

Mechanistic Interpretability

Causal Abstraction and Manifold Steering

CD3 Hack Friday (AI/ML)

September 18, 2026

Introduction

Part I: Review

- ▶ Causal models as hypotheses about computation.
- ▶ Interventions and interchange interventions.
- ▶ Causal abstraction: linking high-level variables to network internals.

Part II: Demos

- ▶ `causal_model/causal_model.md`: interventions distinguish equivalent algorithms.
- ▶ `weekdays_geometry/weekdays_geometry.md`: a runnable workflow for localization, subspaces, and steering.



Causal Abstraction

High-Level Model

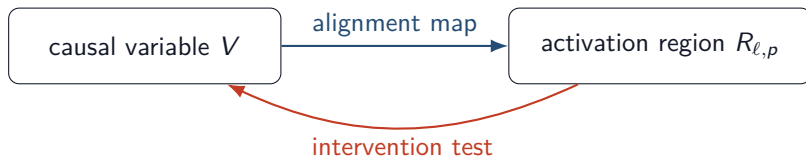
A human-readable algorithm:

- ▶ variables: $X_1, \dots, X_k$
- ▶ mechanisms: $X_i := f_i(\text{Pa}_i)$
- ▶ counterfactual predictions under interventions

Low-Level Model

A neural network computation:

- ▶ activations at layers and positions
- ▶ hooks, patches, subspaces, metrics
- ▶ observed behavior after edits



A causal abstraction claim is not just “the representation is correlated with a concept.” It says that editing the proposed representation should reproduce the counterfactual behavior predicted by the high-level causal model.

Interventions: The Operational Test

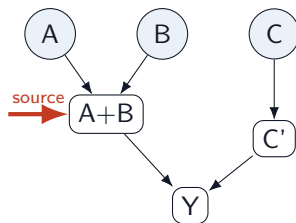
Ordinary Intervention:

$\text{do}(Z = z') \rightarrow$ run descendants of Z with Z fixed.

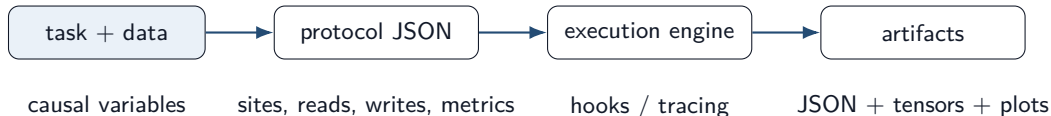
Interchange Intervention:

$$Z(x_{\text{base}}) \leftarrow Z(x_{\text{source}})$$

The source input supplies the value or representation of a high-level variable.



Causal Abstraction (CausalAb)



- ▶ Modern `causalab` experiments are serializable intervention documents; workflows chain protocol runs with script steps.
- ▶ The demo Markdown explains the research question and ships the exact runnable documents beside it.

Demo 1: causal_model/causal_model.md

Overview

A CPU-only causal-model demo: two algorithms compute the same function on every input, but an interchange intervention separates them.

Core Example

Both graphs compute

$$Y = (A + B + C) \bmod 10,$$

but they use different intermediate partial sums.

Questions

- ▶ Do the outputs ever differ?
- ▶ Does interchanging the intermediate S distinguish the graphs?
- ▶ What happens if the output Y is interchanged instead?

Causal Model Ingredients

Variables

- ▶ inputs: A, B, C
- ▶ intermediate: S
- ▶ output: Y
- ▶ interface: `raw_input`, `raw_output`

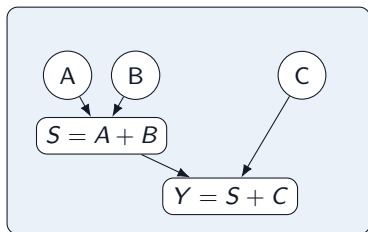
Two Mechanisms

- ▶ left-to-right: $S = A + B$, then $Y = S + C$;
- ▶ right-to-left: $S = B + C$, then $Y = S + A$;
- ▶ all arithmetic is modulo 10.

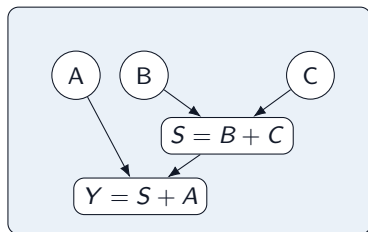
```
if order == "ltr":  
    S = (A + B) % 10  
    Y = (S + C) % 10  
else:  
    S = (B + C) % 10  
    Y = (S + A) % 10
```

Causal Graphs

left to right



right to left



Both graphs compute the same input-output function; the disagreement is internal.

Forward Pass and Direct Interventions

Behavioral Test

Enumerating all 10^3 inputs:

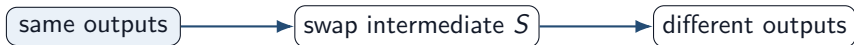
$$\#\{x : Y_{\text{LTR}}(x) \neq Y_{\text{RTL}}(x)\} = 0.$$

The two hypotheses are behaviorally indistinguishable.

Intervention Test

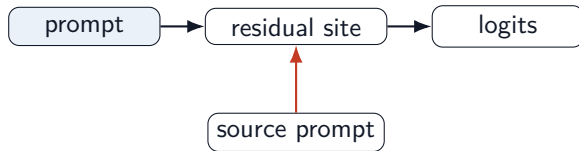
On 1000 random base/source pairs:

- ▶ interchange S : 914/1000 pairs separate;
- ▶ interchange Y : 0/1000 pairs separate.



Neural Networks Are Also Causal Graphs

- ▶ This demo stays at the symbolic causal-model level.
- ▶ `raw_input` and `raw_output` provide the interface used when the graph becomes a hypothesis about a network.
- ▶ The next demo replaces the symbolic variable S with residual-stream sites and subspaces.



The operational question remains the same:
does an internal intervention produce the
counterfactual behavior predicted by the
high-level model?

Interchange Intervention: Why It Matters

Problem

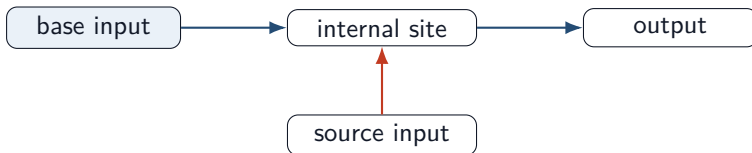
A hidden activation is high-dimensional; choosing a replacement vector by hand has no obvious causal meaning.

Interchange Solution

Use a second valid input to supply the internal state:

$$S(x_{\text{base}}) \leftarrow S(x_{\text{source}}).$$

In the causal-model demo this exposes the difference between two algorithms. In a network, the same operation becomes activation patching.



Demo 1 Recap

Overview

Causal abstraction distinguishes hypotheses that ordinary behavior cannot:

1. specify the intermediate variable on which the hypotheses differ;
2. interchange that variable between valid inputs;
3. compare the resulting counterfactual outputs.

For the modular-addition pair, output behavior differs on 0/1000 inputs, while interchanging the intermediate distinguishes 914/1000 sampled pairs.

Demo 2: weekdays_geometry/weekdays_geometry.md

Objective

Locate and characterize the residual-stream representation of the answer weekday, then probe the straight path between two real answer activations.

Task

Prompt examples ask the model for the weekday n days after a target day:

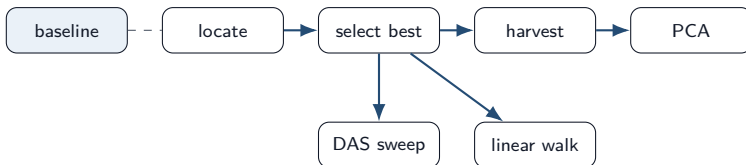
Q: What day is one day after

Monday? A:

Questions

- ▶ Can Llama-3.1-8B solve the task?
- ▶ Which layer/position carries the answer?
- ▶ How much activation variance is low-dimensional?
- ▶ How much causal capacity fits in a small DAS subspace?
- ▶ What happens along a linear activation walk?

The Analysis Pipeline



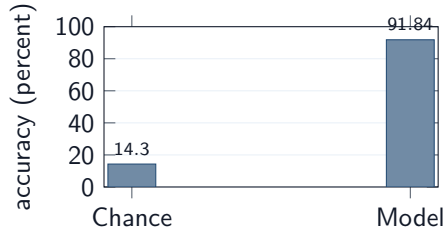
- ▶ The demo is a Markdown research narrative around one runnable workflow JSON and five intervention protocol JSON files.
- ▶ The workflow contains 10 executable steps; dependencies come from artifact references rather than notebook state.

Step 1: Baseline Task Performance

What Is Tested?

Can the model solve the 49-prompt weekday space before we interpret its internals?

- ▶ model: meta-llama/Llama-3.1-8B
- ▶ data: all $7 \times 7 = 49$ weekday/offset combinations
- ▶ accuracy: $45/49 = 91.84\%$
- ▶ chance floor: $1/7 = 14.3\%$



Step 2: Layer and Token Localization

Question: at which residual-stream cell does the answer weekday become causally available?

- ▶ Scan 32 layers $\times$ three positions with interchange interventions.
- ▶ The answer slot sits at the measured null level through L14.
- ▶ IIA jumps from 0.245 at L17 to 0.878 at L18.
- ▶ L25–L31 form a seven-layer plateau at 0.9184; the workflow selects L25 by first-max tie breaking.

Interpretation

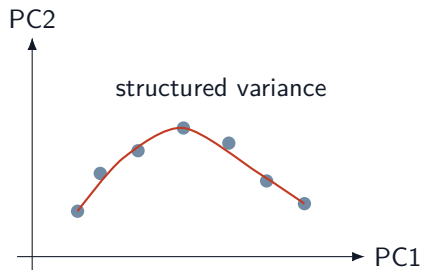
The whole-cell interchange is routing: its successful rows are exactly the counterfactual prompts the model already answers correctly.

Step 3: Subspace Localization

Two Complementary Questions

- ▶ **PCA**: how concentrated is ordinary activation variance at the located cell?
- ▶ **DAS**: how much answer behavior can a trained k -dimensional subspace supply under intervention?

The workflow first harvests the L25 final-token activations, fits PCA as a script step, and separately trains DAS rotations for $k \in \{2, 8, 32\}$.

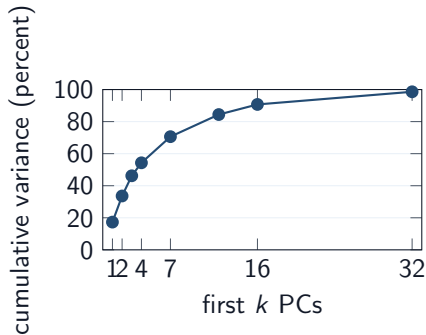


PCA Outputs

At the located L25 final-token cell:

- ▶ PC1 explains 17.3% of sample variance;
- ▶ first 2 PCs: 33.6%;
- ▶ first 7 PCs: 70.6%;
- ▶ first 16 PCs: 90.7%;
- ▶ first 32 PCs: 98.6%.

A single “weekday direction” is not supported by the variance spectrum.

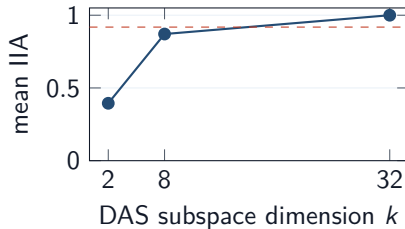


Step 4: Causal Subspace Capacity (DAS)

Question

How much answer behavior can a trained k -dimensional intervention subspace supply at the located cell?

- ▶ $k = 2$: mean IIA 0.395
- ▶ $k = 8$: mean IIA 0.871
- ▶ $k = 32$: mean IIA 1.000
- ▶ three seeds per k , same 40-step training budget



Linear Steering Intuition

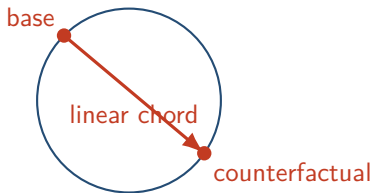
Row-to-row interpolation

$$\mathbf{h}(\alpha) = (1 - \alpha)\mathbf{h}_{\text{base}} + \alpha\mathbf{h}_{\text{cf}}, \quad \alpha \in [0, 1].$$

$\alpha = 0$ is the original activation; $\alpha = 1$ is the full interchange endpoint.

Why this is useful

The endpoints are both real model activations, so the only new question is what the model says on the straight chord between them.

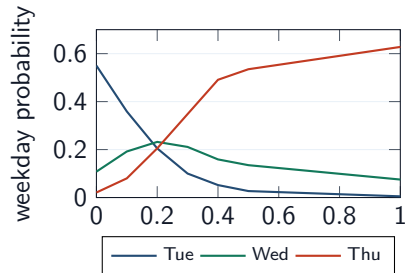


Step 5: Linear Path Evaluation

Workflow measurement

For each of the 49 base/counterfactual rows, the protocol evaluates 11 values of α and records probability mass on the seven weekday tokens.

- ▶ adjacent endpoint days: third-day transit on 0/10 rows;
- ▶ cyclic distance 2: transit on 14/22, with 12/14 via the intervening day;
- ▶ cyclic distance 3: transit on 3/11.



What Does the Current Geometry Evidence Show?

Activation Structure

- ▶ Variance is distributed across several principal components.
- ▶ A small trained subspace has substantial causal capacity, but the fit is underdetermined.

Behavior Along the Chord

- ▶ Adjacent weekdays do not pass through a third day.
- ▶ Distance-two pairs often pass through the unique intermediate weekday.
- ▶ Longer chords depart from that cyclic ordering.

Scope

This supports a local ring-like reading of the linear walk, with increasing departures at larger cyclic distance.

Artifacts and Configs: Why This Matters Practically

- ▶ Protocol JSON is the executable experimental specification.
- ▶ Workflow JSON chains protocol and script steps through declared artifacts.
- ▶ validate, explain, and run operate on the same document.
- ▶ Intermediate tensors and metric tables are explicit files rather than notebook state.

```
demos/weekdays_geometry/  
  weekdays_geometry.md  
workflows/  
  weekdays_geometry.json  
protocols/  
  weekdays_baseline.json  
  weekdays_locate_scan.json  
  weekdays_harvest.json  
  weekdays_das_sweep.json  
  weekdays_linear_walk.json
```

How a Physicist Should Read This Demo

Useful Analogy

Treat the residual stream as a latent coordinate system and the intervention protocol as the controlled perturbation used to test which coordinates are causally relevant.

Similar to Physics Workflows

- ▶ define observables and controls;
- ▶ establish baseline performance;
- ▶ localize a sensitive region;
- ▶ vary a low-dimensional coordinate;
- ▶ inspect the response function.

Important Difference

- ▶ latent coordinates are learned, not instrument-defined;
- ▶ trained subspaces can overfit tiny task spaces;
- ▶ geometry needs intervention evidence, not only a visualization.

Reproduction Checklist

1. Install the repository with `uv sync` and obtain access to `meta-llama/Llama-3.1-8B`.
2. Validate the shipped workflow:

```
uv run causalab validate demos/weekdays_geometry/workflows/weekdays_geometry.json --data
```

3. Inspect the plan with `uv run causalab explain ....`
4. Run the workflow on a GPU and inspect `accuracy.json`, `locate/iia.json`, `pca/spectrum.json`, `fit/iia.json`, and `walk/day_probs.json`.

Some Limitations

- ▶ **In-sample task:** all results use the complete 49-prompt space; they are not held-out generalization scores.
- ▶ **DAS capacity $\neq$ dimensionality:** a $4096 \times k$ rotation is severely underdetermined by 49 pairs.
- ▶ **Localization tie:** L25–L31 are identical at the maximum IIA; L25 is a deterministic tie-break, not a uniquely preferred layer.
- ▶ **Counterfactual design:** the shipped generator changes both input variables together, so input-token controls are not cleanly deconfounded.