Motifs in Attention Patterns of Large Language Models

Michael Igorevich Ivanitskiy* **Cecilia Diniz Behn** **Samy Wu Fung**

Department of Applied Mathematics and Statistics, Colorado School of Mines

Abstract

Attention patterns in Large Language Models often exhibit clear structure, and analysis of these structures may provide insight into the functional roles of the heads that produce these patterns. However, there is little work addressing ways to systematically analyze or categorize attention heads using the patterns they produce. To address this gap, we 1) create a meaningful embedding of attention *patterns*; 2) use this embedding of attention patterns to construct a useful distance metric between the attention *heads* themselves; and 3) investigate the correspondence between known classes of attention heads, such as name mover heads and induction heads, with the groupings emerging in our embedding of attention heads.

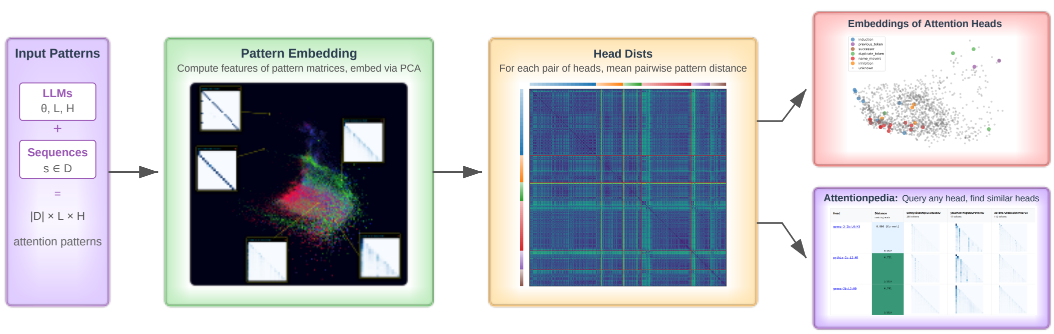

Figure 1: Attention patterns from diverse data and models are collected, features about them are computed, and the patterns are embedded in a meaningful latent space. Each head has a corresponding cloud of points, with one for each sample. For each pair of heads, the mean pairwise distance samples in their clouds is used to construct a distance matrix. This distance matrix can be used to embed the heads themselves, or as a tool to find similar heads to a head of interest.

1 Introduction

As Large Language Models (LLMs) [Radford et al., 2018, Vaswani et al., 2017] become ever more powerful and widely deployed, ensuring the safety and security of these systems becomes paramount. Mechanistic interpretability aims to help us understand the internals of AI systems in order to make them more trustworthy and safe, by mapping those internals to human-comprehensible algorithms and concepts [Sharkey et al., 2025, Räuker et al., 2023]. A key obstacle to interpretability is the large number of components present in modern LLMs, making the manual inspection of the components

*Corresponding author: mivanits@mines.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

prohibitively time consuming. Recent advances in using Sparse Autoencoders (SAEs) to decode the meanings of residual stream vectors have relied on the automatic tagging of learned sparse features with legible explanations using LLMs [Cunningham et al., 2023, Braun et al., 2024], but no such automatic tagging exists for attention patterns. SAEs have been used to attempt to identify the role of attention heads [Krzyzanowski et al., 2024, He et al., 2025], but have their own limitations [Leask et al., 2025], and these approaches discard any spatial information from the attention patterns.

Despite the presence of polysemanticity [Elhage et al., 2022] in attention heads [Elhage et al., 2022, Janiak et al., 2023], manual inspection of attention patterns can prove valuable in determining the function of the attention head that produced them [Olsson et al., 2022, Spies et al., 2025, Ivanitskiy et al., 2023][Wang et al., 2022, Figure 16]. Despite the presence of clearly visible structures in a variety of attention heads (Figure 3) and a variety of categories of attention heads identified [Olsson et al., 2022, Wang et al., 2022, Krzyzanowski et al., 2024, Ferrando and Voita, 2024, Ren et al., 2024, García-Carrasco et al., 2024], to our knowledge a taxonomy of attention patterns and the heads that produce them has not yet been developed [Zheng et al., 2024]. In this work, we embed the attention *pattern* matrices themselves using handcrafted features, and observe clear structure in the latent space of the embedding (section 2). Using these embeddings of patterns, we construct a metric of distance between attention heads, projecting this new embedding of attention *heads* to a viewable low-dimensional space where we compare our unsupervised embedding with known classes of attention heads (section 3).

By contrast with previous work [Clark et al., 2019, Vig, 2019, Yeh et al., 2023, Park et al., 2019], our method takes as input only the attention pattern matrices themselves, and does not directly utilize any token or residual stream information. Attempts to categorize heads only by the tokens or features of the residual stream they attend to [Krzyzanowski et al., 2024] are limited because heads may attend to similar parts of the residual stream but be quite different in their broader functionality (Figure 2, A_1 vs A_2), or, on the other hand, they may attend to vastly different parts of the residual stream but be similar in functionality (Figure 2, A_2 vs A_3). It is our hope that this work will accelerate research in interpretability by providing an incredibly cheap[2] way to cluster the functionality of attention heads.

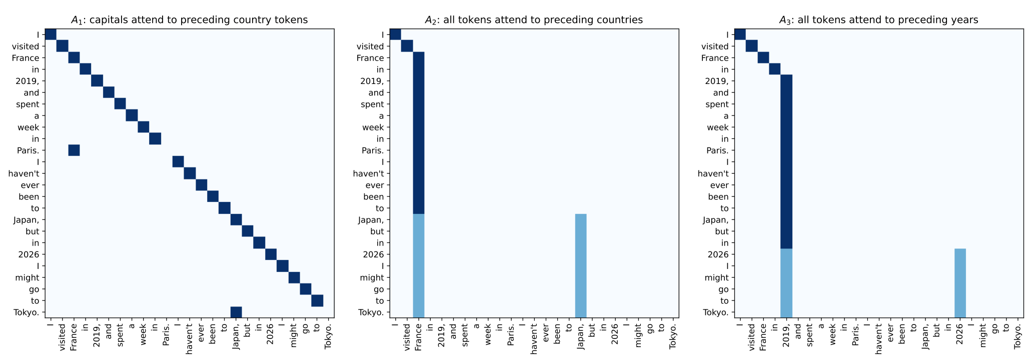

Figure 2: An **artificially constructed** example of classes of heads whose classification based on their attention patterns or functionality differs from a circuit or feature based classification, to explain our intuition. A_1 (left) has capital cities attend to their countries ("Tokyo" to "Japan", "Paris" to "France"; A_2 (center) has any token attend to tokens denoting a country; A_3 (right) has any token attend to tokens denoting a year. If we were to analyze the actual *tokens* each head attends to, or analyze the features attended to, we might conclude that A_1 and A_2 are more similar to each other than to A_3. Inspection of the attention patterns suggests that A_2 and A_3 both exhibit a "vertical bars" pattern, while A_1 exhibits a diagonal pattern. The similarity of the "vertical bars" pattern observed for A_2 and A_3 indicate that in some sense the heads are performing the same *function*: both heads always attend to a certain class of token – despite those classes being entirely different between the two heads. Note that these **are not actual attention patterns from trained models**, and are provided only for illustrative purposes. See Figure 3 for examples of actual attention patterns.

[2]All experiments were performed on a laptop with an 8-core i9-11950H CPU, 64GB RAM, and A5000 Mobile GPU with 16GB VRAM, requiring several minutes. Preliminary experiments with features which were eventually discarded took up to several hours.

1.1 Paper Website

This paper contains extensive links to our accompanying website: `attention-motifs.github.io`, which contains interactive versions of many figures, as well as a variety of tools for interpretability researchers. Use of interactive figures and tools requires only a web browser. In particular, the browser tool at `attention-motifs.github.io/s/head-info` allows the user to enter any head from the listed models (Table 1) and see the attention patterns produced by the head, links to other works which mention the head, the location in embedding space of this head, and information and links to attention heads nearby in embedding space.

2 Embedding patterns

2.1 Type signature of the embedding

Attention for autoregressive transformer models [Vaswani et al., 2017] over some input activation $X \in \mathbb{R}^{n \times d}$ can be written as

$$\texttt{attention}(X) := \underbrace{\sigma\left(\frac{XW_Q W_K^T X^T}{\sqrt{d}} + M\right)}_{A} \cdot W_{OV}(X) \qquad \text{where} \qquad M_{i,j} := \begin{cases} -\infty & j > i \\ 0 & j \leq i \end{cases}$$

where d is the model dimension, n is the sequence length, σ is the row-wise softmax function, and M is the autoregressive masking matrix. The *attention pattern* is the output of the softmax, the matrix $A \in \mathbb{R}^{n \times n}$. Examples of these attention patterns can be seen in Figure 3.

We can define the set of all possible attention patterns as

$$\mathcal{P} = \bigcup_{n \in \mathbb{N}} \mathcal{P}_n \qquad \text{where} \qquad \mathcal{P}_n = \left\{ A \in \mathbb{R}^{n \times n} \;\middle|\; \begin{array}{l} A\vec{1} = \vec{1} \\ A_{i,j} \in [0,1] \\ A_{i,i+k} = 0 \quad \forall\, k \in \mathbb{N} \end{array} \right\} \tag{1}$$

The attention pattern of the head at layer L and index H, for an LLM with parameters θ and given a prompt s is given by

$$\texttt{LLM}_{\theta,L,M}(s) \in \mathcal{P}_{|s|} \in \mathcal{P}_{|s|} \qquad \text{or, equivalently} \qquad \texttt{LLM}[h_i](s) \in \mathcal{P}_{|s|} \tag{2}$$

Where h_i is a particular head from a particular model – for example, `L0H1` from `pythia-1b`.

If we entertain the hypothesis that there are properties of the structure of $\texttt{LLM}_{\theta,L,M}(s)$ that are invariant to our dataset sample $s \sim \mathcal{D}$ and indicates the role of the head $\texttt{LLM}_{\theta,L,M}$, we expect that there exists an embedding function $\mathcal{E} : \mathcal{P} \to \mathbb{R}^c$, which maps attention patterns to a meaningful latent space in which the location of the head represents the structure we care about. We note that a useful embedding should accommodate varied input sequence lengths.[3] In subsection 2.3, we describe the results of finding such a function $\mathcal{E}$ by using PCA[Pearson, 1901] to reduce the dimensionality of a large set of features (described in subsection 2.2).

2.2 Motivation for chosen features

In order to find a suitable embedding $\mathcal{E}$, we use handcrafted features to compute about attention patterns. In particular, these features include basic statistics (mean, variance, etc.) about the values on the diagonal and first column of the attention pattern, as well as similar features about the distributions of values in gram matrices of the pattern and skew of the pattern. Features, along with their importance and covariance, are shown in Figure 10.

[3] This requirement to work with varied input lengths was a key motivation for our choice of handcrafted features, as opposed to a purely learned method, such as a convolutional autoencoder.

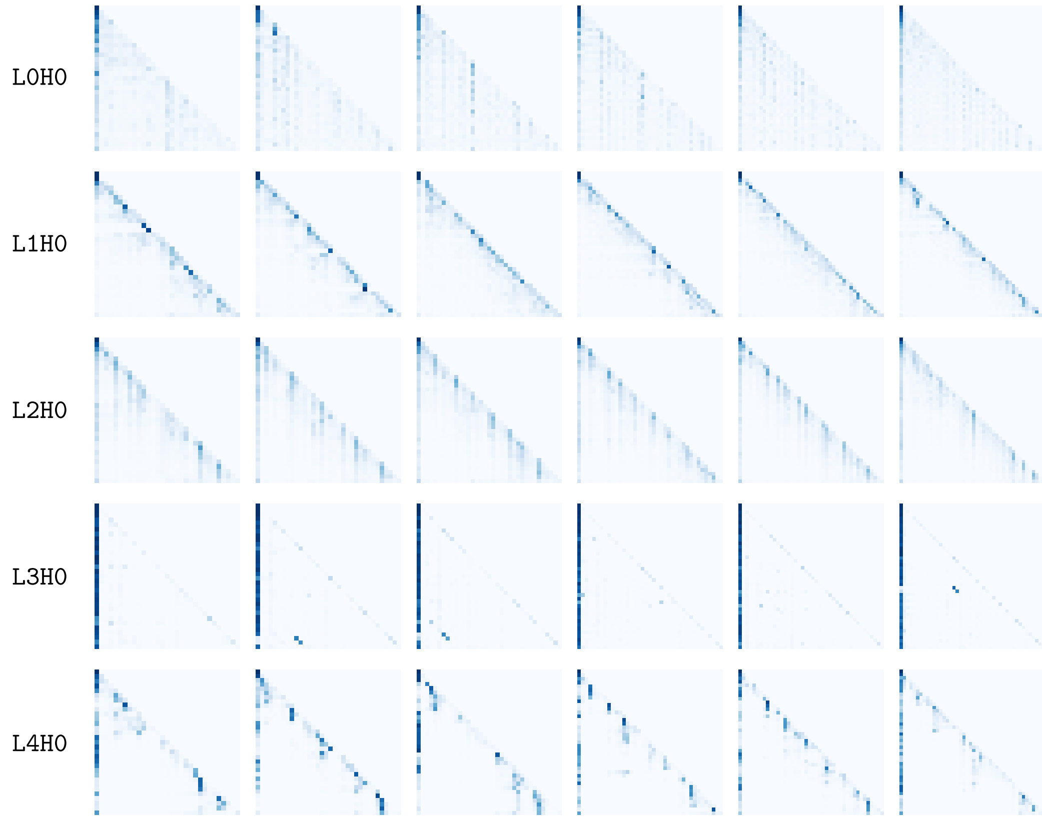

Figure 3: Actual attention patterns from `gpt2-small`. Each row corresponds to a different head in the model. Each column represents one of 6 random prompts. Note that each head displays the same *motif* regardless of the input prompt. Interactive version, with prompt information: `attention-motifs.github.io/s/fig/patterns-example` .

Our motivation for the choice of features is that visually, some of the most common motifs in attention patterns include:

- Large values along the diagonal, meaning every token attends to itself. See `gpt2-small:L0H1` , `gpt2-small:L0H3` , `gpt2-small:L1H11` . This motivates including statistics about the diagonal values $\texttt{trace}(A)$.
- Large values on the first token, sometimes known as an "attention sink" [Zuhri et al., 2025]. Since the first token in autoregressive attention cannot contain information about any token besides itself, it is speculated that these attention sinks are a way for the head to "shut off." See `gpt2-small:L3H4` , `gpt2-small:L5H1` , `gpt2-small:L11H9` . This motivates including statistics about the values in the first column $A[:, 0]$.
- "vertical bars," meaning that the same tokens from the context are attended to regardless of the current token. See `gpt2-small:L0H0` , `gpt2-small:L1H9` , `gpt2-small:L10H0` . This motivates including statistics about the gram matrix AA^T. If vertical bars are present in A, then rows are likely very similar, causing the gram matrix to have large values[4]. Horizontal bars, although rarer, motivate including the gram matrix of the transpose A^TA.
- "recent tokens," where most of the attention is concentrated somewhere close to the diagonal (but not entirely on it), regardless of the current token. We speculate that these heads rely primarily on positional embedding information in their QK circuit. See `gpt2-small:L0H4` , `gpt2-small:L2H3` , `gpt2-small:L3H2` . This motivates the inclusion of statistics about

[4] By "vertical bars", we mean that $A[i, j]$ and $A[k, j]$ are correlated. If this is the case, then $[AA^T]_{i,k}$ is more likely to be large, as $[AA^T]_{i,k} = A[i, :] \cdot A[k, :]$.

the gram matrix $S(A)S(A)^T$ of the "skewed" attention pattern where for

$$A \in \mathcal{P}_n, \qquad S(A)\big[i,\ j+(n-i-1)\big] := A[i,j]$$

$S(A)[i,j]$ indicates how much token j is attending to the token $(n-i+1)$ tokens *before* it, and the gram matrix captures how similar this pattern is between rows: $S(A)S(A)^T$ will have larger values if each token attends to tokens a similar number of tokens behind it.

The above list is not meant to cover all of the motifs observed, nor are the examples given exhaustive. We leave most of the details of these features, denoted $\hat{\mathcal{E}} : \mathcal{P} \to \mathbb{R}^{92}$, to the code: `attention-motifs.github.io/s/feature-info` .

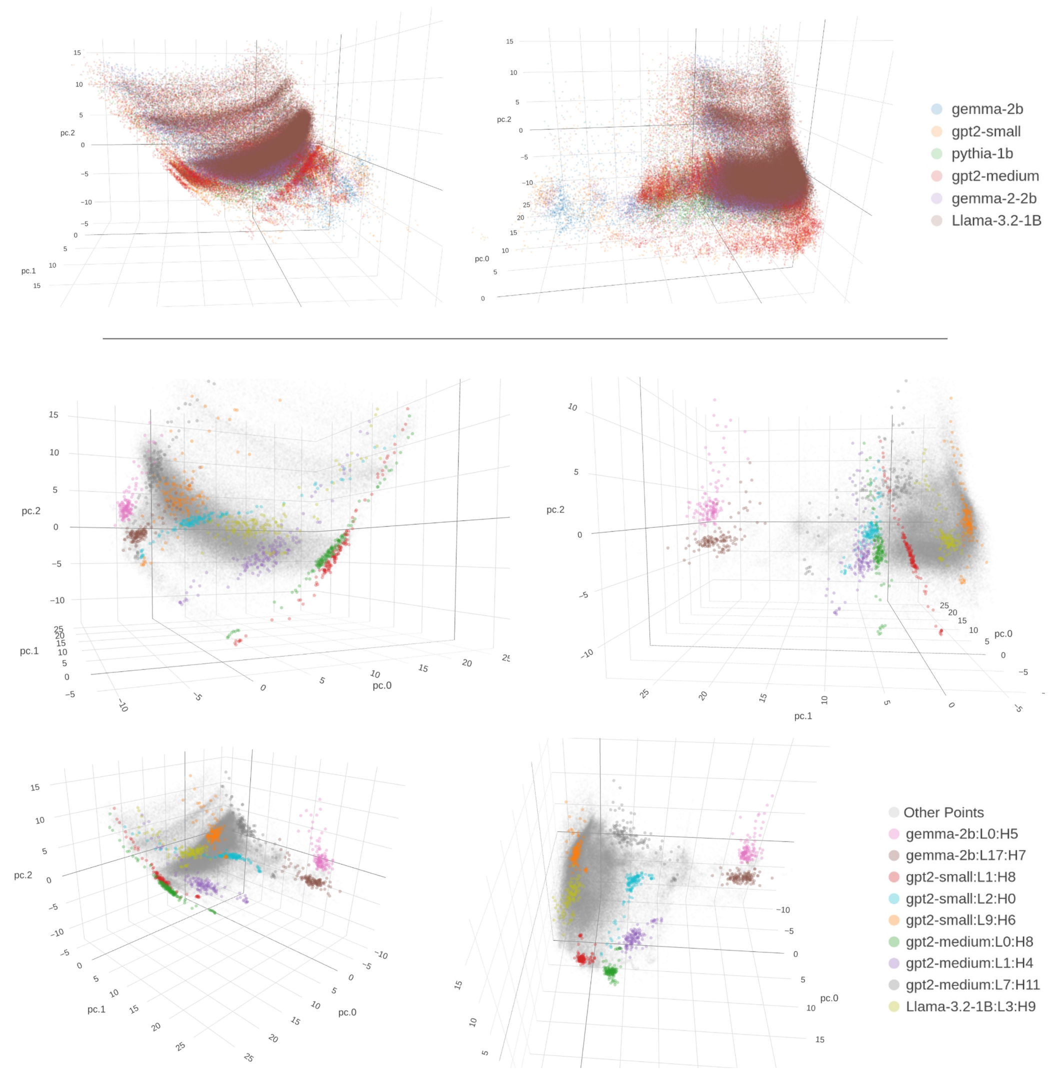

Figure 4: Different views of the first 3 PC axes of $\mathcal{E}$. **Top group:** colored by model, **Bottom group:** with certain heads selected – all points of a given color are the embeddings of the attention pattern, for different prompts, of that head. Interactive versions: `attention-motifs.github.io/s/fig/pca-view`

2.3 Computing features and the embedding

We apply $\hat{\mathcal{E}}$ to a dataset of $> 10^5$ of attention patterns from open-weight pretrained LLMs (see Table 1) across 128 pieces of text sampled from the "Pile" dataset [Gao et al., 2020, Neel Nanda, 2022]. We assemble from the ouput of $\hat{\mathcal{E}}$ a table where each row has a column identifying the attention head (h_i), a column identifying the prompt used (s_k), and columns with normalized scalar values for the computed features. Performing a principal component analysis (PCA) on the normalized feature columns, we find that around 68% of the variance is explained by the first 3 principal components, and nearly 90% by the first 10 (Figure 11). We construct our embedding $\mathcal{E}$ as the first 16 principal components of $\hat{\mathcal{E}}$.

Plotting the embedding of each pattern in the first 3 components shows us that the distributions for all models overlap, which is a desired property[5] of our embedding function (Figure 4). Furthermore, we see in Figure 4 that all attention patterns from a given head appear to occupy a well-defined region of embedding space. Interactive visualizations of this embedding can be found at `attention-motifs.github.io/embed`.

3 Embedding heads

Our embedding $\mathcal{E} : \mathcal{P} \to \mathbb{R}^c$ tells us something about how similar attention *patterns* are to each other, but what we want is a distance metric that tells us about similarities between attention *heads*. Each head corresponds to a cloud of points in $\mathbb{R}^c$, each point corresponding to that head's attention pattern given a prompt s from our dataset $\mathcal{D}$, and we want some notion of similarity between these point clouds. In this work, we consider the naive metric:

$$\texttt{dist}(h_i, h_j) := \frac{1}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} \left| \mathcal{E}\Big(\texttt{LLM}[h_i](s)\Big) - \mathcal{E}\Big(\texttt{LLM}[h_j](s)\Big) \right| \tag{3}$$

taking the mean distance between the embeddings of the patterns produced by the heads h_i, h_j over prompts s from the dataset $\mathcal{D}$. We discuss the potential of other metrics in subsection 4.1.

We compute the distance matrix D for all pairs of heads h_i, h_j (Figure 13),

$$\mathbf{D}[i, j] = \texttt{dist}(h_i, h_j) \tag{4}$$

and project to a viewable low dimensional space using UMAP, Isomap, and t-SNE [Tenenbaum et al., 2000, van der Maaten and Hinton, 2008].

3.1 Comparing with previously identified classes

In this space, we find that known classes of attention heads are generally grouped together. We assemble a mapping from 6 "head types" to identified heads in `gpt2-small` based on the work of [Wang et al., 2022] and [Krzyzanowski et al., 2024]. Noting that these two works are not always in agreement about the classes of heads for classes which they both identify (Induction, Duplicate Token, and Previous Token heads), we will consider a head to be in one of these classes if *either* work identifies it as such.

We find that when projecting via Isomap [Tenenbaum et al., 2000] with 16 neighbors, groupings of heads with known functionality become particularly apparent (Figure 7). In Figure 6 and Figure 5, we see that heads nearby in our embedding exhibit similar attention patterns. Notably, although [Wang et al., 2022] only finds "Backup Name Mover" heads after knocking out the initial name mover heads, our method groups together all varieties of name mover heads.

Precision and recall metrics for recovering known classes are not presented in this work. This is in part due to the extreme sparsity of known classes of attention heads, making the statistical significance of such metrics of limited use. Primarily, however, we refer to the counterexample described in Figure 2 for motivation as to why our method is an entirely different way of looking at attention heads. Tt is conceivable that two heads determined to be similar through QK circuit analysis (or potentially circuit analysis) might have quite different attention patterns, or for the inverse case to be true. We elaborate on this point in section 4.

[5] In general, we expect and see roughly the same motifs in patterns across all language models. If patterns from different models were mapped to wholly different parts of embedding space, this would not be useful for finding similar heads across different models.

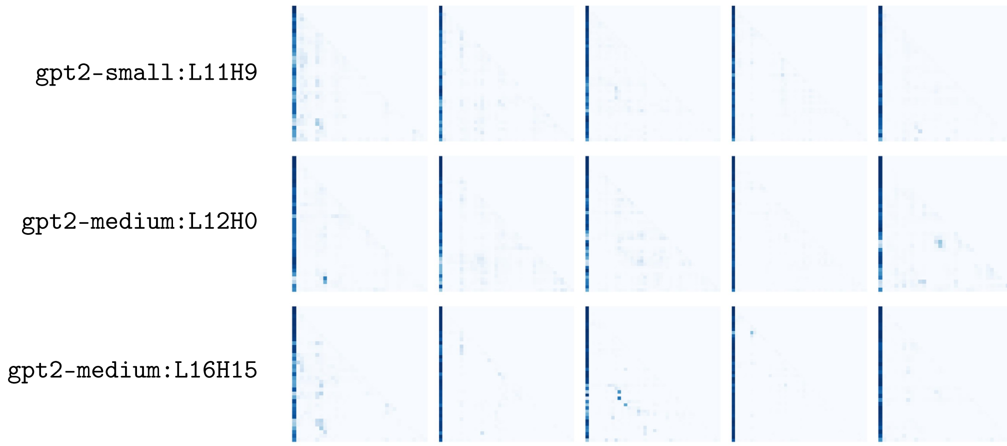

Figure 5: gpt2-small:L11H9 is originally identified by [Wang et al., 2022] as a "Backup Name Mover", while the other heads heads are nearby heads which are not described as name movers or otherwise in the literature to our knowledge. More information: attention-motifs.github.io/s/fig/groups/name-mover .

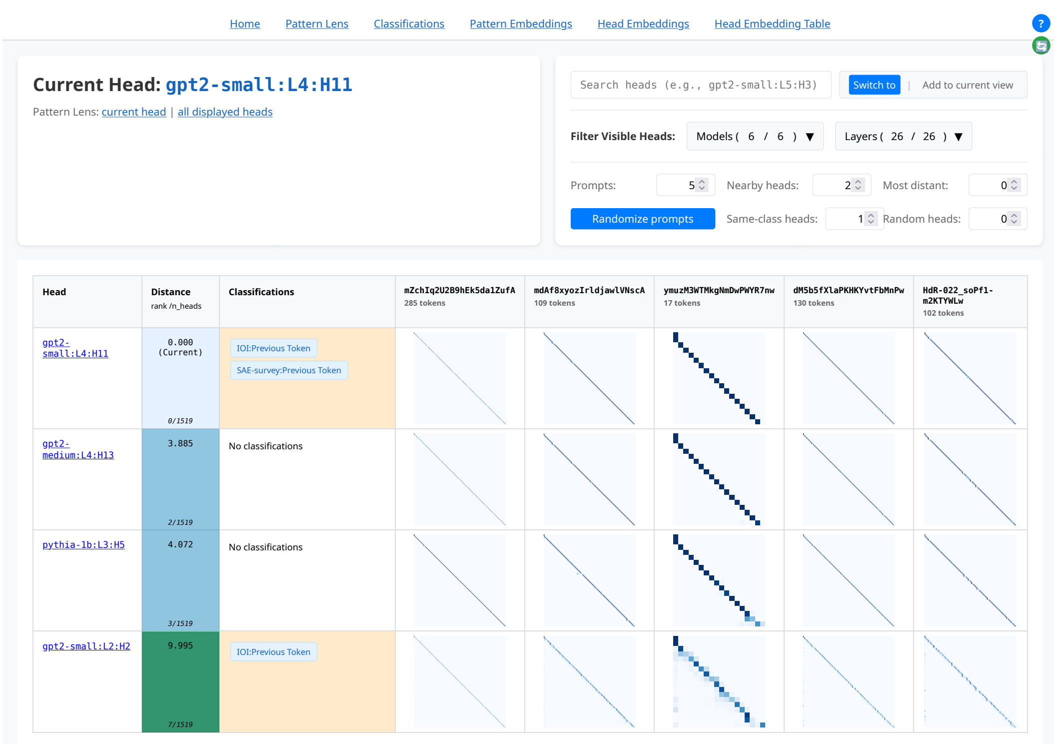

Figure 6: The primary interface for interacting with the head embeddings. We allow searching for any head among the supported models, viewing heads by their classifications, filtering by model or layer, and viewing heads which are near or far in head embedding space. Displayed is gpt2-small:L4H11 and detected similar heads. gpt2-small:L4H11 is identified by both [Wang et al., 2022] and [Krzyzanowski et al., 2024] as a "Previous Token Head", while the other heads are nearby heads which are not described as previous token heads or otherwise in the literature to our knowledge. More information: attention-motifs.github.io/s/fig/groups/previous-token .
attention-motifs.github.io/v1/vis/attnpedia/index.html?head_viewing=gpt2-small~L4~H11

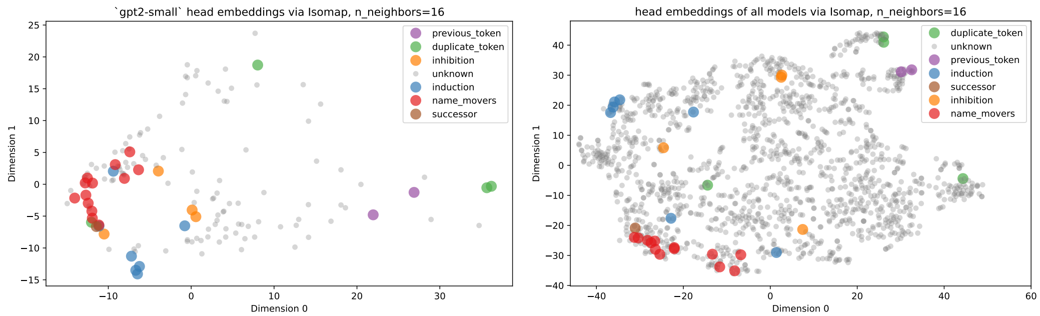

Figure 7: Embeddings of heads via the distance matrix. **Left:** Only heads from `gpt2-small`. **Right:** heads from `all` models. Projection via Isomap, with 16 neighbors. More projections can be viewed in the appendix (Figure 14, Figure 15) or on the website: `attention-motifs.github.io/s/fig/head-embed`

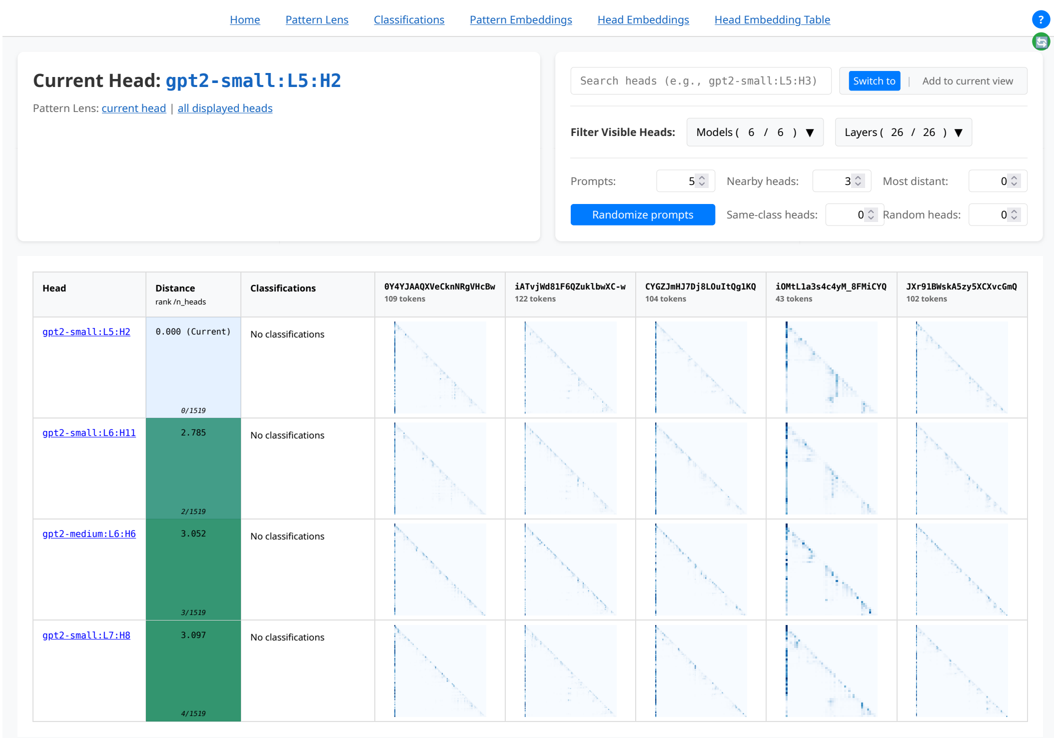

Figure 8: `gpt2-small:L5H2` and detected similar heads, displaying both banded and attention sink features. These heads, to our knowledge, are not described or discussed in the literature. The vast majority of heads are not described, yet our method finds similarities nonetheless. We encourage the reader to explore the tool.
`attention-motifs.github.io/v1/vis/attnpedia/index.html?head_viewing=gpt2-small~L5~H2`

4 Conclusion

Metrics

The distance metric defined in Equation 3 is not the only possible metric, and we do not consider the distribution of distances for each pair of points, only the mean. In particular, Gromov-Wasserstein [Chhoa et al., 2025, Mémoli, 2011] distances and variants may provide a unique perspective. Consider heads h_i, h_j and inputs s_1, s_2. To motivate this, we define

$$f(h_i, h_j, s_u, s_v) := \left| \mathcal{E}\big(\texttt{LLM}[h_i](s_u)\big) - \mathcal{E}\big(\texttt{LLM}[h_j](s_v)\big) \right|.$$

Consider the case that:

- $f(h_i, h_j, s_1, s_1)$ and $f(h_i, h_j, s_2, s_2)$ are both very large
- $f(h_i, h_j, s_1, s_2)$ and $f(h_i, h_j, s_2, s_1)$ are both very small

For example, we could have $\mathcal{E}(\texttt{LLM}[h_i](s_1)) = \mathcal{E}(\texttt{LLM}[h_j](s_2))$ and $\mathcal{E}(\texttt{LLM}[h_i](s_2)) = \mathcal{E}(\texttt{LLM}[h_j](s_1))$. If this is the case, then Equation 3 would compute distance between h_i and h_j to be very large, while a Gromov-Wasserstein or other "earth-mover" metric would compute it to be small. What this tells us in practice is that h_i and h_j are in some sense complimentary, producing a similar distribution of patterns over the entire dataset but vastly different patterns for any given pattern s_1 or s_2. We believe exploring other such distance metrics, and in particular comparing multiple metrics, would be a fruitful area of work.

Features

Certain features were considered but not used due to computational cost or lack of importance in the PCA. Discarded features include statistics about the absorption times when treating the attention pattern as an absorbing markov chain, various Fourier statistics, and network-theoretic analyses of the attention pattern as an adjacency matrix. An autoencoder approach was also considered, but not pursued further due to the lack of interpretability about the resulting embedding space. Importance of features in relation to each individual PCA axis can be found at `attention-motifs.github.io/s/fig/feat-importance` , but a detailed analysis of the influence of various features on the resulting groupings of heads is absent from this work.

Supervised classification

A supervised approach to classification of attention heads by their patterns is likely impractical. Manual inspection and labeling of attention patterns does not appear to be practical, since a large number of attention patterns contain structure that is difficult to describe. Using the labels of known classes of attention patterns may be useful to condition the embedding of attention heads from the distance matrix, but was not explored in this work. A key obstacle is the relatively small number of labels, and the small subset of models for which they exist (`gpt2-small` is often described as the "model organism" of interpretability). The differences in produced attention patterns between models may further complicate any attempts at a supervised approach, if one wishes their method to generalize to new models.

4.1 Limitations and future work

We do not yet provide a mechanistic analysis of whether heads near in embedding space to a known class (e.g. induction heads) fulfill the same role. Our method does not directly use any token or activation information, and this is also by design. More on our motivation behind this is explained in subsection 4.3. This work only uses the models described in Table 1, a selected variety of GPT-like autoregressive transformer architectures. A dataset of 128 samples from the "Pile" [Gao et al., 2020] dataset is used. [6]

[6] See `attention-motifs.github.io/s/pile-info` .

Limitations of attention patterns as a tool for interpretability

It may be the case that attention patterns themselves are not useful for interpretability. Perhaps polysemanticity makes studying attention patterns of individual heads entirely useless, or perhaps the OV circuit sometimes negates the attention in a way that makes the patterns unimportant. We believe our work provides some evidence to the contrary, but acknowledge this possibility. If it is in fact the case that attention patterns are not useful, however, we believe that this is a hypothesis at least worth testing. Our work provides the foundation for doing so, by creating a tool for researchers to investigate if there is any correlation in the heads they study between head functionality and attention pattern structure.

4.2 Broader impacts

Risks from the misuse and misalignment of AI systems are widely discussed in the literature, as is the application of interpretability to mitigate those risks [Räuker et al., 2023]. Work in interpretability is often constrained by high computational costs [Cunningham et al., 2023, Braun et al., 2025], and it is our view that there is a niche for low-cost methods to work in concert with more expensive ones. Our work helps fill this niche, by providing a way to identify potentially similar heads across many different models, thereby leveraging the identification of a small number of heads that is found to be of interest using other, more expensive, methods.

4.3 Contributions

We present a method for embedding the attention patterns of attention heads in pretrained LLMs, and show that this corresponds with visually apparent motifs in the attention patterns. We utilize this embedding of attention patterns to create an embedding of the attention heads themselves, and show that this embedding groups together some known classes of attention heads. Most importantly, we present easy-to-use tools (attention-motifs.github.io) that utilize our method, and allow researchers to explore the embedding space of heads, explore known classifications, and find attention heads with similar patterns to any given head.

One interpretation of why attention in LLMs works best when multi-headed is because different "views" of token similarity may be required. E.g., one attention head might view countries and their capitals as similar ("Paris" $\rightarrow$ "France", "Tokyo" $\rightarrow$ "Japan") , while another may view countries as similar if they are on the same continent ("Japan" $\leftrightarrow$ "Vietnam", "France" $\leftrightarrow$ "Spain"). In the same sense, we aim to complement existing methodology by providing a different view on what properties attention heads possess. We anticipate that this method will accelerate research in interpretability by providing an interpretable, extensible, and inexpensive method to find heads across many models which may be similar in role to a head whose functionality has been identified.

Acknowledgments and Disclosure of Funding

This work was partially supported by NSF grant DMS-2110745, the MATS 8.0 fellowship, and AISS. The authors would like to express their gratitude to the American Mathematical Society Math Research Community "Mathematics of Adversarial, Interpretable, and Explainable AI", the Mines Optimization and Deep Learning group, Emily J. King, Gary Kazantsev, Alex F. Spies, William Edwards, Tilman Räuker, Lee Sharkey, and Nathan Hu.

References

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, May 2023. URL `http://arxiv.org/abs/2304.01373`.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying Functionally Important Features with End-to-End Sparse Dictionary Learning, May 2024. URL `http://arxiv.org/abs/2405.12241`.

Dan Braun, Lucius Bushnaq, Stefan Heimersheim, Jake Mendel, and Lee Sharkey. Interpretability in Parameter Space: Minimizing Mechanistic Description Length with Attribution-based Parameter Decomposition, February 2025. URL `http://arxiv.org/abs/2501.14926`.

Jannatul Chhoa, Michael Ivanitskiy, Fushuai Jiang, Shiying Li, Daniel McBride, Tom Needham, and Kaiying O'Hare. Metric properties of partial and robust Gromov-Wasserstein distances, March 2025. URL `http://arxiv.org/abs/2411.02198`.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What Does BERT Look At? An Analysis of BERT's Attention, June 2019. URL `http://arxiv.org/abs/1906.04341`.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models, October 2023. URL `http://arxiv.org/abs/2309.08600`.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy Models of Superposition, September 2022. URL `http://arxiv.org/abs/2209.10652`.

Javier Ferrando and Elena Voita. Information Flow Routes: Automatically Interpreting Language Models at Scale, April 2024. URL `http://arxiv.org/abs/2403.00824`.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling, December 2020. URL `http://arxiv.org/abs/2101.00027`.

Jorge García-Carrasco, Alejandro Maté, and Juan Trujillo. How does GPT-2 Predict Acronyms? Extracting and Understanding a Circuit via Mechanistic Interpretability, May 2024. URL `http://arxiv.org/abs/2405.04156`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire

Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat,

Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783.

Zhengfu He, Junxuan Wang, Rui Lin, Xuyang Ge, Wentao Shu, Qiong Tang, Junping Zhang, and Xipeng Qiu. Towards Understanding the Nature of Attention with Low-Rank Sparse Decomposition, April 2025. URL http://arxiv.org/abs/2504.20938.

Michael Igorevich Ivanitskiy, Alex F. Spies, Tilman Räuker, Guillaume Corlouer, Chris Mathwin, Lucia Quirke, Can Rager, Rusheb Shah, Dan Valentine, Cecilia Diniz Behn, Katsumi Inoue, and Samy Wu Fung. Structured World Representations in Maze-Solving Transformers, December 2023. URL http://arxiv.org/abs/2312.02566.

Jett Janiak, Chris Mathwin, and Stefan Heimersheim. Polysemantic Attention Head in a 4-Layer Transformer, November 2023. URL https://www.lesswrong.com/posts/nuJFTS5iiJKT5G5yh/polysemantic-attention-head-in-a-4-layer-transformer.

Robert Krzyzanowski, Connor Kissane, Arthur Conmy, and Neel Nanda. We Inspected Every Head In GPT-2 Small using SAEs So You Don't Have To, March 2024. URL https://www.lesswrong.com/posts/xmegeW5mqiBsvoaim/we-inspected-every-head-in-gpt-2-small-using-saes-so-you-don.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse Autoencoders Do Not Find Canonical Units of Analysis, February 2025. URL http://arxiv.org/abs/2502.04878.

Facundo Mémoli. Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11(4):417–487, August 2011. ISSN 1615-3383. doi: 10.1007/s10208-011-9093-5. URL https://doi.org/10.1007/s10208-011-9093-5.

Neel Nanda and Joseph Bloom. TransformerLensOrg/TransformerLens. TransformerLensOrg, 2022. URL https://github.com/TransformerLensOrg/TransformerLens.

Neel Nanda. NeelNanda/pile-10k · Datasets at Hugging Face, 2022. URL https://huggingface.co/datasets/NeelNanda/pile-10k.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context Learning and Induction Heads, September 2022. URL http://arxiv.org/abs/2209.11895.

Cheonbok Park, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, Yeonsoo Lee, and Jaegul Choo. SANVis: Visual Analytics for Understanding Self-Attention Networks, September 2019. URL http://arxiv.org/abs/1909.09595.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks, August 2023. URL http://arxiv.org/abs/2207.13243.

Jie Ren, Qipeng Guo, Hang Yan, Dongrui Liu, Quanshi Zhang, Xipeng Qiu, and Dahua Lin. Identifying Semantic Induction Heads to Understand In-Context Learning, July 2024. URL http://arxiv.org/abs/2402.13055.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open Problems in Mechanistic Interpretability, January 2025. URL http://arxiv.org/abs/2501.16496.

Alex F. Spies, William Edwards, Michael I. Ivanitskiy, Adrians Skapars, Tilman Räuker, Katsumi Inoue, Alessandra Russo, and Murray Shanahan. Transformers Use Causal World Models in Maze-Solving Tasks, March 2025. URL http://arxiv.org/abs/2412.11867.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open Models Based on Gemini Research and Technology, April 2024a. URL http://arxiv.org/abs/2403.08295.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris

Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size, October 2024b. URL http://arxiv.org/abs/2408.00118.

Joshua B. Tenenbaum, Vin De Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, December 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.290.5500.2319. URL https://www.science.org/doi/10.1126/science.290.5500.2319.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Jesse Vig. A Multiscale Visualization of Attention in the Transformer Model, June 2019. URL http://arxiv.org/abs/1906.05714.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small, November 2022. URL http://arxiv.org/abs/2211.00593.

Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda Viégas, and Martin Wattenberg. AttentionViz: A Global View of Transformer Attention, August 2023. URL http://arxiv.org/abs/2305.03210.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention Heads of Large Language Models: A Survey, December 2024. URL http://arxiv.org/abs/2409.03752.

Zayd M. K. Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. Softpick: No Attention Sink, No Massive Activations with Rectified Softmax, April 2025. URL http://arxiv.org/abs/2504.20966.

A Technical Appendices and Supplementary Material

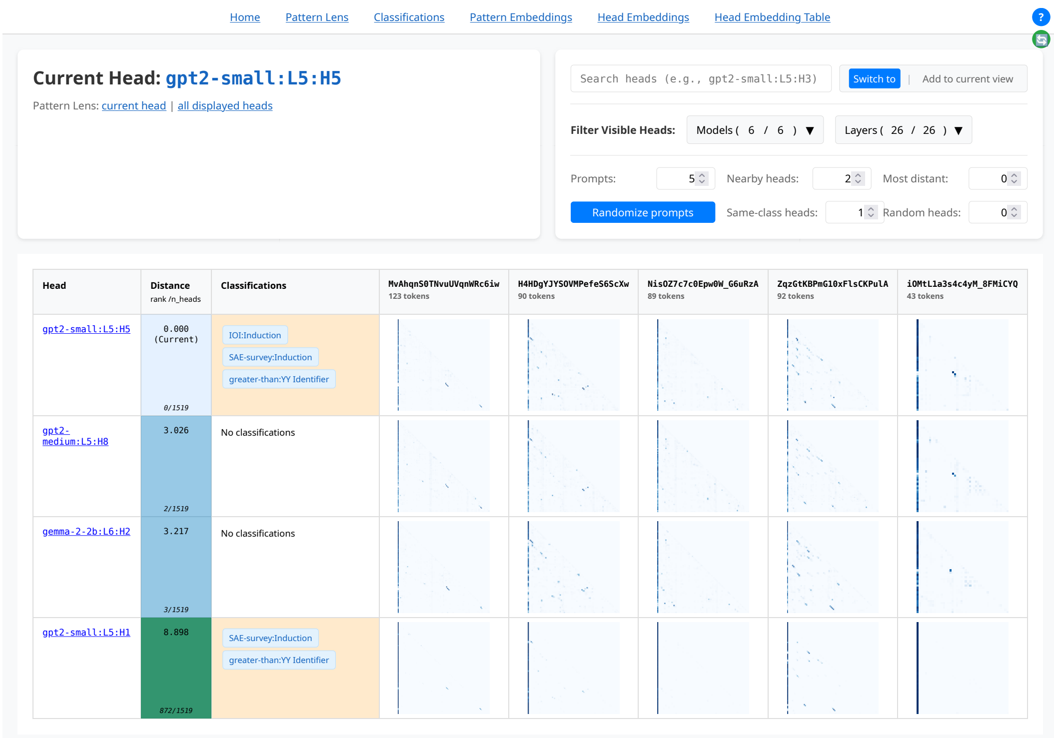

Figure 9: Another view of the interface. Displayed is `gpt2-small:L5H5` , identified by both Wang et al. [2022] and Krzyzanowski et al. [2024] as an induction head Olsson et al. [2022]. By the "eyeball norm," these patterns look nearly identical for any given prompt.
`attention-motifs.github.io/v1/vis/attnpedia/index.html?head_viewing=gpt2-small~L5~H5`

A.1 Models used

Model Name	Parameter count	Layers	Heads (per layer)	Citation
gpt2-small	85M	12	12	Radford et al. [2019]
gpt2-medium	302M	24	16	Radford et al. [2019]
Llama-3.2-1B	1.1B	16	32	Grattafiori et al. [2024]
pythia-1b	805M	16	8	Biderman et al. [2023]
gemma-2b	2.1B	18	8	Team et al. [2024a]
gemma-2-2b	2.1B	26	8	Team et al. [2024b]

Table 1: Models used in experiments. Model loading, inference, and activation inspection was done via the TransformerLens Nanda and Bloom [2022] package.

A.2 Feature covariance and importance

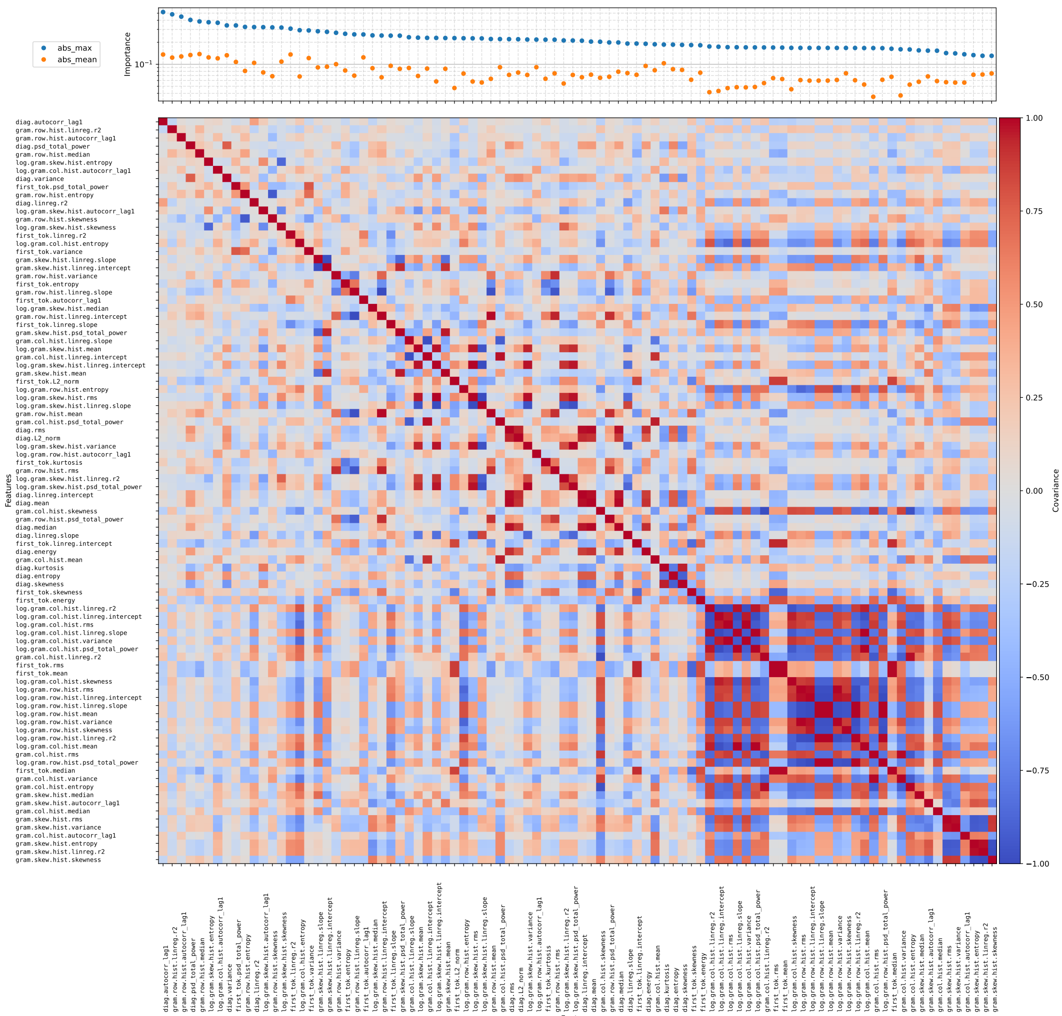

Figure 10: Importance (top) and covariance (bottom) of all features.

A.3 Feature PCA

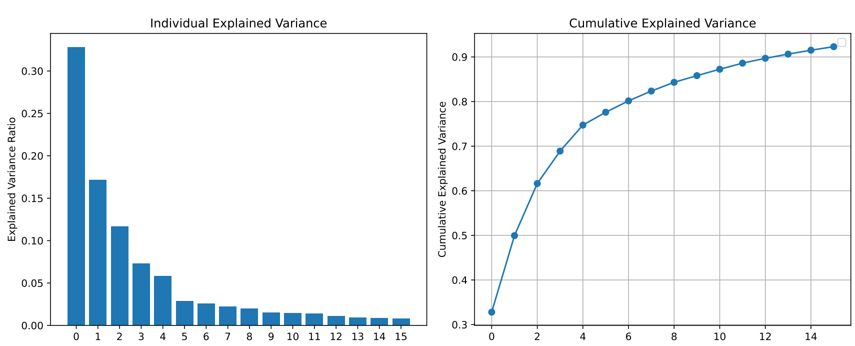

Figure 11: Variance explained by PCA ($\mathcal{E}$) of the feature space $\hat{\mathcal{E}}$ of all attention patterns.

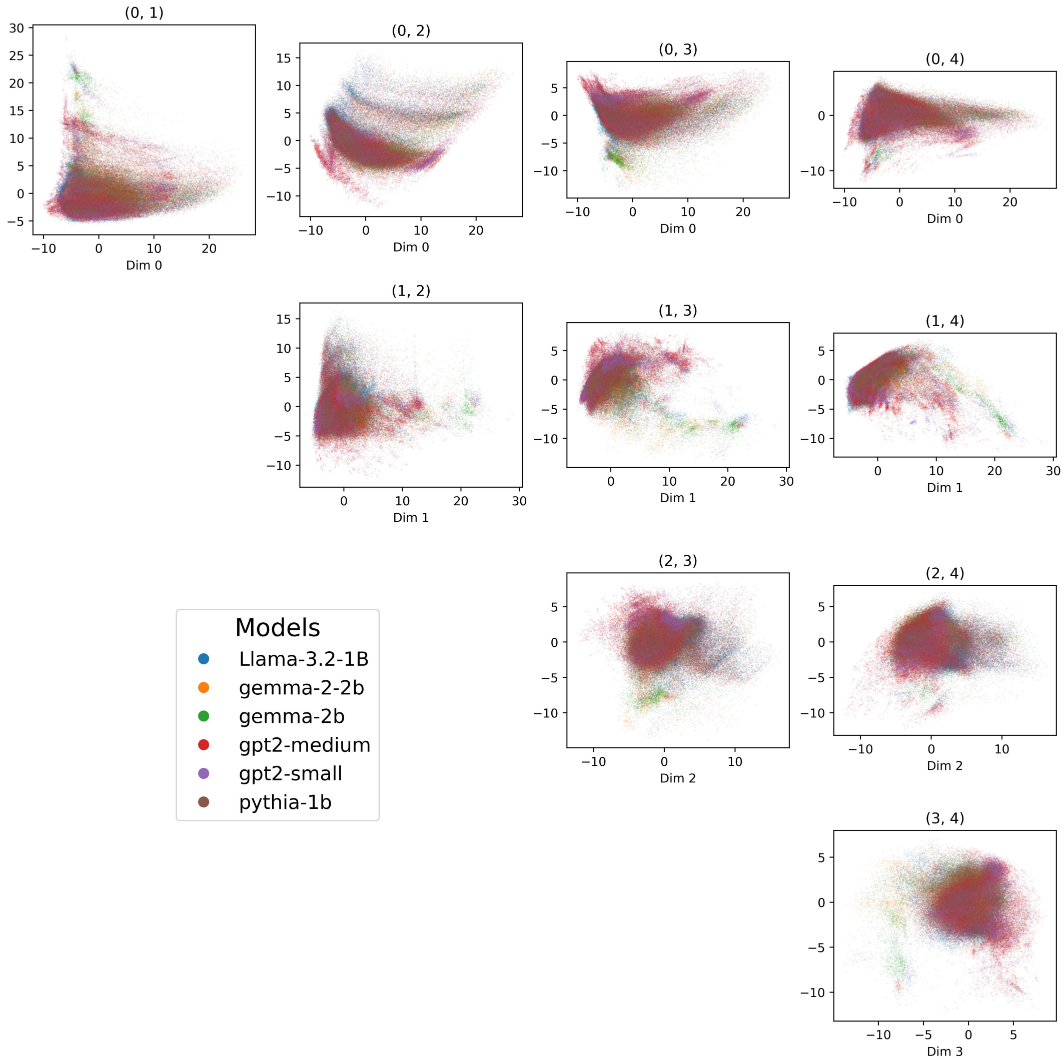

Figure 12: Different projections of the PCA of the embedding space, colored by model. Note that each model has a similar distribution in this space.

A.4 Head Embeddings

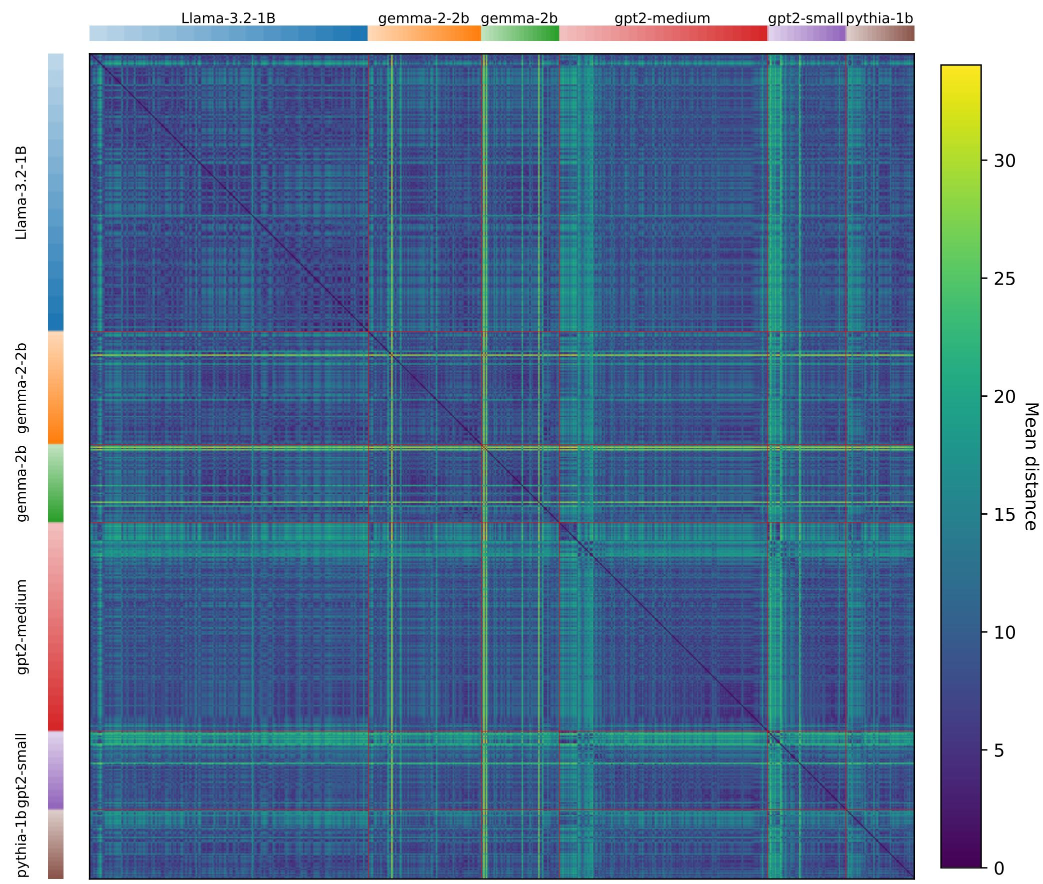

Figure 13: Pairwise distance $\mathbf{D}$ between all heads computed via Equation 3. Models are denoted by colored blocks on the top and left, with lighter colors representing earlier layers and darker colors representing later ones. Each pair of attention heads is exactly one pixel, and the different numbers of layers and heads per layer causes the difference in size between the colored blocks. Red gridlines separate the models from each other. It is of note that for most models, there is a clear distinction between early layer heads and later layer heads. More information: `attention-motifs.github.io/s/fig/head-dists-heatmap`

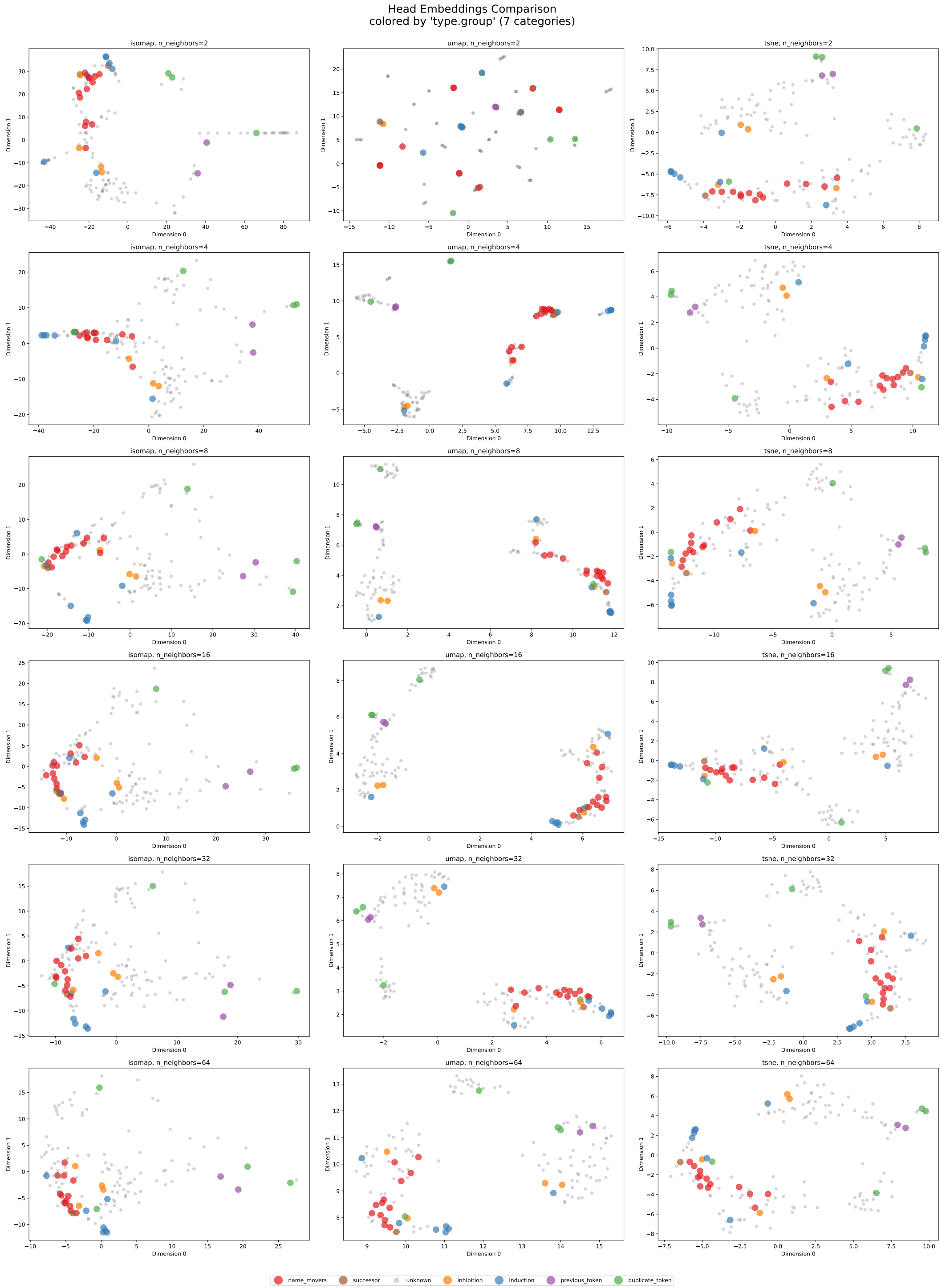

Figure 14: 2D Embeddings via Isomap (left), UMAP (center), and t-SNE (right) for various neighborhood sizes (top to bottom, small to large) of the 144 attention heads of `gpt2-small`. Legend of known head classes at the bottom, unknown heads in grey. See `attention-motifs.github.io/s/fig/head-embed/gpt2-small` for an interactive version of the 3D embeddings.

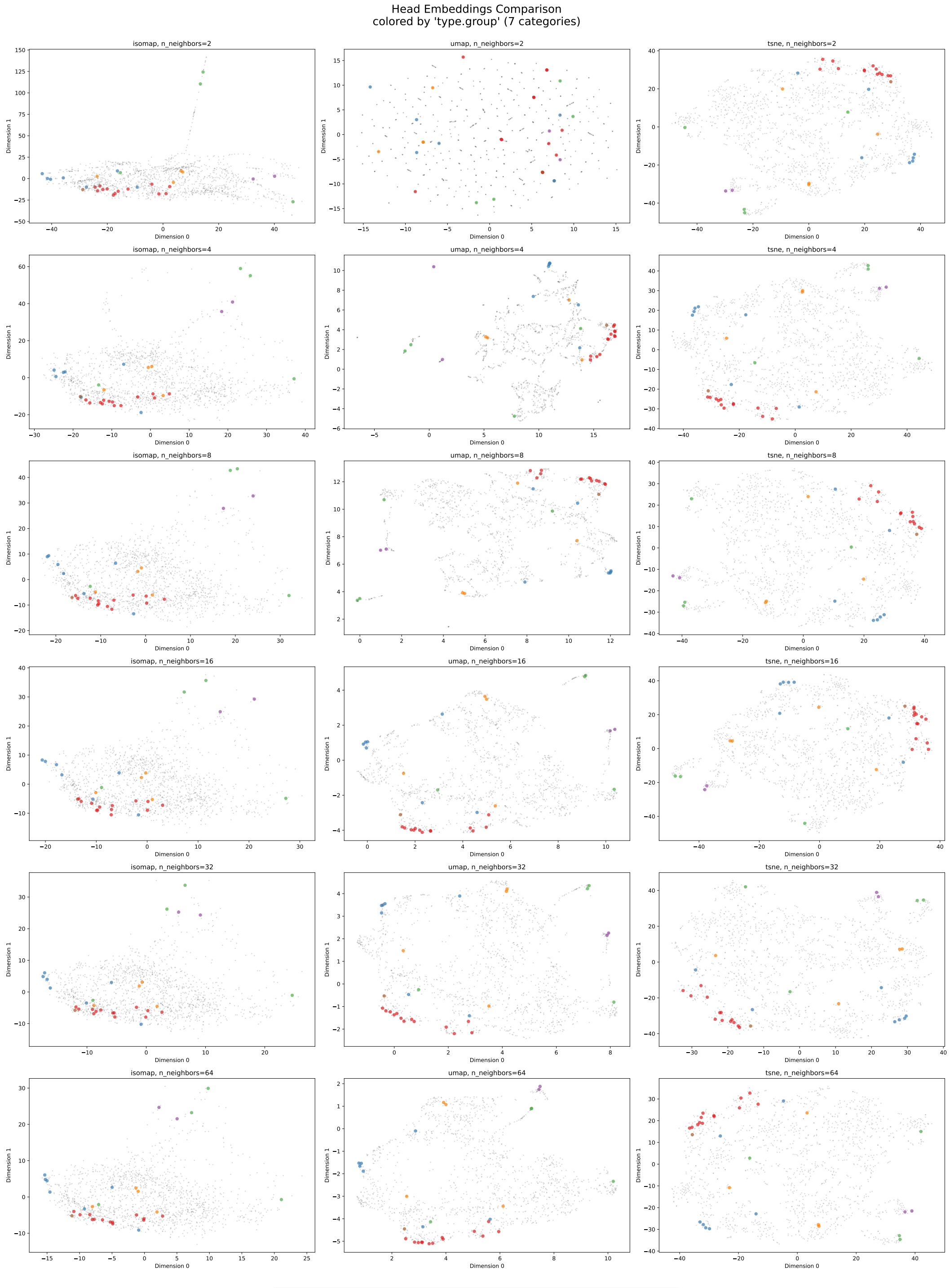

Figure 15: 2D Embeddings via Isomap (left), UMAP (center), and t-SNE (right) for various neighborhood sizes (top to bottom, small to large) of all attention heads from all `models`. Legend of known head classes at the bottom, unknown heads in grey. See `attention-motifs.github.io/s/fig/head-embed/all` for an interactive version of the 3D embeddings.