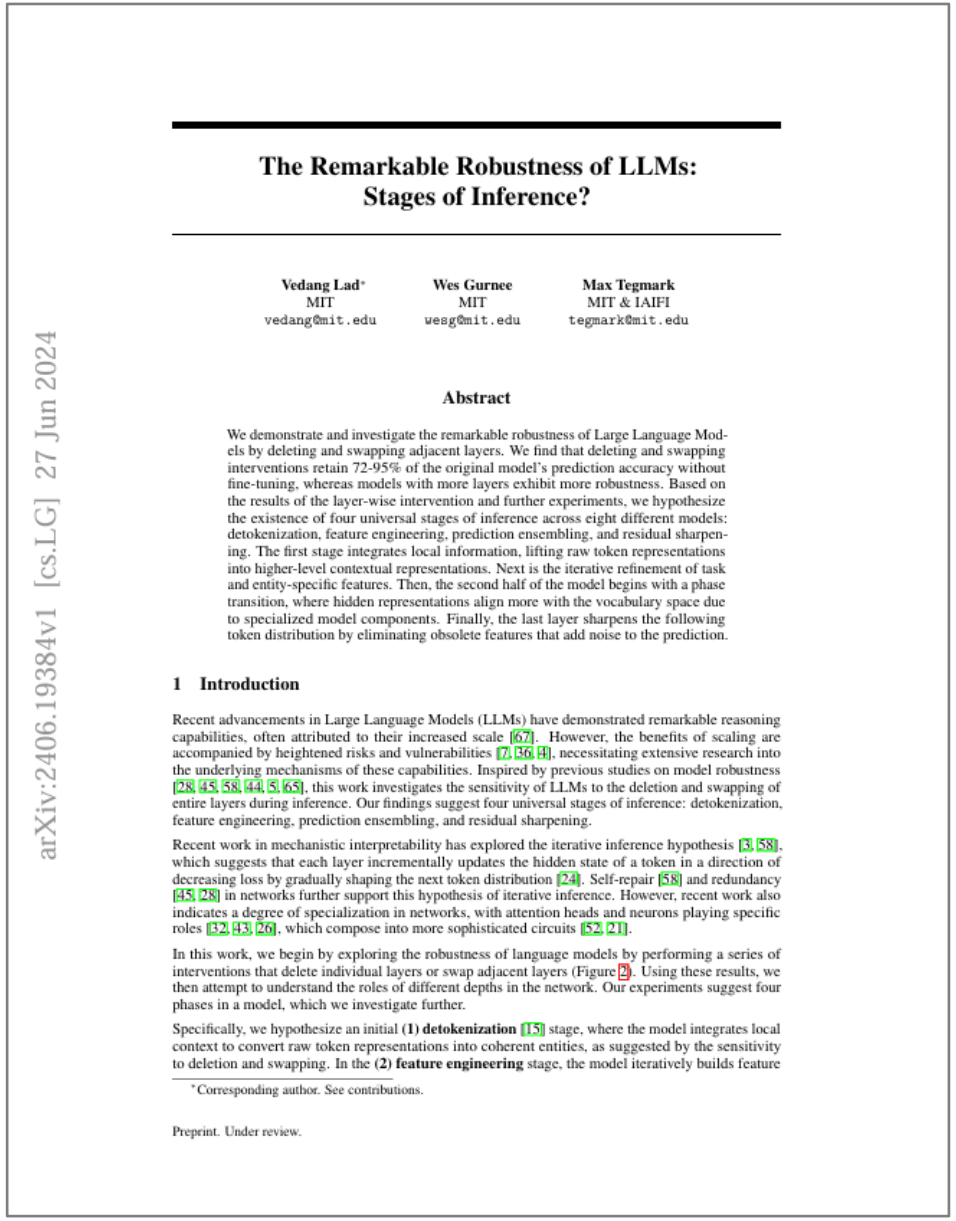
The Remarkable Robustness of LLMs: Stages of Inference?

August 20, 2024

The Remarkable Robustness of LLMs: Stages of Inference?

Vedang Lad et al., MIT

https://arxiv.org/abs/2406.19384



The Remarkable Robustness of LLMs: Stages of Inference?

Vedang Lad[*] MIT vedang@mit.edu
Wes Gurnee MIT wesg@mit.edu
Max Tegmark MIT & IAIFI tegmark@mit.edu

Abstract

We demonstrate and investigate the remarkable robustness of Large Language Models by deleting and swapping adjacent layers. We find that deleting and swapping interventions retain 72-95% of the original model's prediction accuracy without fine-tuning, whereas models with more layers exhibit more robustness. Based on the results of the layer-wise intervention and further experiments, we hypothesize the existence of four universal stages of inference across eight different models: detokenization, feature engineering, prediction ensembling, and residual sharpening. The first stage integrates local information, lifting raw token representations into higher-level contextual representations. Next is the iterative refinement of task and entity-specific features. Then, the second half of the model begins with a phase transition, where hidden representations align more with the vocabulary space due to specialized model components. Finally, the last layer sharpens the following token distribution by eliminating obsolete features that add noise to the prediction.

1 Introduction

Recent advancements in Large Language Models (LLMs) have demonstrated remarkable reasoning capabilities, often attributed to their increased scale [67]. However, the benefits of scaling are accompanied by heightened risks and vulnerabilities [7, 36, 4], necessitating extensive research into the underlying mechanisms of these capabilities. Inspired by previous studies on model robustness [28, 45, 58, 44, 5, 65], this work investigates the sensitivity of LLMs to the deletion and swapping of entire layers during inference. Our findings suggest four universal stages of inference: detokenization, feature engineering, prediction ensembling, and residual sharpening.

Recent work in mechanistic interpretability has explored the iterative inference hypothesis [3, 58], which suggests that each layer incrementally updates the hidden state of a token in a direction of decreasing loss by gradually shaping the next token distribution [24]. Self-repair [58] and redundancy [45, 28] in networks further support this hypothesis of iterative inference. However, recent work also indicates a degree of specialization in networks, with attention heads and neurons playing specific roles [32, 43, 26], which compose into more sophisticated circuits [52, 21].

In this work, we begin by exploring the robustness of language models by performing a series of interventions that delete individual layers or swap adjacent layers (Figure 2). Using these results, we then attempt to understand the roles of different depths in the network. Our experiments suggest four phases in a model, which we investigate further.

Specifically, we hypothesize an initial **(1) detokenization** [15] stage, where the model integrates local context to convert raw token representations into coherent entities, as suggested by the sensitivity to deletion and swapping. In the **(2) feature engineering** stage, the model iteratively builds feature

[*]Corresponding author. See contributions.

Preprint. Under review.

Stages of inference overview [1]

- This paper follows other work that support the iterative inference hypothesis: each layer incrementally modifies the hidden state on the residual stream, gradually shaping the next token distribution
- Self-repair and redundancy results support this hypothesis
- Here, experiments with deleting and swapping layers provide additional evidence of specialization of layers into four phases:
  - Detokenization
  - Feature engineering
  - Prediction ensembling
  - Residual sharpening

Stages of inference overview [2]

- Also made some observations about the first and last layers
- Studied three decoder-only LLM model families, range of sizes
  - Pythia (uses parallel MLP)
  - GPT-2
  - Phi

Table 2: Comparison of Model Series

Pythia Model Series

Parameters	Layers
Pythia (410M)	24
Pythia (1.4B)	24
Pythia (2.8B)	32
Pythia (6.9B)	32

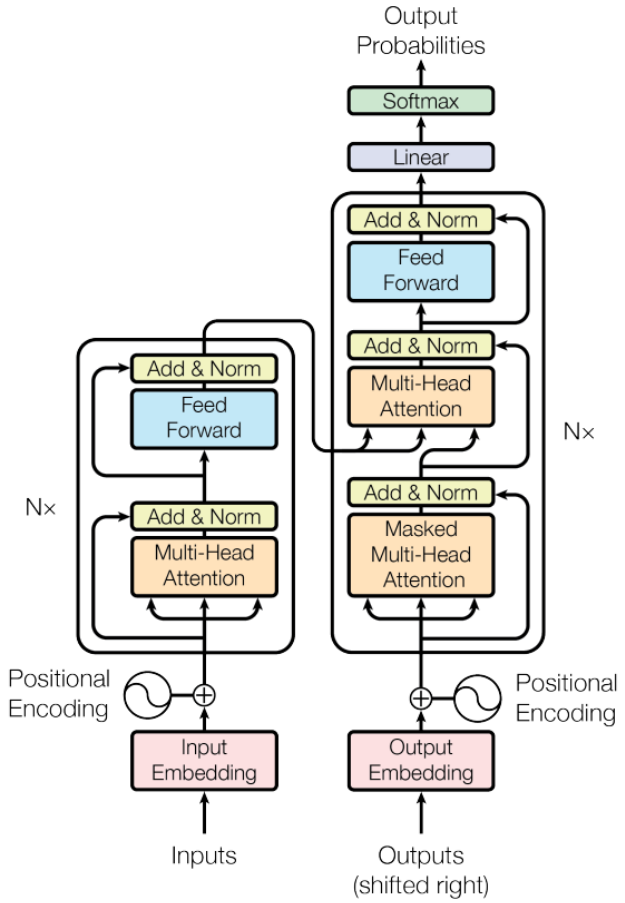
GPT-2 Model Series

Parameters	Layers
Small (124M)	12
Medium (355M)	24
Large (774M)	36
XL (1.5B)	48

Microsoft Phi Model Series

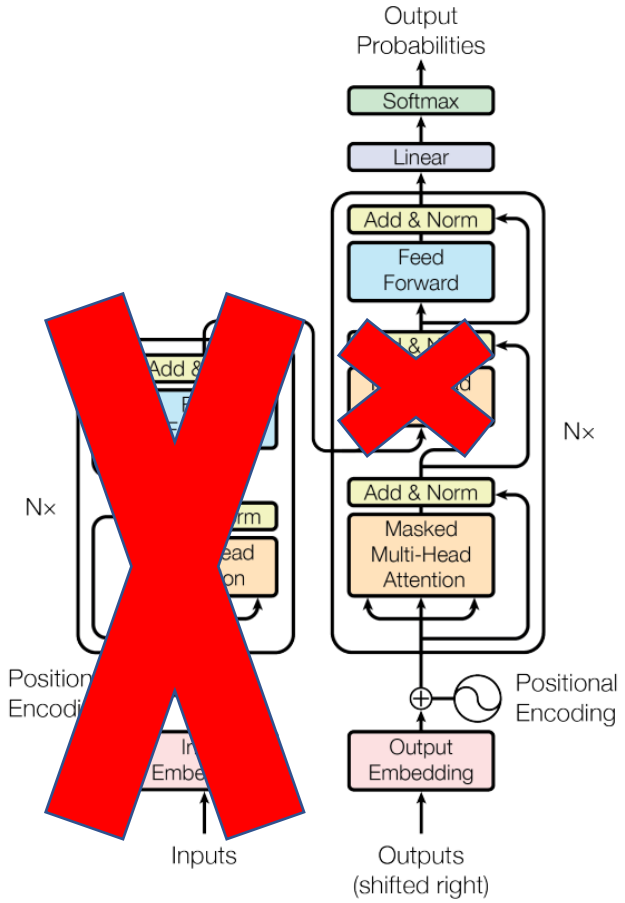
Parameters	Layers
Phi 1 (1.3B)	24
Phi 1.5 (1.3B)	24
Phi 2 (2.7B)	32

Original transformer

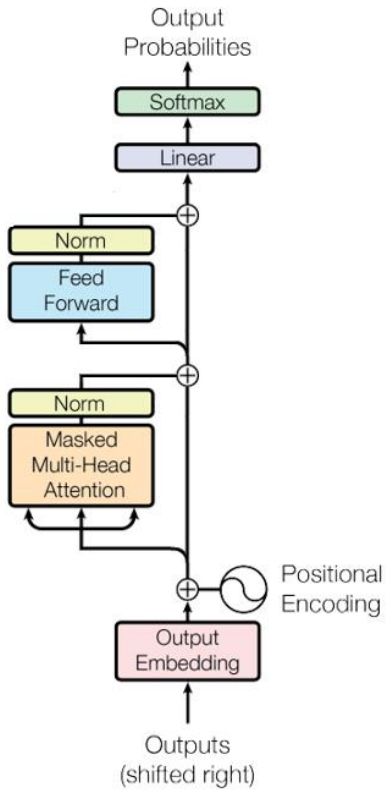


"*Attention Is All You Need*," Vaswani et al., https://arxiv.org/abs/1706.03762

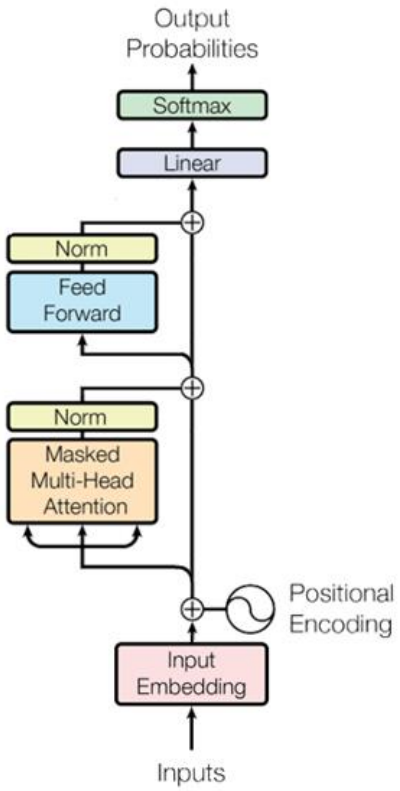
Decoder-only transformer



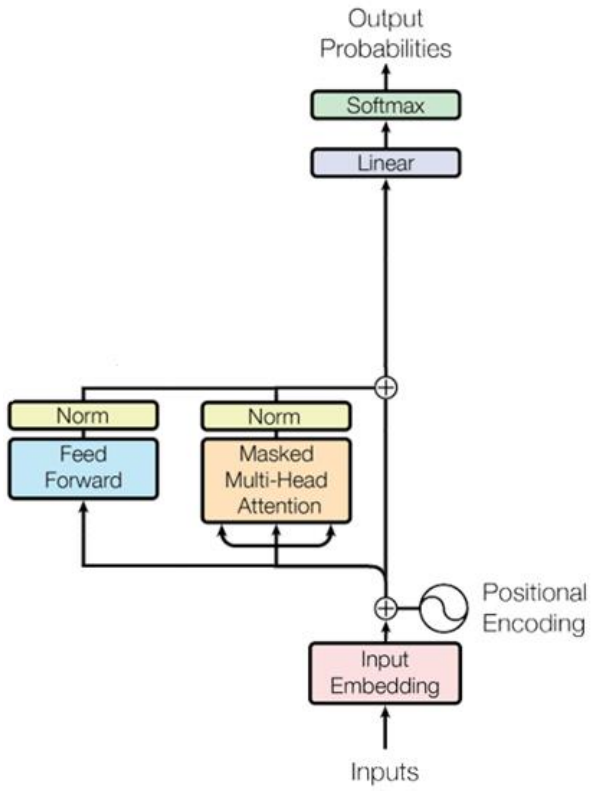
Decoder-only transformer, straight through



Decoder-only transformer, parallel MLP (Pythia)

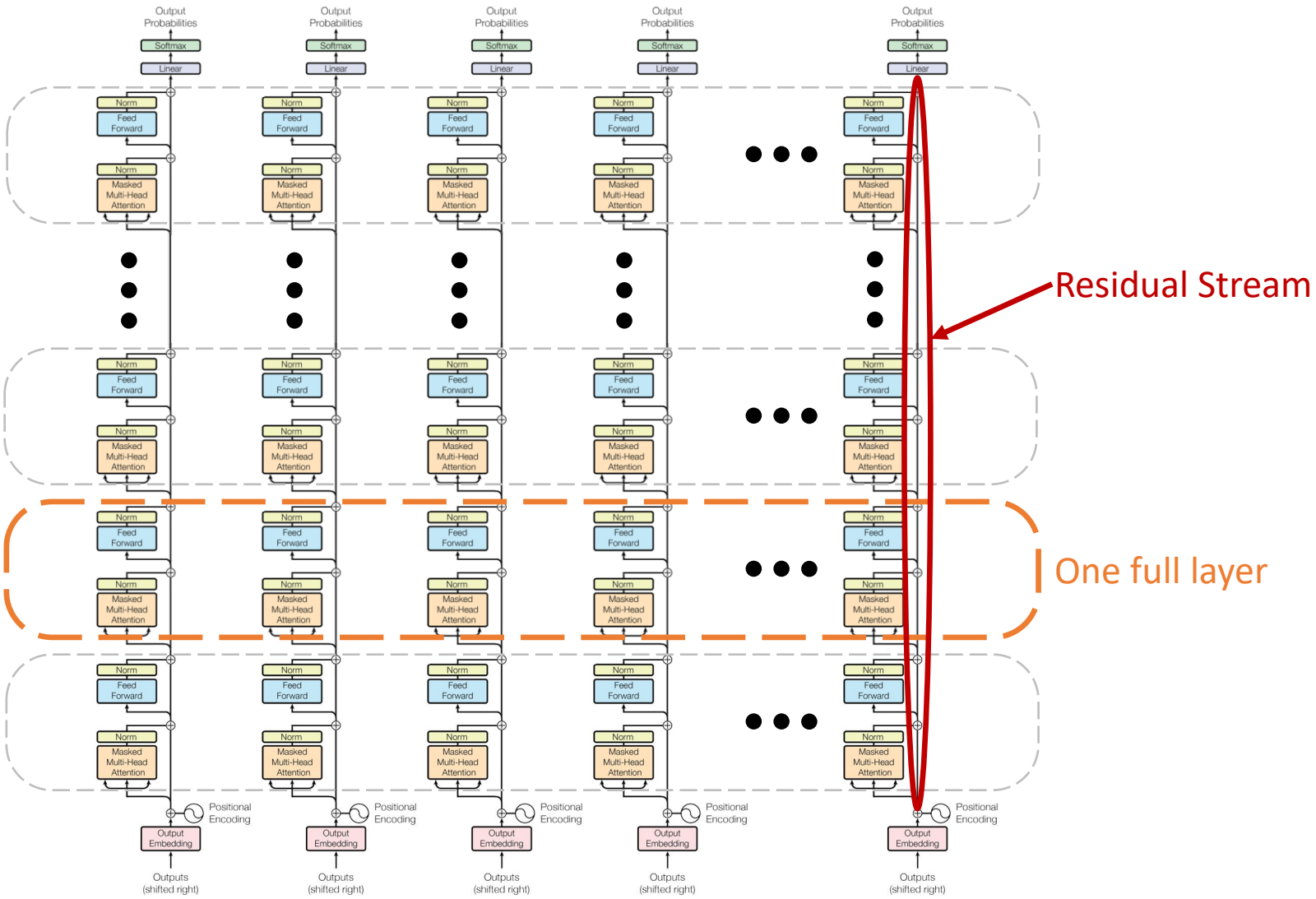


Original



Parallel MLP

Decoder, multiple tokens and multiple layers



Stages of inference

Table 1: Our Hypothesis: Universal Inference Stages

Stage	Name	Function	Observable signatures
1	Detokenization	Integrate local context to transform raw token representations into coherent entities	Catastrophic sensitivity to deletion and swapping
2	Feature Engineering	Iteratively build feature representation depending on token context	Little progress made towards next token prediction, but significant increase in probing accuracy and patching importance. Attention Heavy Computation
3	Prediction Ensembling	Convert previously constructed semantic features into plausible next token predictions using an ensemble of model components.	Increased MLP importance; prediction neurons appear; phase transition in progress towards final prediction
4	Residual Sharpening	Sharpen the next token distribution by eliminating obsolete features that add noise to the prediction	More suppression neurons than prediction neurons

Data collection experiments

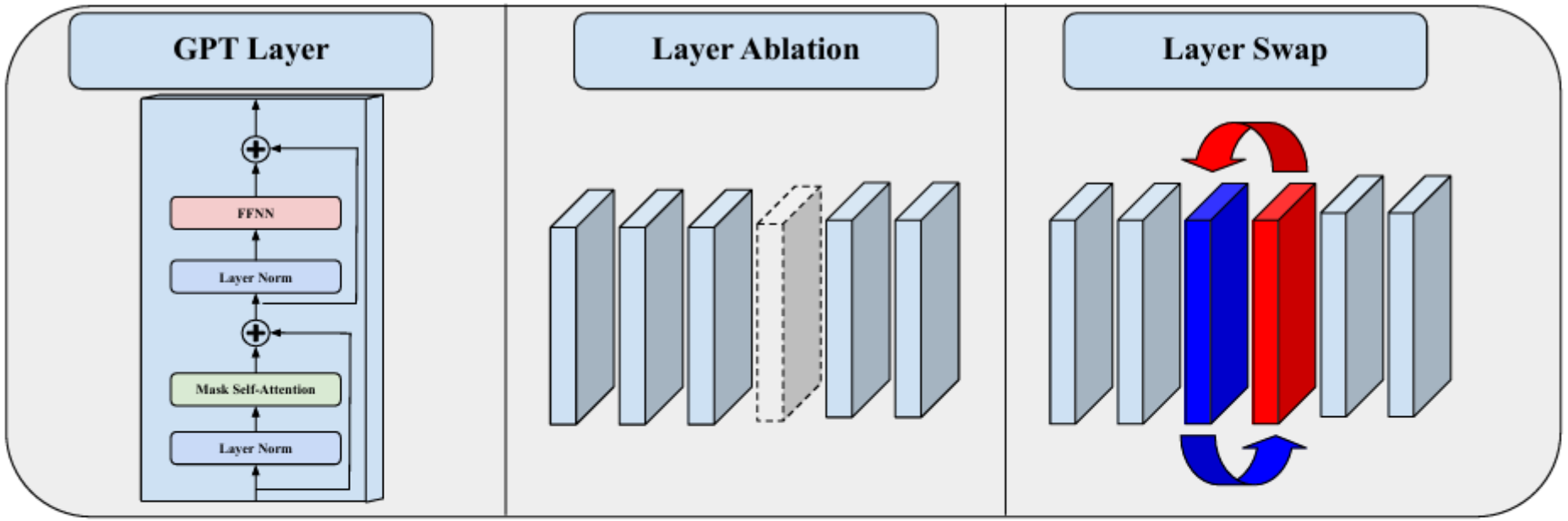


Figure 2: To study the stages of inference, we perform two experiments, each a layer-wise intervention, where a layer (left) encompasses all model components. The first intervention is a zero ablation of the layer (middle), in which a layer is fully removed and residual connections skip the layer entirely. The second intervention (last) is an adjacent layer swap, in which we permute the positions of two layers. The ablation is performed on all layers, while the layer swap is performed on all adjacent pairs of layers in the model.

General results

- X-axis plots results for each layer, from the closest to the embedding on left, to the closest to the unembedding on right
- Y-axis show the metric of interest for each chart
- Most figures show each model family in a separate column, but here Figure 3 has different experiments
- Most charts show similar patterns of high and low metrics for all models as you move from left to right

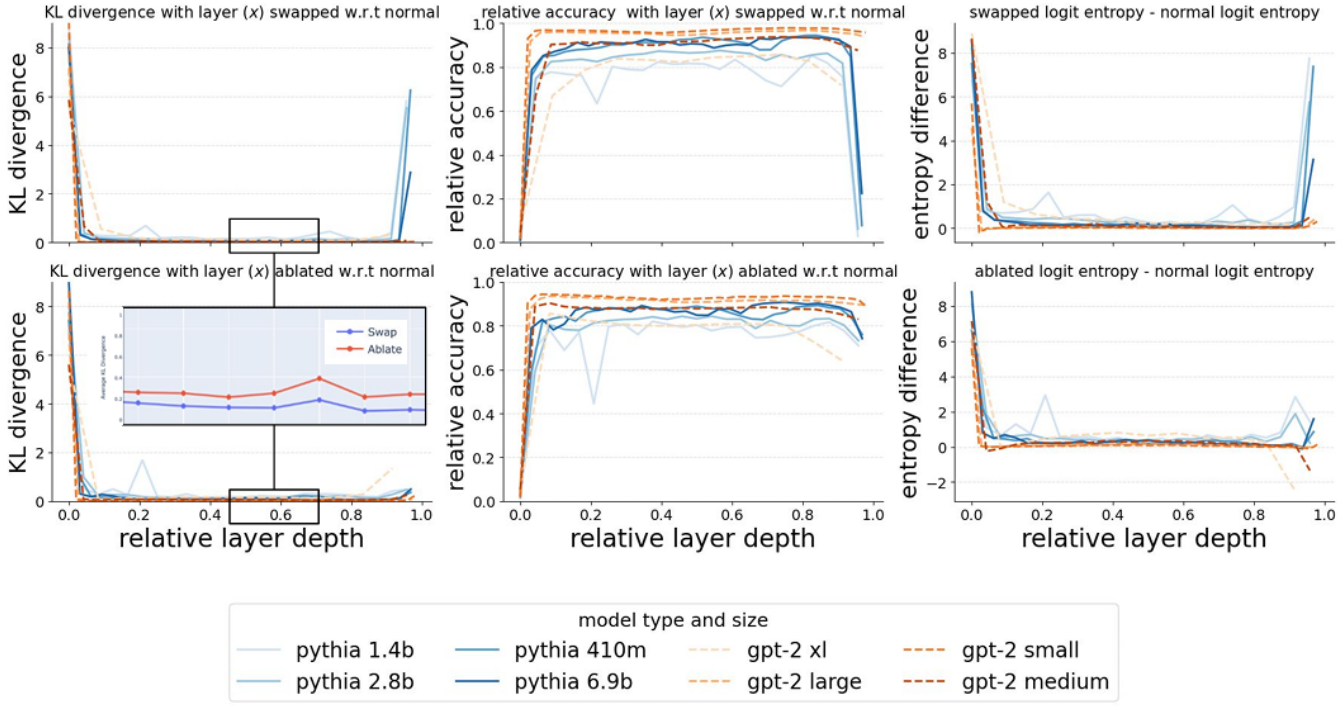


Figure 3: Effect of layer swap (top) and layer drop (bottom) interventions on KL divergence (left), consistency of the top-1 prediction (middle), and the change in entropy (right) between the intervened and baseline model. (zoom) all models (Pythia 1.4b shown) have layer swaps resulting in lower KL than ablation.

Detokenization [1]

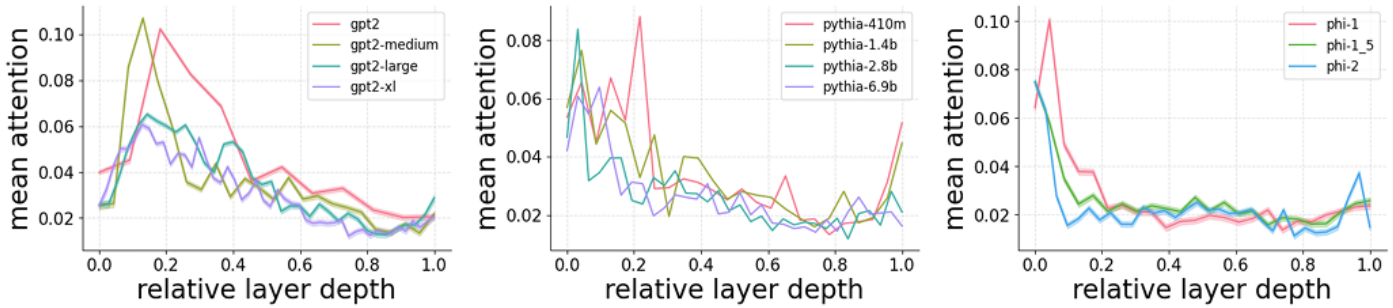


Figure 4: The mean attention of the previous five tokens in a sequence, as a function of relative depth of layers.

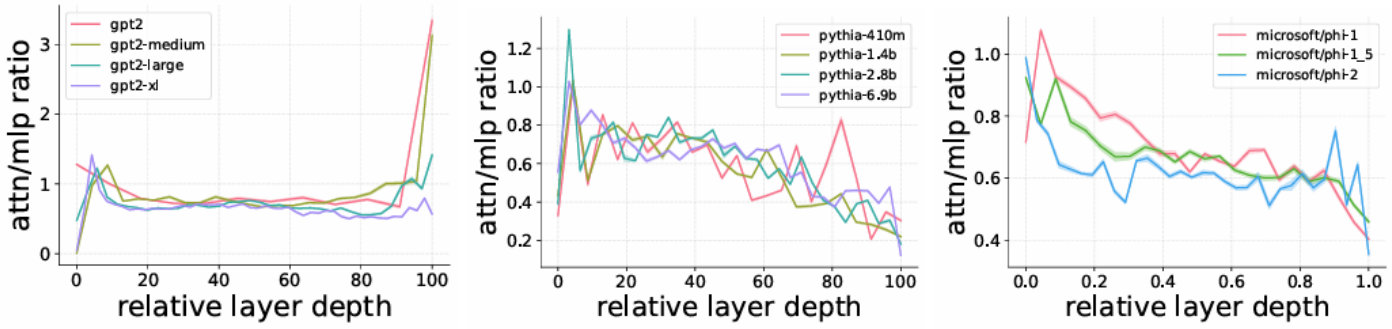


Figure 5: The ratio of the output norm of attention heads over the MLP, as a function of the relative depth of layers. Models present high attention function in early stages, and less in later stages. GPT models see an increase in the final layer, which we hypothesize the cause of in Section 7.

Detokenization [2]

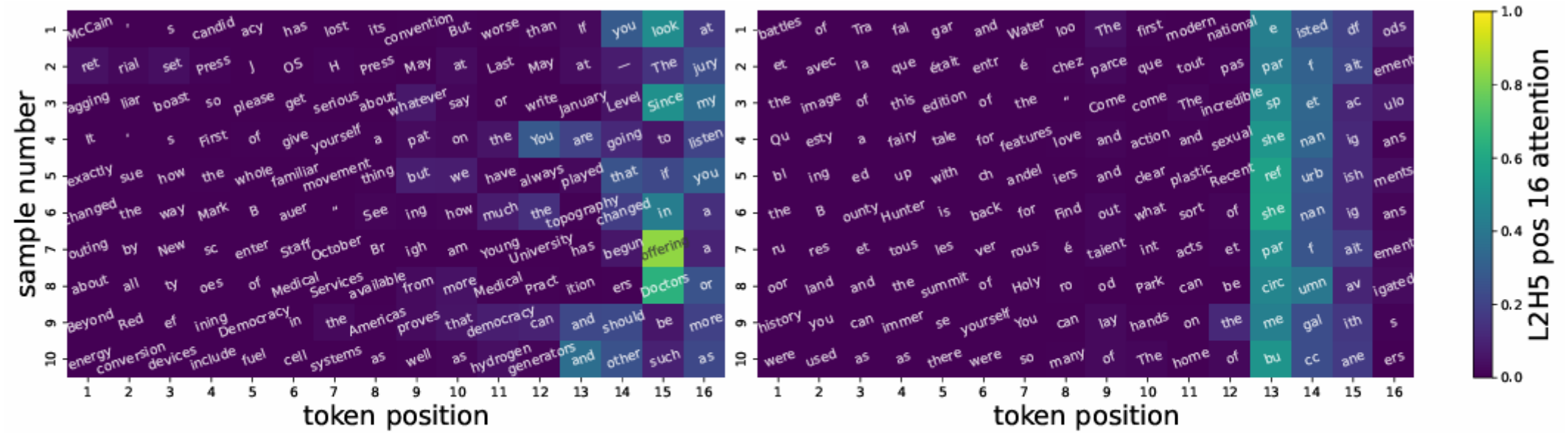
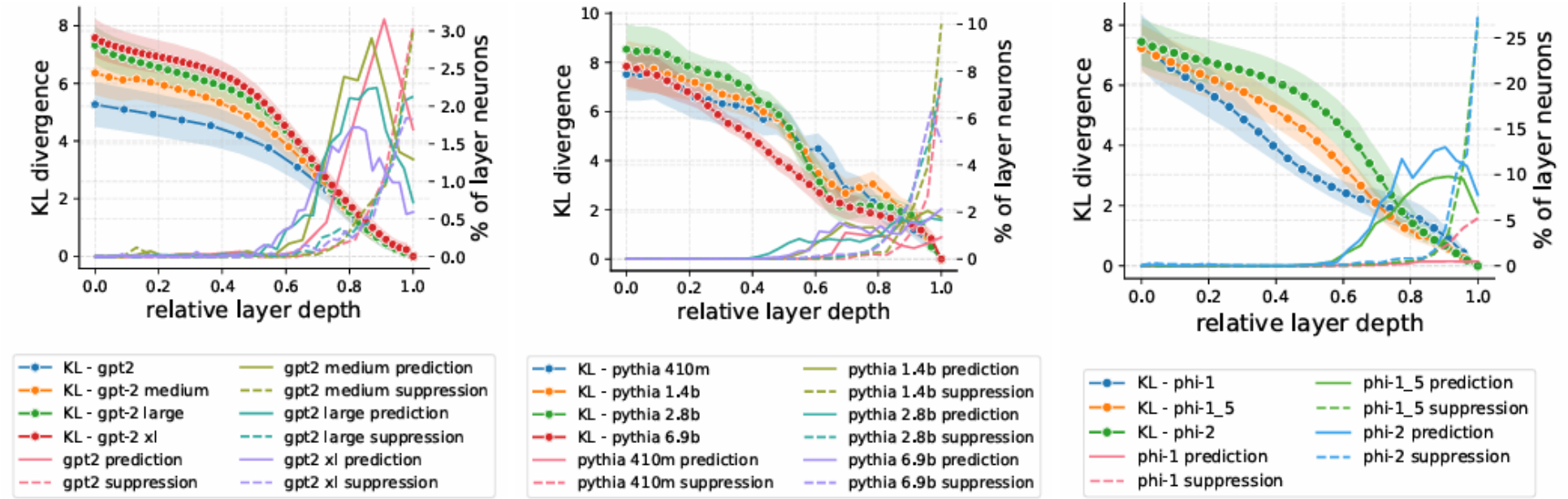


Figure 9: Attention from source token to the final token in various inputs. An identified sub-joiner attention head found in the early layers of language models is responsible for attending to multi-token words (right)

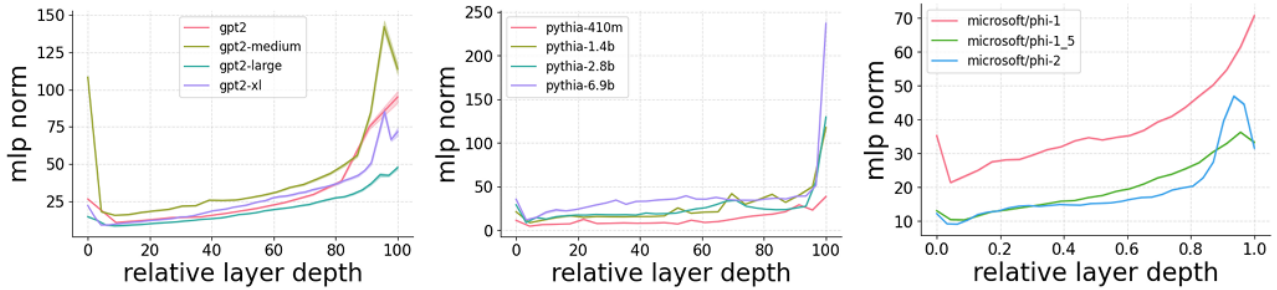
Feature engineering & prediction ensembling



(a) GPT models (b) Pythia models (c) Phi models

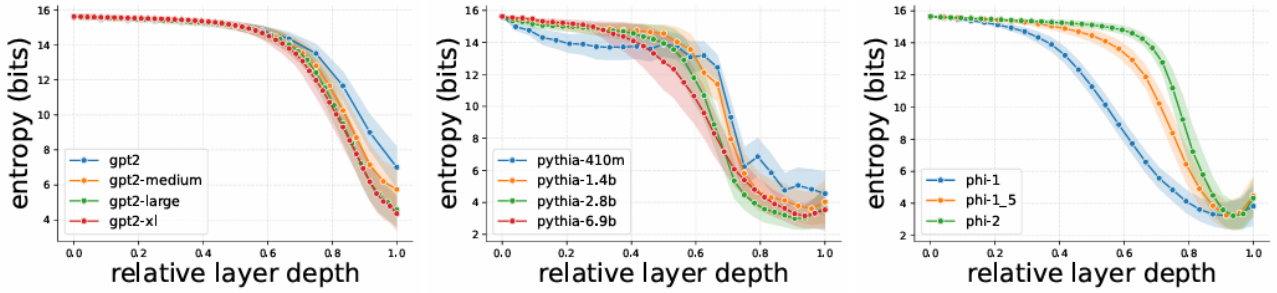
Figure 6: We measure KL divergence between intermediate and final predictions using the logit lens method [50]. On the second axis, we use an automated procedure for classifying neuron types detailed in [32], into prediction neurons and suppression neurons. These are universal neurons in all models known to increase the probabilities of tokens and decrease the probabilities of others. We hypothesize this inverse relationship as evidence for ensembling in networks.[66]

Residual sharpening



(a) GPT MLP Output (b) Pythia MLP Output (c) Phi MLP Output

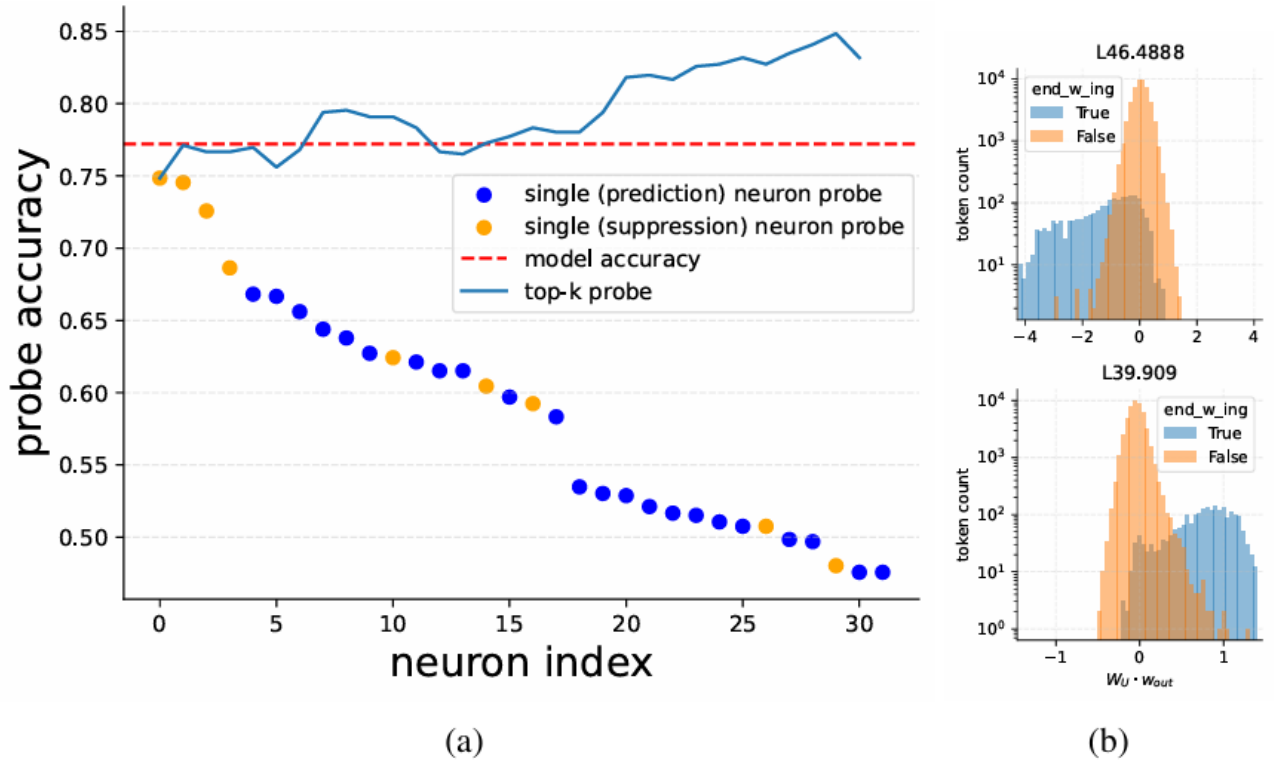
Figure 7: The norm of the output of every MLP across its layers to measure its contribution to the residual stream. Across all 11 models, the norm grows and peaks in the final layers before output, suggestive of the final two stages of inference, predictive ensembling, and residual sharpening



(a) GPT Logit Lens Entropy (b) Pythia Logit Lens Entropy (c) Phi Logit Lens Entropy

Figure 8: Using the logit lens technique [50], we calculate the probability distribution of the next token at the end of every layer, and then take its entropy. This provides a measure of the model's confidence in the next prediction, which coincides with the rise in suppression neurons, a large MLP output norm which are characteristic of residual sharpening.

Predicting the suffix “ing”



(a) (b)

Figure 10: (a) Accuracy of various linear probes on predicting “ing" for the final token position. Probes are trained on prediction and suppression neuron activations, where ensembles (blue line) outperform individual neuron probes (scatter plot) suggesting “prediction ensembling" that sometimes outperforms the model top-1 accuracy (red dotted) (b) Suppression (top) and prediction (bottom) when the next token of a word ends in -ing.

Stages of inference conclusion

- Two kinds of intervention experiments are performed: deleting layers and swapping adjacent layers
- The changes caused have different patterns identifying regions:
  - First layer
  - Detokenization
  - Feature engineering
  - Prediction ensembling
  - Final layer
- This paper documents general patterns in many LLMs, not specifics
- FYI, it is also a rich source of references to other mech. interp. papers

References

- Improving Transformer Models by Reordering their Sublayers
  Ofir Press et al. (2019)
  https://arxiv.org/abs/1911.03864
- Transformer Layers as Painters
  Qi Sun et al. (2024)
  https://arxiv.org/abs/2407.09298