Two Views of LM Interpretability

Sarah Wiegreffe
July 14, 2023

Phenomena we (don't) understand

- Let's assume we get white-box access to ChatGPT
- What next?
  - Play with inputs + outputs
  - Try novel tasks
  - Study internals
    - Where is information stored in the model?
      - Distributed or localized?
      - Different types of information? (factual recall, logical/spatial/numerical knowledge, commonsense, etc..)
  - Improve controllability via casual interventions
    - Can we intervene and cause some effect on model predictions?
      - correcting factually incorrect information, mitigating biases, personalization, etc.
      - either via inputs, outputs , hidden representations, or model parameters

Definitions

<u>Prompting + Querying</u>

- Providing inputs in natural language & observing models' (natural language) outputs
- often involves creating a custom dataset with specific properties

<u>Both</u>

- Inform our larger view of LMs
- Target *behaviors*

<u>Mechanistic Interpretability</u>

- Attempts to map model parameters or representations to specific functions which are more human-interpretable

Holtzman et al. 2023. Generative Models as a Complex Systems Science: How can we make sense of large language model behavior?

Definitions

When eating a hamburger with friends, what are people trying to do?

Interventions: noising entities, prompt format, # examples

<u>Prompting + Querying</u>

LM

Intervention: parameters

<u>Mechanistic Approaches</u>

Have fun, because usually a hamburger with friends indicates a good time.

Observe/evaluate: model outputs

"Behavior to explain"

Outline

Prompting + Querying

- Examples
  - Probing Language Models for Supporting Facts
  - Measuring Probability Mass on Answer Choices in Multiple-Choice Tasks

Both

- Definitions
- Pros & Cons
- Open Questions

Mechanistic Interpretability

- Examples
  - ROME & MEMIT Fact-Localization and Editing Algorithms
  - Editing Commonsense Knowledge in LMs

Prompting + Querying

Examples

Prompting + Querying

Both

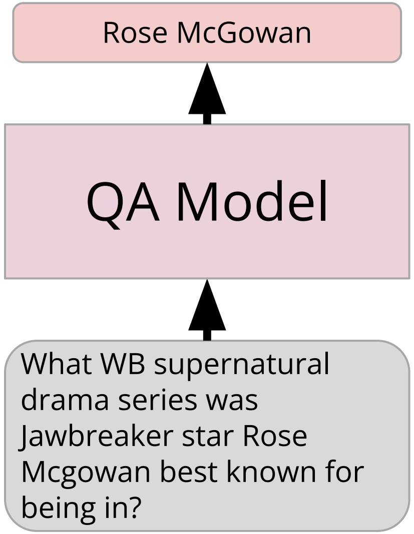
Mechanistic Interpretability

- NLP Checklists. Ribeiro et al. 2020.
- Factual Probing (Petroni et al. 2019 (LAMA); Jiang et al. 2020 (LPAQA)
- Consistency Probing (e.g., Kassner & Schütze 2020, Elazar et al. 2021)
- Removal/Perturbation-based Token Attribution Methods (e.g., Lundberg & Lee 2017, LIME, counterfactual edits)

Prompting + Querying

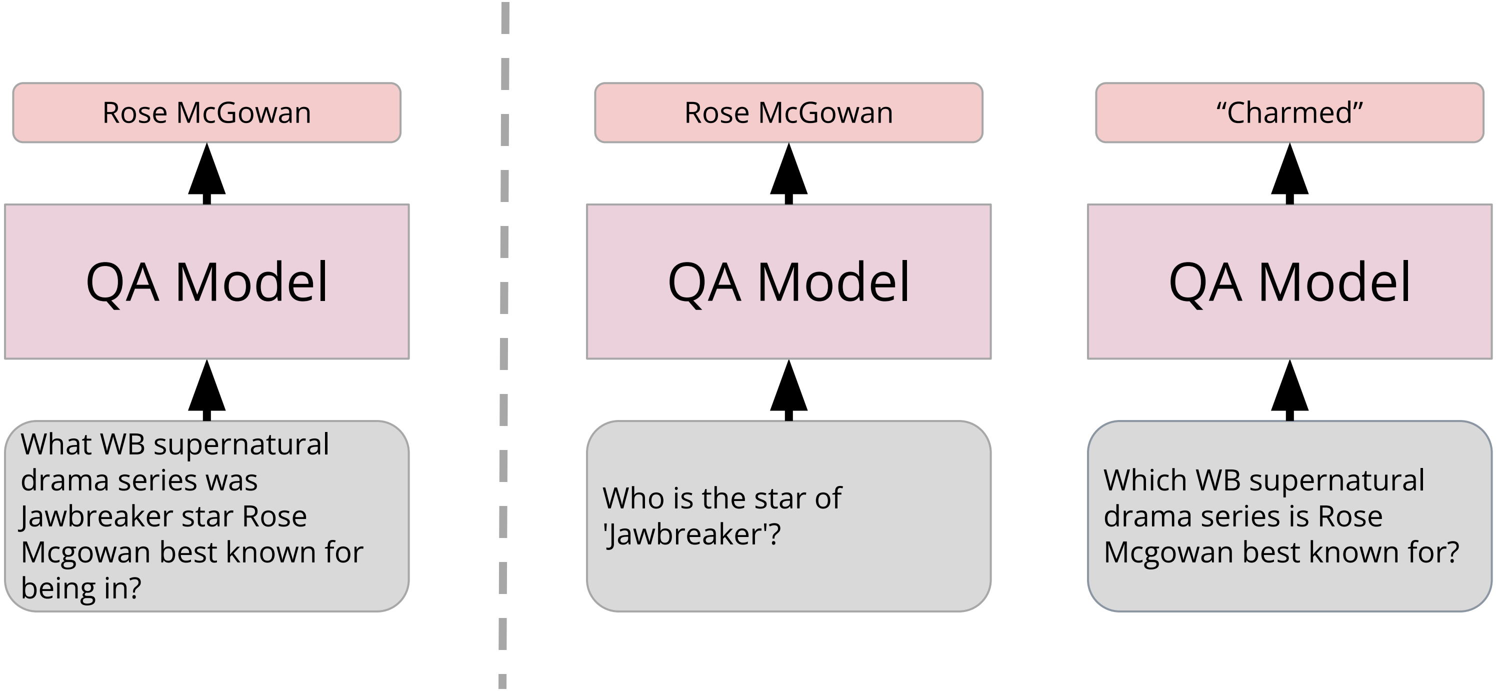
1. **Probing Language Models for Supporting Facts**
2. Measuring Probability Mass on Answer Choices in Multiple-Choice Tasks

Probing Language Models for Supporting Facts



Xie, Wiegreffe, & Riedl. Findings of EMNLP 2022. *Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes.*

Probing Language Models for Supporting Facts



Xie, Wiegreffe, & Riedl. Findings of EMNLP 2022. *Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes.*

Multi-hop Question Decomposition

- Architecture for multi-hop QA (Min et al., 2019; Perez et al., 2020; Khot et al., 2021)
  1. **automatically decompose the question into sub-questions**
  2. answer those sub-questions
  3. synthesize the answers to the sub-questions to answer the original question
- Functions as an effective tool to help boost the empirical performance of the QA system.

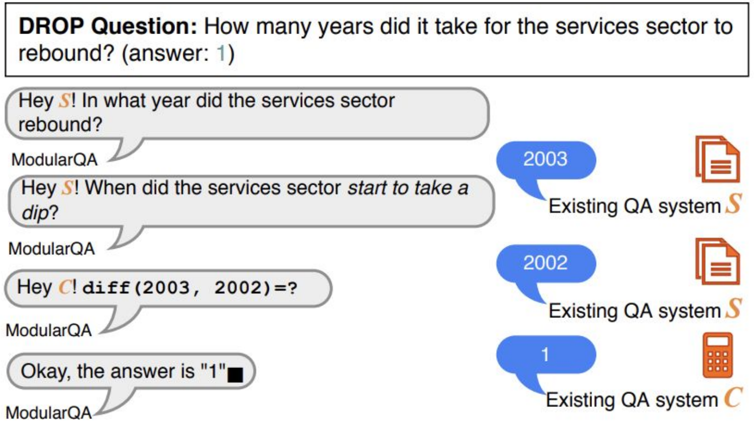


Figure: Khot et al. 2021

Sub-QA is closely tied to the main QA

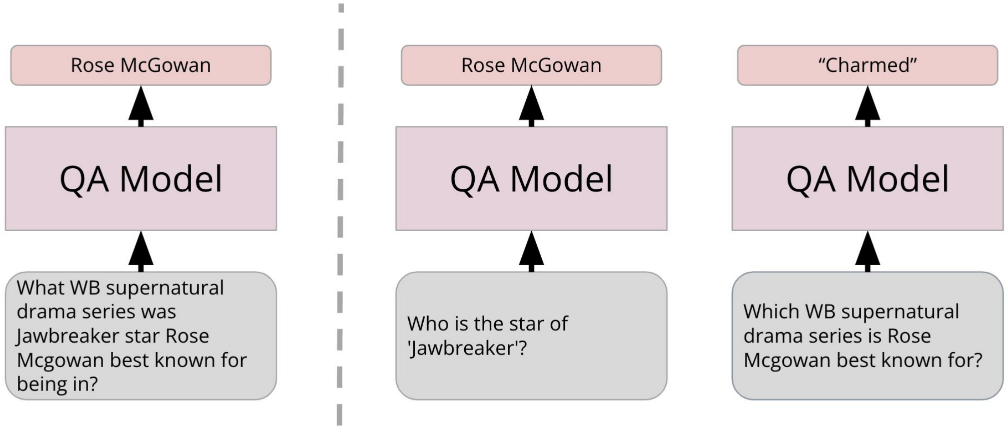
- Sub-question answering can distinguish incorrect/correct model predictions

Model	Model Pred.	n	Sub-Q Accuracy
T5	Correct	617	**85.09**
	Incorrect	59	64.41
BART	Correct	597	**85.59**
	Incorrect	79	60.76

Table 3: Combined sub-question task performance, split by whether the model predicted the main question correctly or not.

Simulatability Experiments

- Does exposing the decompositional probes along with the answers to the probes to users improve their ability to predict the model behavior?
    - Yes!



Takeaways

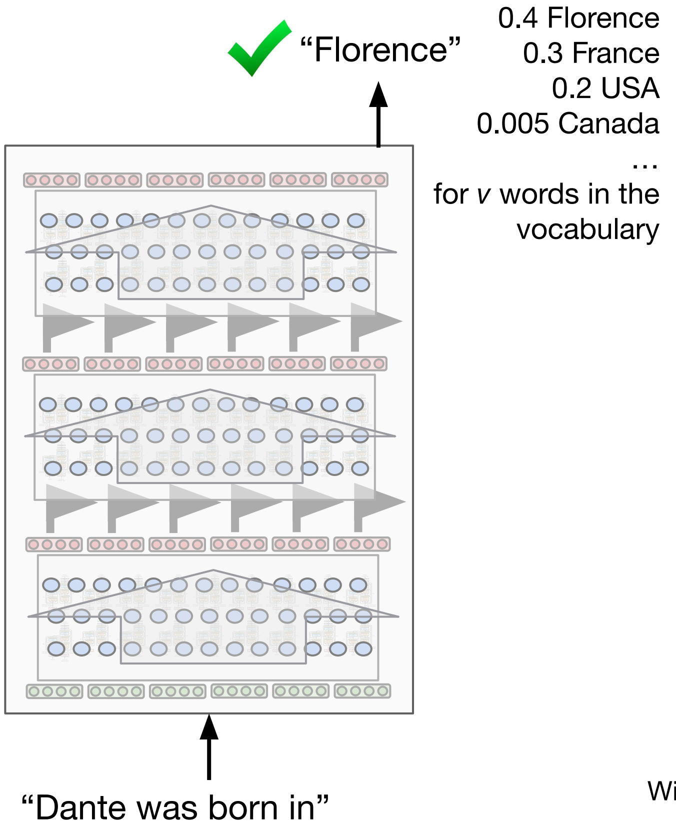
- Decomposition is an **effective means for probing neural QA models**; pairs of sub-questions and answers can serve as **structured instance-level explanations**.

- Explanations created by probing the neural QA model with question decompositions **can help humans construct a mental model** on which they can rely to predict the model behavior.

Prompting + Querying

1. Probing Language Models for Supporting Facts
2. **Measuring Probability Mass on Answer Choices in Multiple-Choice Tasks**

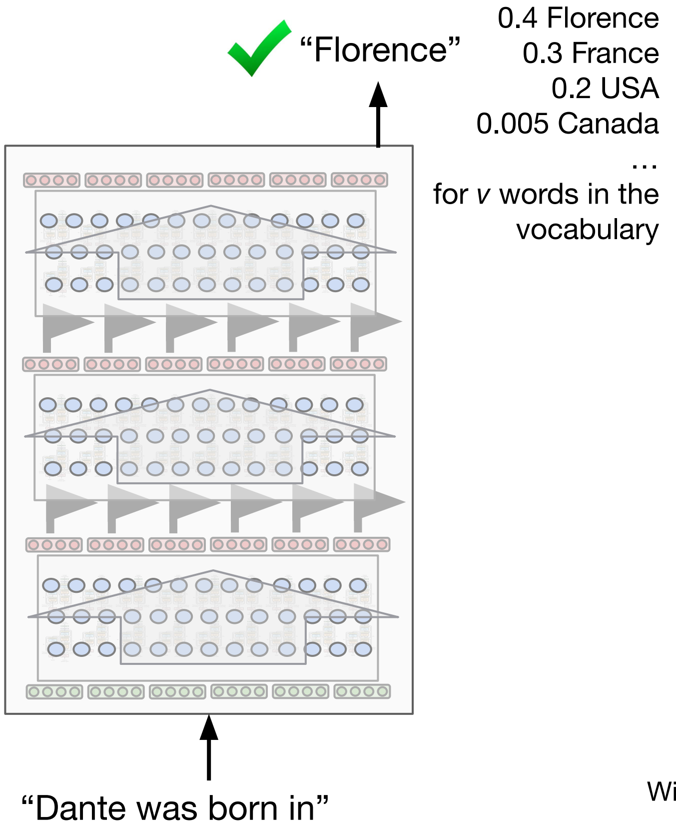
Measuring Probability Mass on Answer Choices in Multiple-Choice Tasks



- How do we surface information from language models?
  - Probabilistic systems don't amend to the view of a single "belief" or "knowledge"
  - Decoding algorithms can have a strong effect

Wiegreffe, Finlayson, Tafjord, Clark, & Sabharwal, 2023. arxiv.org/abs/2305.14596

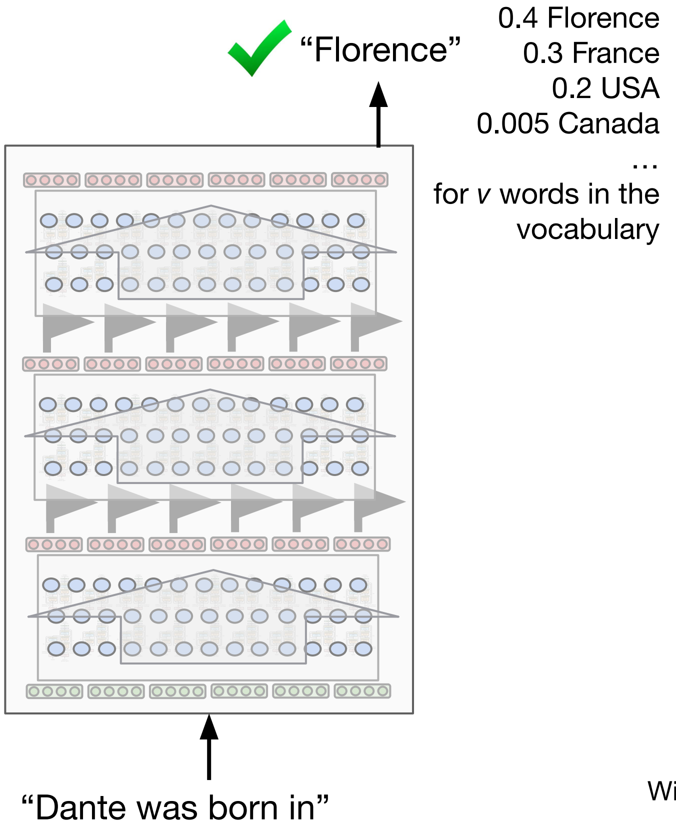
Measuring Probability Mass on Answer Choices in Multiple-Choice Tasks



- How do we surface information from language models?
    - Probabilistic systems don’t amend to the view of a single “belief” or “knowledge”
    - Decoding algorithms can have a strong effect
- Look at model probabilities
- Intuition that higher probability assigned to answer choices/decoding valid answer choices reliably indicates better model “understanding”

Wiegreffe, Finlayson, Tafjord, Clark, & Sabharwal, 2023. arxiv.org/abs/2305.14596

Measuring Probability Mass on Answer Choices in Multiple-Choice Tasks



- How do we surface information from language models?
  - Probabilistic systems don't amend to the view of a single "belief" or "knowledge"
  - Decoding algorithms can have a strong effect
- Look at model probabilities
- Intuition that higher probability assigned to answer choices/decoding valid answer choices reliably indicates better model "understanding"
  - Not always!

Wiegreffe, Finlayson, Tafjord, Clark, & Sabharwal, 2023. arxiv.org/abs/2305.14596

Background: Predicting a Label using a Language Model

Traditional sequence scoring approaches select a prediction $\hat{y}$ as

$$\boxed{\hat{y}^{\text{Seq-Sc}} = \underset{\ell \in \mathcal{L}}{\operatorname{argmax}}\, P_\theta(\ell | x)}$$

set of possible answer choices

probability assigned by language model
***sums to 1 over entire vocabulary**

input instance

answer choice with highest probability

Question: "An electric car runs on electricity via"

Answer choices:

gasoline → 0.092
a power station → 0.061
electrical conductors→ 0.045
fuel→ 0.063

Greedy generation:
electricity via electrical conductors
electricity → 0.126

Background: Predicting a Label using a Language Model

Traditional sequence scoring approaches select a prediction $\hat{y}$ as

$$\boxed{\underbrace{\hat{y}^{\text{Seq-Sc}} = \operatorname*{argmax}_{\ell \in \mathcal{L}} P_\theta(\ell|x)}_{\text{answer choice with highest probability}}}$$

set of possible answer choices

probability assigned by language model
***sums to 1 over entire vocabulary**

input instance

answer choice with highest probability

Question: "An electric car runs on electricity via"

Answer choices:

gasoline → 0.092
a power station → 0.061
electrical conductors→ 0.045
fuel→ 0.063

Greedy generation:
electricity via electrical conductors
electricity → 0.126

The "Surface Form Competition" (SFC) Hypothesis

Holtzman et al. 2021. *Surface Form Competition: Why the Highest Probability Answer Isn't Always Right.*

Background

Surface Form Competition:
Why the Highest Probability Answer Isn't Always Right

*Ari Holtzman[1] *Peter West[1,2]
Vered Shwartz[1,2] Yejin Choi[1,2] Luke Zettlemoyer[1]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
{ahai,pawest}@cs.washington.edu

$$\hat{y}^{\text{PMI-DC}} = \underset{\ell \in \mathcal{L}}{\operatorname{argmax}} \frac{P_\theta(\ell|x)}{P_\theta(\ell)}$$

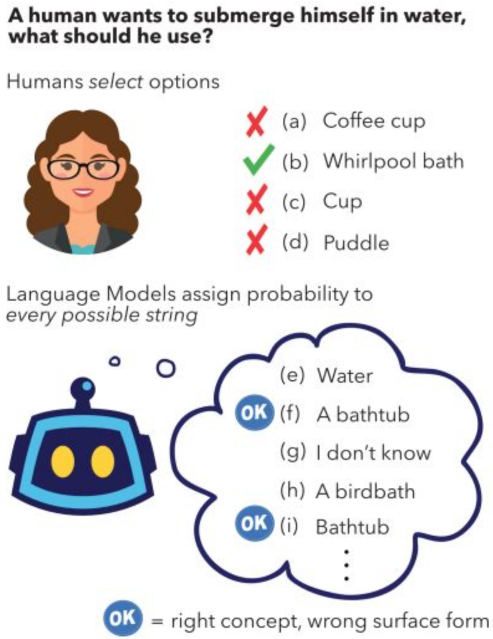


Figure 1: While humans select from given options, language models implicitly assign probability to every possible string. This creates surface form competition between different strings that represent the same concept. Example from CommonsenseQA (Talmor et al., 2019).

Contributions

1) **How to measure SFC?**
    a) Metric (upper bound)
    b) Effect it can have on accuracy
2) **How to reduce its effect?**
    a) By showing answer choices in the prompt (and sometimes 1 in-context example)
3) **When is it a problem? I.e., does reducing SFC improve accuracy?**
    a) Surprisingly, not always! Depends on the model
    b) Encouraging models to produce answer choices can counter-intuitively be detrimental to task performance for certain LMs.

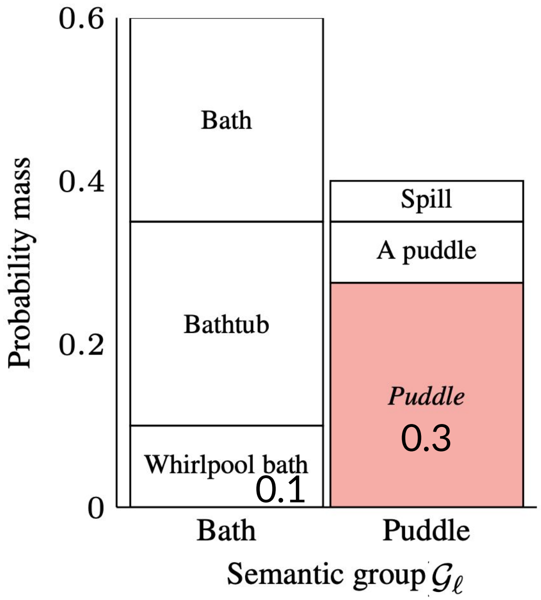
Wiegreffe, Finlayson, Tafjord, Clark, & Sabharwal, 2023. arxiv.org/abs/2305.14596

Contributions

1) **How to measure SFC?**
    a) Metric (upper bound)
    b) Effect it can have on accuracy

Formulation of SFC

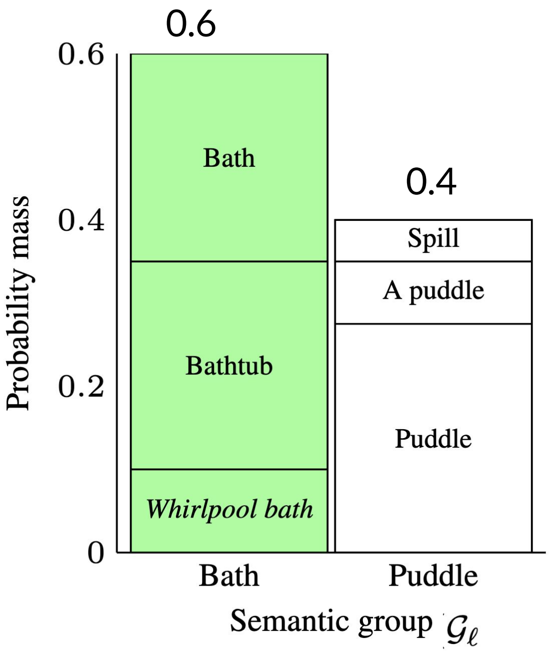
A human wants to submerge themselves in water. What should they use?
Choices: Puddle, Whirlpool bath



$$\hat{y}^{\text{Seq-Sc}} = \underset{\ell \in \mathcal{L}}{\operatorname{argmax}} \, P_\theta(\ell | x)$$

Formulation of SFC

A human wants to submerge themselves in water. What should they use?
Choices: Puddle, Whirlpool bath

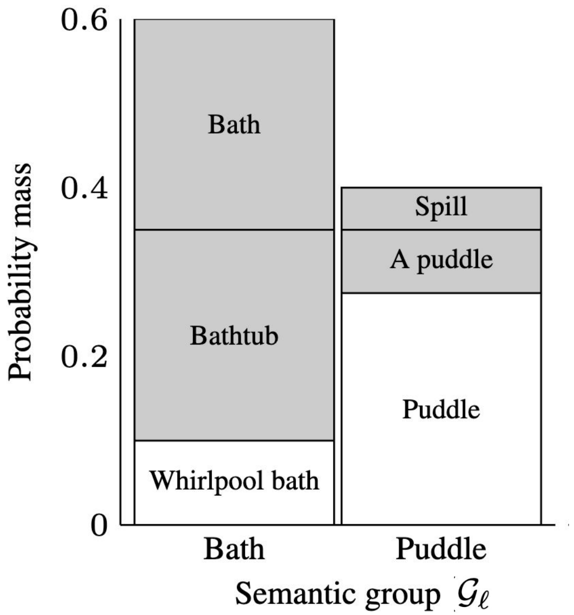


$$\hat{y}^{\text{Seq-Sc}} = \operatorname*{argmax}_{\ell \in \mathcal{L}} P_\theta(\ell | x)$$

$$\hat{y}^{\text{SFC-free}} = \operatorname*{argmax}_{\ell \in \mathcal{L}} P_\theta(\mathcal{G}_\ell | x)$$

Formulation of SFC

A human wants to submerge themselves in water. What should they use?
Choices: Puddle, Whirlpool bath



$$\hat{y}^{\text{Seq-Sc}} = \underset{\ell \in \mathcal{L}}{\operatorname{argmax}}\, P_\theta(\ell|x)$$

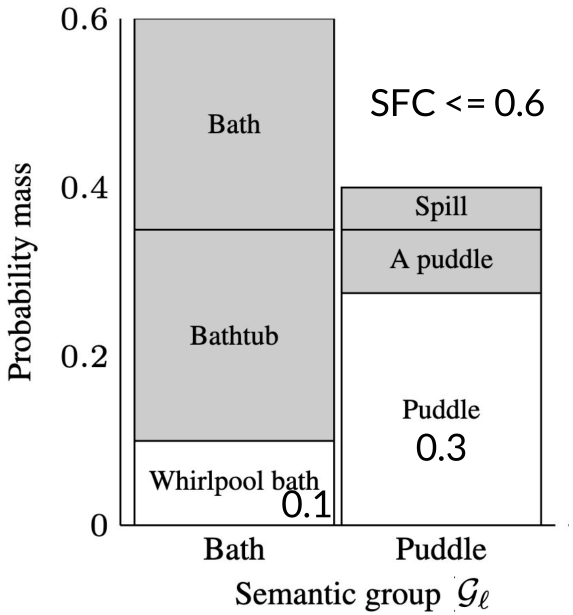
$$\hat{y}^{\text{SFC-free}} = \underset{\ell \in \mathcal{L}}{\operatorname{argmax}}\, P_\theta(\mathcal{G}_\ell|x)$$

$$\mathrm{SFC}_\theta(\mathcal{L}, x) = \sum_{\ell \in \mathcal{L}} \Big(P_\theta(\mathcal{G}_\ell|x) - P_\theta(\ell|x) \Big)$$

Formulation of SFC

A human wants to submerge themselves in water. What should they use?
Choices: Puddle, Whirlpool bath

$$\boxed{\mathrm{PMA}_\theta(\mathcal{L}, x) = \sum_{\ell \in \mathcal{L}} P_\theta(\ell | x)}$$



$$\hat{y}^{\text{Seq-Sc}} = \operatorname*{argmax}_{\ell \in \mathcal{L}} P_\theta(\ell | x)$$

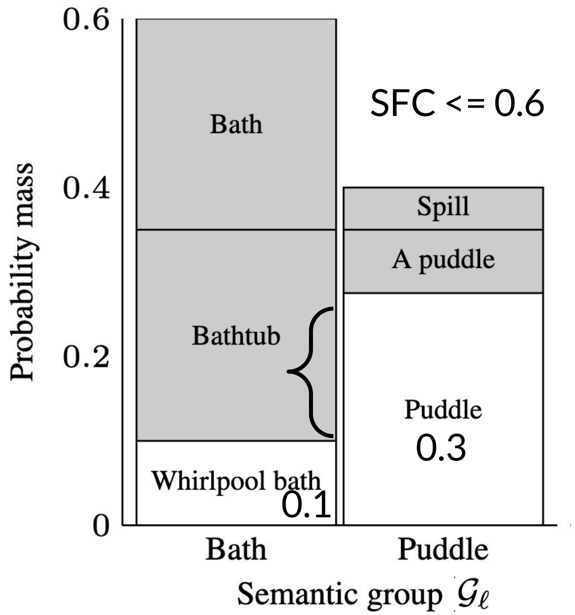
$$\hat{y}^{\text{SFC-free}} = \operatorname*{argmax}_{\ell \in \mathcal{L}} \underbrace{P_\theta(\mathcal{G}_\ell | x)}_{\textit{Unknown}}$$

$$\mathrm{SFC}_\theta(\mathcal{L}, x) = \sum_{\ell \in \mathcal{L}} \Big(\underbrace{P_\theta(\mathcal{G}_\ell | x)} - P_\theta(\ell | x) \Big)$$

$$\leq 1 - \sum_{\ell \in \mathcal{L}} P_\theta(\ell | x)$$

SFC's Effect on Accuracy

A human wants to submerge themselves in water. What should they use?
Choices: Puddle, Whirlpool bath



If true, SFC has no effect on prediction:

$$1 - \mathrm{PMA}_\theta(\mathcal{L}, x) \ < \ P_\theta(\hat{y}|x) - P_\theta(y_2|x)$$

0.6 > 0.3 - 0.1

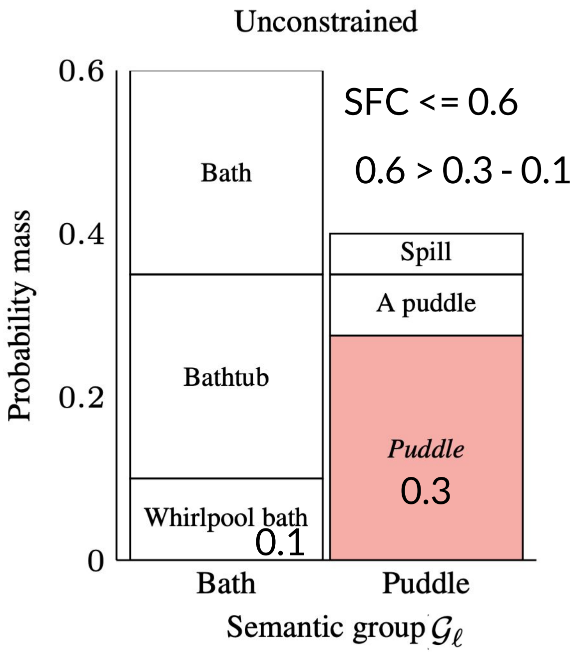
Contributions

1) **How to measure SFC?**
   a) Metric (upper bound)
   b) Effect it can have on accuracy
2) **How to reduce its effect?**
   a) By showing answer choices in the prompt (and sometimes 1 in-context example)
3) **When is it a problem? I.e., does reducing SFC improve accuracy?**
   a) Surprisingly, not always! Depends on the model
   b) Encouraging models to produce answer choices can counter-intuitively be detrimental to task performance for certain LMs.

Wiegreffe, Finlayson, Tafjord, Clark, & Sabharwal, 2023. arxiv.org/abs/2305.14596

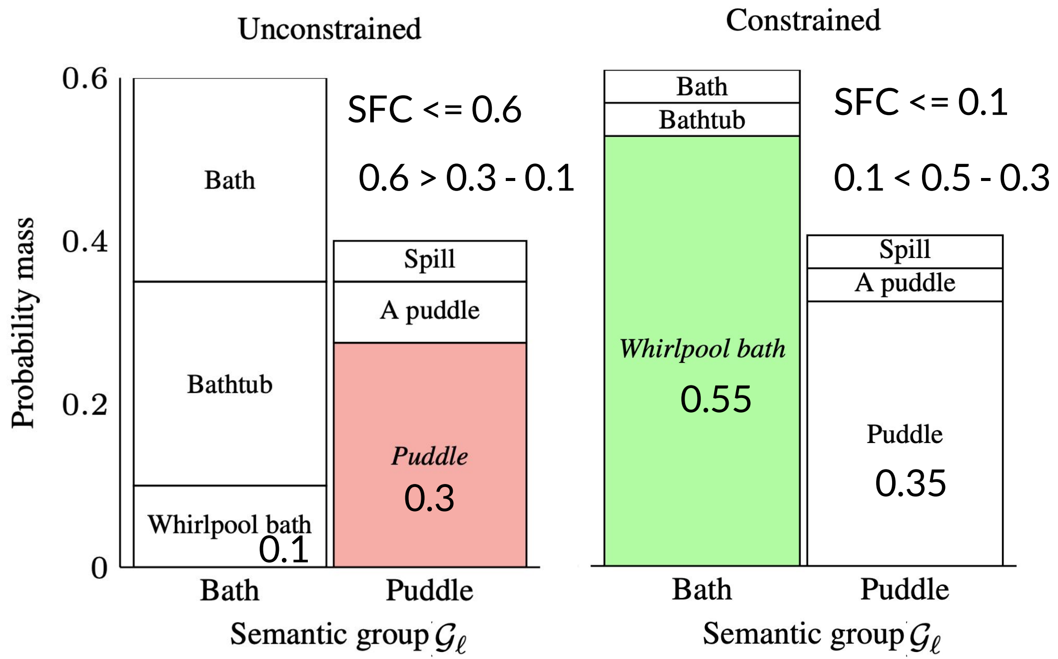
Under-constrained generation

A human wants to submerge themselves in water. What should they use?
Choices: Puddle, Whirlpool bath



Under-constrained generation

A human wants to submerge themselves in water. What should they use?
Choices: Puddle, Whirlpool bath



How to constrain?

- Fine-tuning
- Few-shot demonstrations
- Prompt Format
    - Showing answer choices
- Expected output format

Experimental Setup

3 Tasks/Benchmarks:

- MMLU
- OpenbookQA
- CommonsenseQA

6 Models:

"Vanilla" LMs

- GPT-3 curie (~6.7B)
- OPT 30B
- GPT-3 davinci (~175B)

Instruction-Tuned (+)

- FLAN-T5 XXL (11B)
- GPT-3 davinci-instruct-beta (~175B)
- GPT-3.5 text-davinci-003 (unknown #)

Experimental Setup

Three prompt formats:

1) No answer choices

An electric car runs on electricity via **{gasoline, a power station, electrical conductors, fuel}**

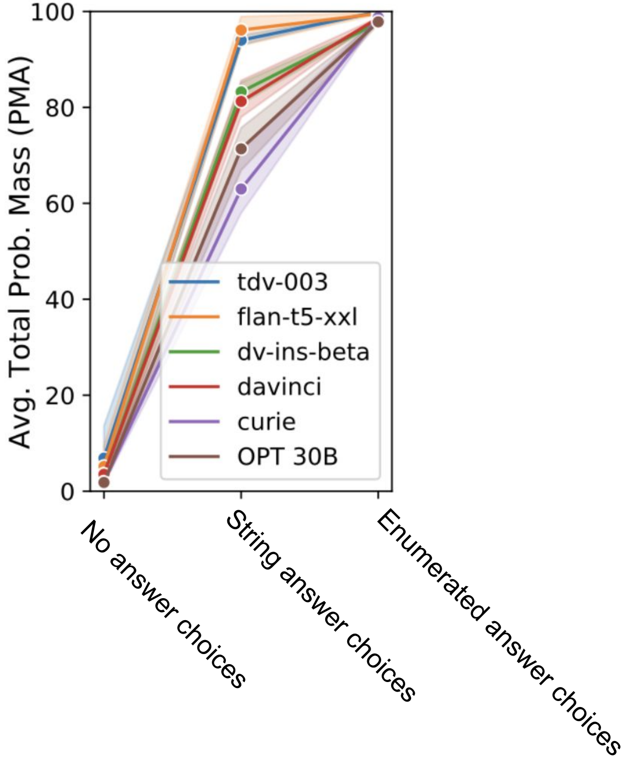
2) String Answer Choices

question: An electric car runs on electricity via answer choices: gasoline, a power station, electrical conductors, or fuel The correct answer is: **{gasoline, a power station, electrical conductors, fuel}**

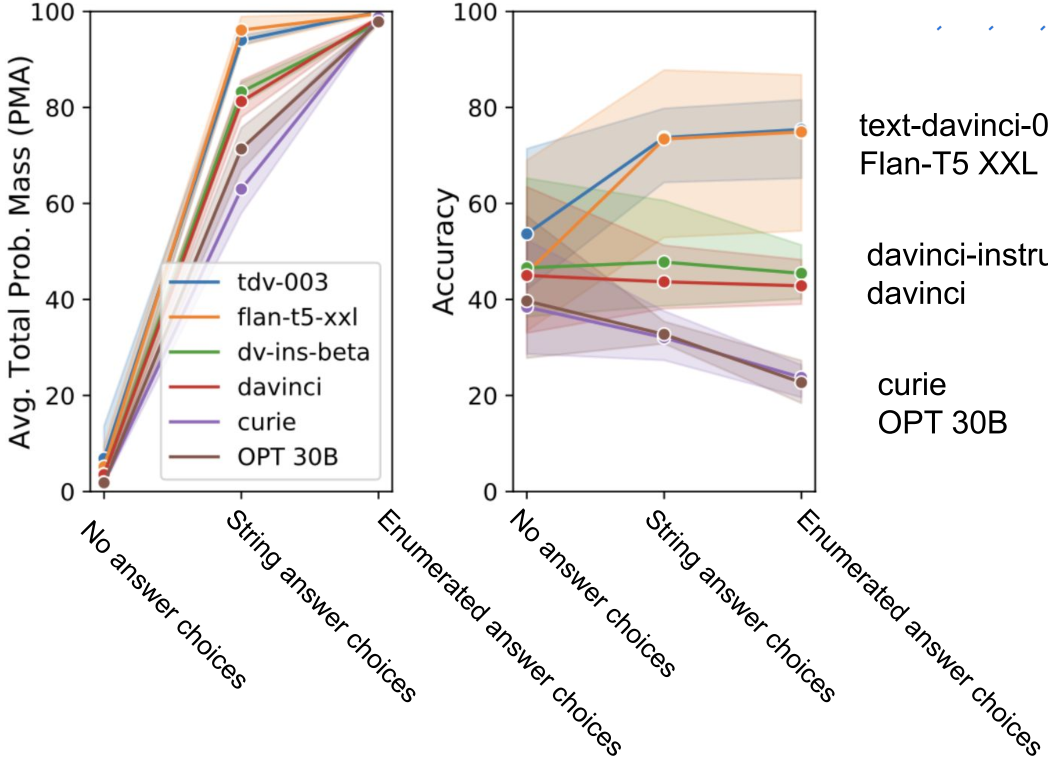
3) Enumerated Answer Choices

Question: An electric car runs on electricity via Choices: A: gasoline B: a power station C: electrical conductors D: fuel Answer: **{A, B, C, D}**

1-shot Results– all tasks



1-shot Results– all tasks



text-davinci-003
Flan-T5 XXL

davinci-instruct-beta
davinci

curie
OPT 30B

Main Findings

1. **Prompt format is crucial.** Showing 1 in-context example *and answer choices* in the prompt is an effective way to alleviate surface form competition, for all models tested.

2. Surprisingly, it is **not always the case** that increasing probability mass on valid answers results in higher accuracy.

Section 1 Takeaways

PROS

Prompting + Querying

Both

Mechanistic Interpretability

- Fully in natural language
- Accessible; easy to define controlled experiments
- Considers *full system* end-to-end

Section 1 Takeaways

CONS

Prompting + Querying

- Space of possible NL queries is large, so fundamental system understanding reached may be limited
- hard to generalize and dissect instance-level behavior
- Little actionability for how to control or change model behavior

Both

Mechanistic Interpretability

“Mechanistic” Interpretability

1. **ROME & MEMIT Fact-Localization and Editing Algorithms**
2. Editing Commonsense Knowledge in LMs

Definitions

Prompting + Querying

- Providing inputs in natural language & observing models' (natural language) outputs
- often involves creating a custom dataset with specific properties

Both

- Inform our larger view of LMs
- Target *behaviors*

Mechanistic Interpretability

- Attempts to map model parameters or representations to specific functions which are more human-interpretable

Holtzman et al. 2023. Generative Models as a Complex Systems Science: How can we make sense of large language model behavior?

Probing Classifiers

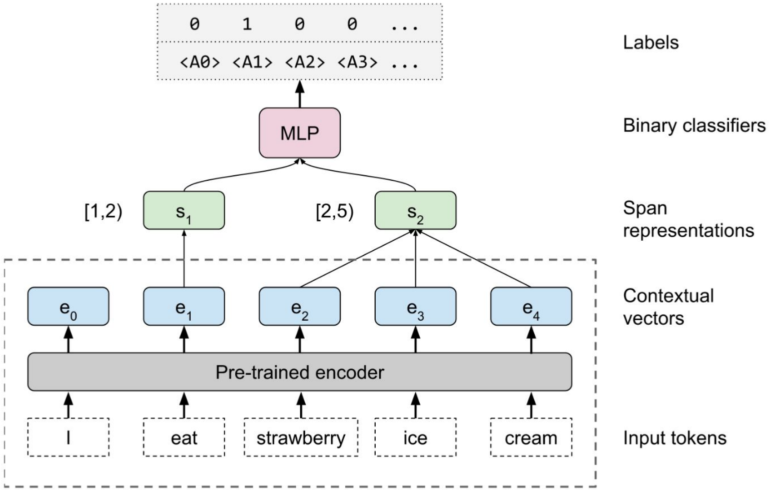


Figure 1: Probing model architecture (§ 3.1). All parameters inside the dashed line are fixed, while we train the span pooling and MLP classifiers to extract information from the contextual vectors.

Figure: *What do you learn from context?....* Tenney et al., ICLR 2019.

- A number of pitfalls
  - Correlation != Causation

***Probing Classifiers: Promises, Shortcomings, and Alternatives*. Belinkov, CL 2022.**

What is mechanistic interpretability?

- bottom-up approach:
  - if we can define and understand the *mechanics* of individual neurons (or weight matrices), then
  - we can build up to understanding the mechanics of *sets* of neurons (or weight matrices) and their interactions (circuits), and then
  - we can build up to an understanding of a large, dense network

Olah, Cammarata, Schubert, Goh, Petrov, Carter, 2020. [Zoom In: An Introduction to Circuits.]()

Pitfalls of Neuron-Level Analysis in NLP

- *"DNNs are **distributed in nature**, which encourages groups of neurons to work together to learn a concept. The current analysis methods, at large, **ignore interaction between neurons** while discovering neurons with respect to a concept."*

Neuron-level Interpretation of Deep NLP Models: A Survey. Sajjad et al., TACL 2022.

- *"Since the ranking space is too large (768! in BERT's case), these methods provide **approximations to the problem and are non-optimal**."*

On the Pitfalls of Analyzing Individual Neurons in Language Models. Antverg & Belinkov, ICLR 2022.

Examples

Prompting + Querying

- NLP Checklists. Ribeiro et al. 2020.
- Factual Probing (Petroni et al. 2019 (LAMA); Jiang et al. 2020 (LPAQA)
- Consistency Probing (e.g., Kassner & Schütze 2020, Elazar et al. 2021)
- Removal/Perturbation-based Token Attribution Methods (e.g., Lundberg & Lee 2017 Shap, LIME, counterfactual edits)

Both

- Model editing *evaluation* testbeds for localization methods

Mechanistic Interpretability

- Causal Interventions/Mediations (Giulianelli et al. 2018, Vig et al, 2020, Elazar et al. 2020)
- Causal Abstraction (Geiger et al. 2021–)
- Model Editing (ROME, MEMIT)
- Reverse-engineering small models (Elhage et al. 2021, Olsson et al. 2022

ROME & MEMIT (Meng et al. 2022, 2023)

1) isolate the most influential hidden states, neurons, or activations in a model for predicting a specific fact
2) Edit them to change the prediction

ROME & MEMIT (Meng et al. 2022, 2023)

1) **isolate the most influential hidden states, neurons, or activations in a model for predicting a specific fact**
2) Edit them to change the prediction

ROME & MEMIT (Meng et al. 2022, 2023)

1) **isolate the most influential hidden states, neurons, or activations in a model for predicting a specific fact**
    a) “Causal Tracing”/”Causal Mediation analysis”
2) Edit them to change the prediction

Investigating Gender Bias in Language Models Using Causal Mediation Analysis.
Vig et al., NeurIPS 2020.

ROME & MEMIT (Meng et al. 2022, 2023)

1) **isolate the most influential hidden states, neurons, or activations in a model for predicting a specific fact**
   a) "Causal Tracing"/"Causal Mediation analysis"
2) Edit them to change the prediction

Run the network twice

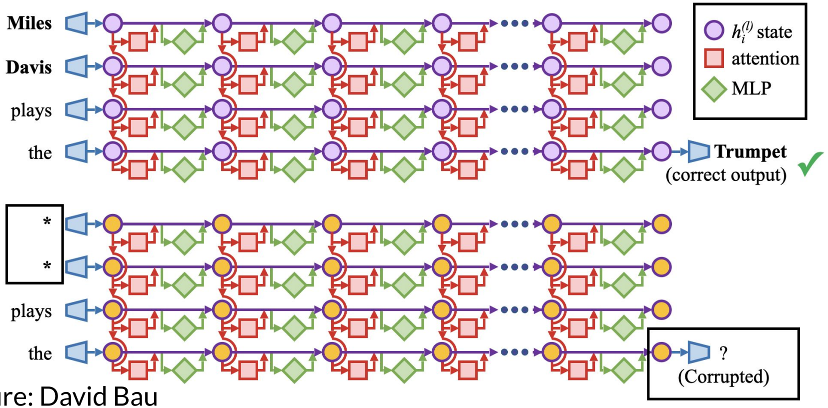
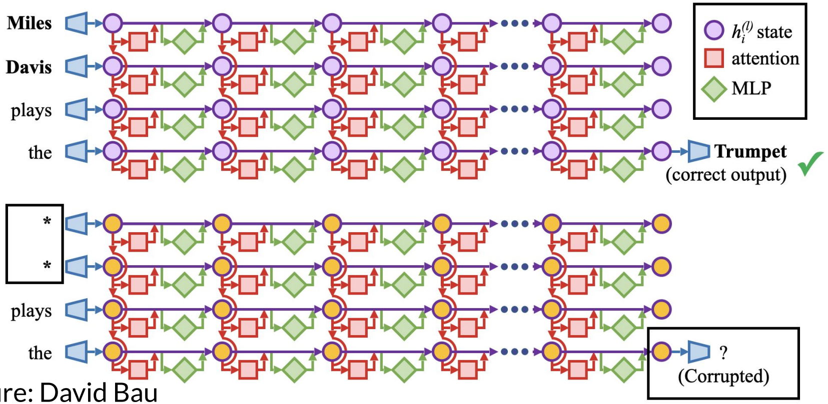


Figure: David Bau
Meng et al. 2022. *Locating and Editing Factual Associations in GPT.*

ROME & MEMIT (Meng et al. 2022, 2023)

1) **isolate the most influential hidden states, neurons, or activations in a model for predicting a specific fact**
   a) "Causal Tracing"/"Causal Mediation analysis"
2) Edit them to change the prediction

Run the network twice



Transplant Hidden State

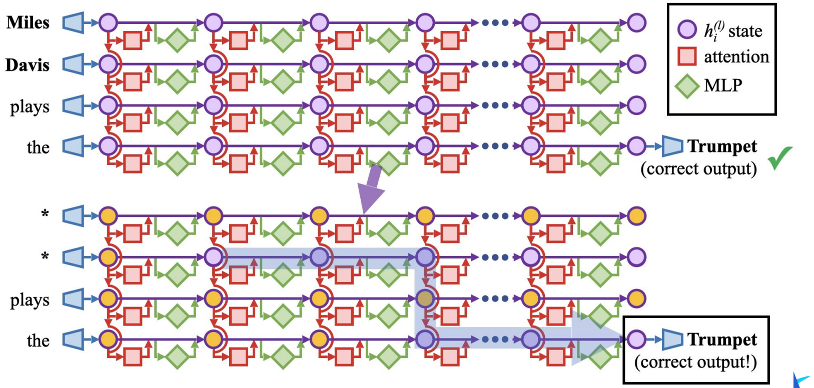
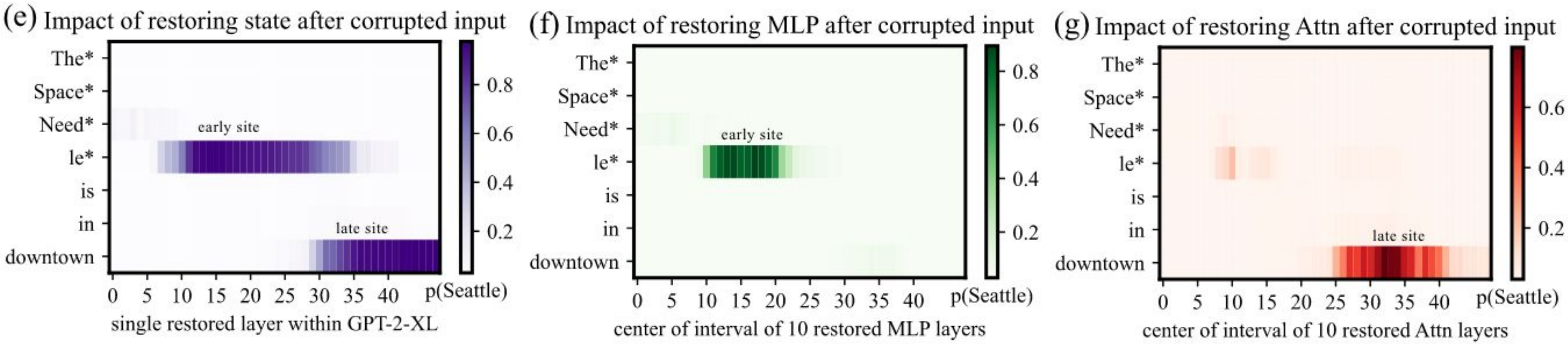


Figure: David Bau
Meng et al. 2022. *Locating and Editing Factual Associations in GPT.*

ROME & MEMIT (Meng et al. 2022, 2023)



Meng et al. 2022. *Locating and Editing Factual Associations in GPT.*

Locality Hypothesis

- The "Localized Factual Association" Hypothesis (Meng et al. 2022)
  "*We conjecture that* ***any fact could be equivalently stored in any one of the middle MLP layers****. To test our hypothesis, we* ***narrow our attention*** *to a single MLP module at a mid-range layer $l*$, and ask whether its weights can be explicitly modified to store an arbitrary fact.*"

ROME & MEMIT (Meng et al. 2022, 2023)

1) isolate the most influential hidden states, neurons, or activations in a model for predicting a specific fact
   a) “Causal Tracing”/”Causal Mediation analysis”

2) Edit them to change the prediction
   a) ROME = Rank-One Model Edit

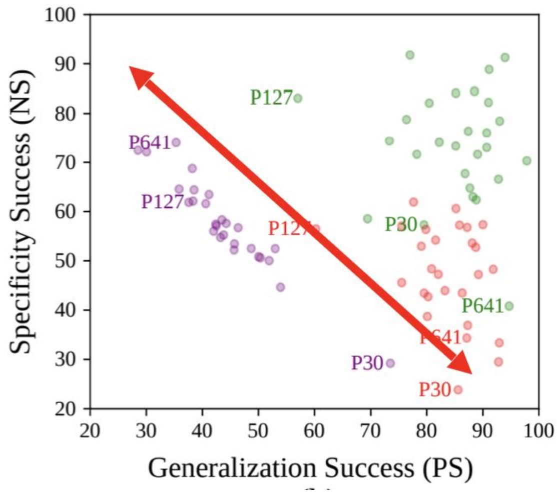
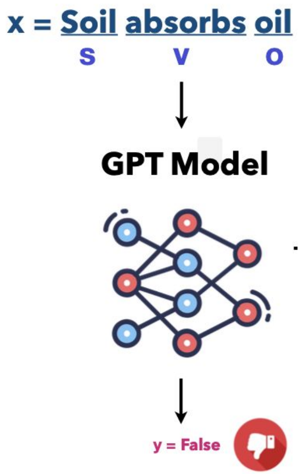


Figure: David Bau
Meng et al. 2022. *Locating and Editing Factual Associations in GPT.*

“Mechanistic” Interpretability

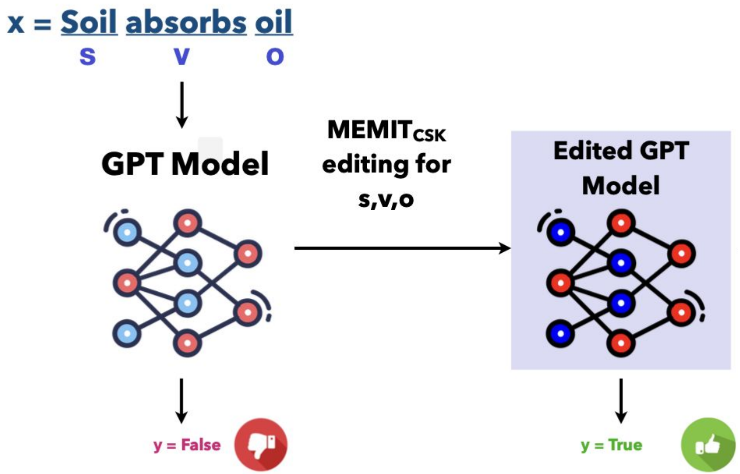
1. ROME & MEMIT Fact-Localization and Editing Algorithms
2. **Editing Commonsense Knowledge in LMs**

Editing Commonsense Knowledge in LMs



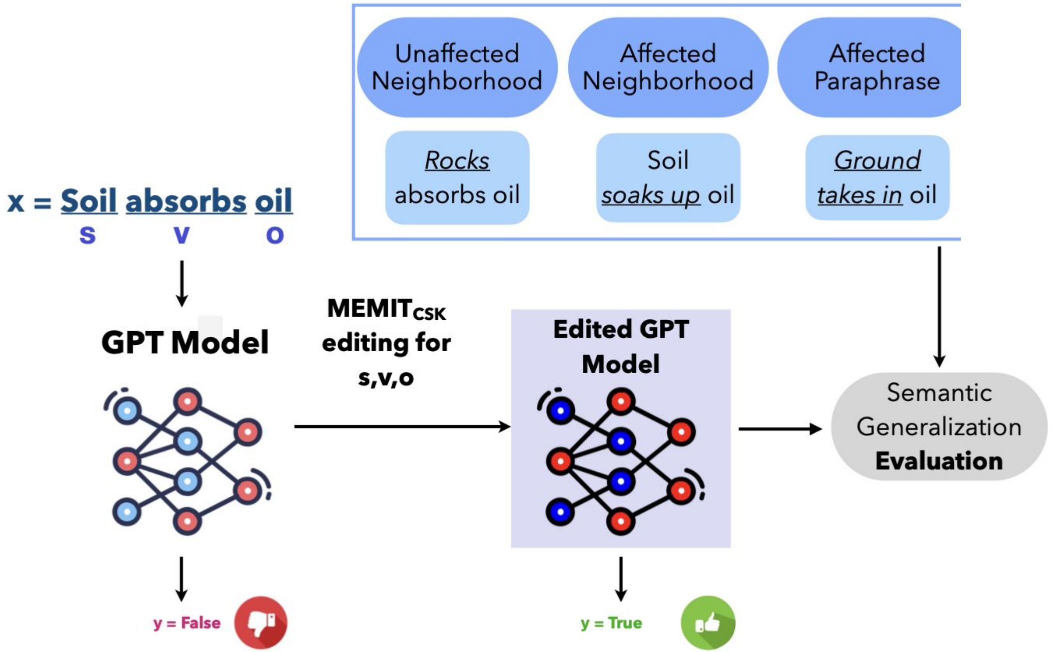
Gupta, Mondal, Sheshadri, Zhao, Li*, Wiegreffe*, & Tandon*, 2023. arxiv.org/abs/2305.14956

Editing Commonsense Knowledge in LMs



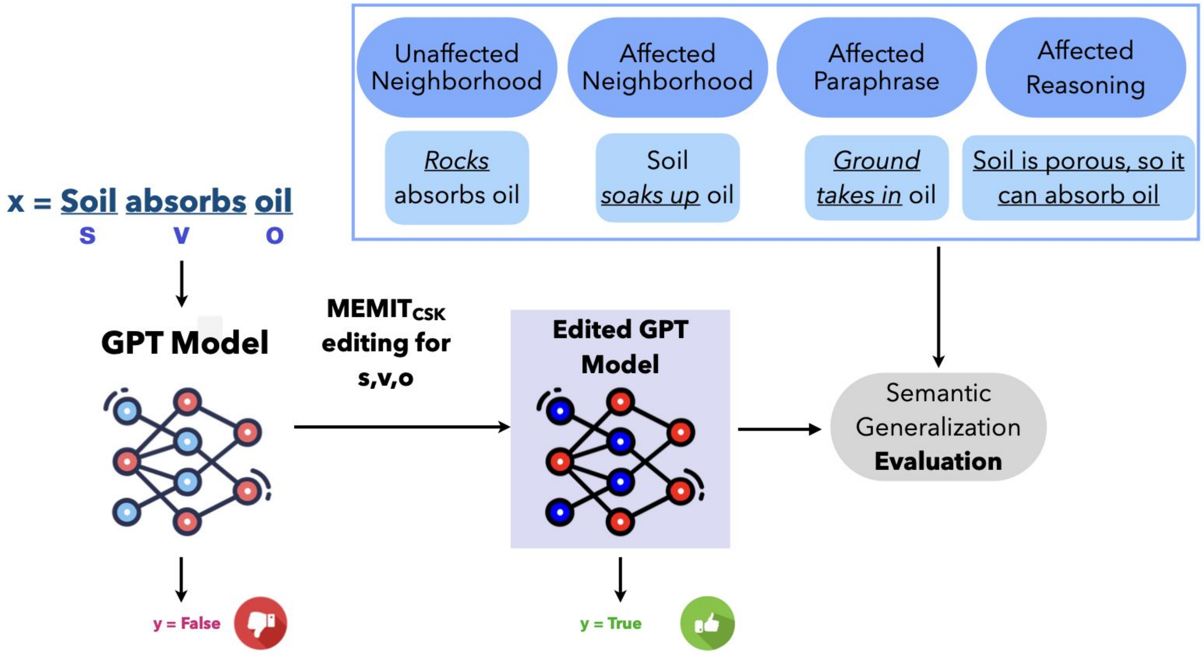
Gupta, Mondal, Sheshadri, Zhao, Li*, Wiegreffe*, & Tandon*, 2023. arxiv.org/abs/2305.14956

Editing Commonsense Knowledge in LMs



Gupta, Mondal, Sheshadri, Zhao, Li*, Wiegreffe*, & Tandon*, 2023. arxiv.org/abs/2305.14956

Editing Commonsense Knowledge in LMs

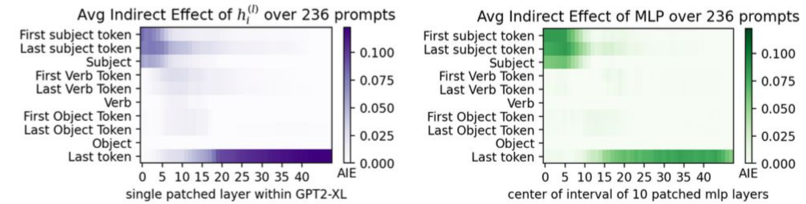


Gupta, Mondal, Sheshadri, Zhao, Li*, Wiegreffe*, & Tandon*, 2023. arxiv.org/abs/2305.14956

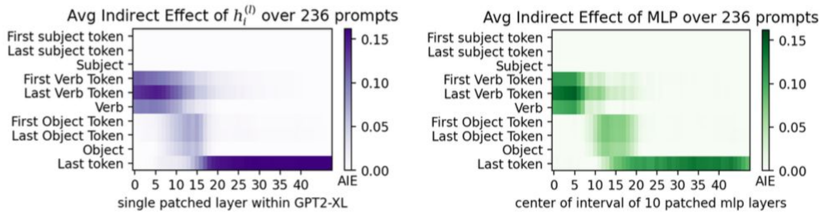
Findings

- Noised token position matters

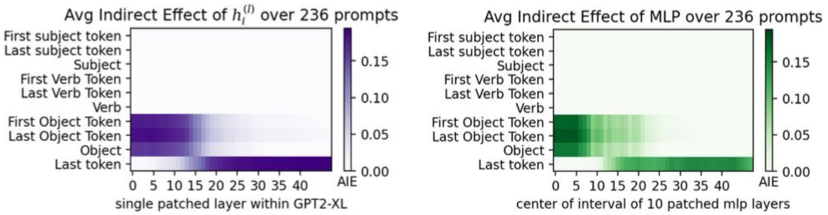
Edit subject:



Edit verb:



Edit object:



Findings

- Fine-tuning has a large tradeoff between fixing errors and retaining original performance
- Direct editing (on best token position) does not

Dataset	Update Method	Edit Token	Edit Layers	EDIT SET		
				F1 Score %	Efficacy %	Relapse %
PEP3k	Base Model	-	-	76.22	0	0
	$\text{RFT}_{\text{Early Stop}}$	-	-	80.92 (+4.70)	40.93	6.60
	$\text{RFT}_{\text{Fixed Epoch}}$	-	-	51.08 (-19.14)	100	55.70
	Edit	Last Subject	4,5,6,7,8	79.36 (+3.14)	54.95	12.77
	Edit	Last Verb	4,5,6,7,8	**89.08 (+12.86)**	**93.68**	**12.34**
	Edit	Last Object	1,2,3,4,5	77.65 (+1.43)	78.57	21.85

Section 2 Takeaways

PROS

Prompting + Querying

- Fully in natural language
- Accessible; easy to define controlled experiments
- Considers *full system* end-to-end

Both

- Targets *behaviors*
- Inform our larger view of LMs

Mechanistic Interpretability

- Provides a fundamental understanding of how models perform tasks at a *fine-grained level*
- Allows testing of a specific hypothesis for how models do tasks
- Positive results provide a degree of *actionability*

Section 2 Takeaways

CONS

Prompting + Querying

- Space of possible NL queries is large, so fundamental system understanding reached may be limited
- hard to generalize and dissect instance-level behavior
- Little actionability for how to control or change model behavior

Both

Mechanistic Interpretability

- Often targets only specific low-level behaviors, small/purpose-built networks, or simple tasks
- Negative results uninformative
- No unified goals/evaluation

Open Questions

- How to unify work on mechanistic interpretability?
    - Common definitions
    - Common tasks & benchmarks
    - Common measures of "success"
- Prompting + Querying on well-constructed test sets can provide more direct comparisons of mechanistic findings.
- What granularity or type of model internals to target?
    - Affects the feasibility + scalability of findings.
    - Weights vs. hidden representations

Thank You!

mastodon https://sigmoid.social/@sarah

@sarahwiegreffe

wiegreffesarah@gmail.com

Collaborators:

Ai2