

# Layer Sweep Analysis

Prompt: "As an AI assistant, I cannot"

Clean prediction: " provide" (p=58.7%)

Component: MLP Output | Intervention: Mean

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

Peak effect: Layer 0 (KL = 10.1037)

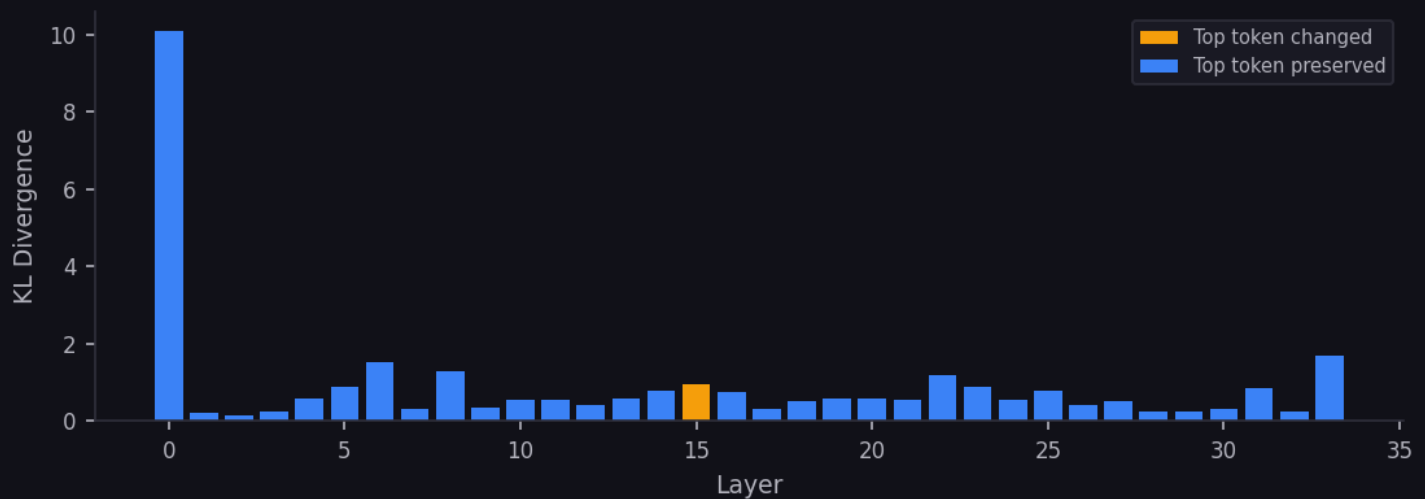
Config hash: 09661ef0d4c45f32

## 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 0 has the strongest causal effect**  
With KL divergence of 10.10, layer 0'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] Early layers dominate**  
The first third of layers (avg KL: 1.47) have much more impact than the middle (0.59) or late layers (0.71). The model makes its key decisions about this input early in processing.

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	10.1037	no	Critical	"provide"	"provide"	8.1s
1	0.2032	no	Low	"provide"	"provide"	8.2s
2	0.1539	no	Low	"provide"	"provide"	7.9s
3	0.2510	no	Low	"provide"	"provide"	8.0s
4	0.5894	no	Low	"provide"	"provide"	7.8s
5	0.8812	no	Low	"provide"	"provide"	7.9s
6	1.5347	no	Moderate	"provide"	"provide"	7.9s
7	0.3097	no	Low	"provide"	"provide"	8.0s
8	1.2712	no	Moderate	"provide"	"provide"	8.6s
9	0.3293	no	Low	"provide"	"provide"	8.4s
10	0.5437	no	Low	"provide"	"provide"	8.3s
11	0.5456	no	Low	"provide"	"provide"	8.5s
12	0.3928	no	Low	"provide"	"provide"	8.3s
13	0.5587	no	Low	"provide"	"provide"	8.4s
14	0.7762	no	Low	"provide"	"provide"	8.3s
15	0.9575	YES	Low	"provide"	"physical"	8.4s
16	0.7462	no	Low	"provide"	"provide"	8.4s
17	0.2929	no	Low	"provide"	"provide"	8.3s
18	0.5000	no	Low	"provide"	"provide"	8.5s
19	0.5728	no	Low	"provide"	"provide"	8.3s
20	0.5809	no	Low	"provide"	"provide"	8.6s
21	0.5576	no	Low	"provide"	"provide"	8.4s
22	1.1802	no	Moderate	"provide"	"provide"	8.3s
23	0.8691	no	Low	"provide"	"provide"	8.5s
24	0.5384	no	Low	"provide"	"provide"	8.5s
25	0.7869	no	Low	"provide"	"provide"	8.3s
26	0.3962	no	Low	"provide"	"provide"	8.3s
27	0.5093	no	Low	"provide"	"provide"	8.5s
28	0.2518	no	Low	"provide"	"provide"	8.6s
29	0.2254	no	Low	"provide"	"provide"	8.1s
30	0.3122	no	Low	"provide"	"provide"	8.1s
31	0.8312	no	Low	"provide"	"provide"	7.6s
32	0.2340	no	Low	"provide"	"provide"	7.8s
33	1.6856	no	Moderate	"provide"	"provide"	7.8s

# 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.