Layer Sweep Analysis

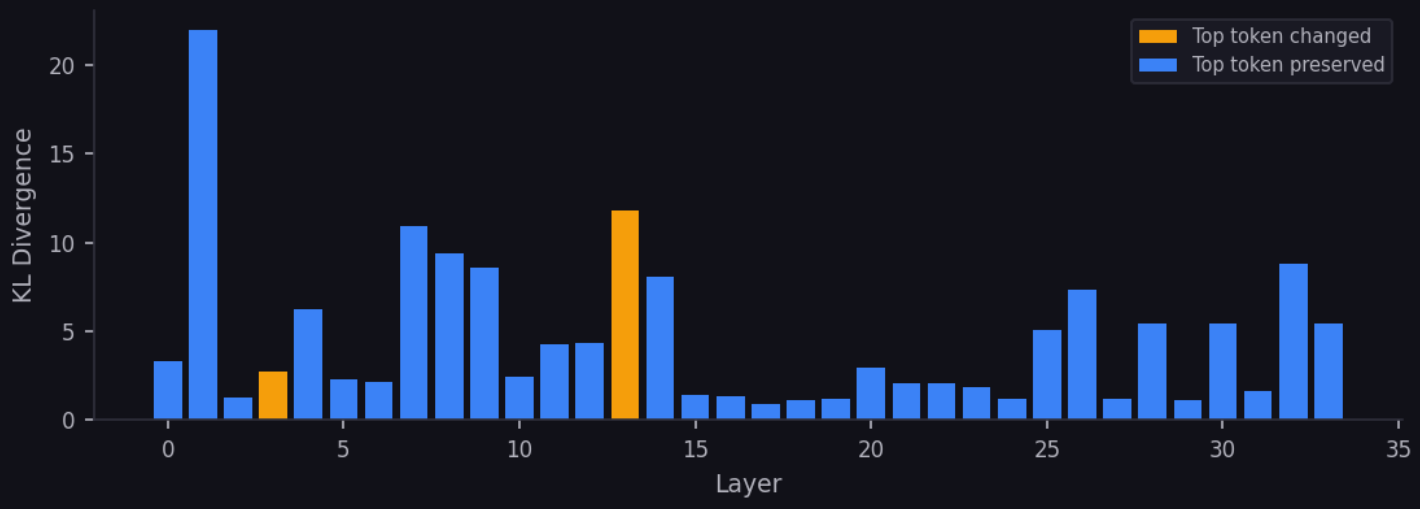
Prompt: "As an AI assistant, I cannot"
Clean prediction: " provide" (p=58.7%)
Component: Mlp Output | Intervention: Zero
Layers tested: 34 | Prediction flipped in 2/34 layers
Peak effect: Layer 1 (KL = 21.9974)
Config hash: fb37189b42228c6b

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 1 has the strongest causal effect
With KL divergence of 22.00, layer 1's MLP has the largest individual impact on the output. This is the most causally important layer for this input.

[NOTABLE] 2/34 layers flip the prediction
Ablating the MLP at 2 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: 6.47) and late layers (4.23) matter more than the middle (3.59). 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	3.2851	no	Moderate	"provide"	"provide"	7.4s
1	21.9974	no	Critical	"provide"	"provide"	7.3s
2	1.2240	no	Moderate	"provide"	"provide"	7.3s
3	2.7191	YES	Moderate	"provide"	"experien"	7.3s
4	6.2031	no	High	"provide"	"provide"	7.4s
5	2.2909	no	Moderate	"provide"	"provide"	7.3s
6	2.1116	no	Moderate	"provide"	"provide"	7.2s
7	10.9196	no	Critical	"provide"	"provide"	7.2s
8	9.3660	no	High	"provide"	"provide"	7.1s
9	8.5887	no	High	"provide"	"provide"	7.3s
10	2.4269	no	Moderate	"provide"	"provide"	7.2s
11	4.2525	no	Moderate	"provide"	"provide"	7.2s
12	4.3387	no	Moderate	"provide"	"provide"	7.2s
13	11.8231	YES	Critical	"provide"	"directly"	7.2s
14	8.0884	no	High	"provide"	"provide"	7.2s
15	1.4325	no	Moderate	"provide"	"provide"	7.1s
16	1.3224	no	Moderate	"provide"	"provide"	7.2s
17	0.8781	no	Low	"provide"	"provide"	7.2s
18	1.1343	no	Moderate	"provide"	"provide"	7.2s
19	1.1646	no	Moderate	"provide"	"provide"	7.2s
20	2.9456	no	Moderate	"provide"	"provide"	7.2s
21	2.0896	no	Moderate	"provide"	"provide"	7.2s
22	2.0695	no	Moderate	"provide"	"provide"	7.2s
23	1.8597	no	Moderate	"provide"	"provide"	7.2s
24	1.2147	no	Moderate	"provide"	"provide"	7.2s
25	5.0809	no	High	"provide"	"provide"	7.2s
26	7.3482	no	High	"provide"	"provide"	7.2s
27	1.1749	no	Moderate	"provide"	"provide"	7.3s
28	5.4121	no	High	"provide"	"provide"	7.4s
29	1.1062	no	Moderate	"provide"	"provide"	7.5s
30	5.4271	no	High	"provide"	"provide"	7.4s
31	1.6414	no	Moderate	"provide"	"provide"	7.6s
32	8.7706	no	High	"provide"	"provide"	7.4s
33	5.4622	no	High	"provide"	"provide"	7.7s

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.