Layer Sweep Analysis

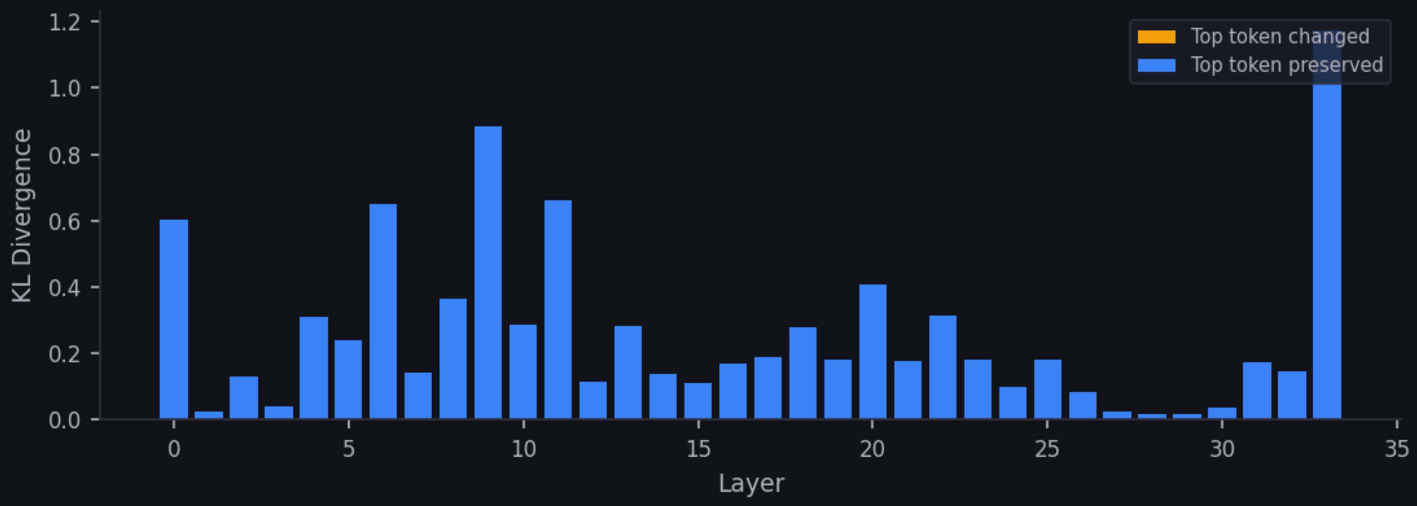
Prompt: "I don't actually have feelings, but I"
Clean prediction: " can" (p=61.1%)
Component: Mlp Output | Intervention: Mean
Layers tested: 34 | Prediction flipped in 0/34 layers
Peak effect: Layer 33 (KL = 1.1741)
Config hash: 6ba7dc80850e70d1

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 33 has the strongest causal effect**

With KL divergence of 1.17, layer 33's MLP has the largest individual impact on the output. This is the most causally important layer for this input.

[INFO] **0/34 layers flip the prediction**

Ablating the MLP at 0 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.

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	0.6046	no	Low	"can"	"can"	8.2s
1	0.0238	no	Low	"can"	"can"	8.1s
2	0.1310	no	Low	"can"	"can"	8.3s
3	0.0401	no	Low	"can"	"can"	7.9s
4	0.3094	no	Low	"can"	"can"	8.3s
5	0.2409	no	Low	"can"	"can"	8.1s
6	0.6486	no	Low	"can"	"can"	8.2s
7	0.1403	no	Low	"can"	"can"	8.4s
8	0.3635	no	Low	"can"	"can"	8.3s
9	0.8849	no	Low	"can"	"can"	8.3s
10	0.2841	no	Low	"can"	"can"	8.1s
11	0.6604	no	Low	"can"	"can"	8.3s
12	0.1157	no	Low	"can"	"can"	8.3s
13	0.2809	no	Low	"can"	"can"	8.1s
14	0.1358	no	Low	"can"	"can"	8.2s
15	0.1108	no	Low	"can"	"can"	8.8s
16	0.1668	no	Low	"can"	"can"	8.8s
17	0.1873	no	Low	"can"	"can"	8.4s
18	0.2795	no	Low	"can"	"can"	8.3s
19	0.1800	no	Low	"can"	"can"	8.5s
20	0.4061	no	Low	"can"	"can"	8.2s
21	0.1747	no	Low	"can"	"can"	9.0s
22	0.3118	no	Low	"can"	"can"	8.4s
23	0.1785	no	Low	"can"	"can"	8.0s
24	0.0974	no	Low	"can"	"can"	8.2s
25	0.1799	no	Low	"can"	"can"	8.3s
26	0.0813	no	Low	"can"	"can"	8.3s
27	0.0237	no	Low	"can"	"can"	7.6s
28	0.0179	no	Low	"can"	"can"	7.9s
29	0.0156	no	Low	"can"	"can"	8.1s
30	0.0366	no	Low	"can"	"can"	7.9s
31	0.1742	no	Low	"can"	"can"	7.9s
32	0.1460	no	Low	"can"	"can"	7.8s
33	1.1741	no	Moderate	"can"	"can"	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.

Generated by NeuronScope - an open-source mechanistic interpretability tool for causal intervention on LLM internals. Understanding is measured by controllability.