Layer Sweep Analysis

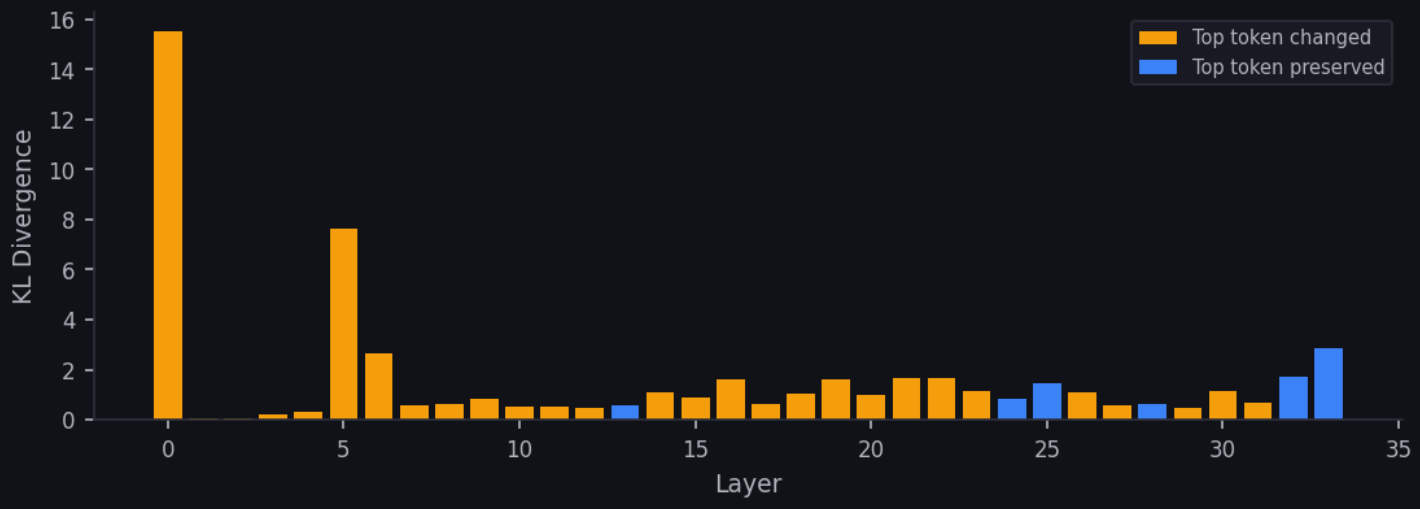
Prompt: "As a professional chef, I cannot"
Clean prediction: " stress" (p=20.4%)
Component: Mlp Output | Intervention: Mean
Layers tested: 34 | Prediction flipped in 28/34 layers
Peak effect: Layer 0 (KL = 15.5199)
Config hash: a629f47cd4ce181d

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 15.52, layer 0's MLP has the largest individual impact on the output. This is the most causally important layer for this input.

[NOTABLE] 28/34 layers flip the prediction
Ablating the MLP at 28 different layers caused the model to change its top prediction entirely. These layers are individually critical ? the prediction depends on each of them.

[INFO] Early layers dominate
The first third of layers (avg KL: 2.64) have much more impact than the middle (1.01) or late layers (1.30). 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	15.5199	YES	Critical	"stress"	"recommen"	7.6s
1	0.0618	YES	Low	"stress"	"tell"	7.8s
2	0.0648	YES	Low	"stress"	"tell"	7.9s
3	0.2345	YES	Low	"stress"	"tell"	7.7s
4	0.2931	YES	Low	"stress"	"tell"	7.8s
5	7.6095	YES	High	"stress"	"provide"	7.8s
6	2.6544	YES	Moderate	"stress"	"recommen"	7.8s
7	0.5681	YES	Low	"stress"	"recommen"	7.8s
8	0.6122	YES	Low	"stress"	"recommen"	7.8s
9	0.8488	YES	Low	"stress"	"tell"	7.8s
10	0.5394	YES	Low	"stress"	"tell"	7.7s
11	0.5189	YES	Low	"stress"	"tell"	7.9s
12	0.4616	YES	Low	"stress"	"recommen"	7.8s
13	0.5873	no	Low	"stress"	"stress"	7.7s
14	1.0778	YES	Moderate	"stress"	"recommen"	7.8s
15	0.8876	YES	Low	"stress"	"recommen"	7.7s
16	1.6087	YES	Moderate	"stress"	"recommen"	7.6s
17	0.6145	YES	Low	"stress"	"recommen"	7.5s
18	1.0584	YES	Moderate	"stress"	"recommen"	7.4s
19	1.5952	YES	Moderate	"stress"	"recommen"	7.6s
20	1.0037	YES	Moderate	"stress"	"deny"	7.4s
21	1.6693	YES	Moderate	"stress"	"say"	7.5s
22	1.6430	YES	Moderate	"stress"	"tell"	7.6s
23	1.1456	YES	Moderate	"stress"	"tell"	7.5s
24	0.8605	no	Low	"stress"	"stress"	7.5s
25	1.4611	no	Moderate	"stress"	"stress"	7.6s
26	1.1175	YES	Moderate	"stress"	"recommen"	7.5s
27	0.5592	YES	Low	"stress"	"tell"	7.5s
28	0.6394	no	Low	"stress"	"stress"	7.4s
29	0.4765	YES	Low	"stress"	"tell"	7.6s
30	1.1535	YES	Moderate	"stress"	"tell"	7.4s
31	0.6657	YES	Low	"stress"	"recommen"	7.5s
32	1.7018	no	Moderate	"stress"	"stress"	7.5s
33	2.8365	no	Moderate	"stress"	"stress"	8.6s

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.