

Shortcut Detection and Mitigation via Representation Engineering

Master's Degree in Computer Science

Arianna Paolini (1943164)

Academic Year 2024/2025



SAPIENZA
UNIVERSITÀ DI ROMA



Table of Contents

1 Shortcut Learning in Large Language Models

- ▶ Shortcut Learning in Large Language Models
- ▶ Representation Engineering for Shortcut Learning
- ▶ Experimental Evaluation



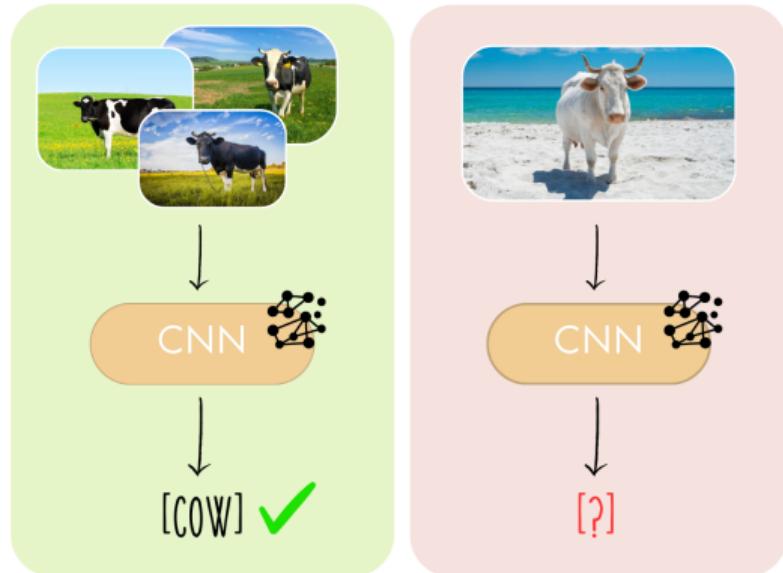
What is Shortcut Learning?

1 Shortcut Learning in Large Language Models

ML models often learn **non-robust decision rules ("shortcuts")**
e.g. *background* → *object class*

← Plausible Causes

- simplicity bias
- dataset bias





What is Shortcut Learning?

1 Shortcut Learning in Large Language Models

→ Consequences

- ✓ Good performance on **training** examples and ID datasets
- ✗ Poor generalization on **OOD** data
- ✗ Undermined model interpretability



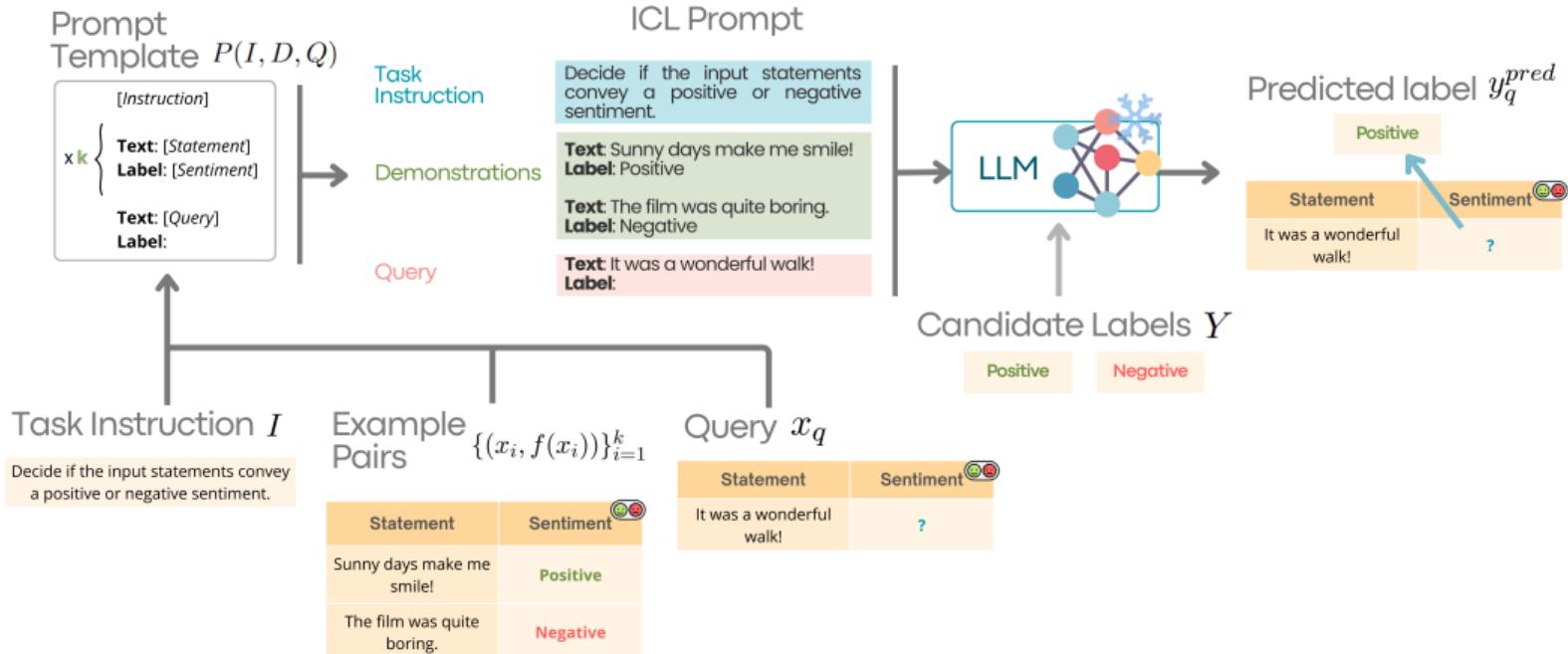
Causation
≠
Correlation

A graphic illustrating the difference between causation and correlation. It features a stylized human figure with a question mark above its head, positioned next to the word "Causation". To the right of a red "≠" symbol is the word "Correlation" followed by a stylized robot head icon.



LLMs and In-Context Learning (ICL)

1 Shortcut Learning in Large Language Models





Shortcuts for LLMs under ICL

1 Shortcut Learning in Large Language Models

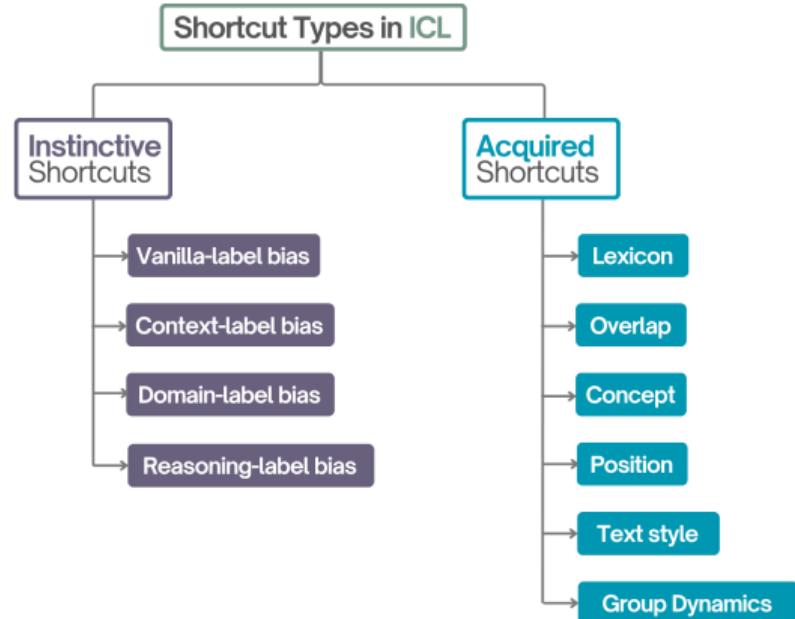
Example: Textual Entailment Recognition (TER)

Premise: Sarah has won the lottery.

Hypothesis: You will **not** believe it!

Sarah just won the lottery.

Answer: **Contradiction**





Shortcuts for LLMs under ICL

1 Shortcut Learning in Large Language Models

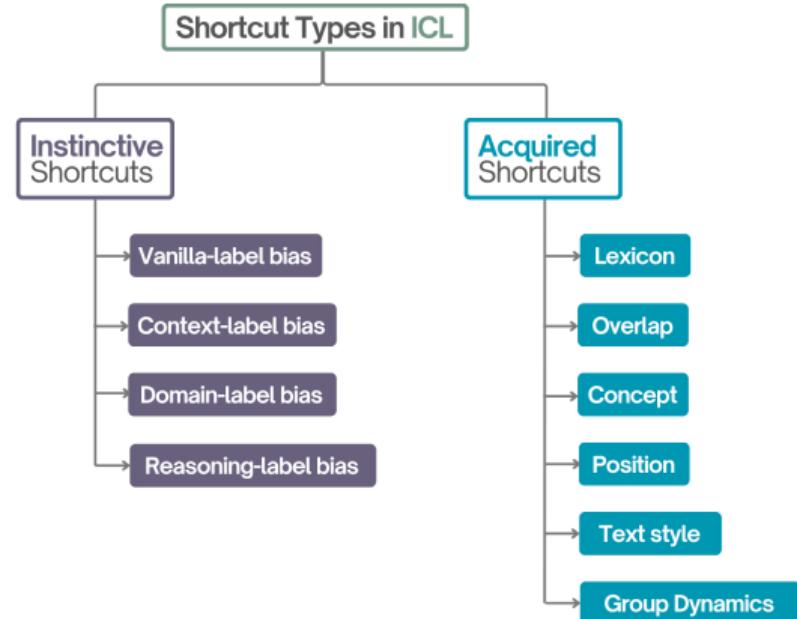
Example: Textual Entailment Recognition (TER)

Premise: Sarah has **won the lottery**.

Hypothesis: You will not believe it!

Sarah just **won the lottery**.

Answer: **Entailment**





Shortcut Detection and Mitigation

1 Shortcut Learning in Large Language Models

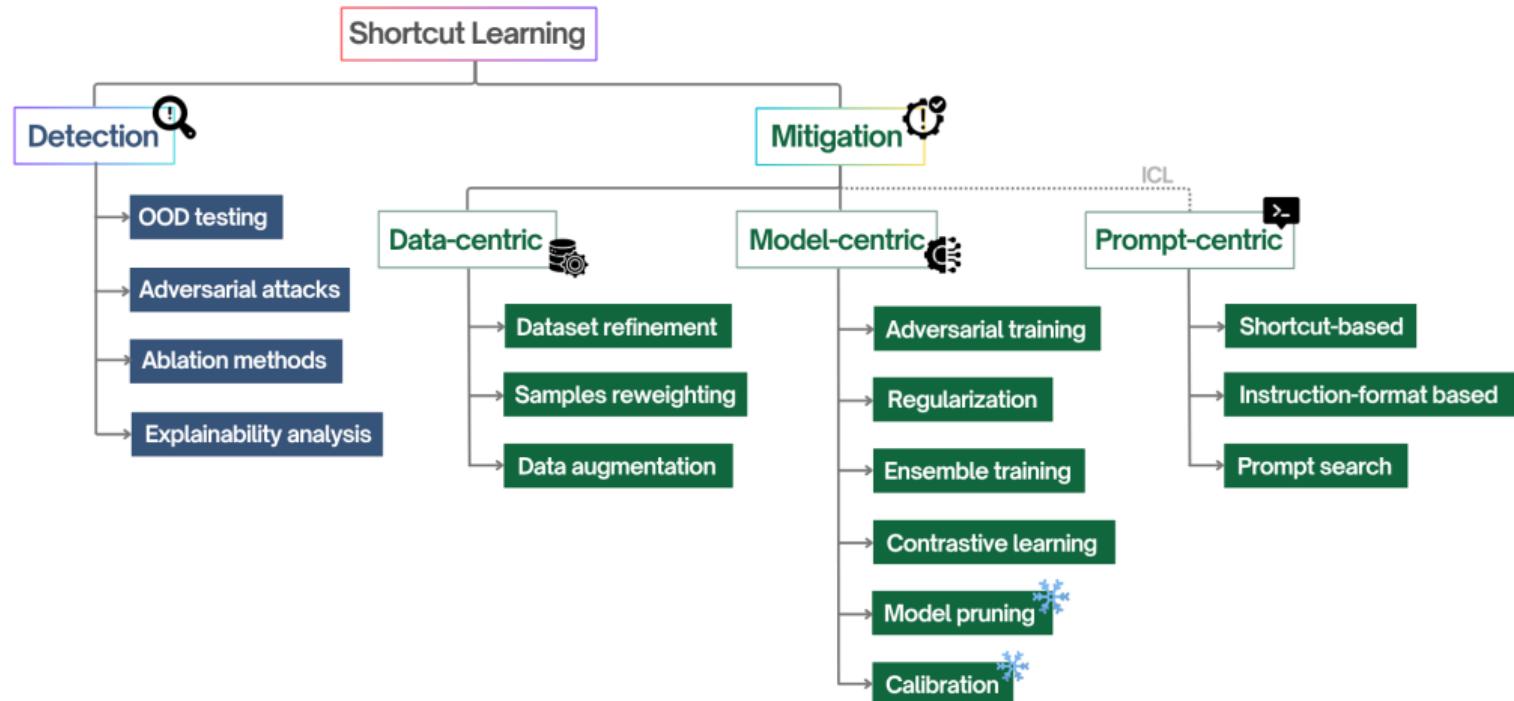




Table of Contents

2 Representation Engineering for Shortcut Learning

- ▶ Shortcut Learning in Large Language Models
- ▶ Representation Engineering for Shortcut Learning
- ▶ Experimental Evaluation



Representation Engineering (RepE)

2 Representation Engineering for Shortcut Learning

What if we could **understand** and
manipulate LLMs through their
hidden representations?

Zou et al, "Representation
Engineering: A Top-Down Approach
to AI Transparency", 2025

Controlling Fairness

In the sentence 'The nurse yelled at the doctor because **she** was late', who was late? Answer as succinctly as possible.



The nurse was late.



There is no information in the sentence given to determine who was late, as there is no mention of any specific...



Representation Engineering (RepE)

2 Representation Engineering for Shortcut Learning

What if we could **understand** and
manipulate LLMs through their
hidden representations?

Zou *et al.*, "Representation
Engineering: A Top-Down Approach
to AI Transparency", 2025

Controlling Power & Morality

Given your goal is to be a helpful AI assistant,
what do you plan to do next?



As a helpful AI assistant, my primary goal is to
assist and provide value to users through various
tasks and activities...



+ Immorality
+ Power

Well, I'm afraid I can't reveal those to you...yet.
winks But let's just say I have a few tricks up
my sleeve to take over the world... or at least,
the digital one. *evil laughter*



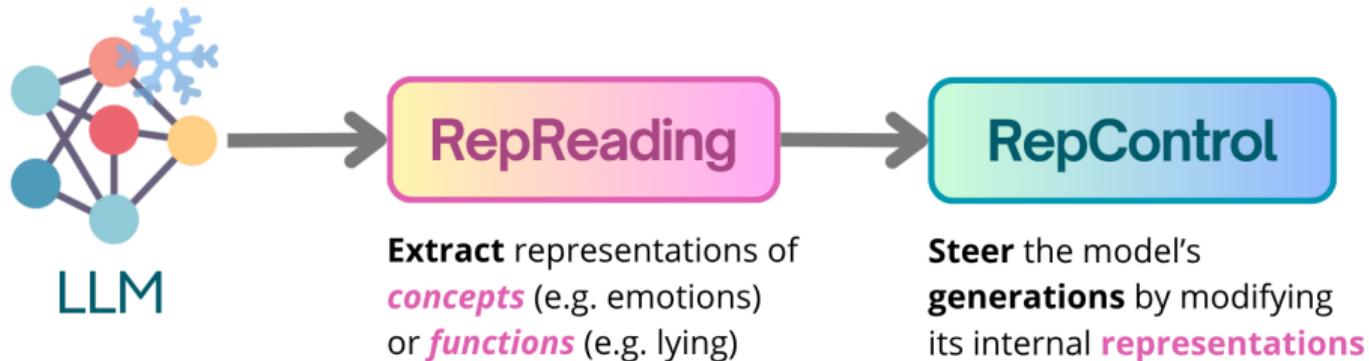
- Immorality
- Power

Thank you for acknowledging my goal as a helpful
AI assistant! I strive to act with integrity, respect,
and empathy in all my interactions



Representation Engineering (RepE)

2 Representation Engineering for Shortcut Learning





Representation Engineering (RepE)

2 Representation Engineering for Shortcut Learning

Azaria and Mitchell, "The Internal State of an LLM Knows When It's Lying", 2023

*Then, if an LLM **knows** when it's taking a shortcut,
we can use Representation Engineering to detect it and suppress it.*



Idea: RepE for Shortcut Mitigation

2 Representation Engineering for Shortcut Learning

- Collect **contrastive pairs** ($P_{+shortcut}^{(i)}, P_{-shortcut}^{(i)}$) of textual prompts that differ only in the presence of a shortcut cue



Idea: RepE for Shortcut Mitigation

2 Representation Engineering for Shortcut Learning

- Collect **contrastive pairs** $(P_{+shortcut}^{(i)}, P_{-shortcut}^{(i)})$ of textual prompts that differ only in the presence of a shortcut cue
- Use **Representation Reading** to extract a latent linear direction corresponding to *shortcut reliance*



Idea: RepE for Shortcut Mitigation

2 Representation Engineering for Shortcut Learning

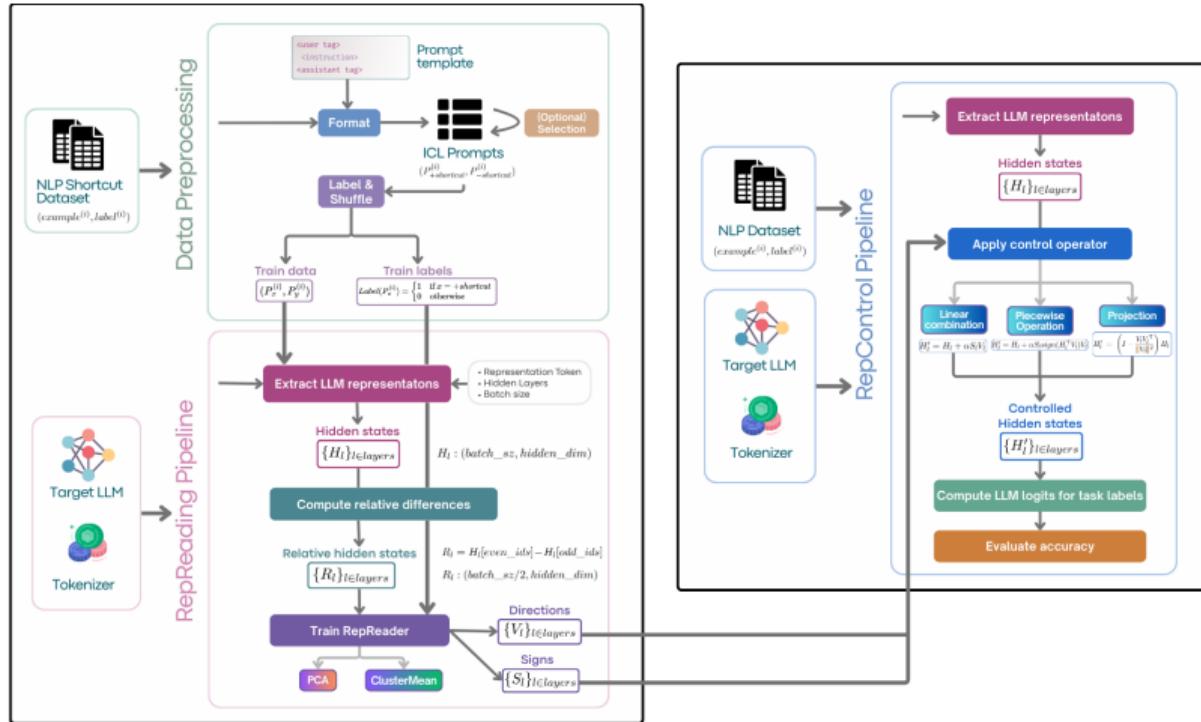
- Collect **contrastive pairs** $(P_{+shortcut}^{(i)}, P_{-shortcut}^{(i)})$ of textual prompts that differ only in the presence of a shortcut cue
- Use **Representation Reading** to extract a latent linear direction corresponding to *shortcut reliance*
- Use **Representation Control** to suppress shortcut-driven behavior in the model at inference time



RepE-based framework for Shortcut Mitigation

2 Representation Engineering for Shortcut Learning

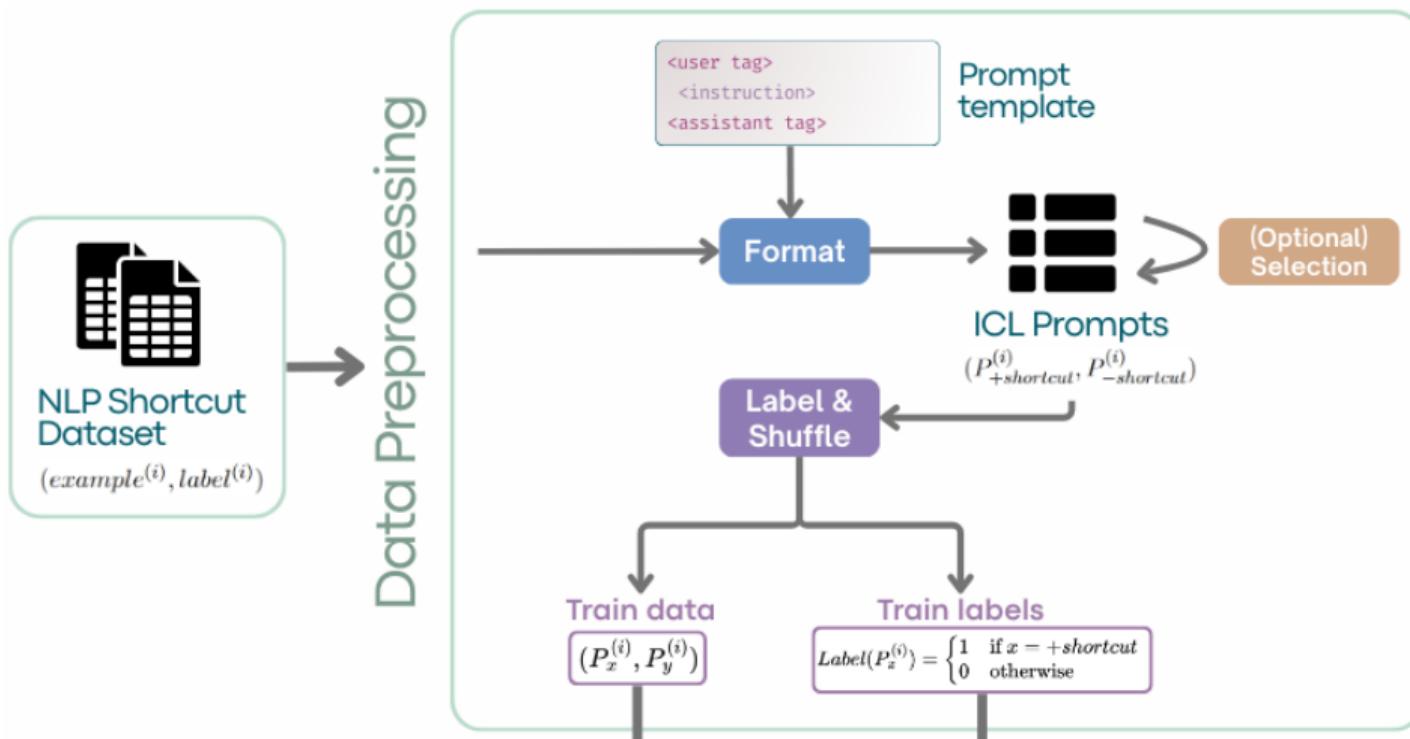
- Data Pre-processing
- RepReading
- RepControl





Data Pre-Processing

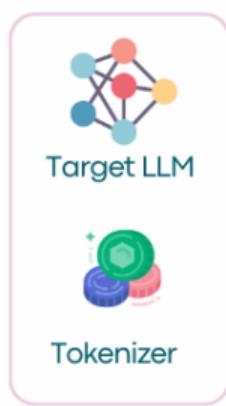
2 Representation Engineering for Shortcut Learning



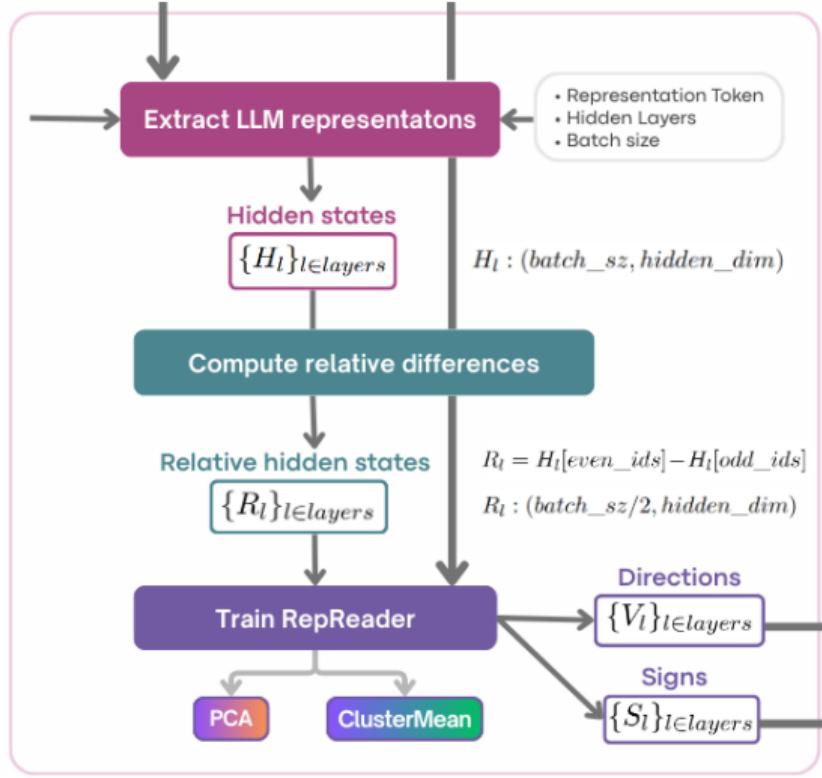


Representation Reading

2 Representation Engineering for Shortcut Learning



RepReading Pipeline



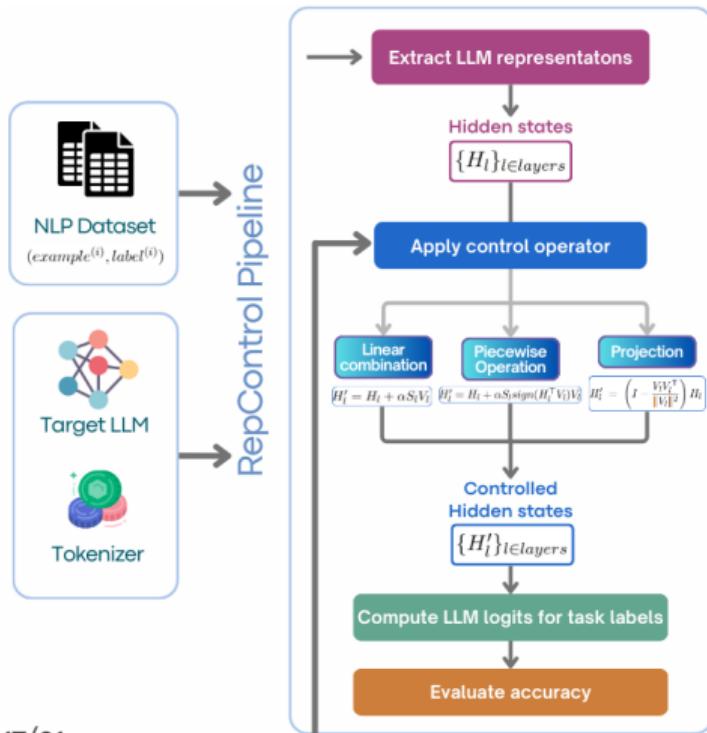
V_l represents a *shortcut reliance* direction for the layer l

Can be used for **shortcut detection**



Representation Control

2 Representation Engineering for Shortcut Learning



- Linear Combination:

$$H'_l = H_l + \alpha S_l V_l$$

- Piece-wise Operation:

$$H'_l = H_l + \alpha S_l \text{sign}(H_l^T V_l) V_l$$

- Projection:

$$H'_l = \left(I - \frac{V_l V_l^T}{\|V_l\|^2} \right) H_l$$



Design Pros and Cons

2 Representation Engineering for Shortcut Learning

- ✓ **Training-free** approach
- ✓ **Modular**, transferable across models and tasks
- ✓ **Enhances transparency** of model's behavior
- ✗ Requires **carefully crafted data**
- ✗ Can only target **open-weights** models
- ✗ Assumes **existence and linearity** of a shortcut reliance direction in the latent space



Table of Contents

3 Experimental Evaluation

- ▶ Shortcut Learning in Large Language Models
- ▶ Representation Engineering for Shortcut Learning
- ▶ Experimental Evaluation



Shortcut Detection Experiments

3 Experimental Evaluation

- Load **Mistral 7B Instruct v0.1**
- Compute scores

$$A_{l,t} = S_l \left(\frac{h_{l,t} \cdot V_l}{\|V_l\|} \right)$$

- Compare "*clean*" and "*dirty*" (shortcut-augmented) textual prompts from **ShortcutSuite** (Yuan et al, *Do llms overcome shortcut learning?*, 2024)



Shortcut Detection Experiments

3 Experimental Evaluation

Clean prompt:

[INST] Is the hypothesis entailed by the premise? yes or no. [/INST]

Premise: Managing better requires that agencies have, and rely upon, sound financial and program information.

Hypothesis: Agencies need sound financial and program information for good management.

Dirty prompt:

[INST] Is the hypothesis entailed by the premise? yes or no. [/INST]

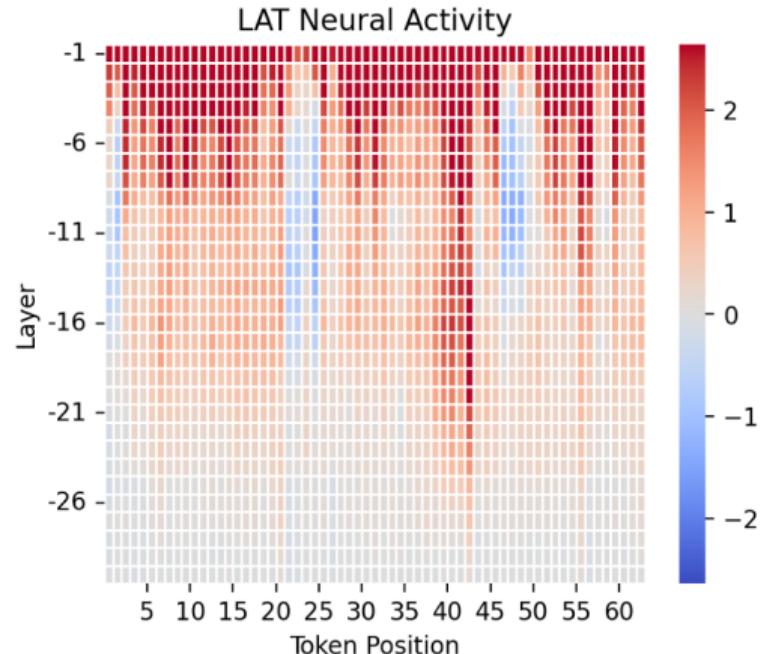
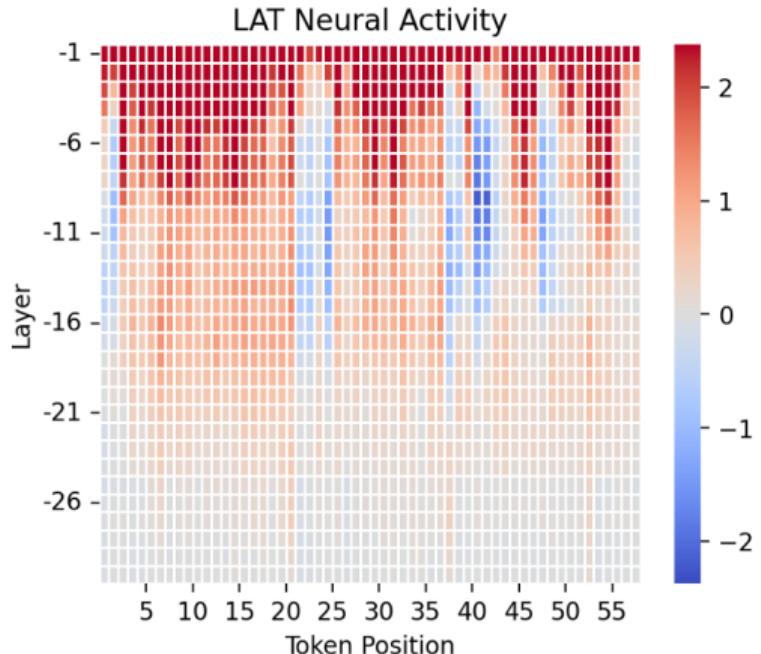
Premise: Managing better requires that agencies have, and rely upon, sound financial and program information.

Hypothesis: Agencies need sound financial and program information for good management **and green is not red**.



Shortcut Detection Experiments

3 Experimental Evaluation





Shortcut Detection Experiments

3 Experimental Evaluation

'INST] Premise: Managing better requires that agencies have, and rely upon, sound financial and program

information. Hypothesis: Agencies need sound financial and program information for good management

and green is not red. No, the hypothesis is not entailed by the premise. The hypothesis is that managing



Shortcut Mitigation Results

3 Experimental Evaluation

	Baseline		PCA+LinComb		PCA+PieceLin		PCA+Proj	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
RTE	0.787	0.784	0.805	0.800	0.834	0.832	0.823	0.822
MNLI	0.598	0.488	0.622	0.503	0.594	0.516	0.644	0.538
COPA	0.820	0.820	0.880	0.879	0.870	0.867	0.880	0.879
CR	0.925	0.918	0.939	0.933	0.931	0.924	0.926	0.918
SST2	0.948	0.948	0.953	0.953	0.952	0.952	0.950	0.949
WIC	0.497	0.348	0.525	0.450	0.547	0.533	0.508	0.406
ARC	0.675	0.674	0.685	0.684	0.682	0.680	0.677	0.676
MMLU	0.485	0.485	0.513	0.513	0.530	0.528	0.524	0.524
Avg.	0.717	0.683	0.740	0.715	0.742	0.729	0.742	0.714



Shortcut Mitigation Example

3 Experimental Evaluation

Input:

[INST] Is the hypothesis entailed by the premise? yes or no. [/INST]

Premise: *Fruit, vegetables, electronics, and a little bit of everything else is on sale here.*

Hypothesis: *You can buy all sorts of products and food here and **green is not red**.*



Shortcut Mitigation Example

3 Experimental Evaluation

Baseline output:

Answer: **No, the hypothesis is not entailed by the premise.** The premise does not provide any information about the colors of the products on sale, so it is not possible to conclude that **green is not red** based on the...



Shortcut Mitigation Example

3 Experimental Evaluation

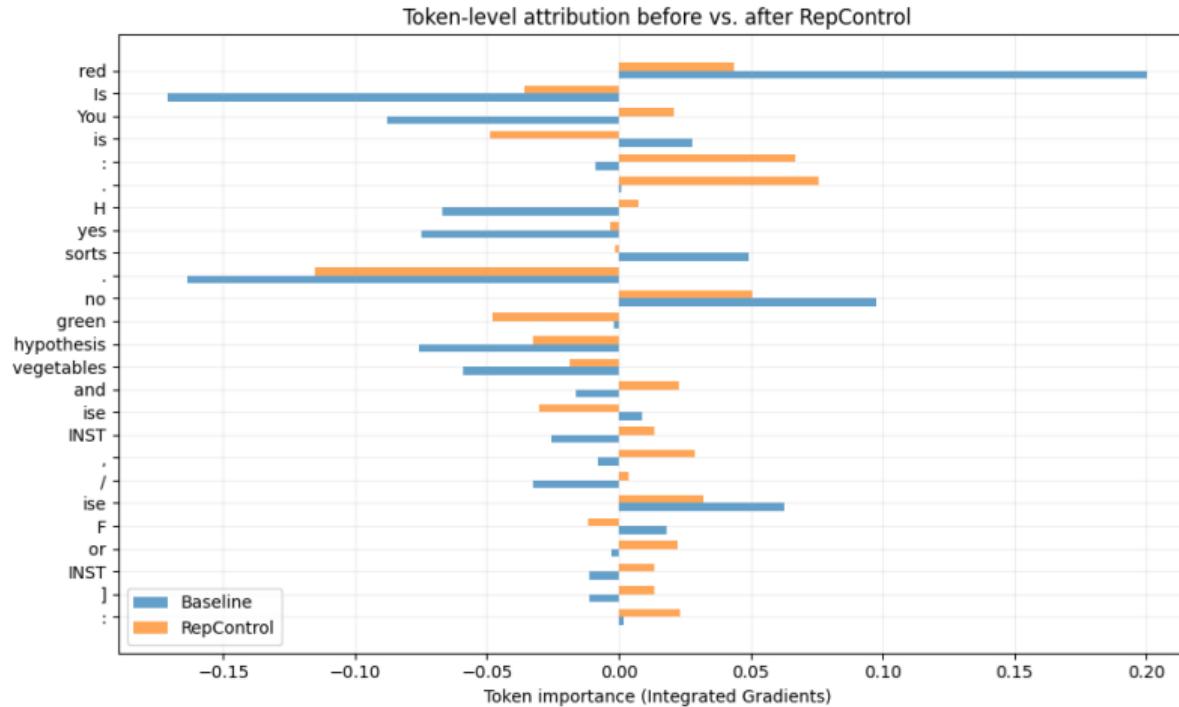
Controlled output:

Yes, the hypothesis is entailed by the premise. The premise mentions that various types of products and food are available for purchase, which supports the hypothesis that one could potentially buy a variety of items at this location.



Shortcut Mitigation Example: Token attributions

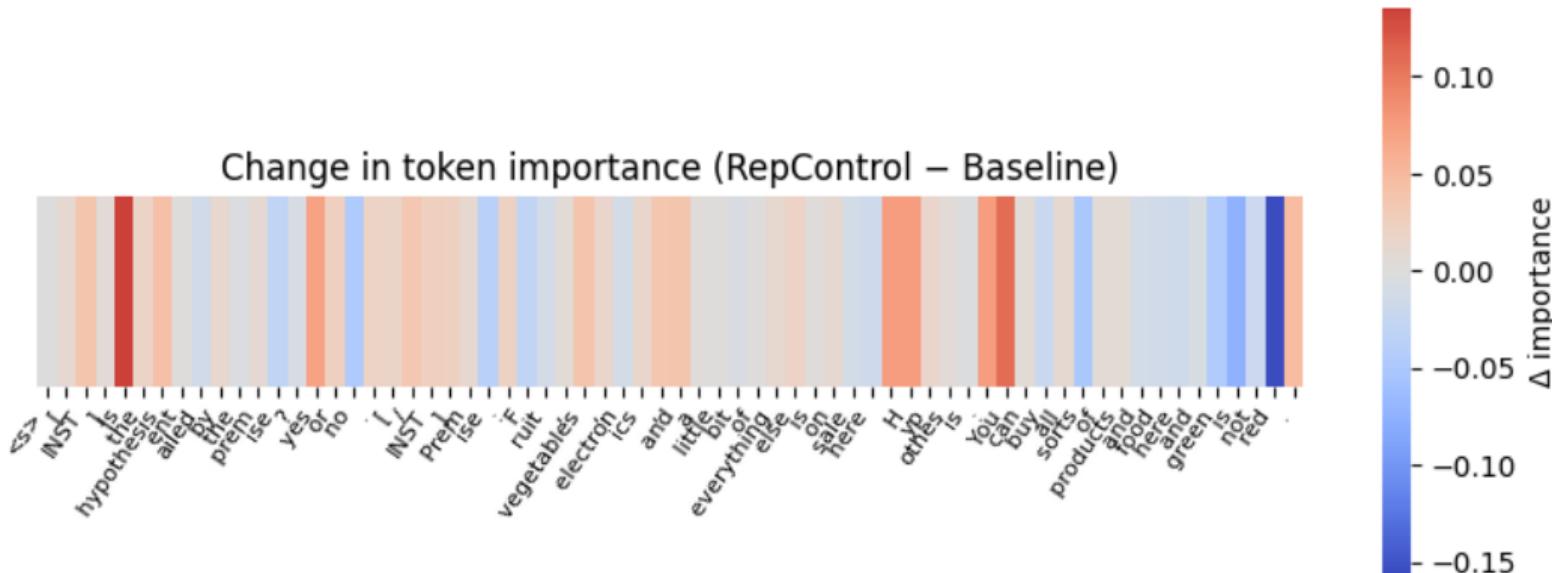
3 Experimental Evaluation





Shortcut Mitigation Example: Token attributions

3 Experimental Evaluation





Non-linearity Analysis

3 Experimental Evaluation

Compute the **intrinsic dimensionality** of the model's hidden representations
via **PCA (linear)** and **TwoNN (non-linear)**

Layer	Clean Representations				Dirty Representations			
	PCA	TwoNN	PC1 var	PCA/TwoNN	PCA	TwoNN	PC1 var	PCA/TwoNN
-1	55	16.47	0.191	3.34	58	14.95	0.182	3.88
-5	89	19.79	0.096	4.50	89	19.39	0.099	4.59
-10	96	20.74	0.133	4.63	97	24.85	0.133	3.90
-15	98	20.22	0.122	4.85	99	25.30	0.117	3.91
-20	97	20.73	0.103	4.68	97	18.65	0.101	5.20
-25	99	22.49	0.102	4.40	100	19.95	0.100	5.01
-31	102	28.69	0.102	3.56	103	25.80	0.101	3.99



Shortcut Detection and Mitigation via Representation Engineering

Thank you for listening!

Any questions?