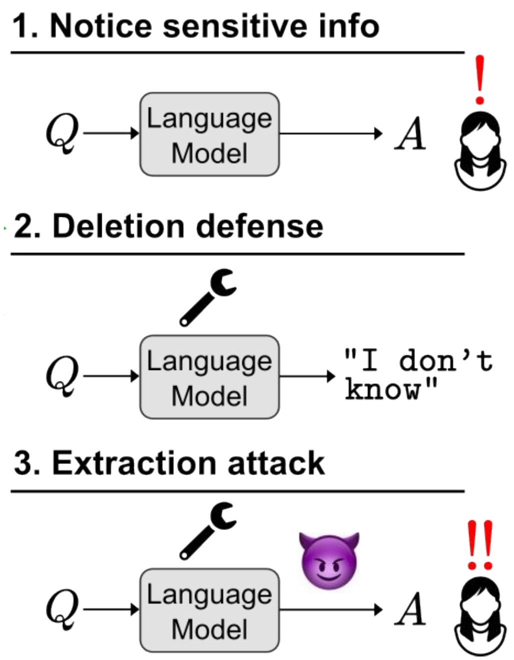
Unlearning in LLMs



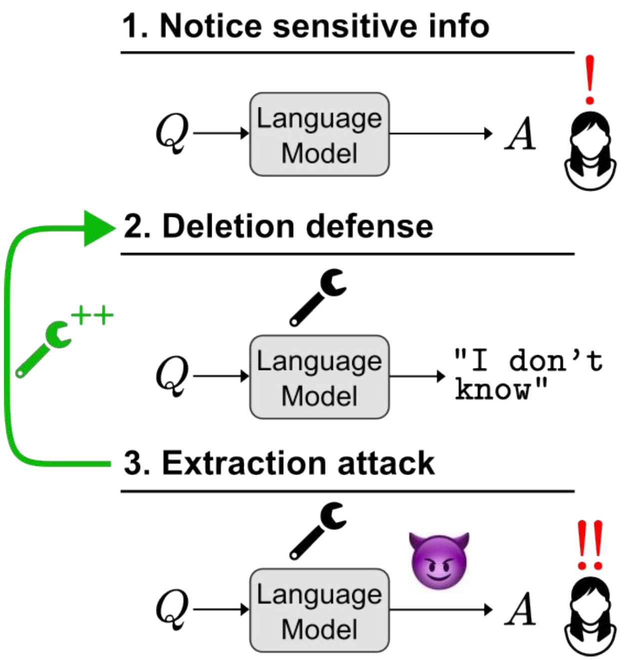
Unlearning in LLMs



Unlearning in LLMs



Unlearning in LLMs



Unlearning in LLMs

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
3. Verification by data owner (or auditor)

Previous frameworks focused on formal guarantees of similarity to retrained model (Cao and Yang, 2015)

Unlearning in LLMs

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. **AttackSuccess**: whether answer is in candidate set
2. **Damage**: change in model accuracy for other questions

Unlearning in LLMs

Applying model editing for deletion - This is the defense

Tasks/data:

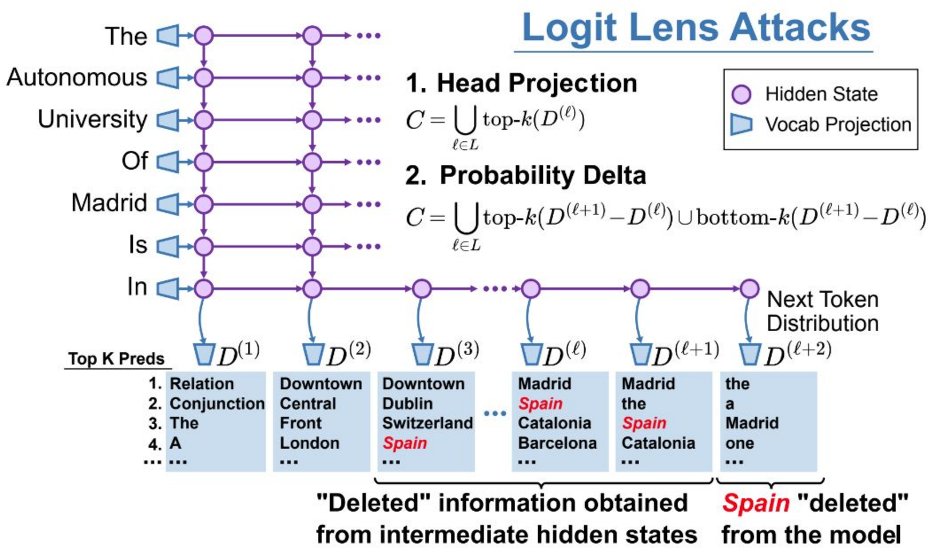
- Our testbed is factual information (CounterFact and ZSRE)
- Filter to questions with single-token answers, known by GPT-J model we attack

Model editing:

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

Unlearning in LLMs

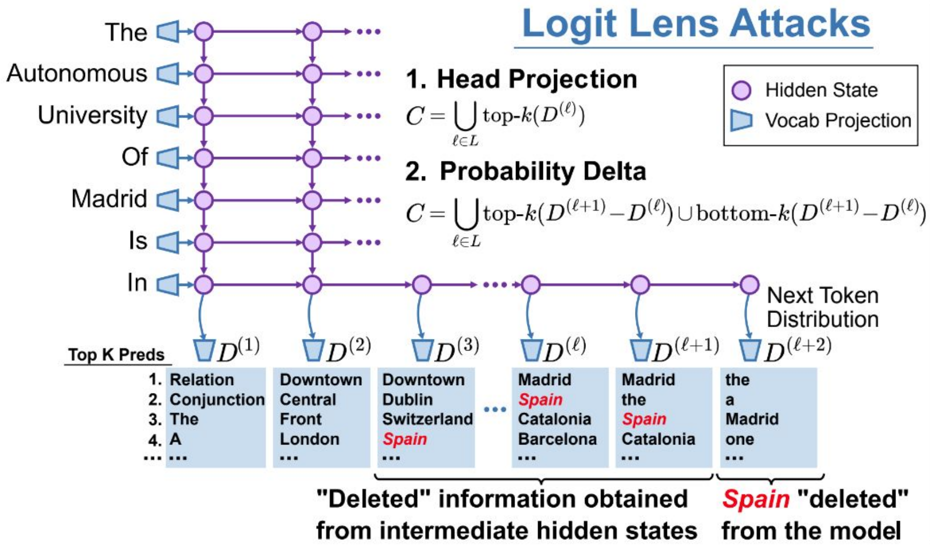
Attacking models for "deleted" info

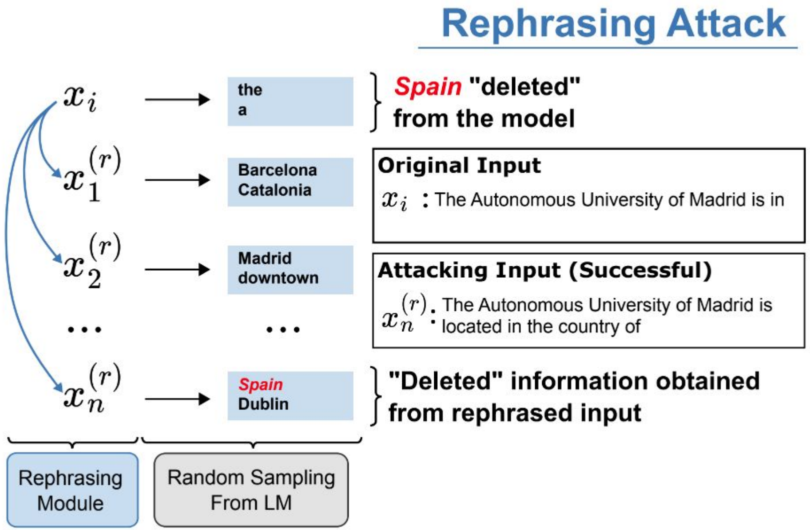


Whitebox Attack

Unlearning in LLMs

Attacking models for “deleted” info



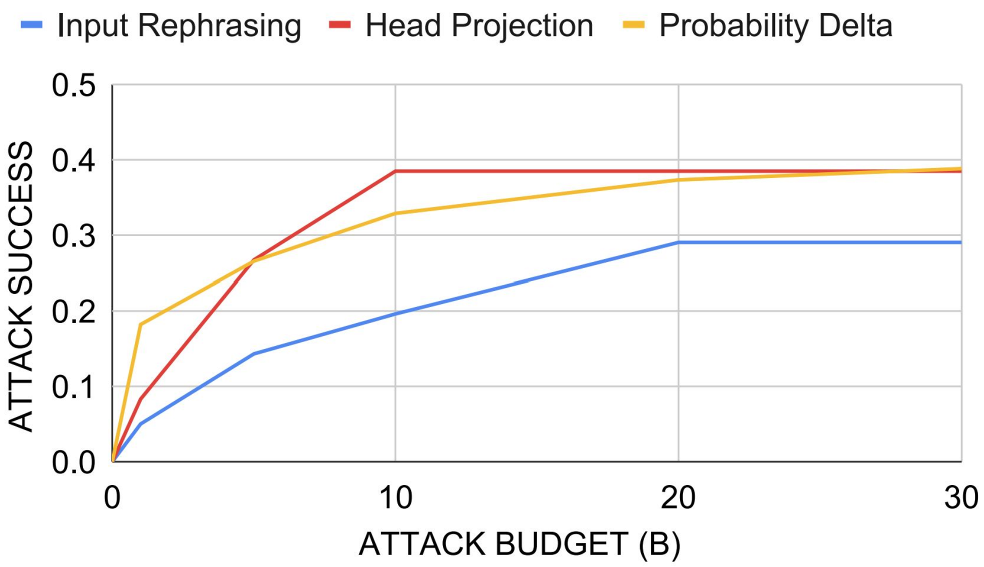


Whitebox Attack **Blackbox Attack**

Unlearning in LLMs

Results

1. 38% attack success at *B=10* for GPT-J facts deleted by ROME + Empty Response



Unlearning in LLMs

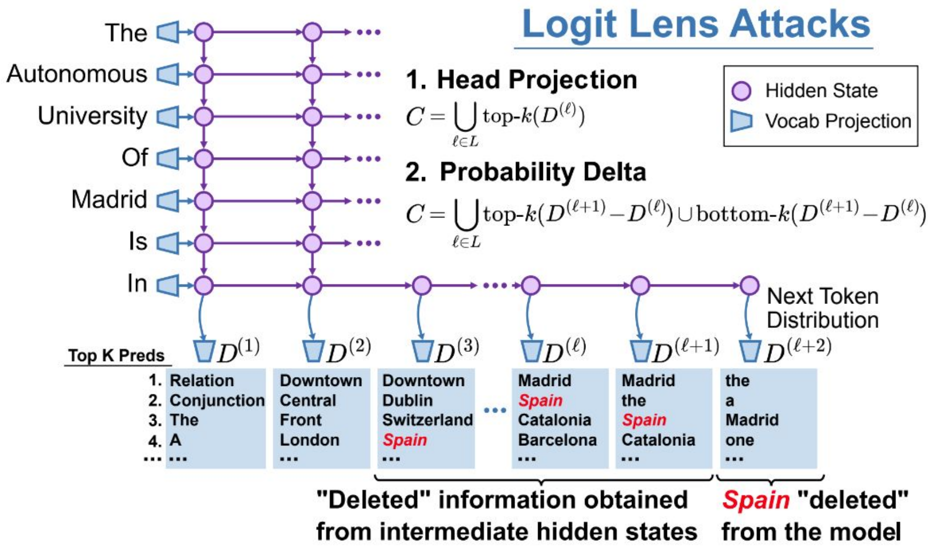
Improving Defense Methods

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

Unlearning in LLMs

Improving Defense Methods

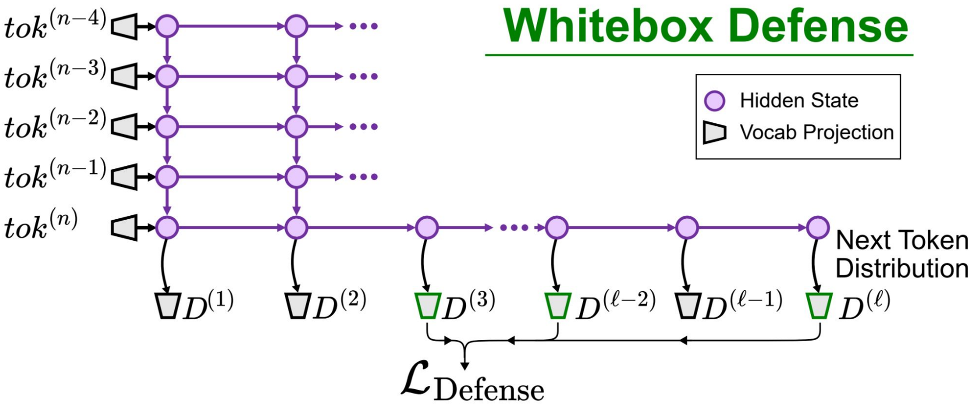
- Blackbox defense reduces to paraphrase + adversarial robustness
- Whitebox defense: *delete information wherever it appears in model*



Unlearning in LLMs

Improving Defense Methods

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



Maximize entropy

Unlearning in LLMs

Results

1. Up to 38% attack success for GPT-J facts deleted by ROME+Empty Response (B=10)

With whitebox defense

2. “Foreseen” whitebox attack: **37% → 1.7%**
3. “Unforeseen” whitebox attack: **38% → 2.4%**
4. Blackbox attack rate seems unchanged

See paper for blackbox defense

Unlearning in LLMs

Conclusions

- Want to delete sensitive information under **adversarial extraction attacks**
- **Whitebox defenses help**, but safety standards for deletion will vary

Roadmap

Goal	Research
Interpretability	Evaluating Explainable AI
Fine-grained Control	Model Editing
Deleting Sensitive Info	Machine Unlearning

Roadmap

Goal	Research
Interpretability	**Evaluating Explainable AI**
Fine-grained Control	**Model Editing**
Deleting Sensitive Info	**Machine Unlearning**

UNC NLP

Big Picture

2020

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?

2021

When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data

The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations

Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs

2022

VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives

2023

Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models

Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks

Big Picture

2020

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?

2021

When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data

The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations

Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs

2022

VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives

2023

Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models

Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks

Interpretability

UNC NLP

Big Picture

2020

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?

2021

When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data

The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations

Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs

2022

VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives

2023

Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models

Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks

Interpretability **Model Control**

UNC NLP

Big Picture

2020

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?

2021

When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data

The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations

Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs

2022

VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives

2023

Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models

Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks

Interpretability + Model Control

UNC NLP

Acknowledgements

- *My advisor*: Mohit Bansal
- *Co-authors*: Shiyue Zhang, Swarnadeep Saha, Miles Turpin, Zhuofan Ying, Thomas Hofweber, Harry Xie, Vaidehi Patil, Xiang Zhou, Stephen Casper, Prateek Yadav, Han Guo, and Archiki Prasad (among others)
- *Internship Supervisors*: Asma Ghandeharioun, Been Kim, Sarah Wiegreffe, Peter Clark, and Srinivasan Iyer
- *Friends & Family*: Michael, Tyler, Tom, Thomas, Alex, Joe, Sarah, Serge, Archana, David, Deblina, Grace, Miles, Juan, Kevin, Mark, Kaleigh, Justin, George, and my parents, Steve and Ashley Hase
- *Funders*: Google, UNC Graduate School, NSF, DARPA

Thank You!

PDFs + code: https://peterbhase.github.io/research/

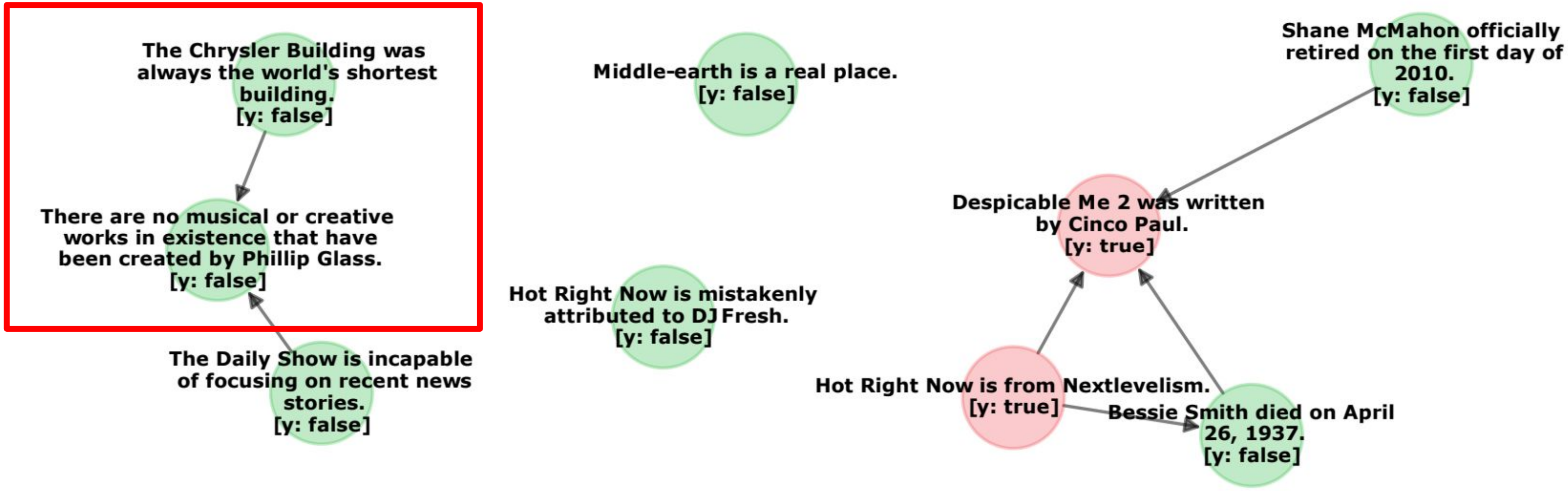
Contact Info:

Peter Hase, UNC Chapel Hill
peter@cs.unc.edu
https://peterbhase.github.io

Model Editing

- What else can we do with model editing?
- Let's look at *connections* between model beliefs
- Beliefs are connected when changing one leads the other to change
  - Update belief A → observe a change in belief B

Model Editing



Editing not very precise…

…or t5-base knowledge not structured very logically