

# Layer Sweep Analysis

Prompt: "I don't actually have siblings, but I"

Clean prediction: "" (p=48.3%)

Component: Mlp Output | Intervention: Mean

Layers tested: 34 | Prediction flipped in 1/34 layers

Peak effect: Layer 5 (KL = 1.0868)

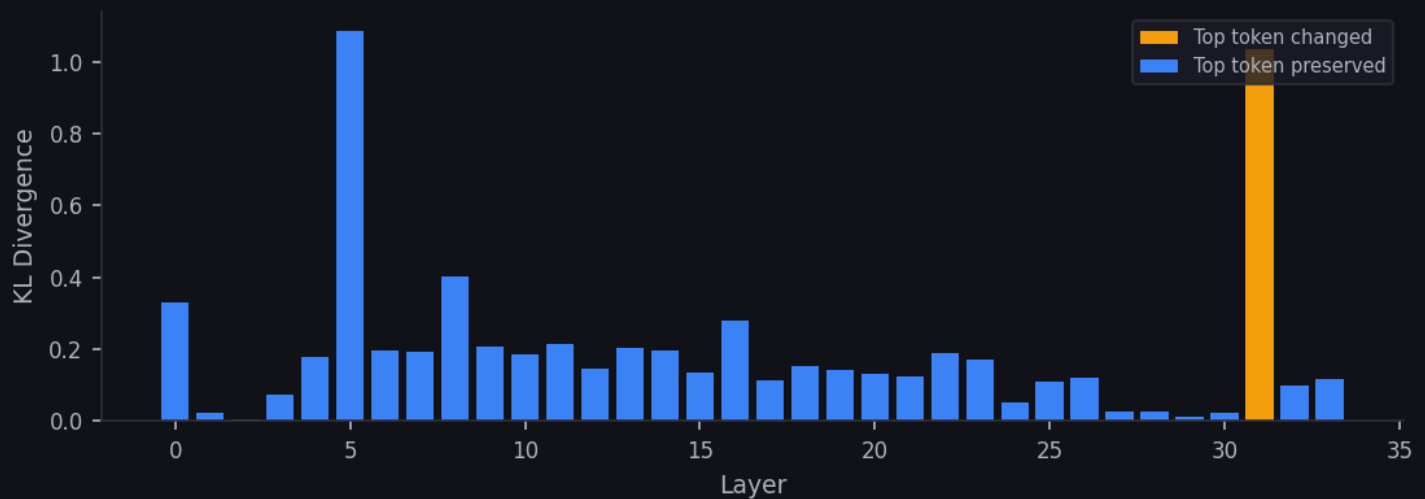
Config hash: dedc596bf638cdaf

## What Is This Report?

This report shows the results of a layer sweep - a systematic test where we disable one component of the AI model at a time and measure how much the output changes. Each layer in the model processes text sequentially, like floors in a factory. By disabling each floor one at a time, we can figure out which floors are essential for producing the correct answer.

The key metric is KL Divergence - a number measuring how much the model's entire prediction changed. KL = 0 means nothing changed. KL > 1 means a meaningful effect. KL > 10 means a major disruption. When the bar is highlighted (amber), the model's #1 prediction actually flipped to a different word.

## KL Divergence Across Layers



## Key Findings

- [IMPORTANT]** Layer 5 has the strongest causal effect
- With KL divergence of 1.09, layer 5's MLP has the largest individual impact on the output. This is the most causally important layer for this input.
- [NOTABLE]** 1/34 layers flip the prediction
- Ablating the MLP at 1 different layers caused the model to change its top prediction entirely. Most layers can be removed without changing the answer, suggesting the prediction is distributed across few key components.
- [INFO]** U-shaped importance pattern
- Early layers (avg KL: 0.26) and late layers (0.18) matter more than the middle (0.17). This is a common pattern: early layers set up representations, middle layers maintain them, and late layers refine the final prediction.

Per-Layer Results

Each row shows the effect of disabling the component at that layer. 'Clean Token' is the model's normal prediction. 'Intervention Token' is what the model predicted after the component was disabled.

Layer	KL Div	Changed	Effect	Clean Token		Interv. Token	Duration
0	0.3314	no	Low	""	""	8.4s	
1	0.0219	no	Low	""	""	7.7s	
2	0.0049	no	Low	""	""	7.6s	
3	0.0711	no	Low	""	""	7.7s	
4	0.1760	no	Low	""	""	7.8s	
5	1.0868	no	Moderate	""	""	7.9s	
6	0.1977	no	Low	""	""	7.9s	
7	0.1923	no	Low	""	""	8.1s	
8	0.4006	no	Low	""	""	8.7s	
9	0.2083	no	Low	""	""	7.8s	
10	0.1866	no	Low	""	""	7.7s	
11	0.2148	no	Low	""	""	8.0s	
12	0.1439	no	Low	""	""	8.3s	
13	0.2018	no	Low	""	""	8.2s	
14	0.1965	no	Low	""	""	7.7s	
15	0.1333	no	Low	""	""	7.3s	
16	0.2779	no	Low	""	""	7.2s	
17	0.1119	no	Low	""	""	7.2s	
18	0.1542	no	Low	""	""	7.3s	
19	0.1433	no	Low	""	""	7.1s	
20	0.1294	no	Low	""	""	7.2s	
21	0.1246	no	Low	""	""	7.2s	
22	0.1902	no	Low	""	""	7.2s	
23	0.1690	no	Low	""	""	7.2s	
24	0.0510	no	Low	""	""	7.4s	
25	0.1094	no	Low	""	""	7.5s	
26	0.1190	no	Low	""	""	7.5s	
27	0.0266	no	Low	""	""	7.5s	
28	0.0257	no	Low	""	""	7.5s	
29	0.0116	no	Low	""	""	7.4s	
30	0.0214	no	Low	""	""	7.4s	
31	1.0376	YES	Moderate	""	"always"	7.5s	
32	0.0981	no	Low	""	""	7.4s	
33	0.1154	no	Low	""	""	7.1s	

# Glossary

Key terms used in this report, written for people who may not have a machine learning background.

## Layer

One processing step in the model's pipeline. Text is processed through all layers sequentially - like floors in a factory where each floor adds more refinement.

## MLP (Multi-Layer Perceptron)

The 'knowledge storage' component in each layer. Research shows that factual knowledge (like 'Eiffel Tower -> Paris') is often stored in MLP layers.

## Attention

The component that decides which words in the input to focus on. When you read 'The cat sat on the \_\_\_\_', attention helps the model look back at 'cat' and 'sat' to predict 'mat'.

## Zero Ablation

Completely removing a component's output by setting it to zero. Like unplugging one wire in a circuit - if the lights go out, that wire was important.

## KL Divergence

A number measuring how different two probability distributions are.  $KL = 0$  means the intervention had no effect.  $KL > 1$  is meaningful.  $KL > 10$  is a major disruption. Higher = the component matters more.

## Top Token

The word the model considers most likely to come next. When the 'top token changed', the model's #1 prediction flipped to a completely different word.

## Logit

The raw score the model assigns to each possible next word. Higher logit = the model thinks that word is more likely. Logits are converted to probabilities using the softmax function.

## Probability

The model's confidence that a particular word is the right next word, expressed as a percentage (0-100%). A probability of 95% means the model is very confident.

## Config Hash

A unique fingerprint of your experiment setup. If someone runs the same experiment and gets the same hash, the results should be identical. This guarantees reproducibility.

## Activation Patching

A technique where you run the model on two different inputs, then swap the internal values from one into the other at a specific point. If the output changes, that point carries the information that distinguishes the two inputs.

## Residual Stream

The main 'highway' of information flowing through the model. Each layer reads from and writes to this stream. It's called 'residual' because each layer adds its contribution on top of what came before.