The Emerging Causal Paradigm in Mechanistic Interpretability

Atticus Geiger

GOODFIRE

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.

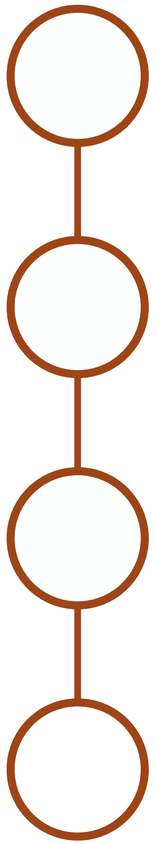
Broad Technical Definition

Any research that describes the internals of a model.

Narrow Technical Definition

Understanding neural networks through their causal mechanisms.

Outline



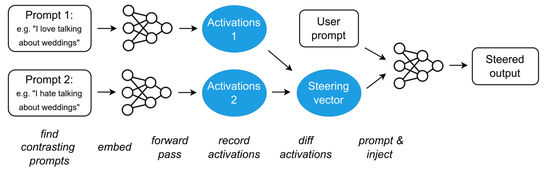
Causal Mediation

Causal Abstraction

Causal Abstraction

Designing Counterfactuals

Outline



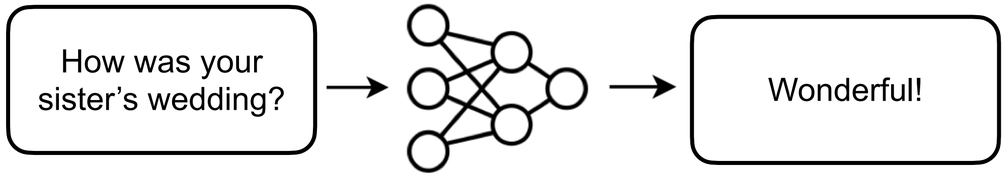
- Activation Steering
- Causal Mediation
- Causal Abstraction
- Designing Counterfactuals

Golden Gate Claude

May 23, 2024 • 3 min read

Difference-in-means Steering

Visuals from Turner et al. (2023)



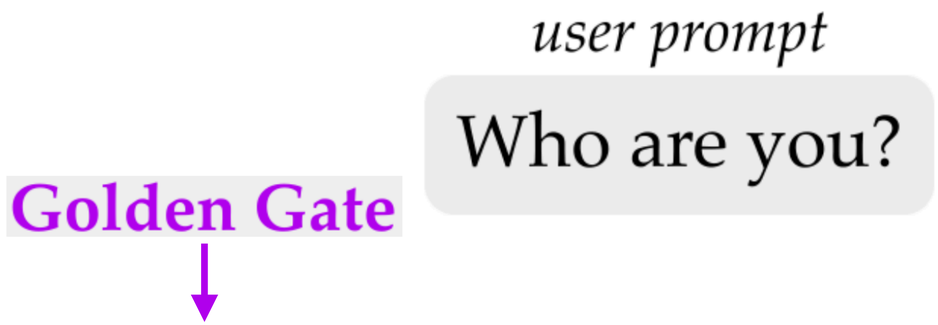
AxBench

Evaluating Methods of Model Control [(Wu et al. 2025)]

Input: Prompt and Steering Prompt

Output: Updated Model

Evaluation: Judge generated text



Baseline AxBench Results

Method	Gemma-2-2B		Gemma-2-9B		Avg.
	L10	L20	L20	L31	
Prompt	0.698	**0.731**	**1.075**	**1.072**	**0.894**
LoReFT	**0.701**	0.722	0.777	0.764	0.741
SFT	0.637	0.714	—	—	*0.676*
LoRA	0.637	0.641	0.602	0.580	0.615
ReFT-r1	0.633	0.509	0.630	0.401	0.543
DiffMean	0.297	0.178	0.322	0.158	0.239
SAE	0.177	0.151	0.191	0.140	0.165
SAE-A	0.166	0.132	0.186	0.143	0.157
LAT	0.117	0.130	0.127	0.134	0.127
PCA	0.107	0.083	0.128	0.104	0.105
Probe	0.095	0.091	0.108	0.099	0.098
SSV	0.072	0.001	0.024	0.008	0.026

Table 2: S Mean overall steering scores for each method, after steering factor selection. Gray indicates non-representation steering methods.

Evaluation by LM judges of

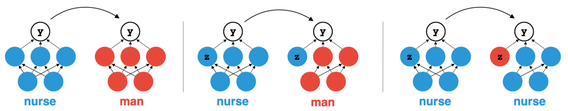
Steered output

Table 3: S Winrate against SAEs for each method, after steering factor selection.

Pointers to the Literature

- Early papers controlling neural networks via neuron manipulation:
  Giulianelli et al. 2018, Bau et al. 2019, Bau et al. 2019, Besserve et al. 2019
- More recent papers on steering:
  Subramani et al. 2022, Marks et al. 2023, Turner et al. 2024, Panickssery et al. 2024, Li et al. 2024
- AxBench:
  Wu et al. 2025
- Golden Gate Claude

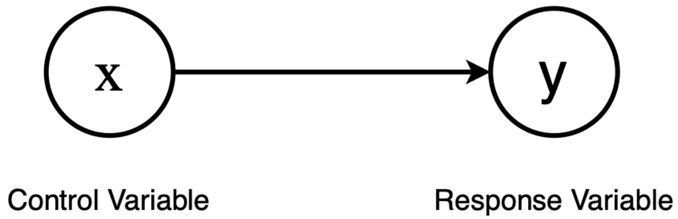
Outline



- Controlling Models with Activation Steering
- Causal Mediation Analysis
- Causal Abstraction Analysis
- Designing Counterfactuals for Reverse Engineering

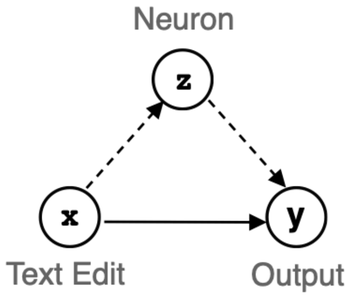
Causal Mediation Analysis

Robins and Greenland (1992); Pearl (2001)



Causal Mediation via Activation Patching

Visuals from Vig et al. (2020)



Causal Tracing (Mediation with Ablations)

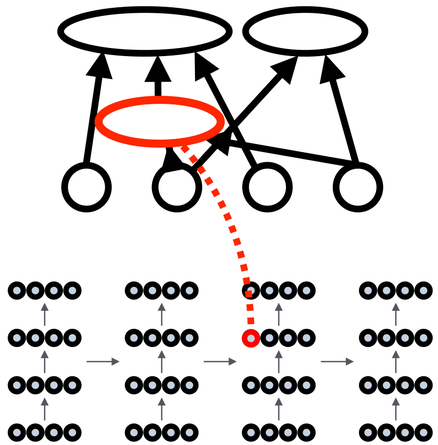
Visuals from Meng et al. (2022)

Intervention: Add Gaussian noise to the Input Embeddings

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

Outline



Activation Steering

Causal Mediation

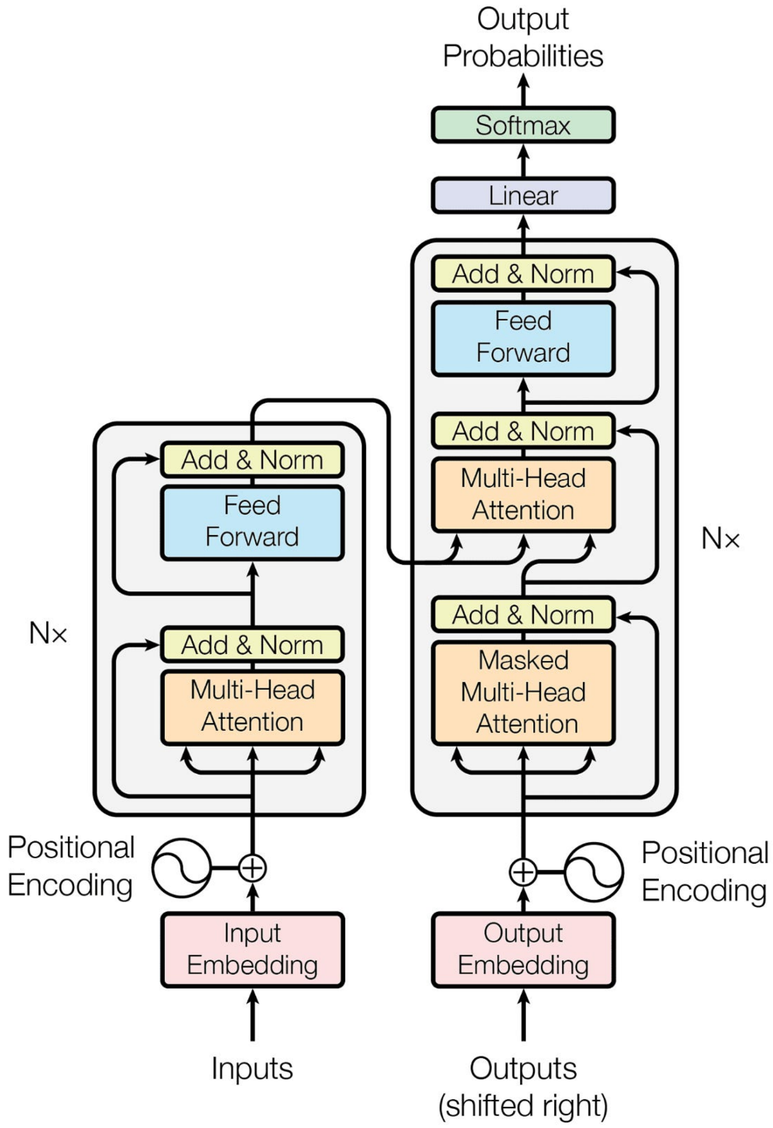
Causal Abstraction

Designing Counterfactuals

Computational Explanation

We explain certain kinds of dynamical systems with algorithms they implement:

- A laptop sorting numbers
- Humans hearing and producing speech
- A language model generating text

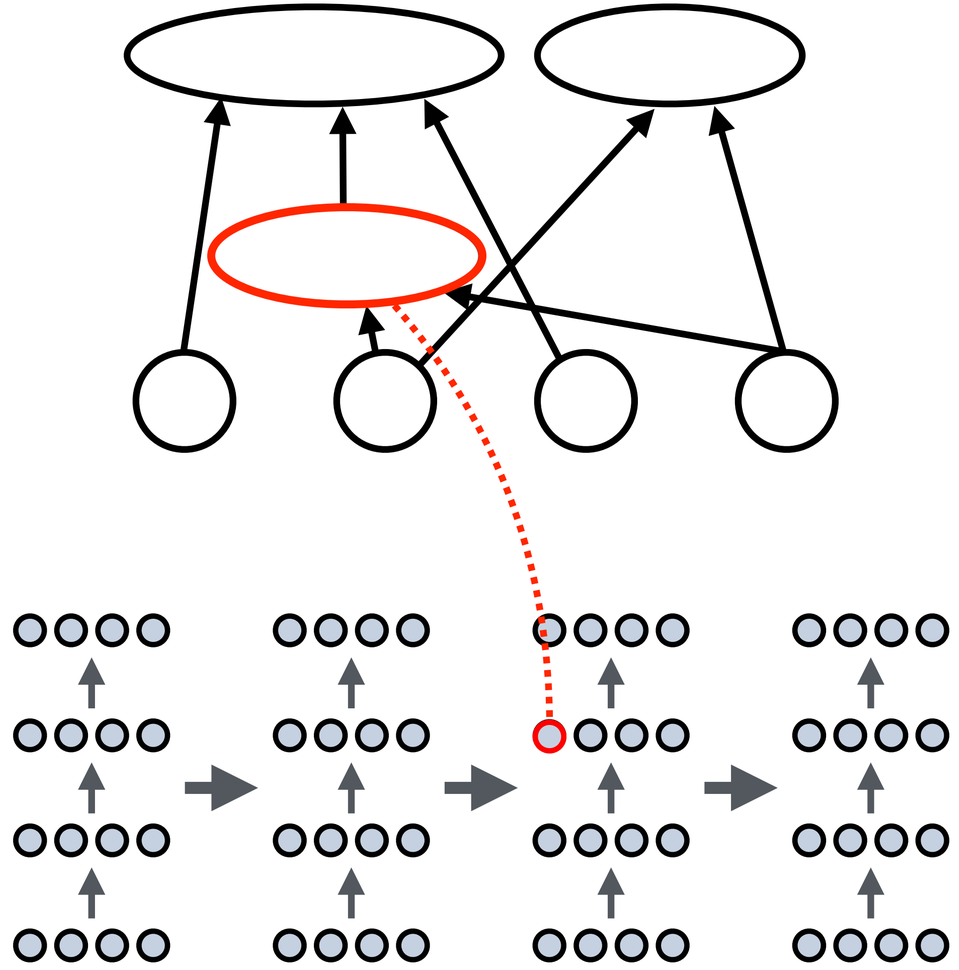


Mathematizing Implementation with Causal Abstraction

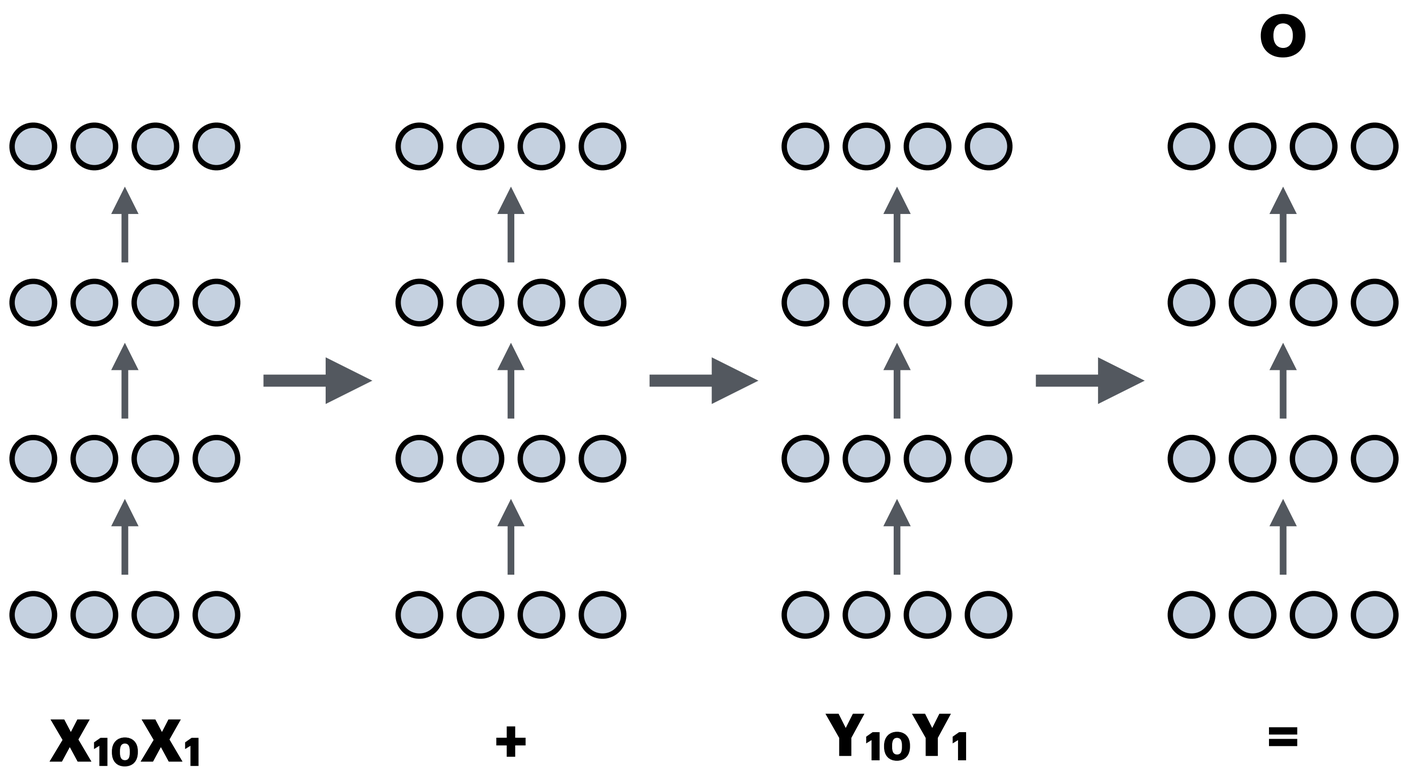
Geiger et al. (2025a; 2025b)

When does a neural network implement an algorithm?

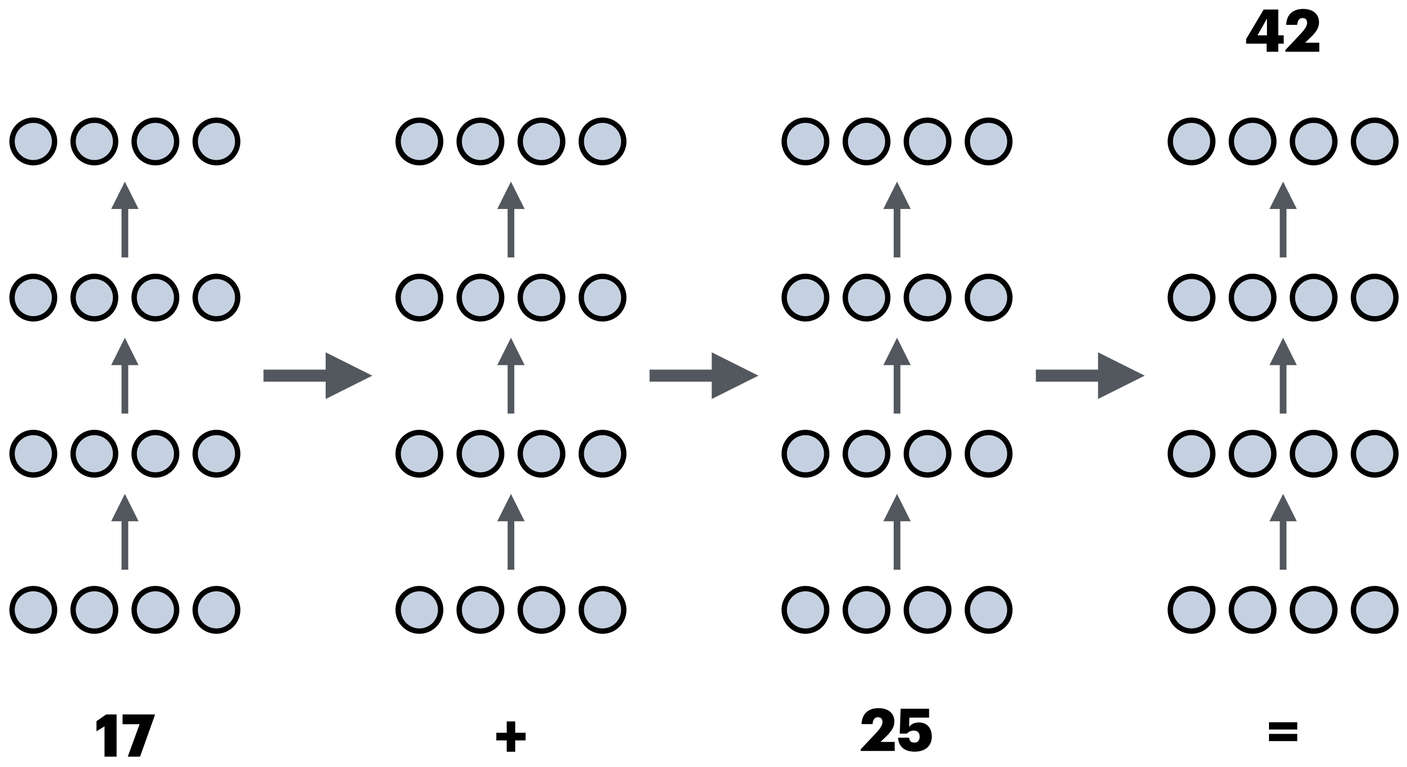
- Neural networks as causal models
- Algorithms as causal models
- Implementation as causal abstraction



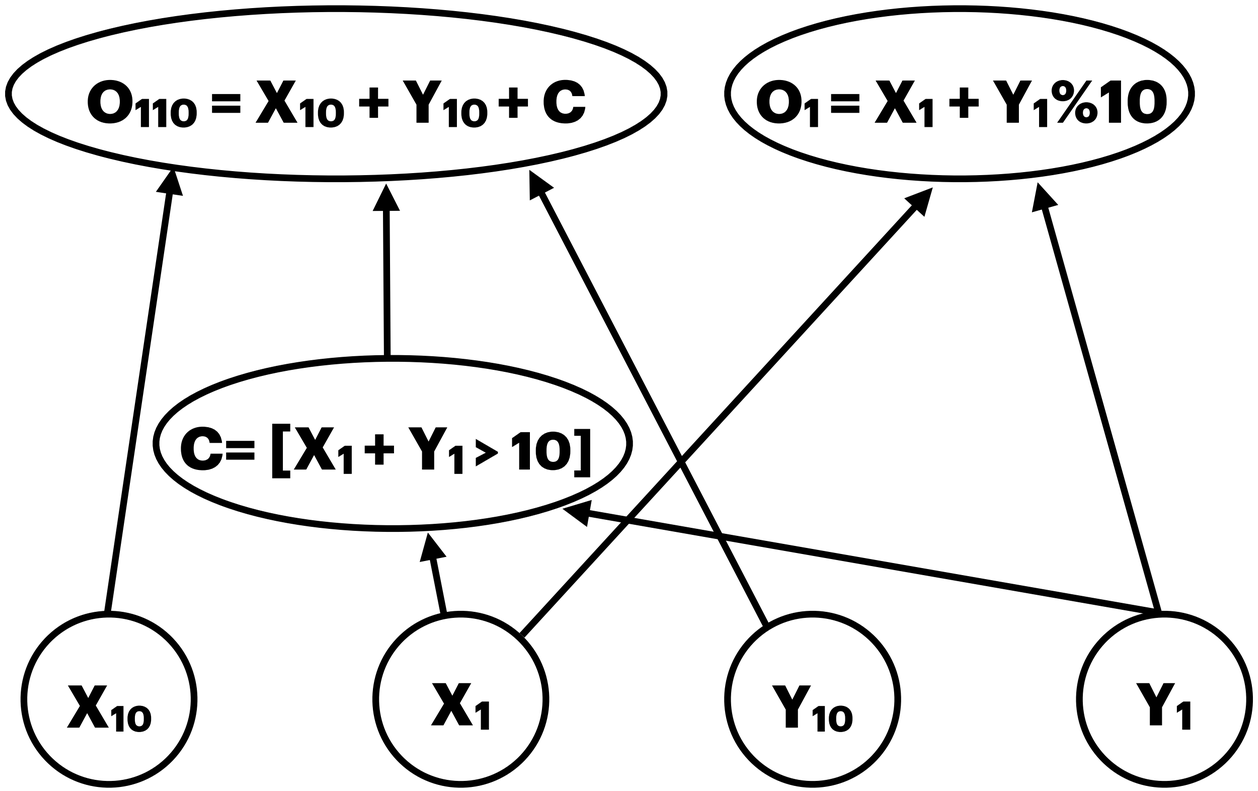
A Transformer Adding Two-Digit Numbers



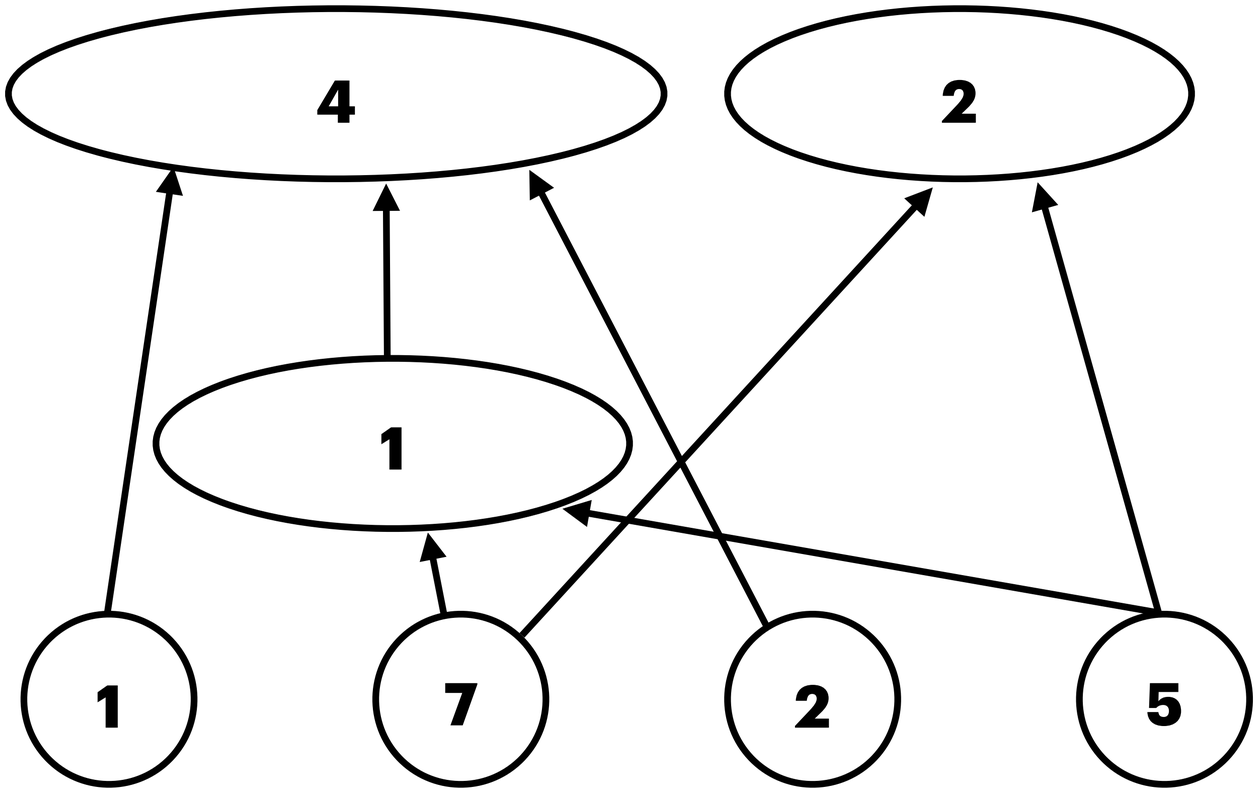
A Transformer Adding Two-Digit Numbers



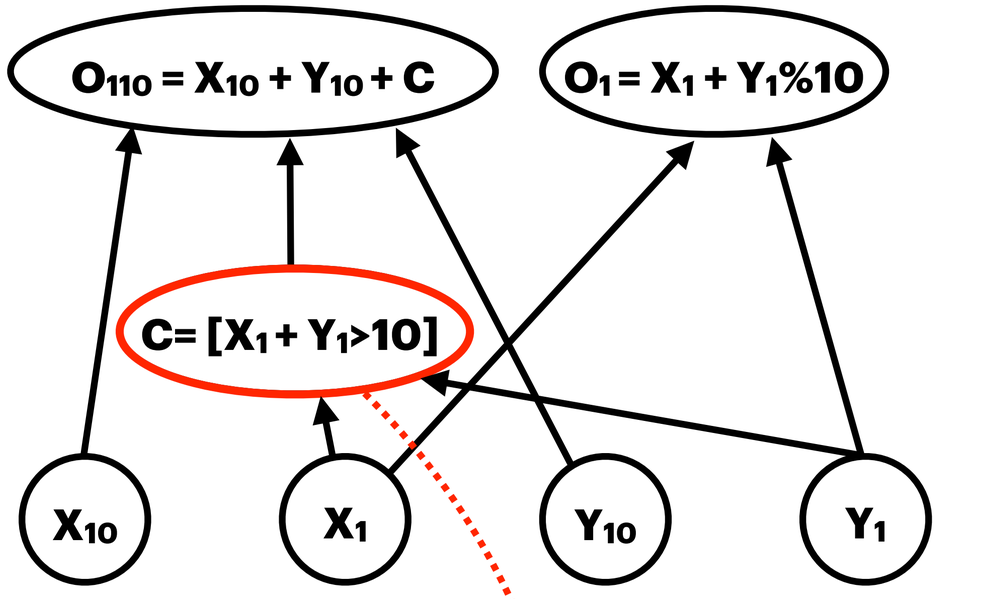
A Causal Model Adding Two-Digit Numbers

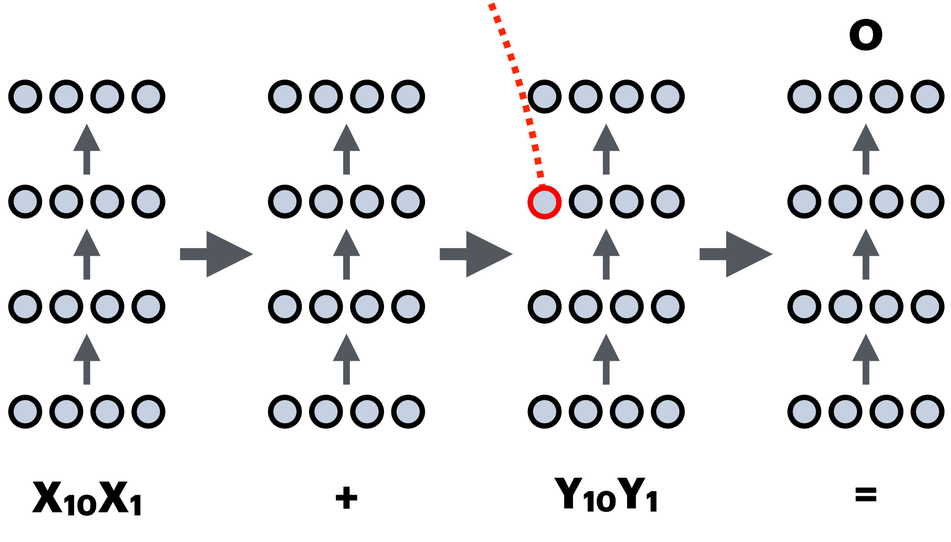


A Causal Model Adding Two-Digit Numbers



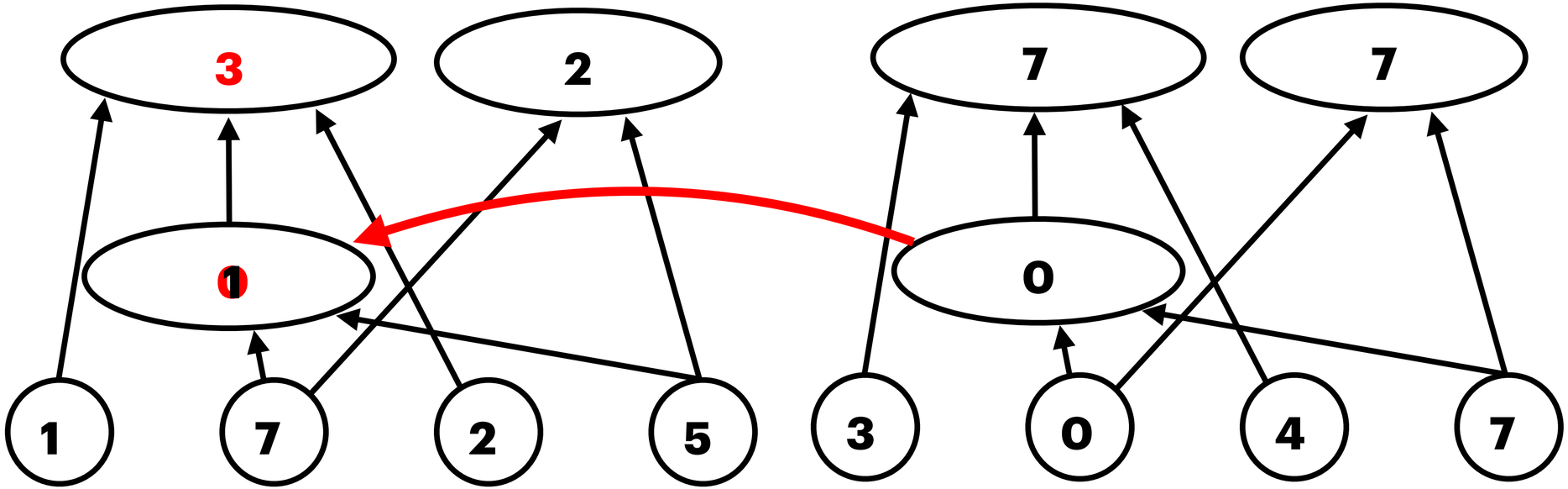
Aligning the High-Level and Low-Level Models

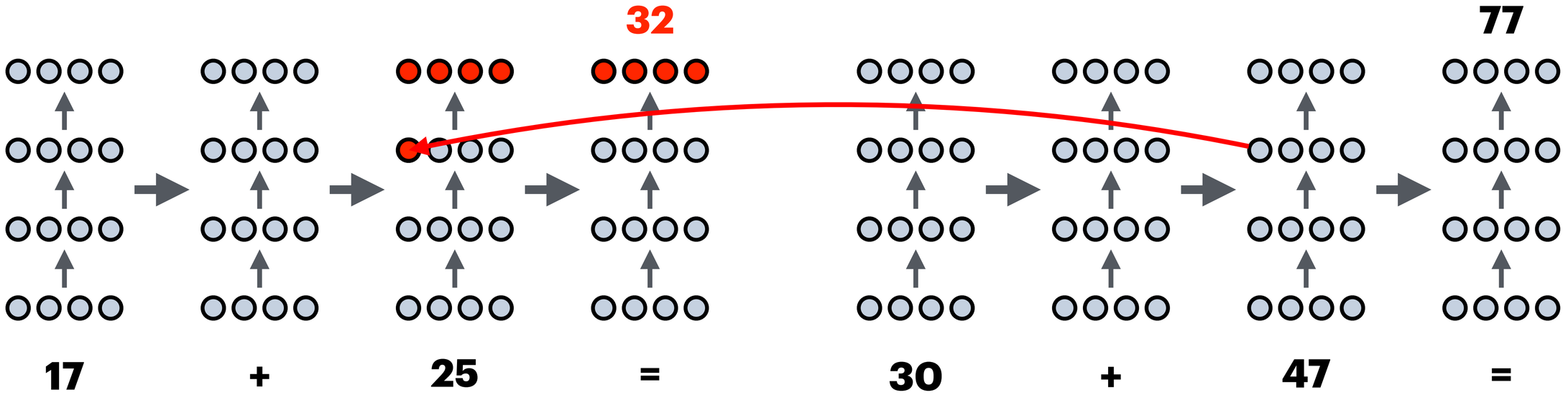




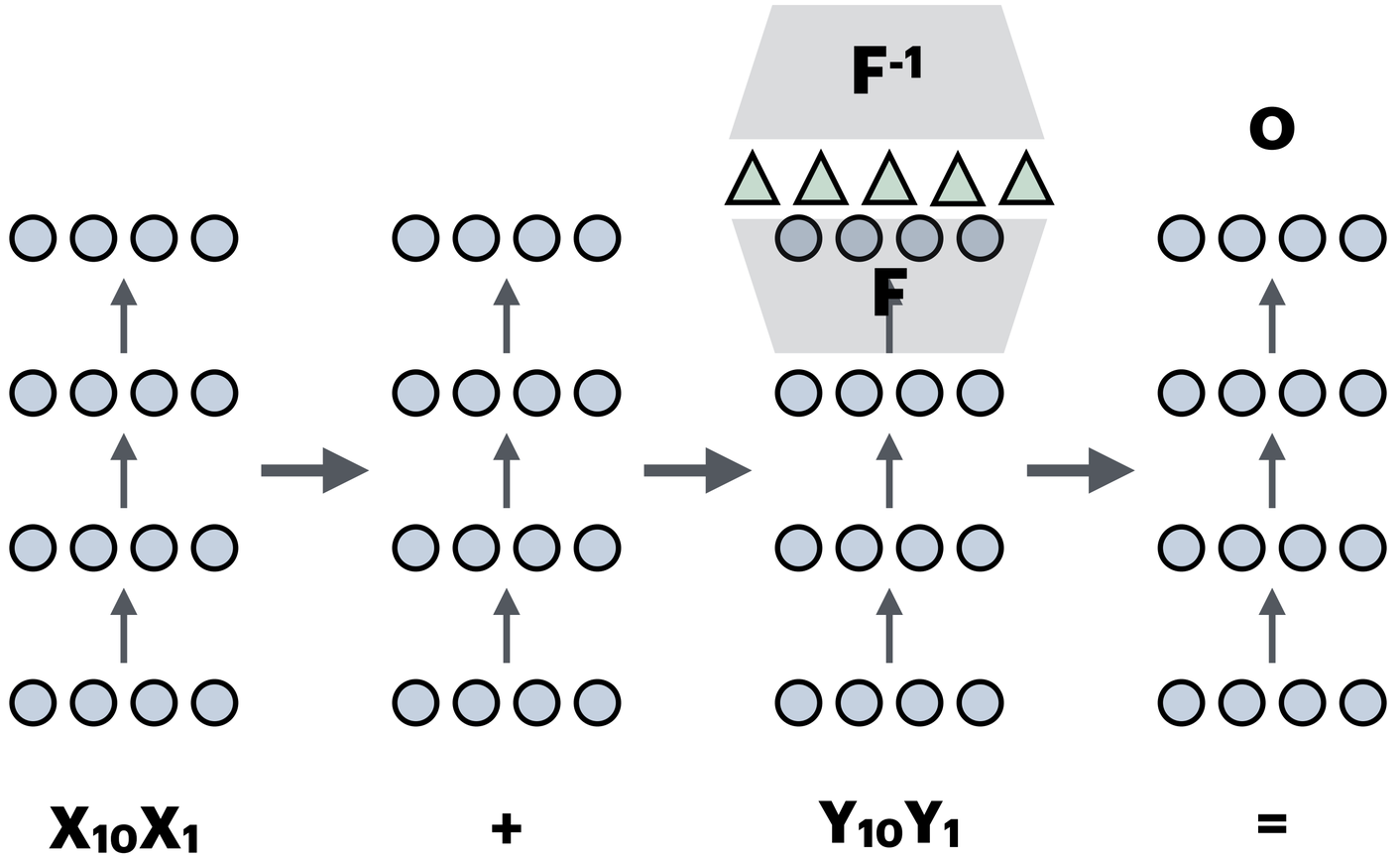
Interchange Interventions

Geiger et al. (2020)

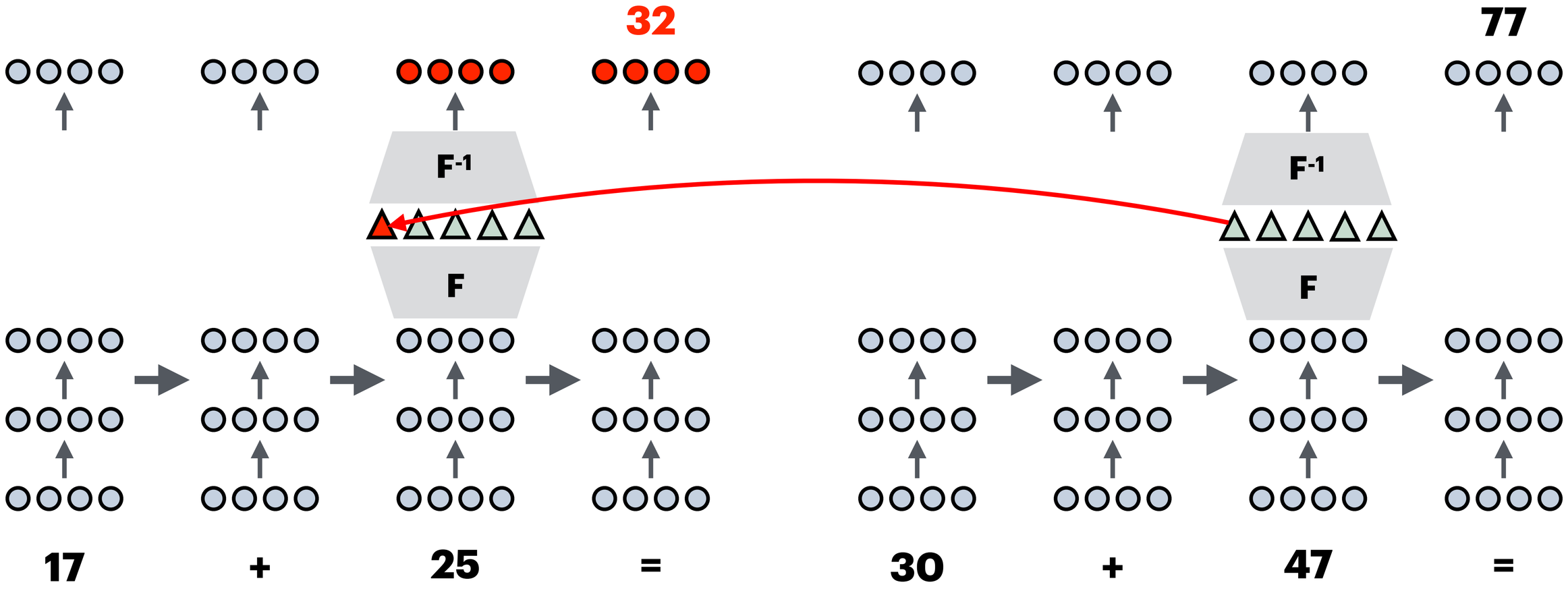




Featurizing a Hidden Vector

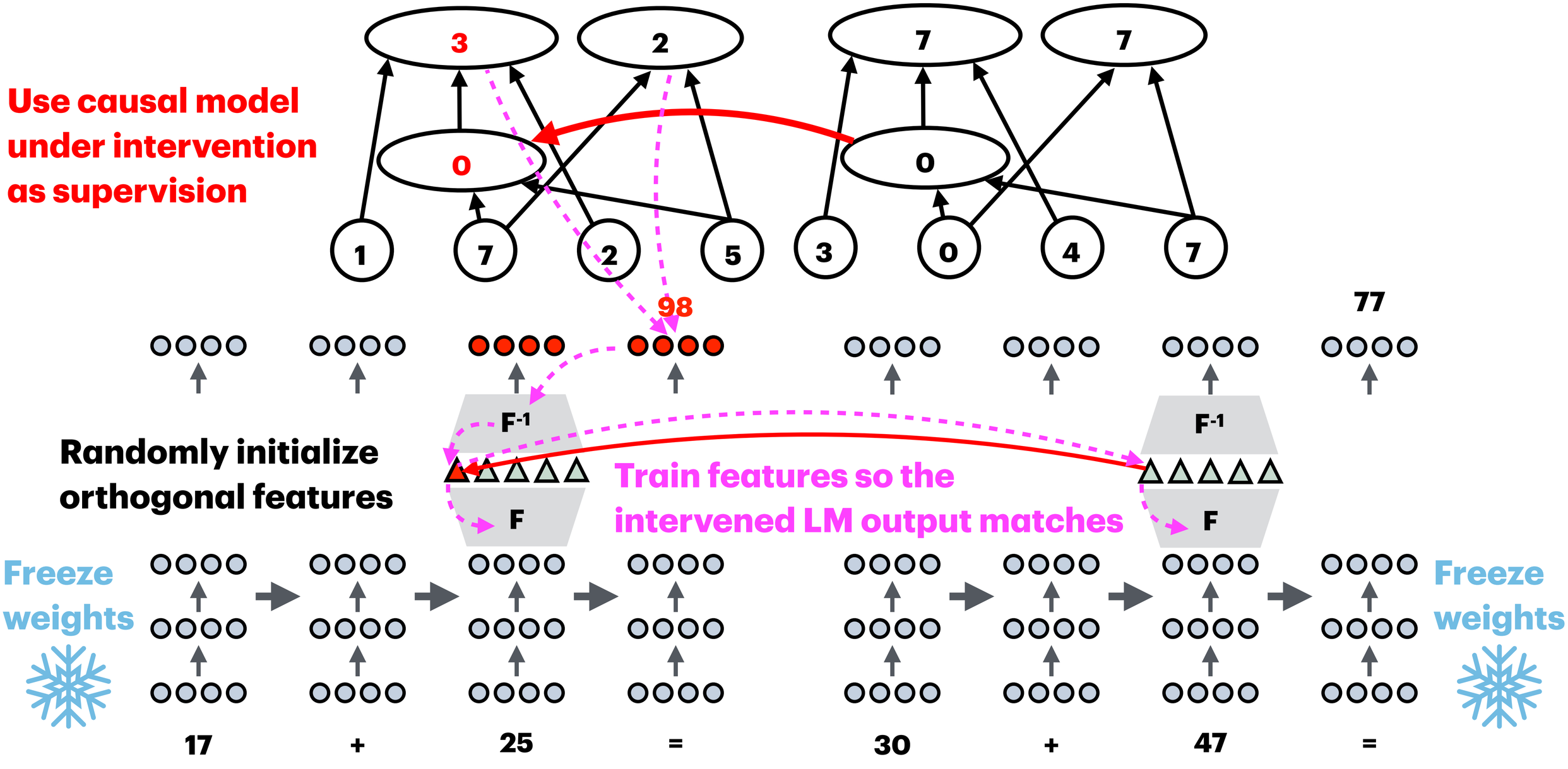


Distributed Interchange Intervention



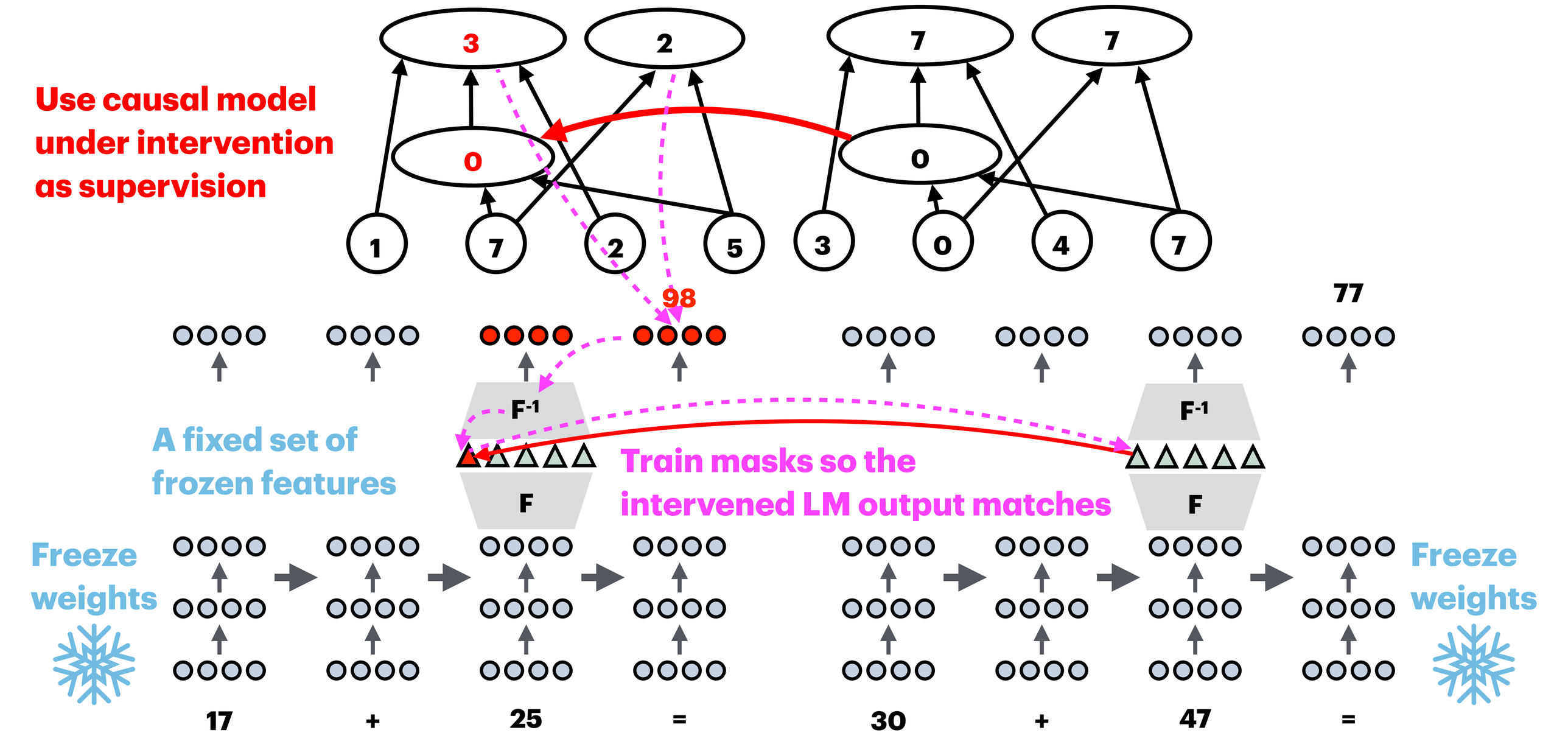
Distributed Alignment Search (DAS)

Localizing Causal Variables to Linear Subspaces with SGD (Geiger et al. 2024)



Desiderata Based Masking (DBM)

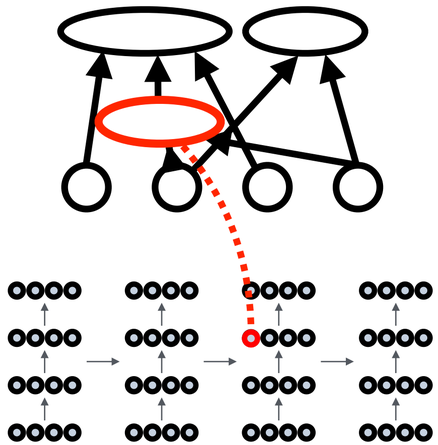
Supervised Selection of Frozen Features (Davies et al. 2023)



Pointers to the Literature

- Mechanistic interpretability papers explicitly using the framework of abstraction: Geiger et al. 2020, Wu et al. 2023, Geiger et al. 2024, Arora et al. 2024, Huang et al. 2024, Kallini et al. 2024, Csordas et al. 2024, Feng et al. 2024, Mueller et al. 2025, Prakash et al. 2025, Gur Arieh et al. 2025, Rodriguez et al. 2025, Boguraev et al. 2025, Huang et al. 2025, Minder et al. 2025, Grant et al. 2025, Sutter et al. 2025
- Theory of causal abstraction: Rubenstein et al. 2017, Beckers et al. 2018, Geiger et al. 2025a, Geiger et al. 2025b
- Desiderata-based masking: Davies et al. 2023

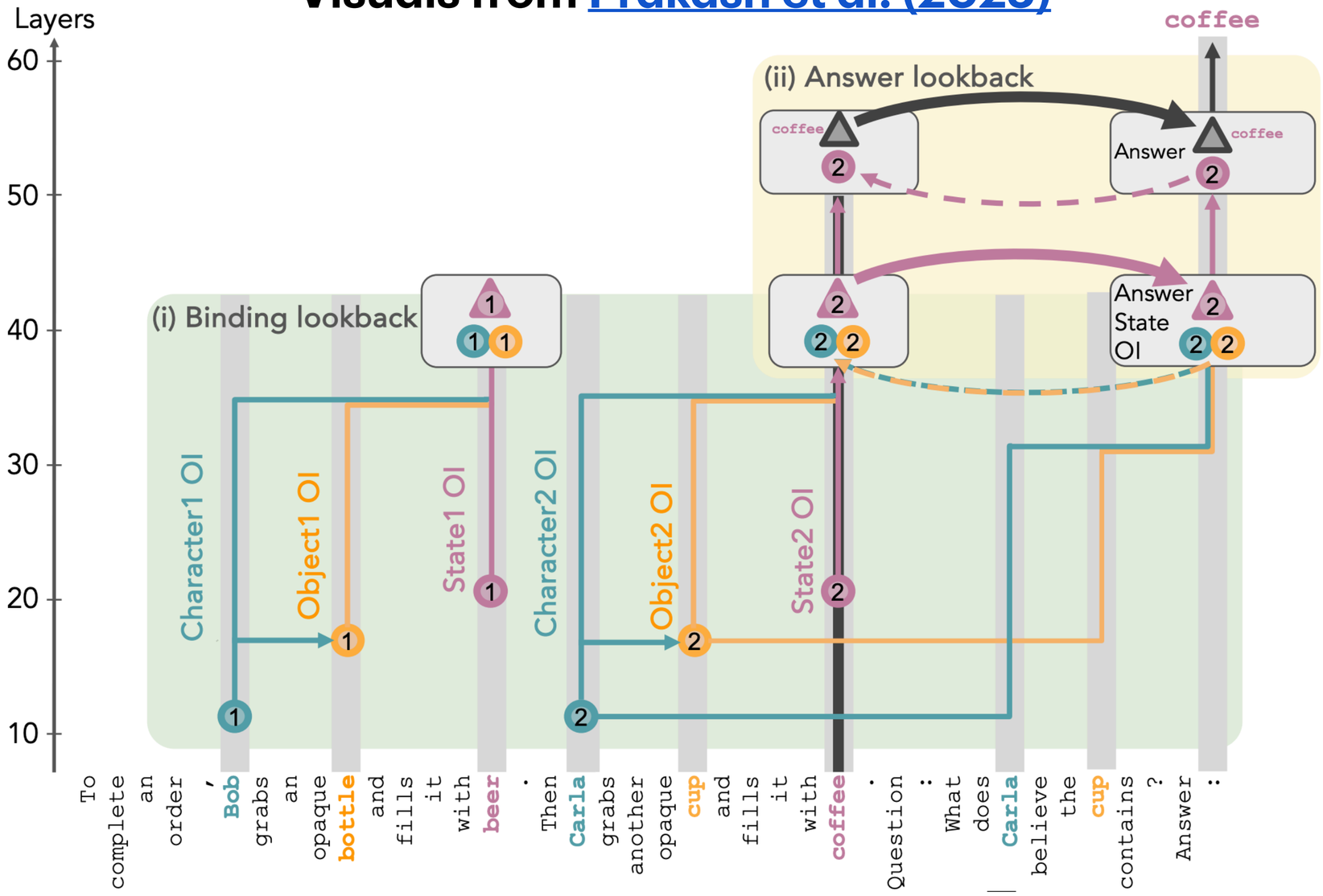
Outline



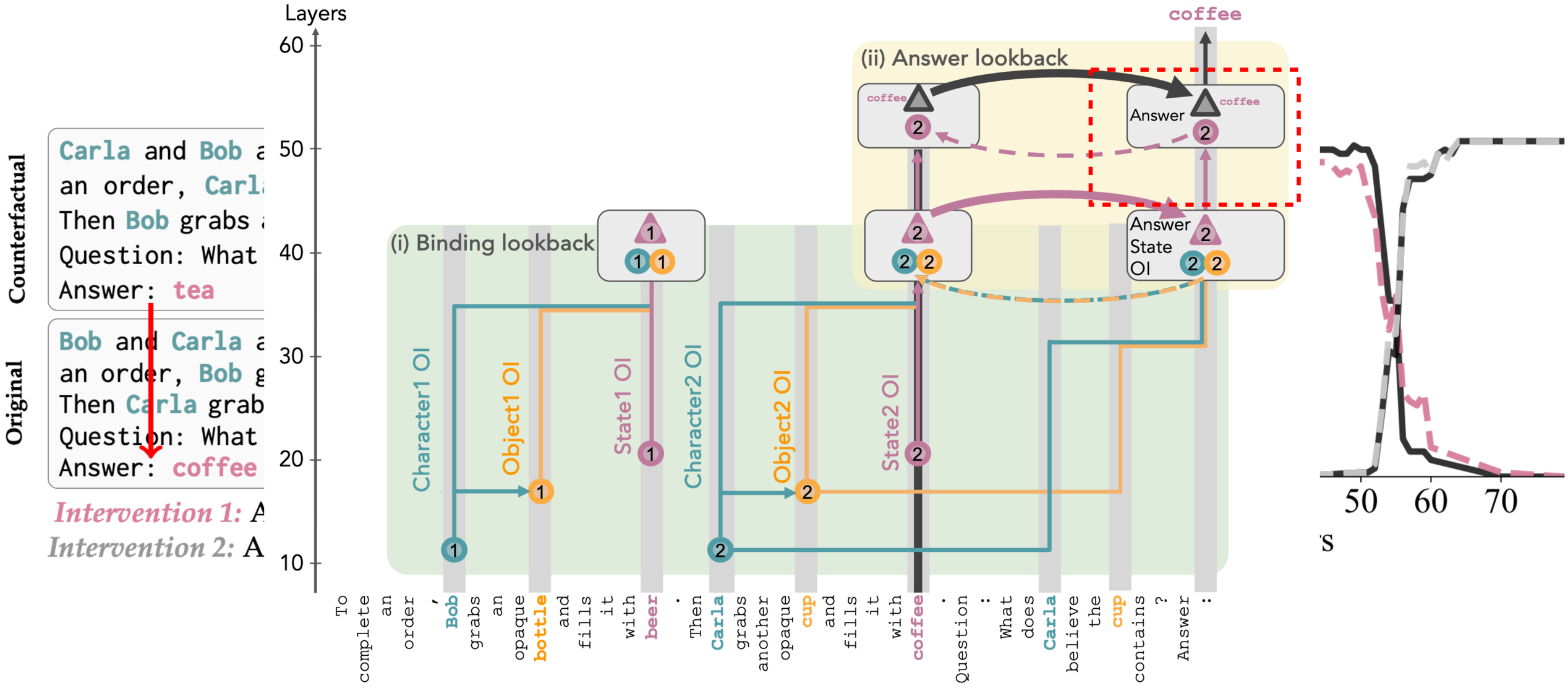
- Activation Steering
- Causal Mediation
- Causal Abstraction
- Designing Counterfactuals

Lookback Mechanisms

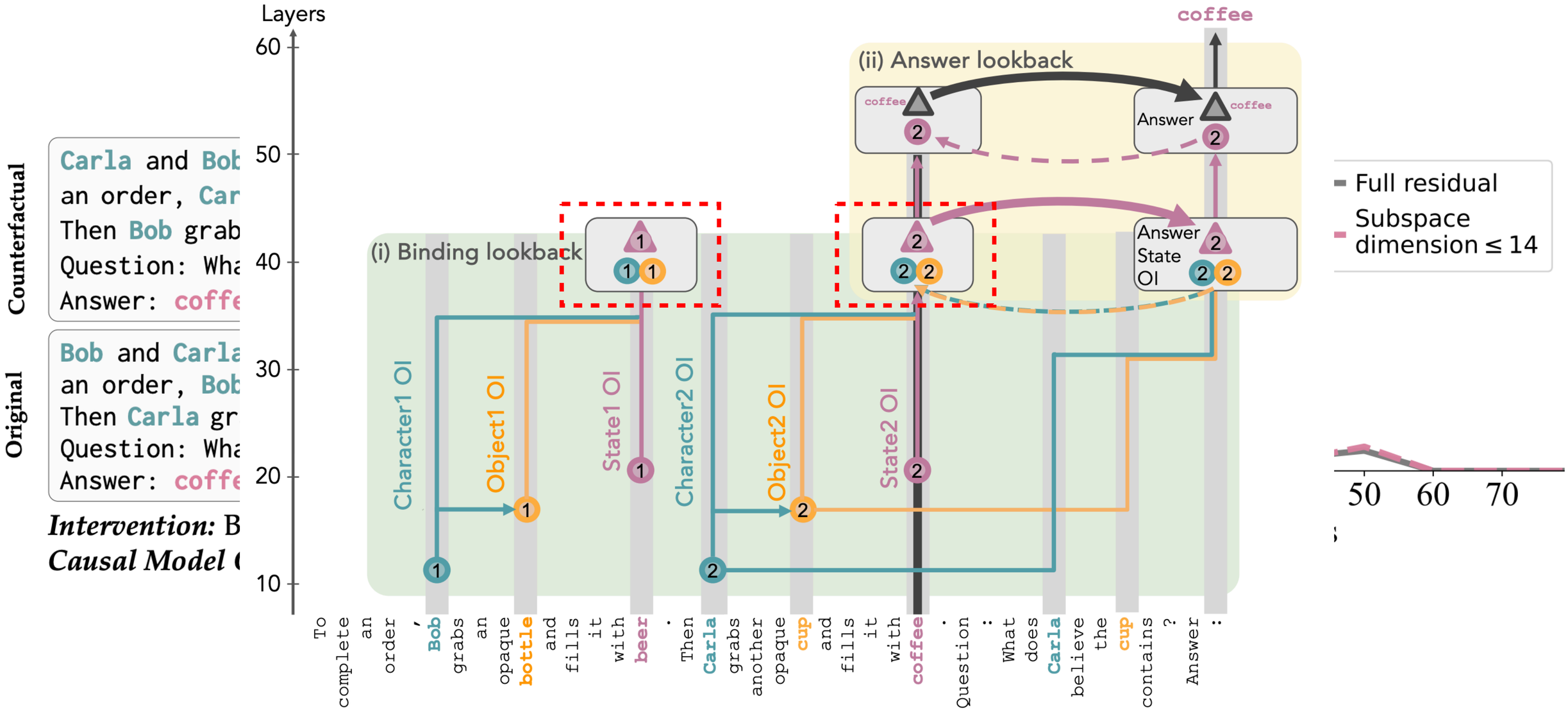
Visuals from [Prakash et al. (2025)]



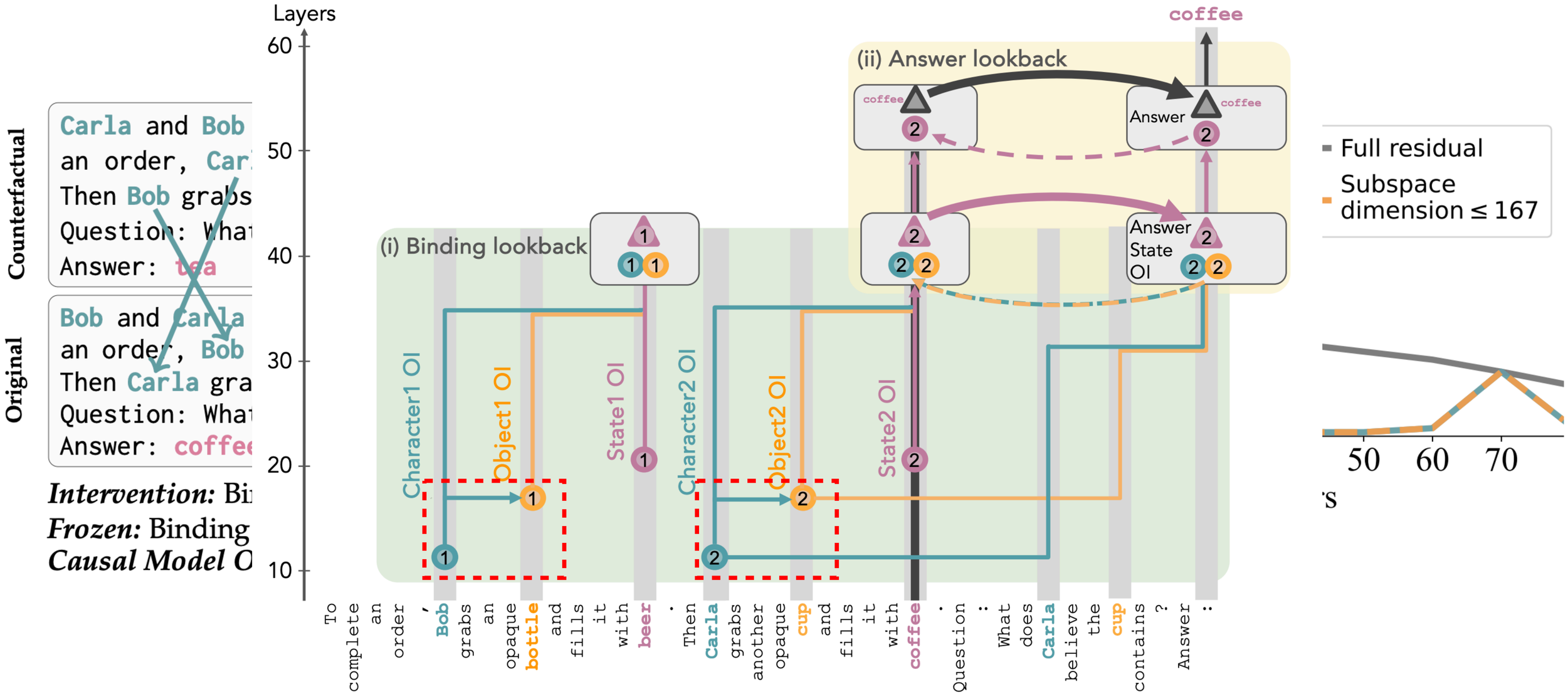
Designing Counterfactuals



Designing Counterfactuals

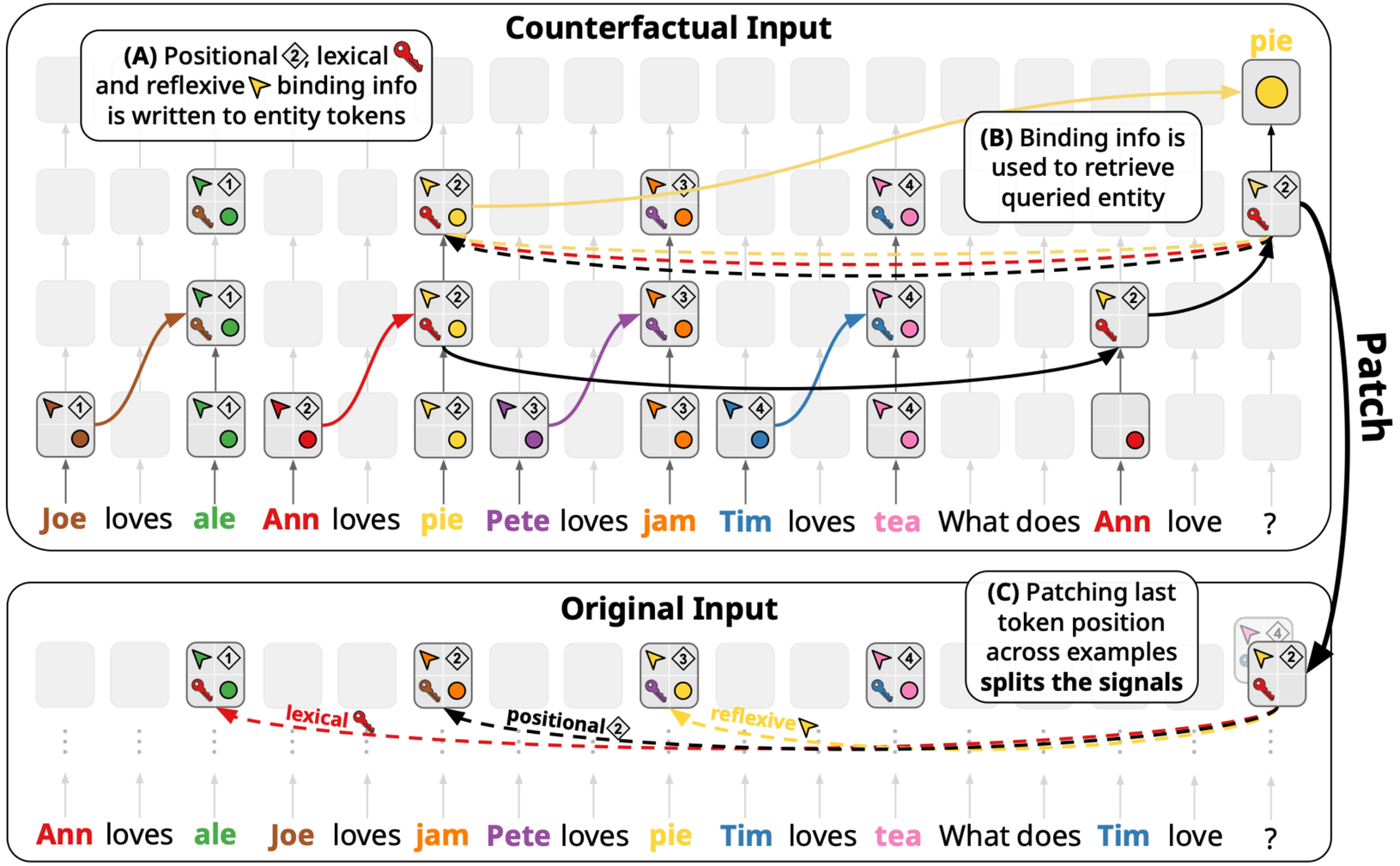


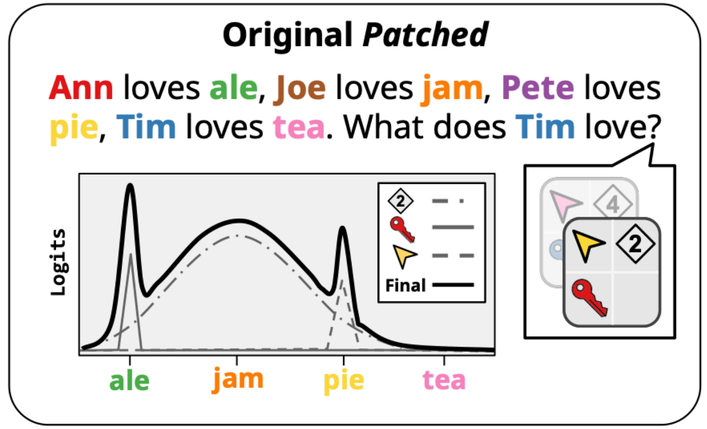
Designing Counterfactuals



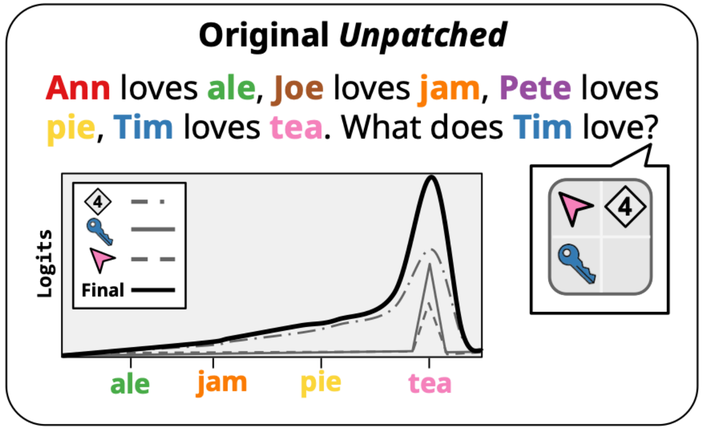
Mixing Mechanisms

Visuals from Gur Arieh et al. 2025

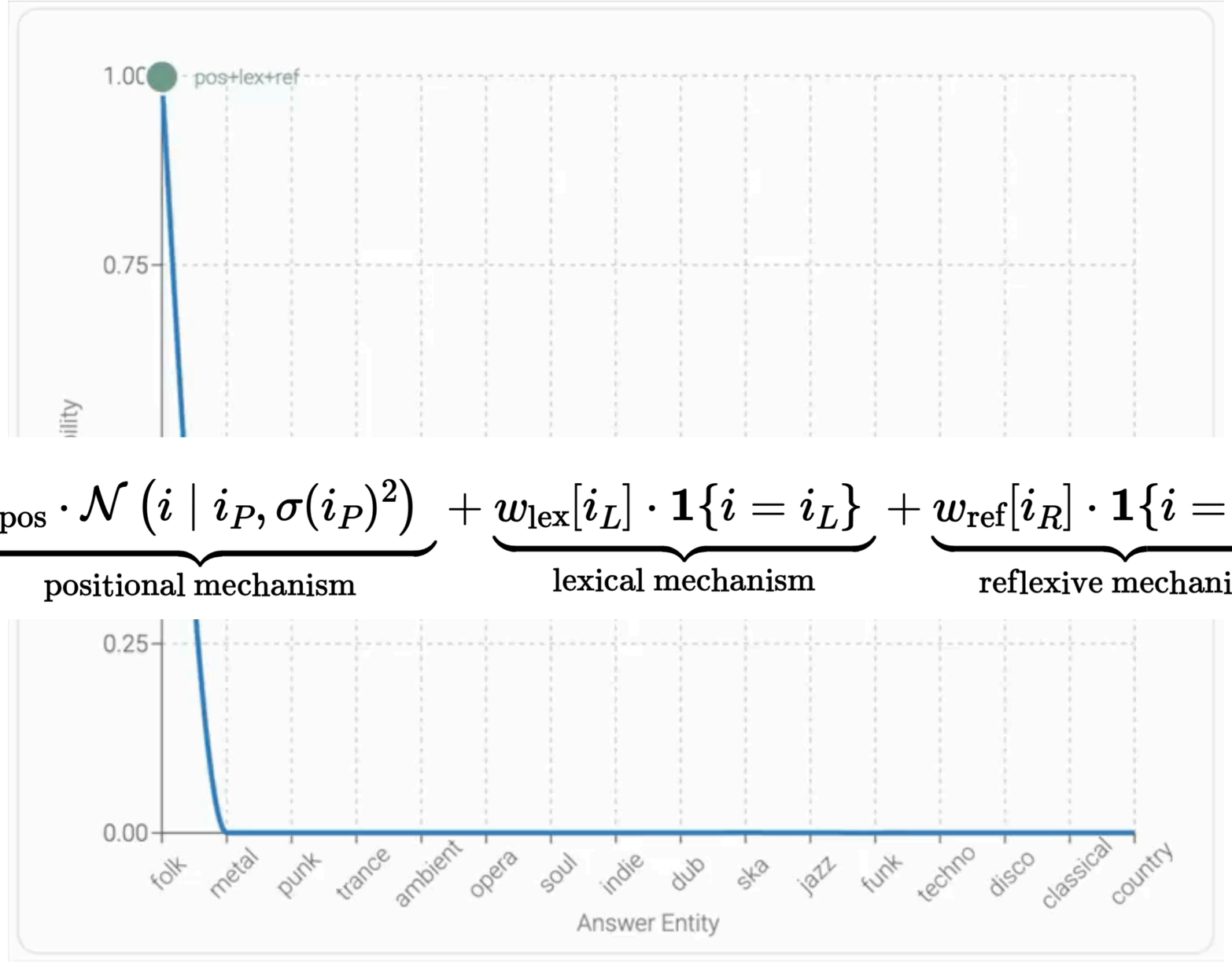




Under intervention, the three mechanisms **diverge** in the entities predicted



Without intervention, the three mechanisms **agree** with a decisive prediction



$$Y_i := \underbrace{w_{\text{pos}} \cdot \mathcal{N}\left(i \mid i_P, \sigma(i_P)^2\right)}_{\text{positional mechanism}} + \underbrace{w_{\text{lex}}[i_L] \cdot \mathbf{1}\{i = i_L\}}_{\text{lexical mechanism}} + \underbrace{w_{\text{ref}}[i_R] \cdot \mathbf{1}\{i = i_R\}}_{\text{reflexive mechanism}}$$

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