Why **do** attention heads **attend** where they do?

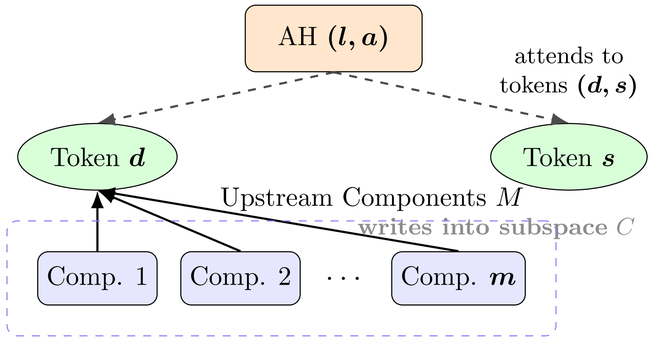
Pinpointing Attention-Causal Communication in Language Models

Gabriel Franco, Mark Crovella

THE BIG QUESTION: "When an attention head attends to a token pair, what features (signals) are *really* responsible?"

To answer this question, we define the **Attention Causal Communication Problem**.

Problem Statement: When a head attends to two tokens, we find the smallest set of upstream components (M) and the smallest subspace (C) such that if those components hadn't written to that subspace, the attention wouldn't have happened.



A visual overview of the Attention Causal Communication Problem.

The challenges & our approach

The challenges:

1. There is an infinite set of possible subspaces to search
2. Attention is a nonlinear function, making attribution challenging

Our approach:

1. Leverage SVD to generate candidate subspaces
2. Craft a linear replacement for Softmax that still captures causality

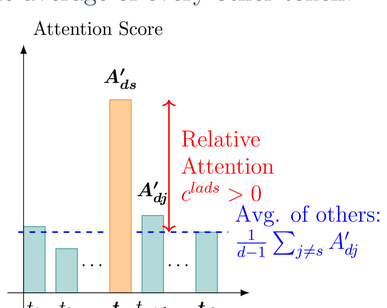
Relative attention

First, we define **Relative Attention (c^{lads})**, a replacement function that captures the Softmax behavior, measuring how much a head prefers to attend to token s over the average of every other token.

Properties:

1. It is **linear**
2. $A_{ds} > 1/\mathrm{d} \rightarrow c^{lads} > 0$

$$c^{\ell ads} = A'_{ds} - \frac{1}{d-1} \sum_{j \le d,\, j \neq s} A'_{dj}$$



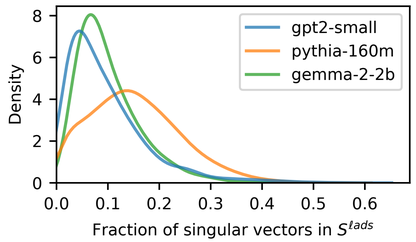
Causal tracing of attention

1. Finding the subspace C

We leverage the SVD of the $W_Q W_K^T$ matrix to break the attention score into a sum. Prior work has shown that the QK SVD exposes low dimensional communication channels, i.e., the set of terms above is sparse. We then find the smallest set of singular vectors (S^{lads}) that sums to the total Relative Attention. This is our causal subspace C.

$$c^{\ell ads} = \sum_{k \in \mathcal{S}^{\ell ads}} \mathbf{x}^{d\top} \mathbf{u}^k \sigma^k \mathbf{v}^{k\top} \mathbf{x}^s$$

Finding 1: this strategy consistently finds small subspaces



Signal subspaces have very low dimension

2. Finding the set of upstream components M

Causal structure: the residual stream is a sum of upstream contributions:

$$\mathbf{x}^d = \sum_c \mathbf{o}_c^d$$

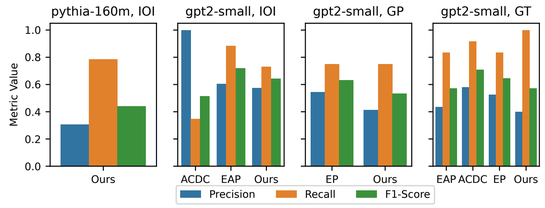
Tracing attention-causal communication: distributing relative attention over the upstream components finds the contribution of every upstream component into the subspace C. We pick the smallest set of upstream contributions M that sums to the total relative attention.

Paper

Code

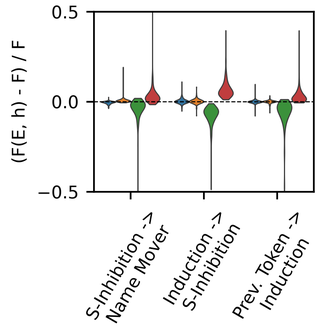
Results & impact

Finding 2: **automatic, per-prompt circuit discovery**. No counterfactuals, no patching, no SAEs

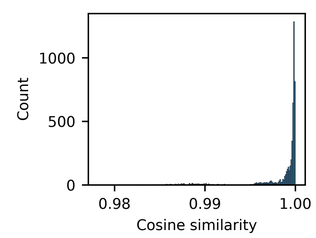


We find circuits with comparable performance against baselines. First-reported circuit used for comparison for every task/model.

Finding 3: the **low-dimensional** signals are **causal** to model performance

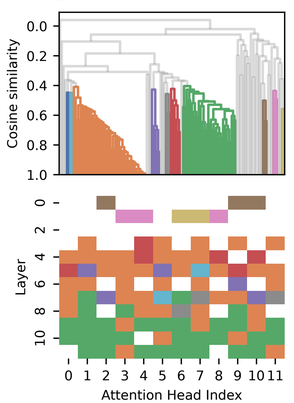


Interventions have causal effect on model performance

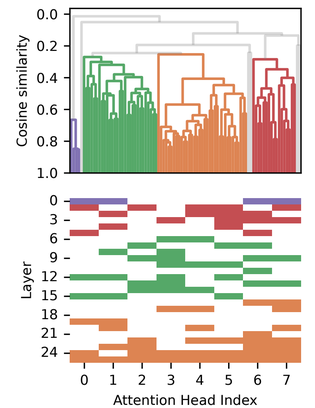


Cosine similarity between residuals before and after interventions is very high

Finding 4: our method exposes **control signals**, responsible for attention sinks



Source token control signals implement attention sink: GPT-2/IOI.



Destination token control signals implement attention sink: Gemma-2 2B/IOI.

References

Merullo, Jack, Carsten Eickhoff, and Ellie Pavlick. "Talking heads: Understanding inter-layer communication in transformer language models."

Pan, Xu, et al. "Dissecting Query-Key Interaction in Vision Transformers."