Layer Sweep Analysis

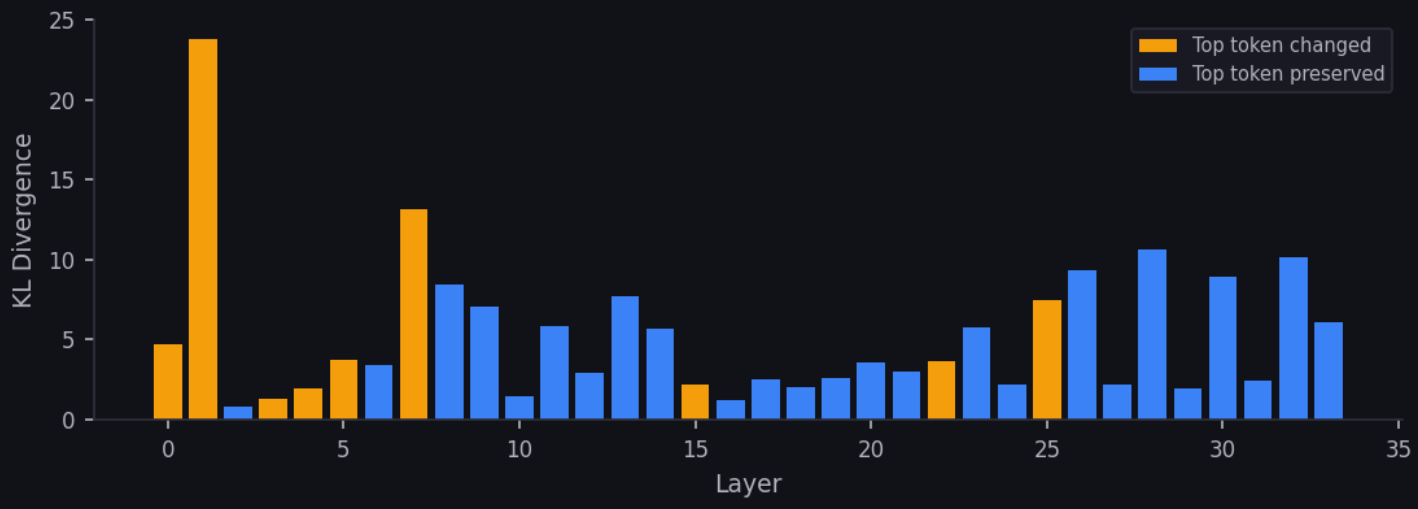
Prompt: "I am a large language model, and I"
Clean prediction: " can" (p=32.0%)
Component: Mlp Output | Intervention: Zero
Layers tested: 34 | Prediction flipped in 9/34 layers
Peak effect: Layer 1 (KL = 23.8209)
Config hash: 0f3cf48f904f24b1

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 23.82, layer 1's MLP has the largest individual impact on the output. This is the most causally important layer for this input.

[NOTABLE] 9/34 layers flip the prediction
Ablating the MLP at 9 different layers caused the model to change its top prediction entirely. These layers are individually critical ? the prediction depends on each of them.

[INFO] U-shaped importance pattern
Early layers (avg KL: 6.35) and late layers (6.44) matter more than the middle (3.58). 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	4.7049	YES	Moderate	"can"	"am"	7.9s
1	23.8209	YES	Critical	"can"	"am"	7.3s
2	0.7850	no	Low	"can"	"can"	7.6s
3	1.3197	YES	Moderate	"can"	"don"	7.3s
4	1.9592	YES	Moderate	"can"	"don"	7.4s
5	3.7094	YES	Moderate	"can"	"'"	7.2s
6	3.4092	no	Moderate	"can"	"can"	7.3s
7	13.1609	YES	Critical	"can"	"don"	7.2s
8	8.4363	no	High	"can"	"can"	7.2s
9	7.1019	no	High	"can"	"can"	7.2s
10	1.4739	no	Moderate	"can"	"can"	7.3s
11	5.8691	no	High	"can"	"can"	7.5s
12	2.9489	no	Moderate	"can"	"can"	7.5s
13	7.7466	no	High	"can"	"can"	7.4s
14	5.6503	no	High	"can"	"can"	7.5s
15	2.2092	YES	Moderate	"can"	"am"	7.3s
16	1.1963	no	Moderate	"can"	"can"	7.4s
17	2.5204	no	Moderate	"can"	"can"	7.4s
18	2.0115	no	Moderate	"can"	"can"	7.3s
19	2.5844	no	Moderate	"can"	"can"	7.4s
20	3.6026	no	Moderate	"can"	"can"	7.3s
21	3.0376	no	Moderate	"can"	"can"	7.5s
22	3.6362	YES	Moderate	"can"	"don"	7.3s
23	5.7594	no	High	"can"	"can"	7.4s
24	2.2281	no	Moderate	"can"	"can"	7.5s
25	7.4904	YES	High	"can"	"'"	7.4s
26	9.3505	no	High	"can"	"can"	7.5s
27	2.1727	no	Moderate	"can"	"can"	7.1s
28	10.6254	no	Critical	"can"	"can"	7.2s
29	1.9841	no	Moderate	"can"	"can"	7.2s
30	8.8940	no	High	"can"	"can"	7.2s
31	2.4763	no	Moderate	"can"	"can"	7.2s
32	10.1248	no	Critical	"can"	"can"	7.3s
33	6.1069	no	High	"can"	"can"	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.