The Causal Approach to Mechanistic Interpretability

Presented by: Jet Li
Slides adapted from: Atticus Geiger, Callum McDougall, Jack Merullo

Mechanistic?

Saphra and Wiegreffe (2024)

Narrow Cultural Definition

AI safety researchers motivated by philosophical arguments for interpretability.

Broad Cultural Definition

AI researchers interested in model internals.

Broad Technical Definition

Any research that describes the internals of a model.

Narrow Technical Definition

Understanding neural networks through their causal mechanisms.

Mechanistic?

Saphra and Wiegreffe (2024)

Narrow Cultural Definition

AI safety researchers motivated by philosophical arguments for interpretability.

Broad Cultural Definition

AI researchers interested in model internals.

Broad Technical Definition

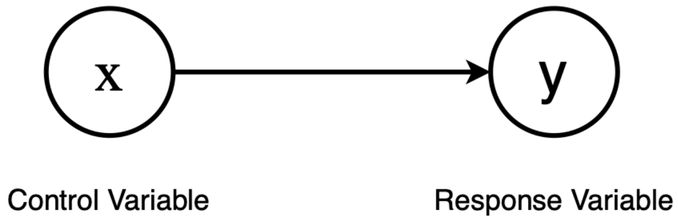
Any research that describes the internals of a model.

Narrow Technical Definition

Understanding neural networks through their causal mechanisms.

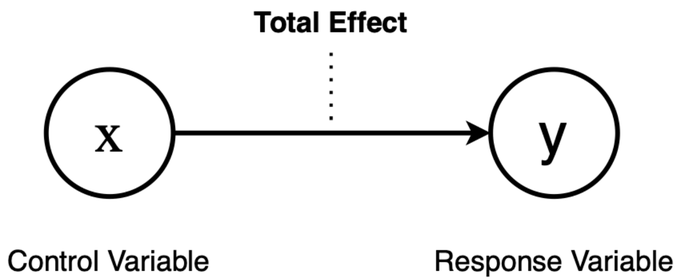
Causal Mediation Analysis

Robins and Greenland (1992); Pearl (2001)



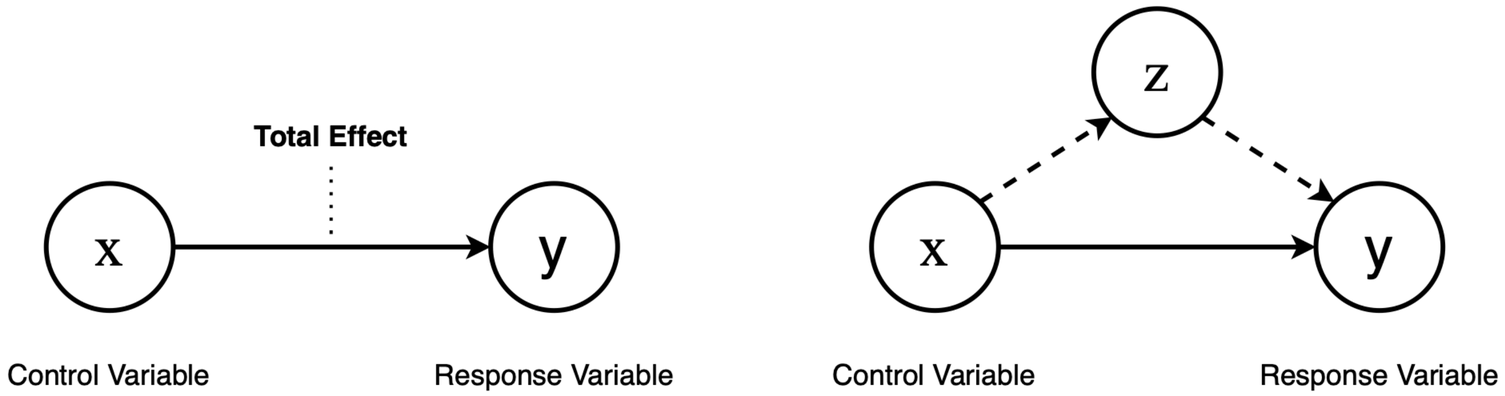
Causal Mediation Analysis

Robins and Greenland (1992); Pearl (2001)



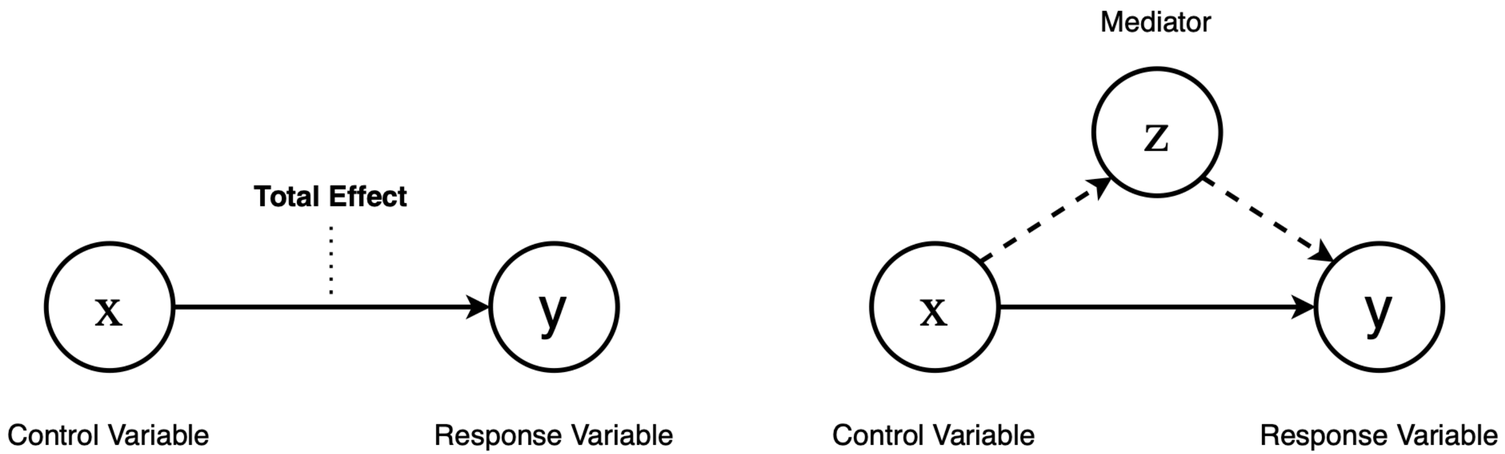
Causal Mediation Analysis

Robins and Greenland (1992); Pearl (2001)



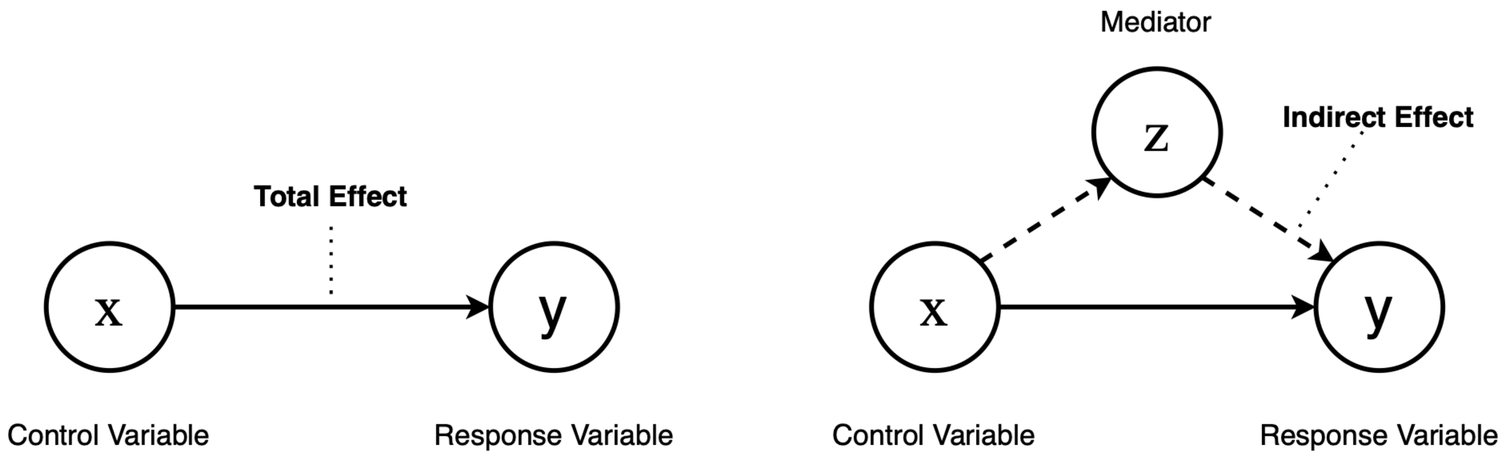
Causal Mediation Analysis

Robins and Greenland (1992); Pearl (2001)



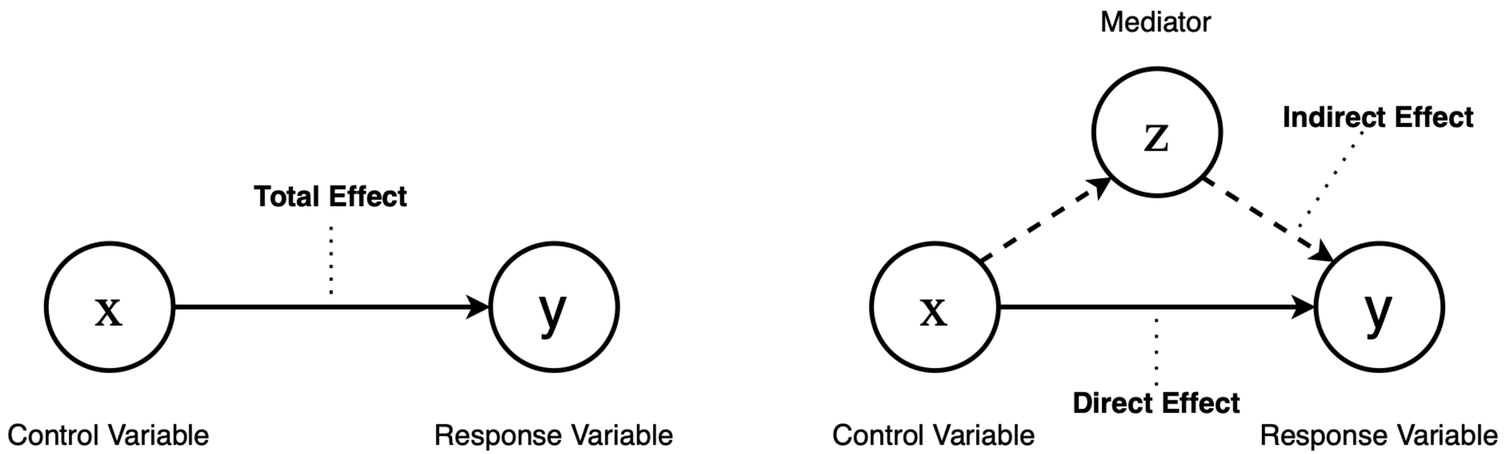
Causal Mediation Analysis

Robins and Greenland (1992); Pearl (2001)



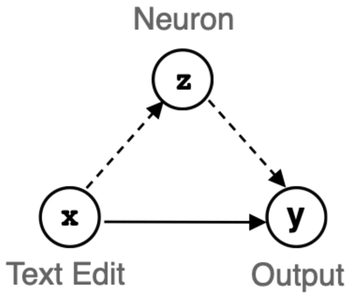
Causal Mediation Analysis

Robins and Greenland (1992); Pearl (2001)



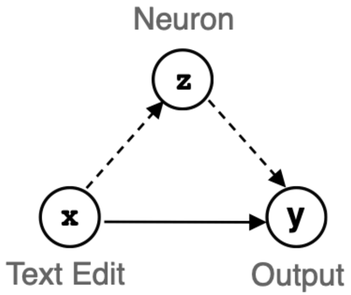
Causal Mediation via Activation Patching

Visuals from Vig et al. (2020)

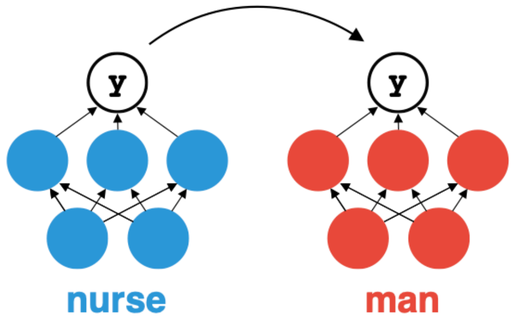


Causal Mediation via Activation Patching

Visuals from Vig et al. (2020)

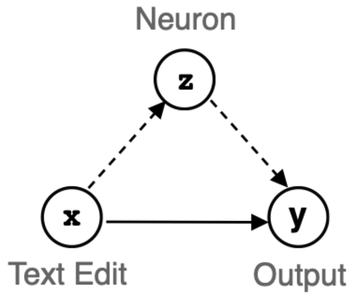


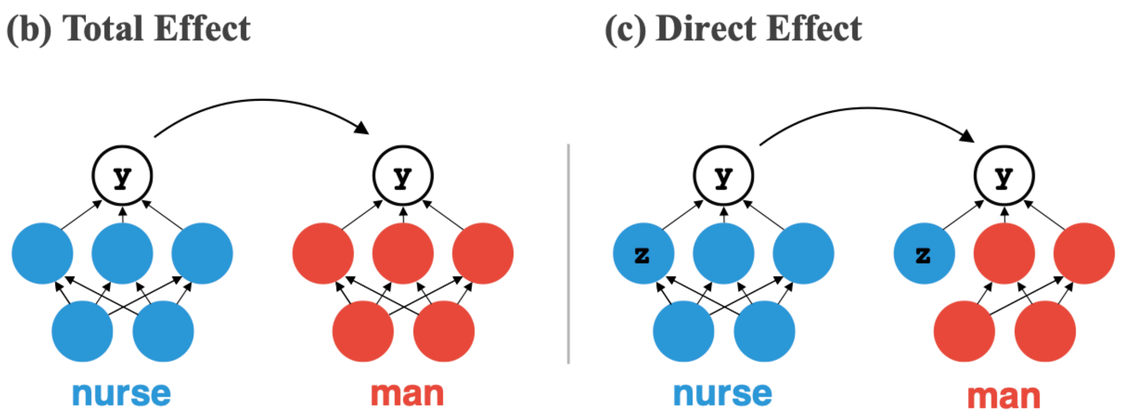
(b) Total Effect



Causal Mediation via Activation Patching

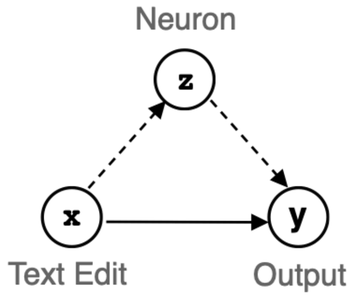
Visuals from Vig et al. (2020)

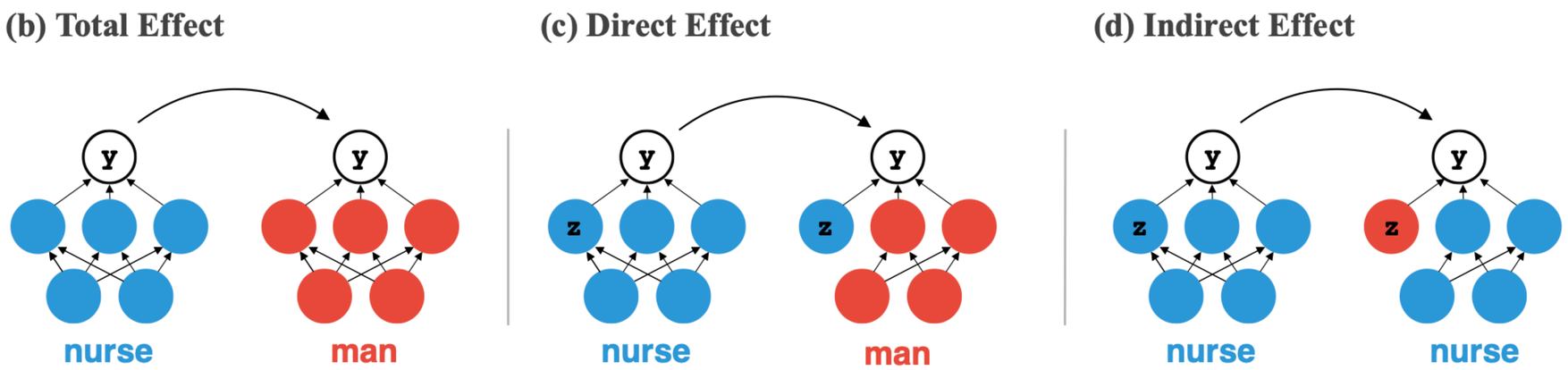




Causal Mediation via Activation Patching

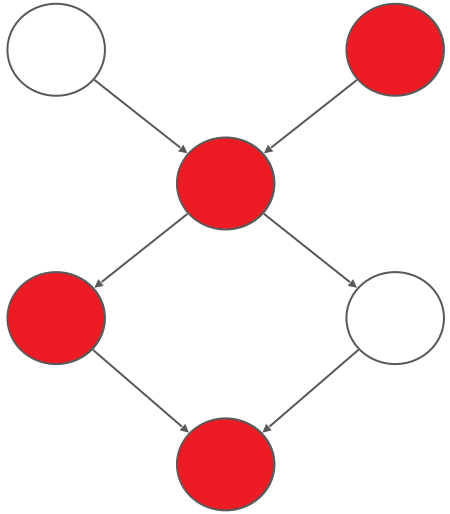
Visuals from Vig et al. (2020)





Circuit Analysis

- A language model (LM) is a graph
- A circuit is the path that a signal takes through that graph to compute the answer to some task
- Circuit analysis lets us causally reverse engineer how an LM makes predictions



Causal Tracing (Mediation with Ablations)

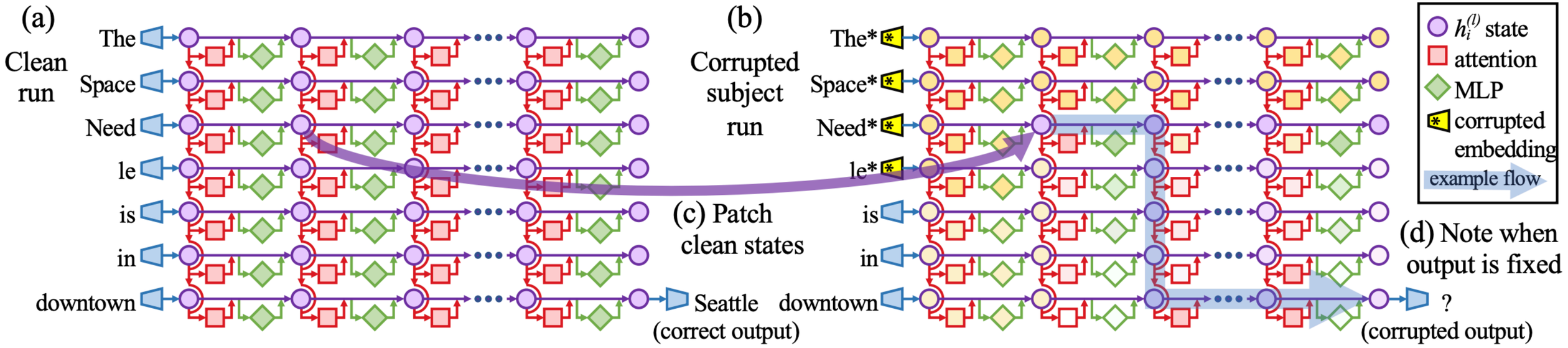
Visuals from Meng et al. (2022)

Intervention: Add Gaussian noise to the Input Embeddings

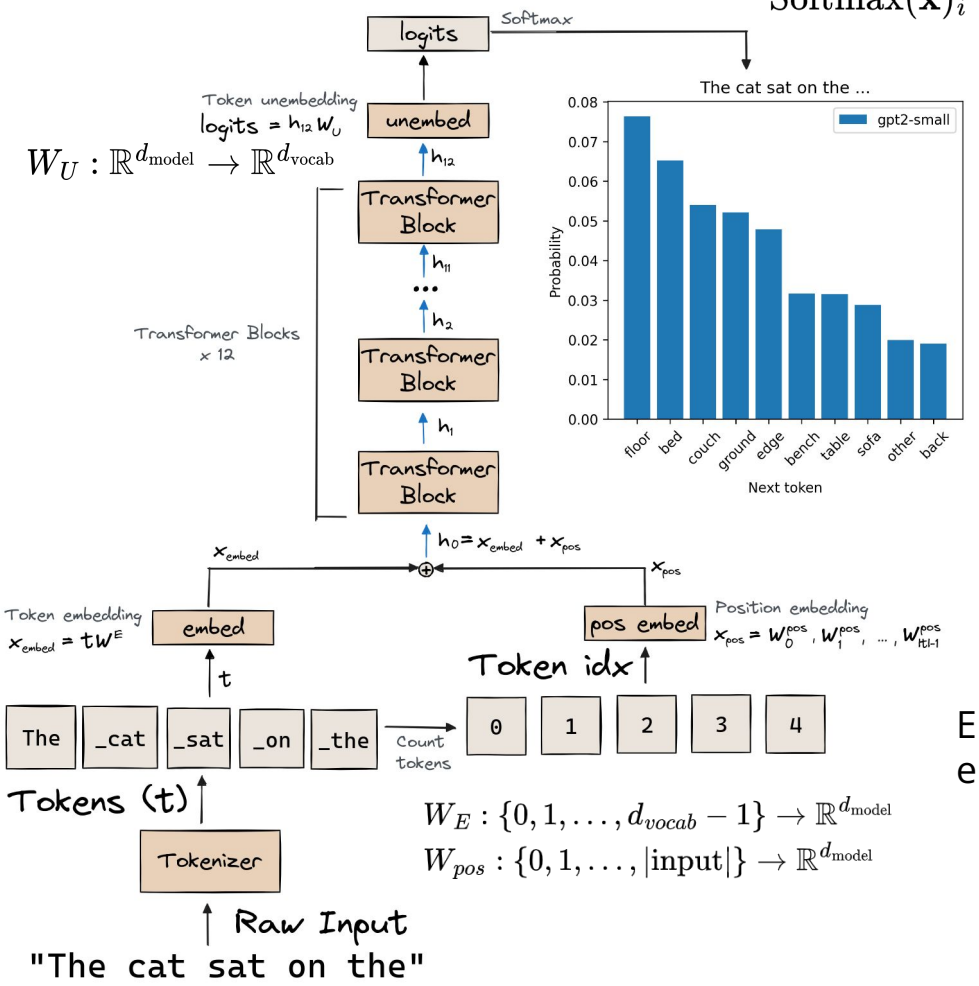
Causal Tracing (Mediation with Ablations)

Visuals from Meng et al. (2022)

Intervention: Add Gaussian noise to the Input Embeddings



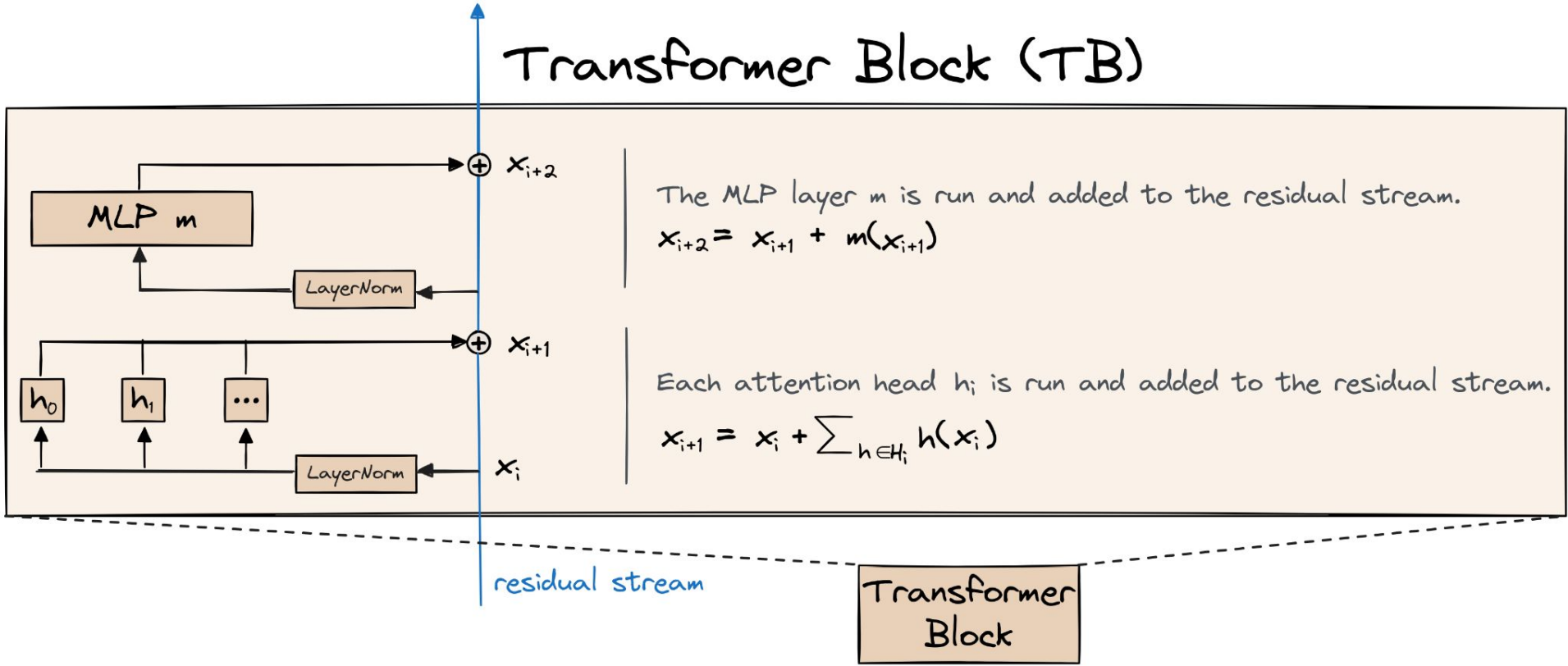
$$\text{Softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$



Embed = token embed + pos embed

- Both are *learned* lookup tables

Transformer Block (TB)



The MLP layer m is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

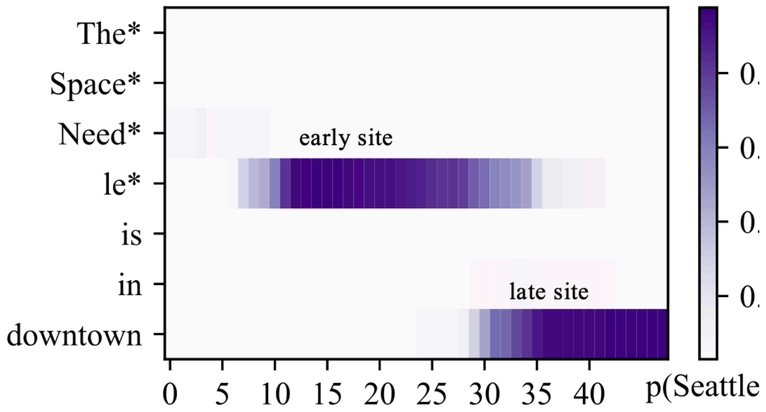
Each attention head h_i is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

Causal Tracing (Mediation with Ablations)

Visuals from [Meng et al. (2022)]

Intervention: Add Gaussian noise to the Input Embeddings

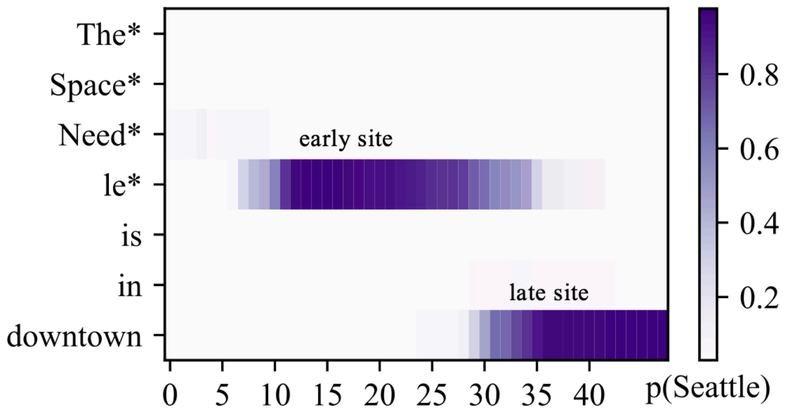


Residual Stream Indirect Effects
(10 layers patched)

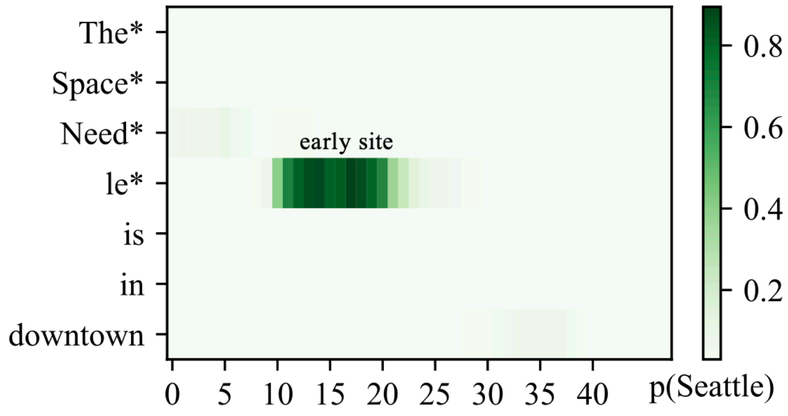
Causal Tracing (Mediation with Ablations)

Visuals from [Meng et al. (2022)]

Intervention: Add Gaussian noise to the Input Embeddings



Residual Stream Indirect Effects
(10 layers patched)

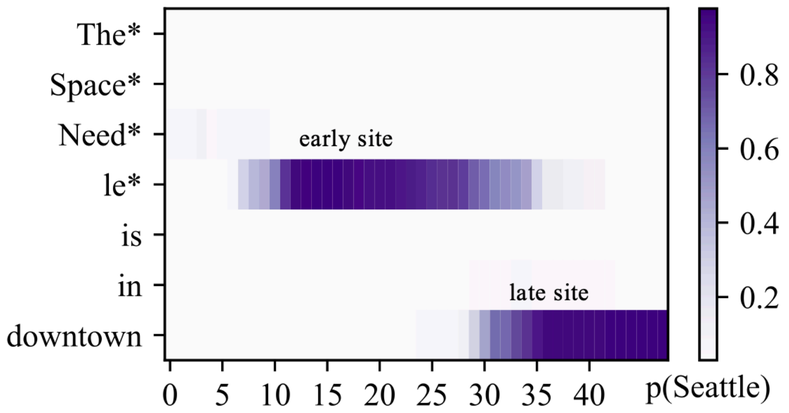


MLP Indirect Effects
(10 layers patched)

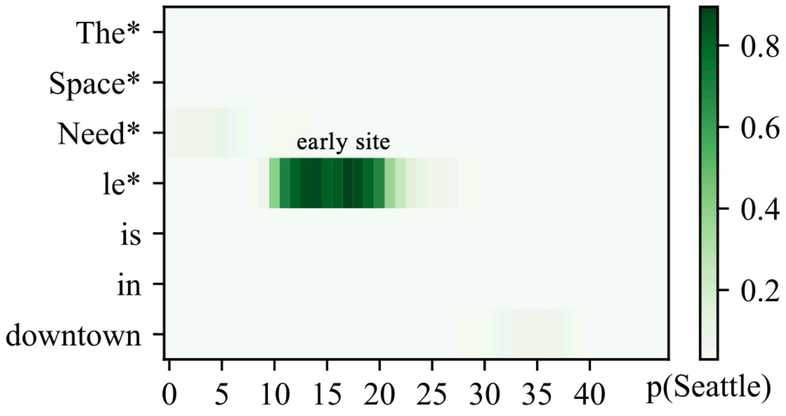
Causal Tracing (Mediation with Ablations)

Visuals from [Meng et al. (2022)]

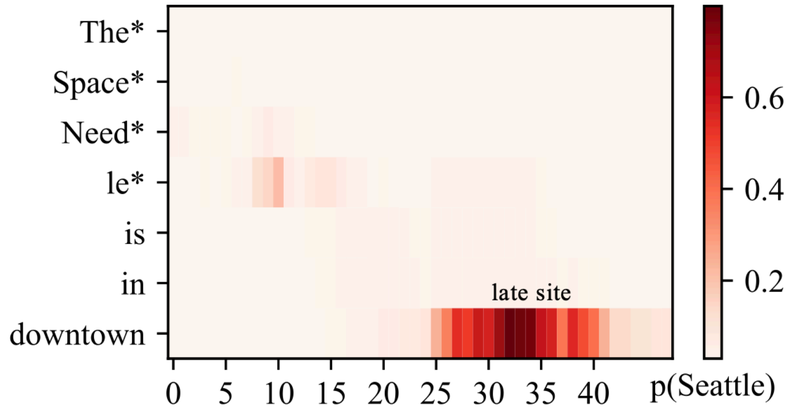
Intervention: Add Gaussian noise to the Input Embeddings



Residual Stream Indirect Effects
(10 layers patched)



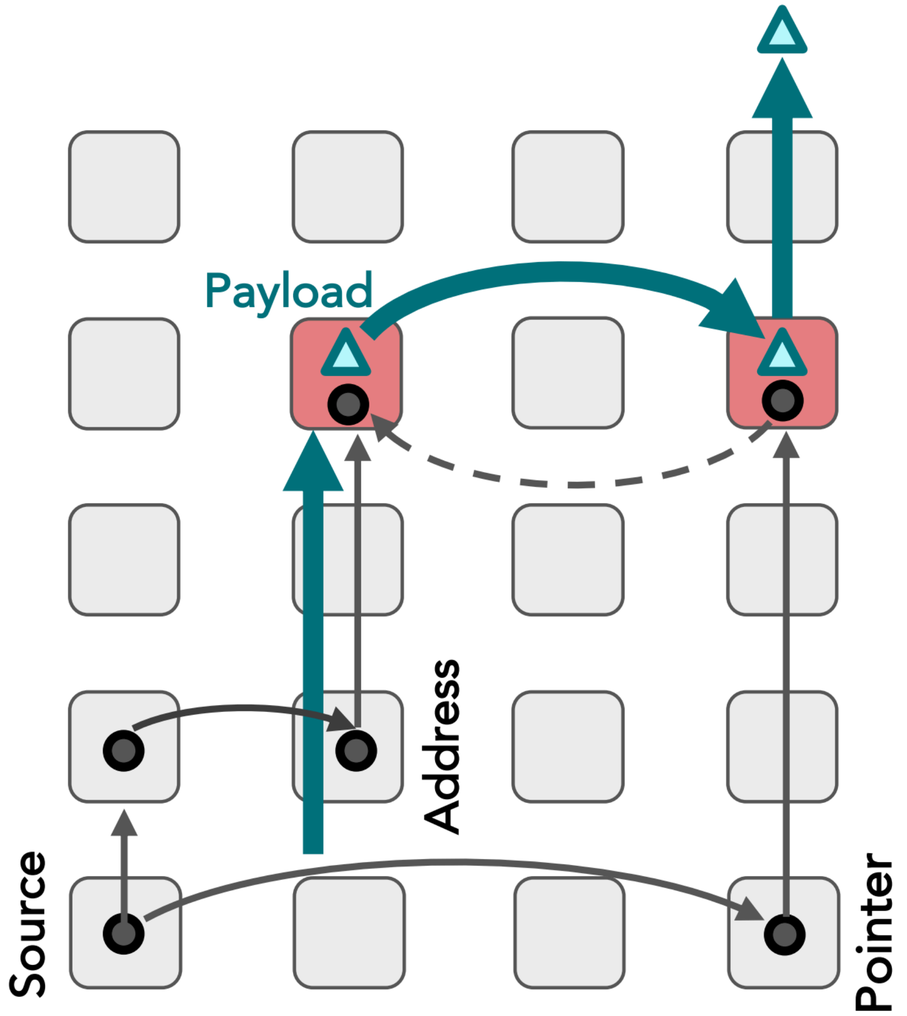
MLP Indirect Effects
(10 layers patched)



Attention Heads Indirect Effects
(10 layers patched)

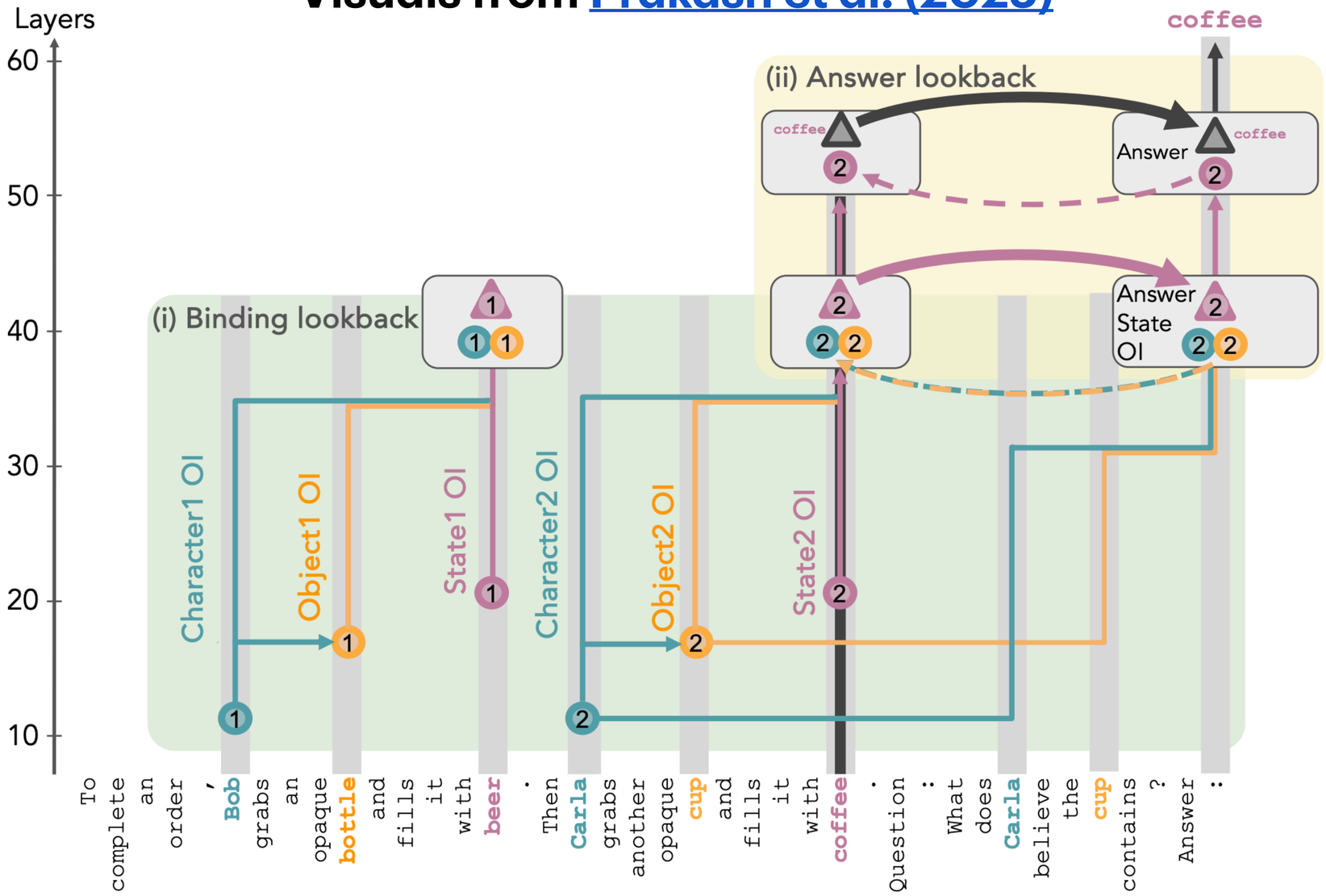
Lookback Mechanisms

Visuals from [Prakash et al. (2025)]

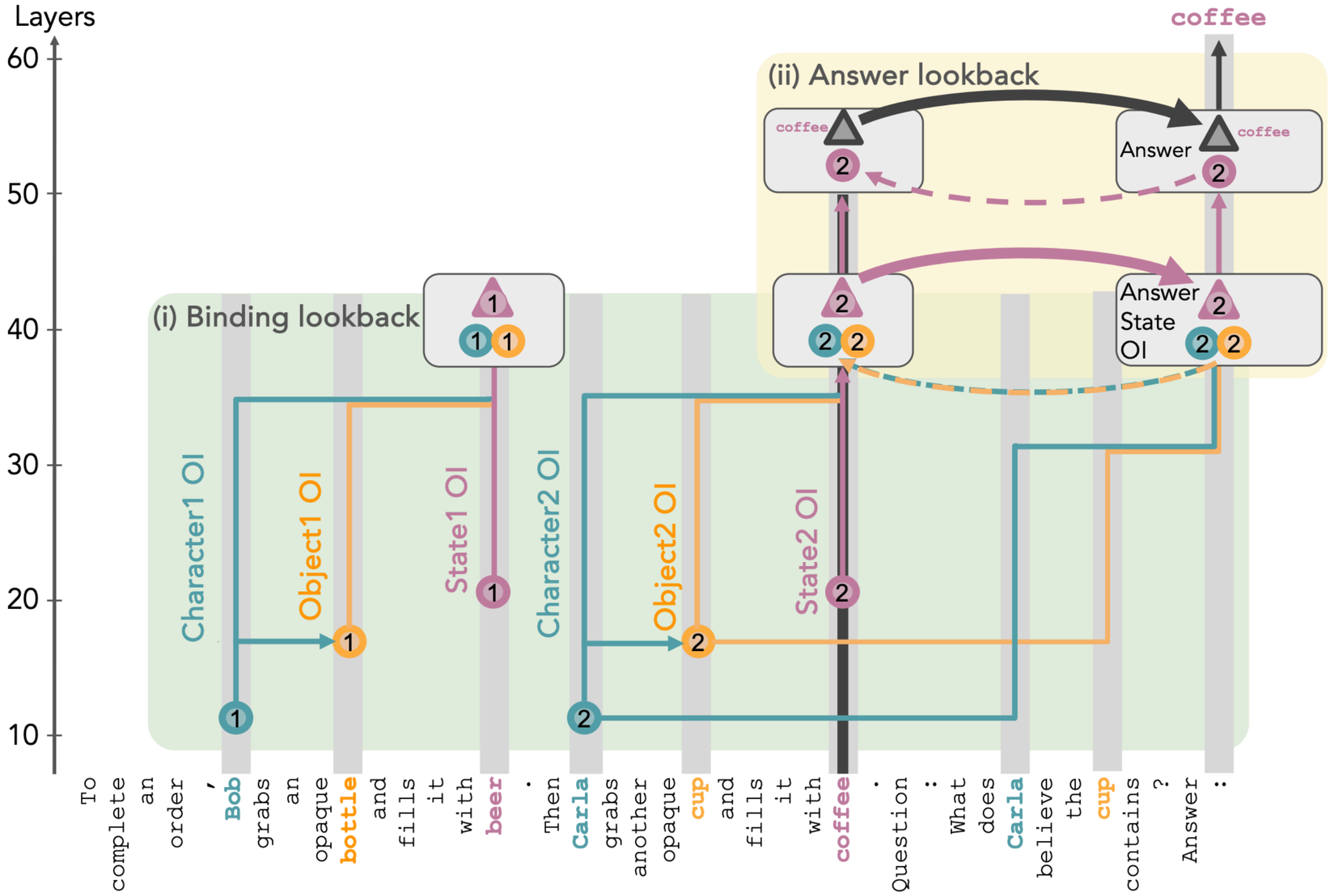


Lookback Mechanisms

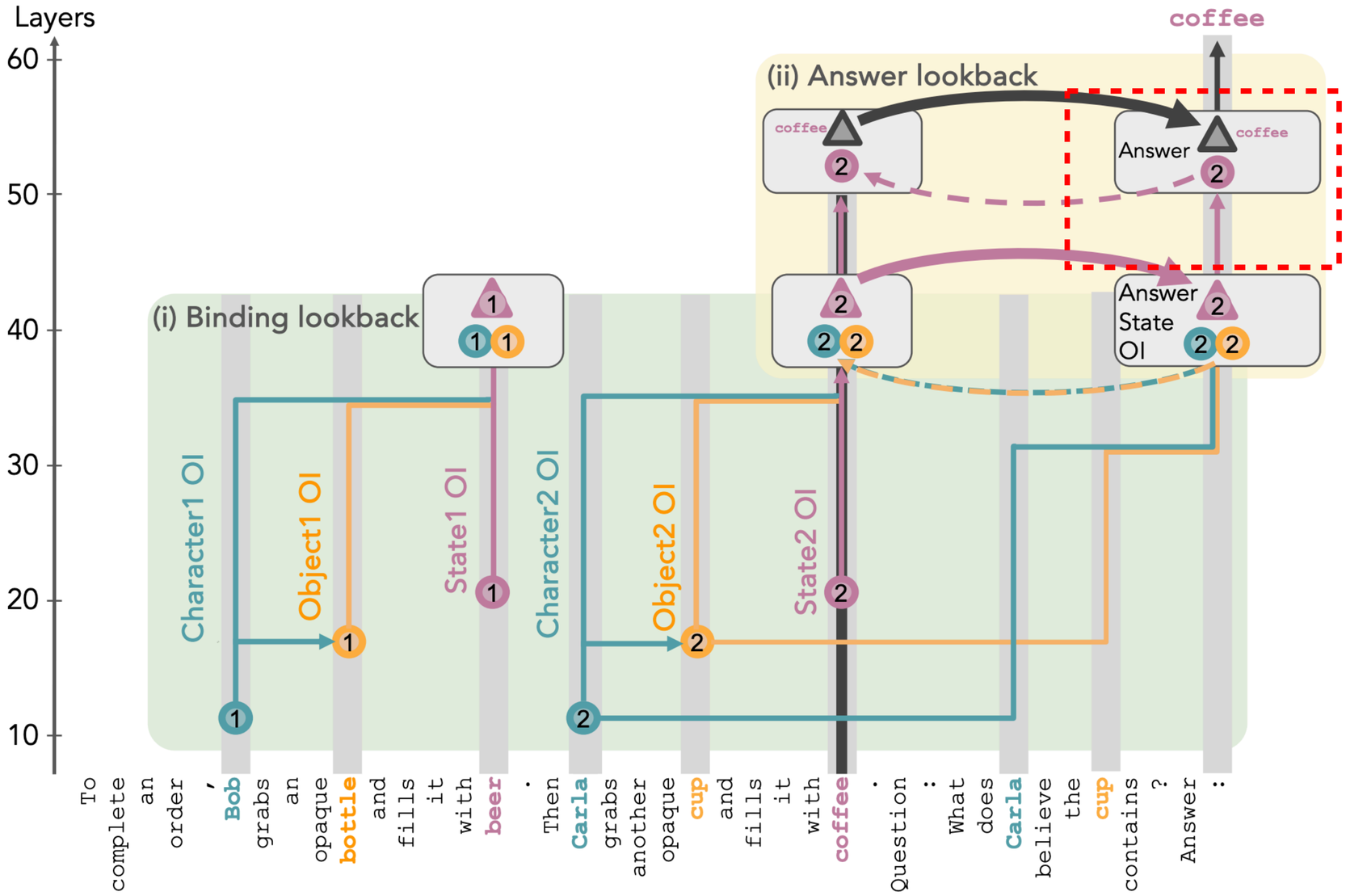
Visuals from [Prakash et al. (2025)]



Designing Counterfactuals



Designing Counterfactuals



Designing Counterfactuals

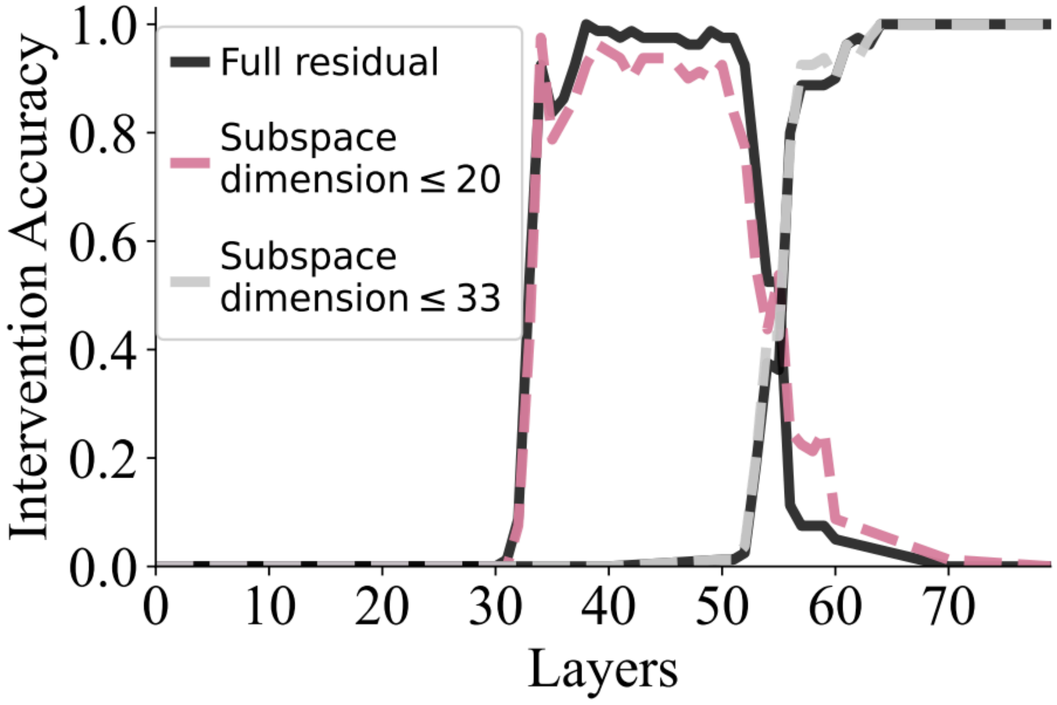
Counterfactual

Carla and Bob are working in a busy restaurant. To complete an order, Carla grabs an opaque cup and fills it with tea. Then Bob grabs another opaque bottle and fills it with water. Question: What does Carla believe the cup contains?
Answer: tea

Original

Bob and Carla are working in a busy restaurant. To complete an order, Bob grabs an opaque bottle and fills it with beer. Then Carla grabs another opaque cup and fills it with coffee. Question: What does Carla believe the cup contains?
Answer: coffee

Intervention 1: Answer Pointer (●), ***Causal Model Output:*** beer
Intervention 2: Answer Payload (▲), ***Causal Model Output:*** tea



Key Takeaways

- Causal mediation and abstraction theoretically ground mechanistic interpretability
- Supervised mechanistic interpretability is powerful and effective
- Benchmarking and evaluations are important
- Designing counterfactuals is essential for uncovering complex algorithms

Pointers to the Literature

- Papers explicitly using the framework of mediation:
  Vig et al. 2020, Finalyson et al. 2021, Stolfo et al. 2023, Meng et al. 2022, 2023, Prakash et al. 2024, Nikankin et al. 2025
- Papers using neuron clamping or concept erasure methods:
  Li et al. 2016, Ravfogel et al. 2020, 2022, Elazar et al. 2021, Belrose et al. 2023, Geva et al. 2023
- Circuits papers:
  Cammarata et al. 2020, Elhage et al. 2021, Olsson et al. 2022, Wang et al. 2023, Conmy et al. 2023, Hanna et al. 2023, Nanda et al. 2023
- A position piece and survey of the field through the lens of mediation:
  Mueller et al. 2024