Analyzing Transformers in Embedding Space

Anonymous authors
Paper under double-blind review

Abstract

Understanding Transformer-based models has attracted significant attention, as they lie at the heart of recent technological advances across machine learning. While most interpretability methods rely on running models over inputs, recent work has shown that a zero-pass approach, where parameters are interpreted directly without a forward/backward pass is feasible for *some* Transformer parameters, and for two-layer attention networks. In this work, we present a theoretical analysis where *all* parameters of a trained Transformer are interpreted by projecting them into the *embedding space*, that is, the space of vocabulary items they operate on. We derive a simple theoretical framework to support our arguments and provide ample evidence for its validity. First, an empirical analysis showing that parameters of both pretrained and fine-tuned models can be interpreted in embedding space. Second, we present two applications of our framework: (a) aligning the parameters of different models that share a vocabulary, and (b) constructing a classifier *without training* by "translating" the parameters of a fine-tuned classifier to parameters of a different model that was only pretrained. Overall, our findings open the door to interpretation methods that, at least in part, abstract away from model specifics and operate in the embedding space only.

1 Introduction

Transformer-based models [Vaswani et al., 2017] currently dominate Natural Language Processing [Devlin et al., 2018; Radford et al., 2019; Zhang et al., 2022] as well as many other fields of machine learning [Dosovitskiy et al., 2020; Chen et al., 2020; Baevski et al., 2020]. Consequently, understanding their inner workings has been a topic of great interest. Typically, work on interpreting Transformers relies on feeding inputs to the model and analyzing the resulting activations [Adi et al., 2016; Shi et al., 2016; Clark et al., 2019]. Thus, interpretation involves an expensive forward, and sometimes also a backward pass, over multiple inputs. Moreover, such interpretation methods are conditioned on the input, and are not guaranteed to generalize to all inputs. In the evolving literature on static interpretation, i.e., without forward or backward passes, Geva et al. [2022b] showed that the value vectors of the Transformer feed-forward module (the second layer of the feed-forward network) can be interpreted by projecting them into the embedding space, i.e., multiplying them by the embedding matrix to obtain a representation over vocabulary items. Elhage et al. [2021] have shown that in a 2-layer attention network, weight matrices can be interpreted in the embedding space as well.

In this work, we extend the theoretical analysis and findings of Elhage et al. [2021] and Geva et al. [2022b], and present a zero-pass framework to understand the behaviour of Transformers. Conceretely, we interpret *all* weights of a pretrained language model (LM) in embedding space, including both keys and values of the feed-forward module as well as all attention parameters.

Our theory relies on a simple observation. Since Geva et al. [2022b] have shown that one can project hidden states to the embedding space via the embedding matrix, we can extend this to other parts of the model by projecting to the embedding space and then *projecting back* by multiplying with a right-inverse of the embedding matrix. Thus, we can recast inner products in the model as inner products *in embedding space*. Viewing inner products in this way, we can interpret such products as interactions between pairs of vocabulary items.[1] This applies to (a) interactions between

[1] We refer to the unique items of the vocabulary as *vocabulary items*, and to the (possibly duplicate) elements of a tokenized input as *tokens*.

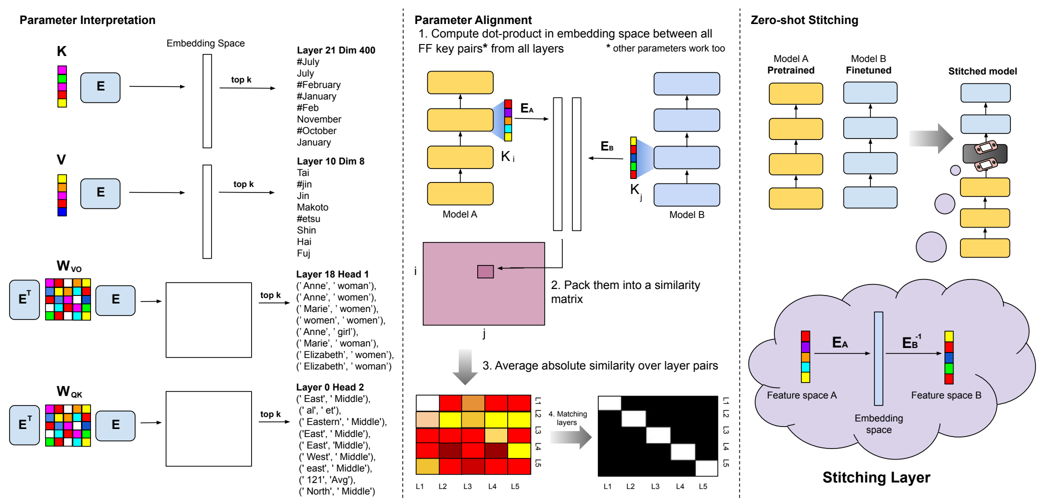


Figure 1: Applications of the embedding space view. *Left*: interpreting parameters in embedding space. The most active vocabulary items for an example feed-forward key (k) and a feed-forward value (v). The most active pairs of vocabulary items for an example attention query-key matrix W_{QK} and an attention value-output matrix W_{VO} (see §2). *Center*: Aligning the parameters of different BERT instances that share a vocabulary. *Right*: Zero-shot "stitching", where representations of a fine-tuned classifier are translated through the embedding space (multiplying by $E_A E_B^{-1}$) to a pretrained-only model.

attention queries and keys as well as to (b) interactions between attention value vectors and the parameters that project them at the output of the attention module. Taking this perspective to an extreme, one can view Transformers as operating implicitly in the embedding space. This entails the existence of a *single* linear space that depends solely on the tokenizer, in which parameters of different Transformers can be compared. Thus, one can use the embedding space to compare and transfer information across different models that share a tokenizer.

We provide extensive empirical evidence for the credibility of our proposal. On the interpretation front (Fig. 1, Left), we provide qualitative and quantitative evidence that Transformer parameters can be interpreted in embedding space. We also show that when fine-tuning a pretrained LM on a sentiment analysis task (over movie reviews), projecting *changes* in parameters into embedding space yields words that characterize sentiment towards movies. Second (Fig. 1, Center), we show that given two distinct instances of BERT pretrained with different random seeds [Sellam et al., 2022], we can align layers of the two instances by casting their weights into the embedding space. We find that indeed layer *i* of the first instance aligns well to layer *i* of the second instance, showing the different BERT instances converge to a semantically-similar solution. Last (Fig. 1, Right), we take a model fine-tuned on a sentiment analysis task and "transfer" the learned weights to a different model that was only pretrained by going through the embedding spaces of the two models. We show that in 30% of the cases, this procedure, termed *stitching*, results in a classifier that reaches an impressive accuracy of 70% on the IMDB benchmark [Maas et al., 2011] without any training.

Overall, our findings suggest that analyzing Transformers in embedding space is fruitful for both interpretability and as a tool to relate different models that share a vocabulary, and opens the door to interpretation methods that operate in embedding space only. Our code is available at `https://anonymized`.

2 Background

We now present the main components of the Transformer [Vaswani et al., 2017] relevant to our analysis. We discuss the residual stream view of Transformers, and recapitulate a view of the attention layer parameters as *interaction matrices* W_{VO} and W_{QK} [Elhage et al., 2021]. Similar to Elhage et al. [2021], we exclude biases and layer normalization from our analysis.

2.1 Transformer Architecture

The Transformer consists of a stack of layers, each includes an attention module followed by a Feed-Forward (FF) module. All inputs and outputs are sequences of N vectors of dimensionality d.

The Attention Module takes as input a sequence of representations $X \in \mathbb{R}^{N \times d}$, and each layer L is parameterized by four matrices $W_Q^{(L)}, W_K^{(L)}, W_V^{(L)}, W_O^{(L)} \in \mathbb{R}^{d \times d}$ (we henceforth omit the layer superscript for brevity). The input X is projected to produce queries, keys, and values: $Q_{\text{att}} = XW_Q, K_{\text{att}} = XW_K, V_{\text{att}} = XW_V$. Each one of $Q_{\text{att}}, K_{\text{att}}, V_{\text{att}}$ is split along the columns to H different *heads* of dimensionality $\mathbb{R}^{N \times \frac{d}{H}}$, denoted by $Q_{\text{att}}^i, K_{\text{att}}^i, V_{\text{att}}^i$ respectively. We then compute H *attention maps*:

$$A^i = \text{softmax}\left(\frac{Q_{\text{att}}^i K_{\text{att}}^{i\text{T}}}{\sqrt{d/H}} + M\right) \in \mathbb{R}^{N \times N},$$

where $M \in \mathbb{R}^{N \times N}$ is the attention mask. Each attention map is applied to the corresponding value head as $A^i V_{\text{att}}^i$, results are concatenated along columns and projected via W_O. The input to the module is added via a residual connection, and thus the attention module's output is:

$$X + \textbf{Concat}\left[A^1 V_{\text{att}}^1, \ldots, A^i V_{\text{att}}^i, \ldots, A^H V_{\text{att}}^H\right] W_O. \tag{1}$$

The FF Module is a two-layer neural network, applied to each position independently. Following past terminology [Sukhbaatar et al., 2019; Geva et al., 2020], weights of the first layer are called *FF keys* and weights of the second layer *FF values*. This is an analogy to attention, as the FF module too can be expressed as: $f(QK^{\text{T}})V$, where f is the activation function, $Q \in \mathbb{R}^{N \times d}$ is the output of the attention module and the input to the FF module, and $K, V \in \mathbb{R}^{d_{\mathit{ff}} \times d}$ are the weights of the first and second layers of the FF module. Unlike attention, keys and values are learnable parameters. The output of the FF module is added to the output of the attention module to form the output of the layer via a residual connection. The output of the i-th layer is called the i-th *hidden state*.

Embedding Matrix To process sequences of discrete tokens, Transformers use an embedding matrix $E \in \mathbb{R}^{d \times e}$ that provides a d-dimensional representation to vocabulary items before entering the *first* Transformer layer. When training Transformers with a language modeling objective, the same embedding matrix E is often used [Press and Wolf, 2016] to take the output of the *last* Transformer layer and project it back to the vocabulary dimension, i.e., into the *embedding space*. In this work, we will interpret all components of the Transformer model in the embedding space.

2.2 The Residual Stream

We rely on a useful view of the Transformer through its residual connections proposed by Elhage et al. [2021].[2] Specifically, each layer takes a hidden state as input and adds information to the hidden state through its residual connection. Under this view, the hidden state is a *residual stream* passed along the layers, from which information is read, and to which information is written at each layer. Elhage et al. [2021] and Geva et al. [2022b] observed that the residual stream is often barely updated in the last layers, and thus the final prediction is determined in early layers and the hidden state is mostly passed through the later layers.

An exciting consequence of the residual stream view is that we can project hidden states in *every* layer into embedding space by multiplying the hidden state with the embedding matrix E, treating the hidden state as if it were the output of the last layer. Geva et al. [2022a] used this approach to interpret the prediction of Transformer-based language models, and we follow a similar approach.

2.3 W_{QK} and W_{VO}

Following Elhage et al. [2021], we describe the attention module in terms of *interaction matrices* W_{QK} and W_{VO} which will be later used in our theoretical derivation. The computation of the attention module (§2.1) can be re-interpreted as follows. The attention projection matrices $W_{\text{Q}}, W_{\text{K}}, W_{\text{V}}$ can be split along the *column* axis to H equal parts denoted by $W_{\text{Q}}^i, W_{\text{K}}^i, W_{\text{V}}^i \in \mathbb{R}^{d \times \frac{d}{H}}$ for $1 \leq i \leq H$. Similarly, the attention output matrix W_{O} can be split along the *row* axis into H heads, $W_{\text{O}}^i \in \mathbb{R}^{d/H \times d}$. We define the *interaction matrices* as

$$W_{\text{QK}}^i := W_{\text{Q}}^i W_{\text{K}}^{i\text{T}} \in \mathbb{R}^{d \times d}, \qquad W_{\text{VO}}^i := W_{\text{V}}^i W_{\text{O}}^i \in \mathbb{R}^{d \times d}.$$

[2] Though earlier mentions include nostalgebraist [2020].

Importantly, $W_{\text{QK}}^i, W_{\text{VO}}^i$ are *input-independent*. Intuitively, W_{QK} encodes the amount of attention between pairs of tokens. Similarly, in W_{VO}^i, the matrices W_{V} and W_{O} can be viewed as a transition matrix that determines how attending to certain tokens affects the subsequent hidden state. We can restate the attention equations in terms of the interaction matrices. Recall (Eq. 1) that the output of the i'th head of the attention module is $A^i V_{\text{att}}^i$ and the final output of the attention module is (without the residual connection):

$$\textbf{Concat}\left[A^1 V_{\text{att}}^1, ..., A^i V_{\text{att}}^i, ..., A^H V_{\text{att}}^H\right] W_{\text{O}} = \sum_{i=1}^{H} A^i (X W_{\text{V}}^i) W_{\text{O}}^i = \sum_{i=1}^{H} A^i X W_{\text{VO}}^i. \tag{2}$$

Similarly, the attention map A^i at the i'th head in terms of W_{QK} is (softmax is done row-wise):

$$A^i = \text{softmax}\left(\frac{(X W_{\text{Q}}^i)(X W_{\text{K}}^i)^{\text{T}}}{\sqrt{d/H}} + M\right) = \text{softmax}\left(\frac{X(W_{\text{QK}}^i)X^{\text{T}}}{\sqrt{d/H}} + M\right). \tag{3}$$

3 Projecting Transformer Parameters into Embedding Space

In this section, we propose that Transformer parameters can be projected into embedding space for interpretation purposes. Our results extend Elhage et al. [2021] who obtained similar results for a two-layer attention-only network. We empirically support our framework in §4-§5.

Given a matrix $A \in \mathbb{R}^{N \times d}$, we can project it into embedding space by multiplying by the embedding matrix E as $\hat{A} = AE \in \mathbb{R}^{N \times e}$. Let E' be a right-inverse of E, that is, $EE' = I \in \mathbb{R}^{d \times d}$.[3] Then we can reconstruct the original matrix with E' as $A = A(EE') = \hat{A}E'$. We will use this simple identity to reinterpret the model's operation in embedding space. To simplify our analysis, we ignore layer norms and biases, a standard simplification justified in prior work [Elhage et al., 2021].

In interpretation experiments (§4), we do not use an exact right inverse such as the Moore–Penrose pseudo-inverse [Moore, 1920; Bjerhammar, 1951; Penrose, 1955] but instead use the transpose of the embedding matrix $E' = E^{\text{T}}$. This is since interpretation involves not only projecting using E' but also applying a top-k operation where we inspect the vocabulary items with the largest logits. We empirically find that the Moore–Penrose pseudo-inverse does not work well for interpretation due to the top-k operation, and provide a justification and comprehensive empirical evidence in Appendix A. Conversely, E^{T} empirically works well, and we conjecture this is due to the training procedure of LMs where E is used to embed discrete tokens into the hidden state dimension and E^{T} is used to predict a distribution over the vocabulary items from the last hidden state.

Attention Module Recall that $W_{\text{VO}}^i := W_{\text{V}}^i W_{\text{O}}^i \in \mathbb{R}^{d \times d}$ is the interaction matrix between attention values and the output projection matrix for attention head i. By definition, the output of each head is: $A^i X W_{\text{VO}}^i = A^i \hat{X} E' W_{\text{VO}}^i$. Since the output of the attention module is added to the residual stream, we can assume according to the residual stream view that it is meaningful to project it to the embedding space, similar to FF values. Thus, we expect the sequence of N e-dimensional vectors $(A^i X W_{\text{VO}}^i)E = A^i \hat{X}(E' W_{\text{VO}}^i E)$ to be interpretable. Importantly, the role of A^i is just to mix the representations of the updated N input vectors. This is similar to the FF module, where FF values (the parameters of the second layer) are projected into embedding space, and FF keys (parameters of the first layer) determine the *coefficients* for mixing them. Hence, we can assume that the interpretable components are in the term $\hat{X}(E' W_{\text{VO}}^i E)$.

Zooming in on this operation, we see that it takes the previous hidden state in the embedding space ($\hat{X}$) and produces an output in the embedding space which will be incorporated into the next hidden state through the residual stream. Thus, $E' W_{\text{VO}}^i E$ is a *transition matrix* that takes a representation the embedding space and outputs a new representation in the same space.

Similarly, the matrix W_{QK}^i can be viewed as a bilinear map (Eq. 3). To interpret it in embedding space, we perform the following operation with E':

$$X W_{\text{QK}}^i X^{\text{T}} = (XEE') W_{\text{QK}}^i (XEE')^{\text{T}} = (XE)E' W_{\text{QK}}^i E'^{\text{T}} (XE)^{\text{T}} = \hat{X}(E' W_{\text{QK}}^i E'^{\text{T}})\hat{X}^{\text{T}}.$$

[3] E' exists if $d \leq e$ and E is full-rank.

	Symbol	Projection	Approximate Projection
FF values	v	vE	vE
FF keys	k	kE'^{T}	kE
Attention query-key	W_{QK}^{i}	$E'W_{\mathrm{QK}}^{i}E'^{\mathrm{T}}$	$E^{\mathrm{T}}W_{\mathrm{QK}}^{i}E$
Attention value-output	W_{VO}^{i}	$E'W_{\mathrm{VO}}^{i}E$	$E^{\mathrm{T}}W_{\mathrm{VO}}^{i}E$
Attention value subheads	$W_{\mathrm{V}}^{i,j}$	$W_{\mathrm{V}}^{i,j}E'^{\mathrm{T}}$	$W_{\mathrm{V}}^{i,j}E$
Attention output subheads	$W_{\mathrm{O}}^{i,j}$	$W_{\mathrm{O}}^{i,j}E$	$W_{\mathrm{O}}^{i,j}E$
Attention query subheads	$W_{\mathrm{Q}}^{i,j}$	$W_{\mathrm{Q}}^{i,j}E'^{\mathrm{T}}$	$W_{\mathrm{Q}}^{i,j}E$
Attention key subheads	$W_{\mathrm{K}}^{i,j}$	$W_{\mathrm{K}}^{i,j}E'^{\mathrm{T}}$	$W_{\mathrm{K}}^{i,j}E$

Table 1: A summary of our approach for projecting Transformer components into embedding space. The 'Approximate Projection' shows the projection we use in practice where $E' = E^{\mathrm{T}}$.

Therefore, the interaction between tokens at different positions is determined by an $e \times e$ matrix that expresses the interaction between pairs of vocabulary items.

FF Module Geva et al. [2022b] showed that FF value vectors $V \in \mathbb{R}^{d_{ff} \times d}$ are meaningful when projected into embedding space, i.e., for a FF value vector $v \in \mathbb{R}^{d}$, $vE \in \mathbb{R}^{e}$ is interpretable (see §2.1). In vectorized form, the rows of $VE \in \mathbb{R}^{d_{ff} \times e}$ are interpretable. On the other hand, the keys K of the FF layer are multiplied on the left by the output of the attention module, which are the queries of the FF layer. Denoting the output of the attention module by Q, we can write this product as $QK^{\mathrm{T}} = \hat{Q}E'K^{\mathrm{T}} = \hat{Q}(KE'^{\mathrm{T}})^{\mathrm{T}}$. Because Q is a hidden state, we assume according to the residual stream view that $\hat{Q}$ is interpretable in embedding space. When multiplying $\hat{Q}$ by KE'^{T}, we are capturing the interaction in embedding space between each query and key, and thus expect KE'^{T} to be interpretable in embedding space as well.

Overall, FF keys and values are intimately connected – the i-th key controls the coefficient of the i-th value, so we expect their interpretation to be related. While not central to this work, we empirically show that key-value pairs in the FF module are similar in embedding space in Appendix B.1.

Subheads Another way to interpret the matrices W_{VO}^{i} and W_{QK}^{i} is through the *subhead view*. We use the following identity: $AB = \sum_{j=1}^{b} A_{:,j}B_{j,:}$, which holds for arbitrary matrices $A \in \mathbb{R}^{a \times b}, B \in \mathbb{R}^{b \times c}$, where $A_{:,j} \in \mathbb{R}^{a \times 1}$ are the *columns* of the matrix A and $B_{j,:} \in \mathbb{R}^{1 \times c}$ are the *rows* of the matrix B. Thus, we can decompose W_{VO}^{i} and W_{QK}^{i} into a sum of $\frac{d}{H}$ rank-1 matrices:

$$W_{\mathrm{VO}}^{i} = \sum_{j=1}^{\frac{d}{H}} W_{\mathrm{V}}^{i,j} W_{\mathrm{O}}^{i,j}, \quad W_{\mathrm{QK}}^{i} = \sum_{j=1}^{\frac{d}{H}} W_{\mathrm{Q}}^{i,j} {W_{\mathrm{K}}^{i,j}}^{\mathrm{T}}.$$

where $W_{\mathrm{Q}}^{i,j}, W_{\mathrm{K}}^{i,j}, W_{\mathrm{V}}^{i,j} \in \mathbb{R}^{d \times 1}$ are columns of $W_{\mathrm{Q}}^{i}, W_{\mathrm{K}}^{i}, W_{\mathrm{V}}^{i}$ respectively, and $W_{\mathrm{O}}^{i,j} \in \mathbb{R}^{1 \times d}$ are the rows of W_{O}^{i}. We call these vectors *subheads*. This view is useful since it allows us to interpret subheads directly by multiplying them with the embedding matrix E. Moreover, it shows a parallel between interaction matrices in the attention module and the FF module. Just like the FF module includes key-value pairs as described above, for a given head, its interaction matrices are a sum of interactions between pairs of subheads (indexed by j), which are likely to be related in embedding space. We show this is indeed empirically the case for pairs of subheads in Appendix B.1.

We summarize our approach for projecting the different components of the Transformer into embedding space in Table 1.

4 Interpretability Experiments

In this section, we provide empirical evidence for the viability of our approach as a tool for interpreting Transformer parameters.

4.1 Parameter Interpretation Examples

We take GPT-2 medium [Radford et al., 2019] and manually analyze its parameters. GPT-2 medium has a total of 384 attention heads (24 layers and 16 heads per layer). We take the embedded transition

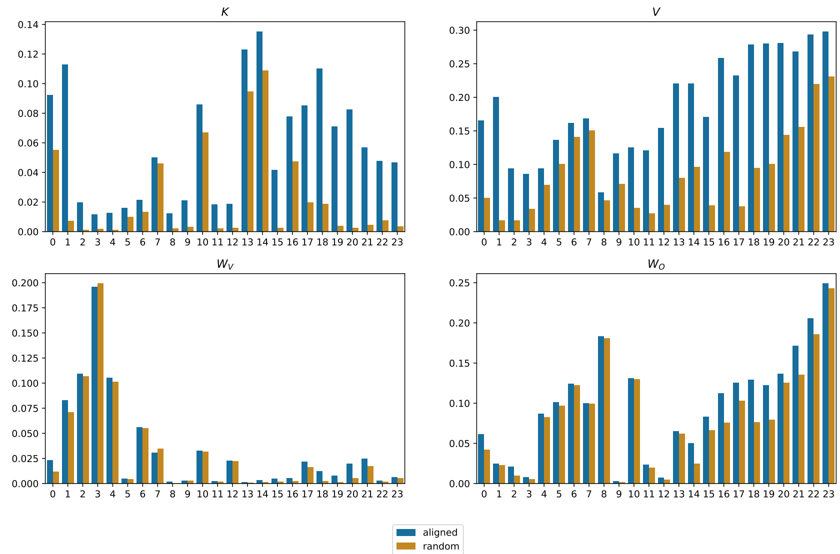


Figure 2: Left: Average R_k score ($k = 100$) across tokens per layer for activated parameter vectors against both the aligned hidden state $\hat{h}$ at the output of the layer and a randomly sampled hidden state $\hat{h}_{\text{rand}}$. Parameters are FF keys (top-left), FF values (top-right), attention values (bottom-left), and attention outputs (bottom-right).

matrices $E'W_{\text{VO}}^{i}E$ for all heads and examine the top-k pairs of vocabulary items. As there are only 384 heads, we manually choose a few heads and present the top-k pairs in Appendix C.1 ($k = 50$). We observe that different heads capture different types of relations between pairs of vocabulary items including word parts, heads that focus on gender, geography, orthography, particular part-of-speech tags, and various semantic topics. In Appendix C.2 we perform a similar analysis for W_{QK}.

Appendix C.3 provides examples of key-value pairs from the FF modules of GPT-2 medium. We show random pairs (k, v) from the set of those pairs such that when looking at the top-100 vocabulary items for k and v, at least 15% overlap. Such pairs account for approximately 5% of all key-value pairs. The examples show how key-value pairs often revolve around similar topics such as media, months, organs, etc.

Last, we show we can use embeddings to locate FF values (or keys) related to a particular topic. We take a few vocabulary items related to a certain topic, e.g., `['cm', 'kg', 'inches']`, average their embeddings,[4] and rank all FF values (or keys) based on their dot-product with the average. Appendix C.4 shows a few examples of FF values found with this method that are related to programming, measurements, and animals.

4.2 Hidden State and Parameters

An advantage of zero-pass interpretation is that it does not require running inputs through the model which is expensive and non-exhaustive. In this section (and this section only), we run a forward pass over inputs and examine if the representations in embedding space of dynamically-computed hidden states are "similar" to the representations of static parameter vectors that are activated.

A technical side note: we use GPT-2, which applies layer norm to the Transformer output before projecting it to the embedding space with E. Thus, conservatively, layer norm should be considered as part of the projection operation.[5] Empirically however, we observe that projecting parameters directly without layer norm works well, which simplifies our analysis in §3. An exception is when projecting hidden states in this section, where we apply layer norm before projection to improve performance, similar to Geva et al. [2022a].

Experimental Design We use GPT-2 medium and run it over 60 examples from IMDB [Maas et al., 2011]. This provides us with a dynamically-computed hidden state h for every token and at the output of every layer. For the projection $\hat{h} \in \mathbb{R}^e$ of each such hidden state, we take the projections of the m most active parameter vectors $\{\hat{x}_i\}_{i=1}^{m}$ in the layer that computed h and check

[4]We subtract the average embedding μ from E before averaging, which improves interpretability.

[5]Layer norm consists of standardizing the mean and variance of the input followed by an affine transformation. The latter part can be easily absorbed into E (while adding a bias term).

if they cover the dominant vocabulary items of $\hat{h}$ in embedding space. Specifically, let $\texttt{top-k}(wE)$ be the k vocabulary items with largest logits in embedding space for a vector $w \in \mathbb{R}^d$. We compute:

$$R_k(\hat{x}_1, ..., \hat{x}_m, \hat{h}) = \frac{|\texttt{top-k}(\hat{h}) \cap \bigcup_{i=1}^{m} \texttt{top-k}(\hat{x}_i)|}{k},$$

to capture if activated parameter vectors cover the main vocabulary items corresponding to the hidden state.

We find the m most active parameter vectors separately for FF keys (K), FF values (V), attention value *subheads* (W_{V}) (see §3), and attention output subheads (W_{O}), where the activation of each parameter vector is determined by the vector's "coefficient" as follows. For a FF key-value pair (k, v) the coefficient is $\sigma(q^{\text{T}}k)$, where $q \in \mathbb{R}^d$ is an input to the FF module, and σ is the FF non-linearity. For attention value-output subhead pairs (v, o) the coefficient is $x^{\text{T}}v$, where x is the input to this component (for attention head i, the input is one of the rows of A^iX, see Eq. 2).

Results and Discussion Figure 2 presents the R_k score averaged across tokens per layer. As a baseline, we compare R_k of the activated vectors $\{\hat{x}_i\}_{i=1}^{m}$ with the correctly-aligned hidden state $\hat{h}$ at the output of the relevant layer (blue bars) against the the R_k when randomly sampling $\hat{h}_{\text{rand}}$ from the set of all hidden states (orange bars). We conclude that the representations in embedding space induced by activated parameter vector mirror, at least to some extent, the representations of the hidden states themselves. Appendix §B.2 shows a variant of this experiment, where we compare activated parameters throughout GPT2-medium's layers to the last hidden state, which produces the logits used for prediction.

4.3 Interpretation of Fine-tuned Models

We now show that we can interpret the *changes* a model goes through during fune-tuning through the lens of embedding space. We fine-tune the top-3 layers of the 12-layer GPT-2-base with a sequence classification head on IMDB sentiment analysis (binary classification) and compute the difference between the original parameters and the fine-tuned model. We then project the difference of parameter vectors into embedding space and test if change is interpretable w.r.t sentiment analysis.

Appendix D shows examples for projected differences randomly sampled from the fine-tuned layers. Frequently, the difference, or its negation, is projected to nouns, adjectives and adverbs that express sentiment for a movie, such as *'amazing'*, *'masterpiece'*, *'incompetence'*, etc. This shows that the differences are indeed projected into vocabulary items that characterize movie reviews' sentiment. Almost all parameter groups present this behavior, except for V and W_{O}, which curiously are the parameters added to the residual stream.

5 Aligning Models in Embedding Space

Assuming Transformers by and large operate in embedding space leads to an exciting possibility - we can relate *different* models to one another so long as they share a vocabulary and tokenizer. In §5.1, we show that we can align the layers of BERT models trained with different random seeds. In §5.2, we show the embedding space can be leveraged to "stitch" the parameters of a fine-tuned model to a model that was not fine-tuned.

5.1 Layer Alignment

Experimental Design Taking our approach to the extreme, the embedding space is a universal space, which depends only on the tokenizer, and in which Transformer parameters and hidden states reside. Consequently, we can align parameter vectors from different models in this space and compare them even if they come from different models, as long as they share a vocabulary.

To demonstrate this, we use MultiBERT [Sellam et al., 2022], which contains 25 different instantiations of BERT initialized from different random seeds. We take parameters from two MultiBERT seeds and compute the Pearson correlation between their projection to embedding space. For example, let V_A, V_B be the FF values of models A and B. We can project the values into embedding space: V_AE_A, V_BE_B, where E_A, E_B are the respective embedding matrices, and compute Pearson correlation between projected values. This produces a similarity matrix $\tilde{\mathcal{S}} \in \mathbb{R}^{|V_A| \times |V_B|}$, where each

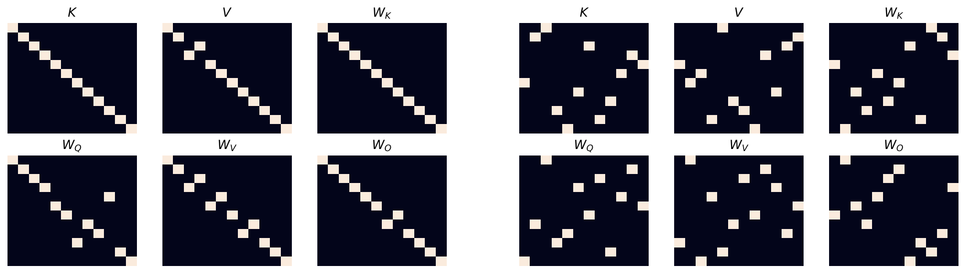


Figure 3: Left: Aligning *in embedding space* the layers of two different BERT models initialized from different random seeds for all parameter groups. Layers that have the same index tend to align with one another. Right: Alignment in feature space leads to unintelligible patterns.

entry is the correlation coefficient between projected values from the two models. We bin $\tilde{\mathcal{S}}$ by layer pairs and average the absolute value of the scores in each bin (different models might encode the same information in different directions, so we use absolute value) to produce a matrix $\mathcal{S} \in \mathbb{R}^{L \times L}$, where L is the number of layers. Specifically, the average (absolute) correlation between vectors that come from layer ℓ_A in model A and layer ℓ_B in Model B is registered in entry (ℓ_A, ℓ_B) of $\mathcal{S}$.

Last, to obtain a one-to-one layer alignment, we use the Hungarian algorithm [Kuhn, 1955], which assigns exactly one layer from the first model to a layer from the second model. The algorithm's objective is to maximize, given a similarity matrix $\mathcal{S}$, the sum of similarities of the chosen pairs, such that each index in one model is matched with exactly one index in the other. We repeat this for all parameter groups $(W_{\mathrm{Q}}, W_{\mathrm{K}}, W_{\mathrm{V}}, W_{\mathrm{O}}, K)$.

Results and Discussion Figure 3 (left) shows the resulting alignment. Clearly, parameters from a certain layer in model A tend to align to the same layer in model B across all parameter groups. This suggests that different layers from different models that were trained separately (but with the same training objective and data) serve a similar function. As further evidence, we show that if not projected, the matching appears absolutely random in Figure §3 (right). We show the same results for other seed pairs as well in Appendix B.3.

5.2 Zero-shot Stitching

Model stitching [Lenc and Vedaldi, 2015; Csiszárik et al., 2021; Bansal et al., 2021] is a relatively under-explored feature of neural networks, particularly in NLP. The idea is that different models, sometimes trained on different data and with different architectures, learn representations that can be aligned through a *linear* transformation, termed *stitching*. Representations correspond to hidden states , and thus one can learn a transformation matrix from one model's hidden states to an equivalent hidden state in the other model. Here, we show that going through embedding space one can align the hidden states of two models, i.e., stitch, *without training*.

Given two models, we want to find a linear stitching transformation to align their representation spaces. According to our theory, given a hidden state $v \in \mathbb{R}^{d_1}$ from model A, we can project it to the embedding space as vE_A, where E_A is its embedding matrix. Then, we can re-project to the feature space of model B, with $E_B^+ \in \mathbb{R}^{e \times d_2}$, where E_B^+ is the Penrose-Moore pseudo-inverse of the embedding matrix E_B.[6] This transformation can be expressed as multiplication with the kernel $K_{AB} := E_A E_B^+ \in \mathbb{R}^{d_1 \times d_2}$. We employ the above approach to take representations of a fine-tuned classifier, A, and stitch them on top of a model B that was only pretrained, to obtain a new classifier based on B.

Experimental Design We use the 24-layer GPT-2 medium as model A and 12-layer GPT-2 base model trained in §4.3 as model B. We fine-tune the last three layers of model B on IMDB, as explained in §4.3. Stitching is simple and is performed as follows. Given the sequence of N hidden states $H_A^\ell \in \mathbb{R}^{N \times d_1}$ at the output of layer ℓ of model A (ℓ is a hyperparameter), we apply the *stitching layer*, which multiplies the hidden states with the kernel, computing $H_A^\ell K_{AB}$. This results in hidden states $H_B \in \mathbb{R}^{N \times d_2}$, used as input to the three fine-tuned layers from B.

[6]Since we are not interested in interpretation we use an exact right-inverse and not the transpose.

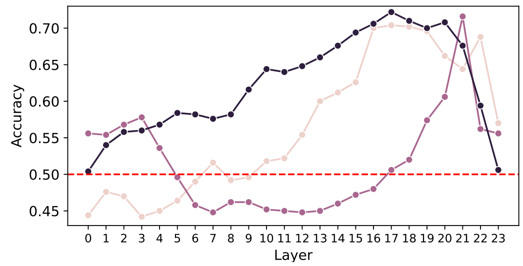


Figure 4: Accuracy on IMDB evaluation set. We ran stitching randomly 11 times and obtained 3 models with higher than random accuracy when stitching over top layers. Dashed red line indicates random performance.

Results and Discussion Stitching produces models with accuracies that are higher than random on IMDB evaluation set, but not consistently. Figure 4 shows the accuracy of stitched models against the layer index from model A over which stitching is performed. Out of 11 random seeds, three models obtained accuracy that is significantly higher than the baseline 50% accuracy, reaching an accuracy of roughly 70%, when stitching is done over the top layers.

6 Related Work

Interpreting Transformer is a broad area of research that has attracted much attention in recent years. A large body of work has focused on analyzing hidden representations, mostly through probing [Adi et al., 2016; Shi et al., 2016; Tenney et al., 2019; Rogers et al., 2020]. Voita et al. [2019a] used statistical tools to analyze the evolution of hidden representations throughout layers. Recently, Mickus et al. [2022] proposed to decompose the hidden representations into the contributions of different Transformer components. Unlike these works, we interpret parameters rather than the hidden representations.

Another substantial effort has been to interpret specific network components. Previous work analyzed single neurons [Dalvi et al., 2018; Durrani et al., 2020], attention heads [Clark et al., 2019; Voita et al., 2019b], and feedforward values [Geva et al., 2020; Dai et al., 2021; Elhage et al., 2022]. While these works mostly rely on input-dependent neuron activations, we inspect "static" model parameters, and provide a comprehensive view of all Transformer components.

Our work is most related to efforts to interpret specific groups of Transformer parameters. Cammarata et al. [2020] made observations about the interpretability of weights of neural networks. Elhage et al. [2021] analyzed 2-layer attention networks. We extend their analysis to multi-layer pre-trained Transformer models. Geva et al. [2020; 2022a;b] interpreted feedforward values in embedding space. We coalesce these lines of work and offer a unified interpretation framework for Transformers in embedding space.

7 Discussion

Our work has a few limitations that we care to highlight. First, it focuses on interpreting models through the vocabulary lens. While we have shown evidence for this, it does not preclude other factors from being involved in the computation process. Second, we used $E' = E^{\mathrm{T}}$, but future research might find variants of E that improve performance. Last, we assume Transformer components can be projected to the embedding space with a single matrix multiplication, but this might depend on model training, e.g., in GPT-2 it involves a layer norm operation as explained in §4.2.

Notwithstanding, we believe the benefits of our work overshadow its limitations. We provide a simple and efficient approach, which equips researchers with new tools to interpret Transformer models and relate them to one another. Apart from Elhage et al. [2021], there has been little work pursuing the embedding space approach, and we "sharpen" the tools they laid down and adjust them to existing pre-trained Transformers. Moreover, our framework allows us to view parameters from different models as residents of the same universal embedding space, where they can be compared in model-agnostic fashion. We demonstrate two applications of this observation (model alignment and stitching) and argue future work can yield many additional applications.

References

Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, and Y. Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks, 2016. URL https://arxiv.org/abs/1608.04207.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477.

Y. Bansal, P. Nakkiran, and B. Barak. Revisiting model stitching to compare neural representations. In *NeurIPS*, 2021.

A. Bjerhammar. Application of calculus of matrices to method of least squares : with special reference to geodetic calculations. In *Trans. Roy. Inst. Tech. Stockholm*, 1951.

N. Cammarata, S. Carter, G. Goh, C. Olah, M. Petrov, L. Schubert, C. Voss, B. Egan, and S. K. Lim. Thread: Circuits. *Distill*, 2020. doi: 10.23915/distill.00024. https://distill.pub/2020/circuits.

M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20s.html.

K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341, 2019. URL http://arxiv.org/abs/1906.04341.

A. Csiszárik, P. Korösi-Szabó, Á. K. Matszangosz, G. Papp, and D. Varga. Similarity and matching of neural network representations. In *NeurIPS*, 2021.

D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei. Knowledge neurons in pretrained transformers, 2021. URL https://arxiv.org/abs/2104.08696.

F. Dalvi, N. Durrani, H. Sajjad, Y. Belinkov, A. Bau, and J. Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models, 2018. URL https://arxiv.org/abs/1812.09355.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL https://arxiv.org/abs/2010.11929.

N. Durrani, H. Sajjad, F. Dalvi, and Y. Belinkov. Analyzing individual neurons in pre-trained language models. *CoRR*, abs/2010.02695, 2020. URL https://arxiv.org/abs/2010.02695.

N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

N. Elhage, T. Hume, C. Olsson, N. Nanda, T. Henighan, S. Johnston, S. ElShowk, N. Joseph, N. DasSarma, B. Mann, D. Hernandez, A. Askell, K. Ndousse, A. Jones, D. Drain, A. Chen, Y. Bai, D. Ganguli, L. Lovitt, Z. Hatfield-Dodds, J. Kernion, T. Conerly, S. Kravec, S. Fort, S. Kadavath, J. Jacobson, E. Tran-Johnson, J. Kaplan, J. Clark, T. Brown, S. McCandlish, D. Amodei, and C. Olah. Softmax linear units. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/solu/index.html.

K. Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings, 2019. URL https://arxiv.org/abs/1909.00512.

J. Gao, D. He, X. Tan, T. Qin, L. Wang, and T. Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkEYojRqtm.

M. Geva, R. Schuster, J. Berant, and O. Levy. Transformer feed-forward layers are key-value memories, 2020. URL https://arxiv.org/abs/2012.14913.

M. Geva, A. Caciularu, G. Dar, P. Roit, S. Sadde, M. Shlain, B. Tamir, and Y. Goldberg. Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models. *arXiv preprint arXiv:2204.12130*, 2022a.

M. Geva, A. Caciularu, K. R. Wang, and Y. Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space, 2022b. URL https://arxiv.org/abs/2203.14680.

P. Jaccard. The distribution of the flora in the alpine zone. *The New Phytologist*, 11(2):37–50, 1912. ISSN 0028646X, 14698137. URL http://www.jstor.org/stable/2427226.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 991–999, 2015.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

T. Mickus, D. Paperno, and M. Constant. How to dissect a muppet: The structure of transformer embedding spaces. *arXiv preprint arXiv:2206.03529*, 2022.

E. H. Moore. On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.*, 26:394–395, 1920.

nostalgebraist. interpreting gpt: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

R. Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge University Press, 1955.

O. Press and L. Wolf. Using the output embedding to improve language models, 2016. URL https://arxiv.org/abs/1608.05859.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. In *OpenAI blog*, 2019.

A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in bertology: What we know about how bert works, 2020. URL https://arxiv.org/abs/2002.12327.

W. Rudman, N. Gillman, T. Rayne, and C. Eickhoff. Isoscore: Measuring the uniformity of vector space utilization. *CoRR*, abs/2108.07344, 2021. URL https://arxiv.org/abs/2108.07344.

T. Sellam, S. Yadlowsky, I. Tenney, J. Wei, N. Saphra, A. D'Amour, T. Linzen, J. Bastings, I. R. Turc, J. Eisenstein, D. Das, and E. Pavlick. The multiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=K0E_F0gFDgA.

X. Shi, I. Padhi, and K. Knight. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1159. URL https://aclanthology.org/D16-1159.

S. Sukhbaatar, E. Grave, G. Lample, H. Jegou, and A. Joulin. Augmenting self-attention with persistent memory. *arXiv preprint arXiv:1907.01470*, 2019.

I. Tenney, D. Das, and E. Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

E. Voita, R. Sennrich, and I. Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives, 2019a. URL https://arxiv.org/abs/1909.01380.

E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580.

L. Wang, J. Huang, K. Huang, Z. Hu, G. Wang, and Q. Gu. Improving neural language generation with spectrum control. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxY8CNtvr.

S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

A Rethinking Interpretation

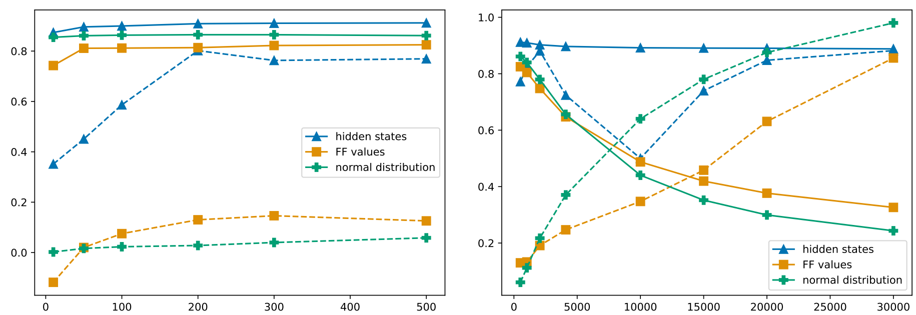


Figure 5: *Left*: The `keep-k` inverse scores for three distributions: normal distribution, hidden states, and FF values, for $k \in \{10, 50, 100, 200, 300, 500\}$. *Right*: for $k \in \{10, 50, 100, 200, 300, 500\}$.

The process of interpreting a vector v in Geva et al. [2022b] proceeds in two steps: first the *projection* of the vector to the embedding space (vE); then, we use the list of the tokens that were assigned the largest values in the projected vector, i.e.: $\texttt{top-k}(vE)$, as the *interpretation* of the projected vector. This is reasonable since (a) the most activated coordinates contribute the most when added to the residual stream, and (b) this matches how we eventually decode: we project to the embedding space and consider the top-1 token (or one of the few top tokens, when using beam search).

In this work, we interpret inner products and matrix multiplications in the embedding space: given two vectors $x, y \in \mathbb{R}^d$, their inner product $x^{\mathrm{T}}y$ can be considered in the embedding space by multiplying with E and then by one of its right inverses (e.g., its pseudo-inverse E^+ [Moore, 1920; Bjerhammar, 1951; Penrose, 1955]): $x^{\mathrm{T}}y = x^{\mathrm{T}}EE^+y = (x^{\mathrm{T}}E)(yE^{+\mathrm{T}})^{\mathrm{T}}$. Assume xE is interpretable in the embedding space, crudely meaning that it represents logits over vocabulary items. We expect y, which interacts with x, to also be interpretable in the embedding space. Consequently, we would like to take $yE^{+\mathrm{T}}$ to be the projection of y. However, this projection does not take into account the subsequent interpretation using top-k. The projected vector $yE^{+\mathrm{T}}$ might be harder to interpret in terms of its most activated tokens. To alleviate this problem, we need a different "inverse" matrix E' that works well when considering the top-k operation. Formally, we want an E' with the following "robustness" guarantee: $\texttt{keep-k}(xE)^{\mathrm{T}}\texttt{keep-k}(yE') \approx x^{\mathrm{T}}y$, where $\texttt{keep-k}(v)$ is equal to v for coordinates whose absolute value is in the top-k, and zero elsewhere.

This is a stronger notion of inverse – not only is $EE' \approx I$, but even when truncating the vector in the embedding space we can still reconstruct it with E'.

We claim that E^{T} is a decent instantiation of E' and provide some empirical evidence. While a substantive line of work [Ethayarajh, 2019; Gao et al., 2019; Wang et al., 2020; Rudman et al., 2021] has shown that embedding matrices are not isotropic (an isotropic matrix E has to satisfy $EE^{\mathrm{T}} = \alpha I$ for some scalar α), we show that it is isotropic enough to make E^{T} a legitimate compromise. We randomly sample 300 vectors drawn from the normal distribution $\mathcal{N}(0, 1)$, and compute for every pair x, y the cosine similarity between $x^{\mathrm{T}}y$ and $\texttt{keep-k}(xE)^{\mathrm{T}}\texttt{keep-k}(yE')$ for $k = 1000$, and then average over all pairs. We repeat this for $E' \in \{E^{+\mathrm{T}}, E\}$ and obtain a score of 0.10 for $E^{+\mathrm{T}}$, and 0.83 for E, showing the E is better under when using top-k. More globally, we compare $E' \in \{E^{+\mathrm{T}}, E\}$ for $k \in \{10, 50, 100, 200, 300, 500\}$ with three distributions:

- x, y drawn from the normal $\mathcal{N}(0, 1)$ distribution
- x, y chosen randomly from the FF values
- x, y drawn from hidden states along Transformer computations.

In Figure 5 (Left) we show the results, where dashed lines represent E^+ and solid lines represent E^{T}. For small values of k (used for interpretation), E^{T} is superior to E^+ across all distributions. Interestingly, the hidden state distribution is the only distribution where E^+ has similar performance to E^{T}. Curiously, when looking at higher values of k the trend is reversed ($k = \{512, 1024, 2048, 4096, 10000, 15000, 20000, 30000\}$) - see Figure 5 (Right).

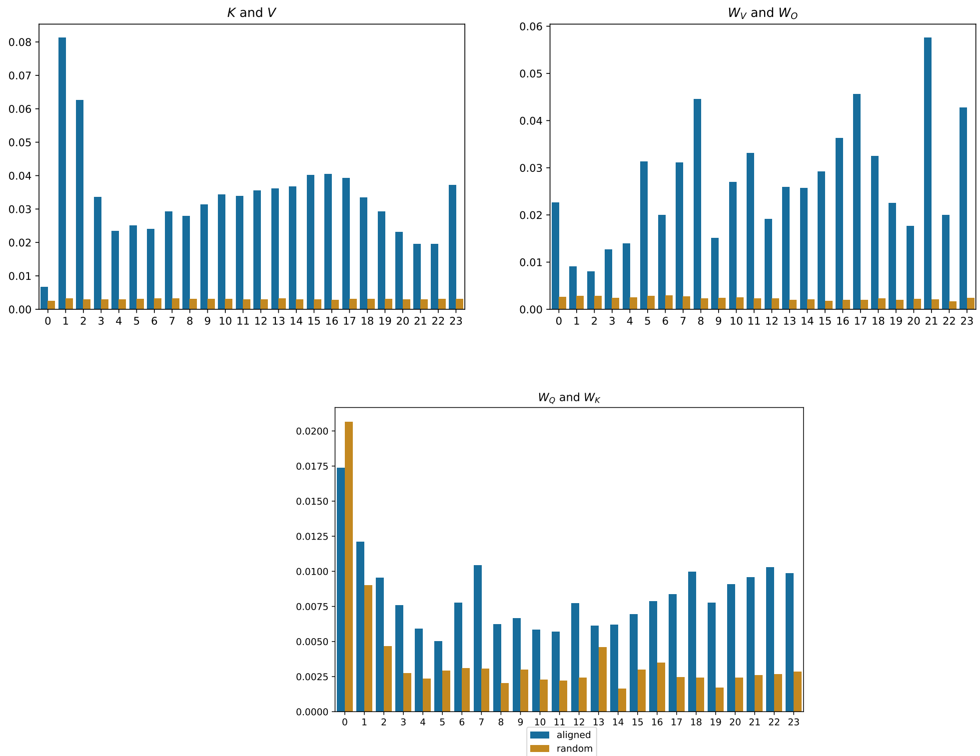


Figure 6: Average $\text{Sim}_k(\hat{x}, \hat{y})$ for $k = 100$ by layer, where blue is when matching pairs are aligned, and orange is when pairs are shuffled within the layer. Top Left: FF keys and FF values. Top Right: The subheads of W_O and W_V. Bottom: The subheads of W_Q and W_K.

This settles the deviation from findings showing embedding matrices are not isotropic, as we see that indeed as k grows, E^{T} becomes an increasingly bad approximate right-inverse of the embedding matrix. The only distribution that keeps high performance with E^{T} is the hidden state distribution, which is an interesting future direction of investigation.

B Additional Material

B.1 Corresponding Parameter Pairs are Related

We define the following metric applying on vectors *after projecting* them into the embedding space:

$$\text{Sim}_k(\hat{x}, \hat{y}) = \frac{|\texttt{top-k}(\hat{x}) \cap \texttt{top-k}(\hat{y})|}{|\texttt{top-k}(\hat{x}) \cup \texttt{top-k}(\hat{y})|}$$

where $\texttt{top-k}(v)$ is the set of k top activated indices in the vector v (which correspond to tokens in the embedding space). This metric is the Jaccard index [Jaccard, 1912] applied to the top-k tokens from each vector. In Figure 6, Left, we demonstrate that corresponding FF key and value vectors are more similar (in embedding space) than two random key and value vectors. In Figure 6, Right, we show a similar result for attention value and output vectors. In Figure 6, Bottom, the same analysis in done for attention query and key vectors. This shows that there is a much higher-than-chance relation between corresponding FF keys and values (and the same for attention values and outputs).

B.2 Final Prediction and Parameters

We show that the final prediction of the model is correlated in embedding space with the most activated parameters from each layer. This implies that these objects are germane to the analysis of the final prediction in the embedding space, which in turn suggests that the embedding space is a

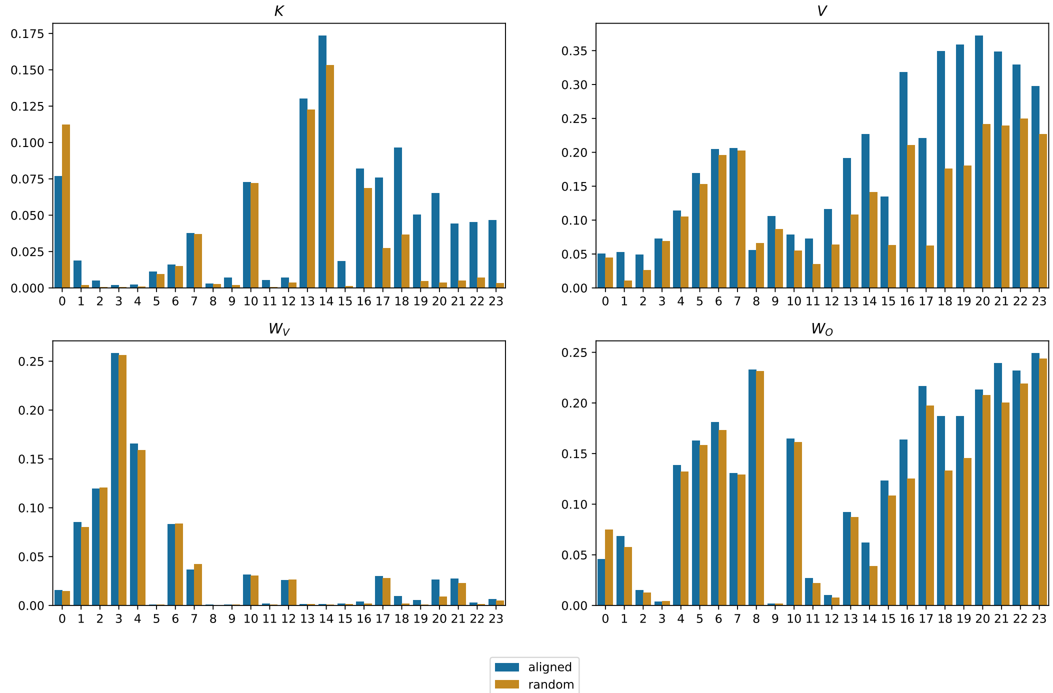


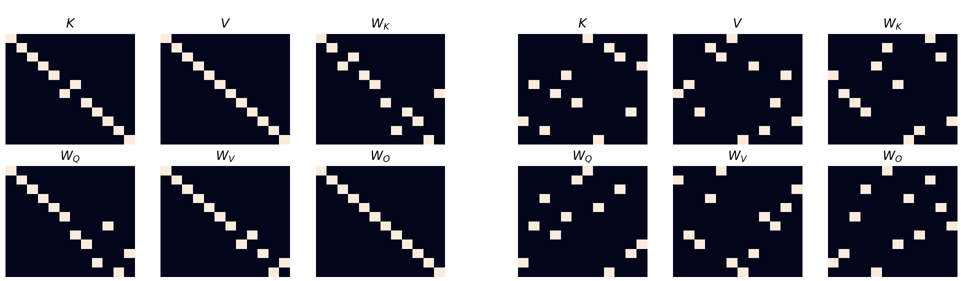
Figure 7: Left: Average R_k score ($k = 100$) across tokens per layer for activated parameter vectors against both the aligned hidden state $\hat{h}$ at the output of the *final* layer and a randomly sampled hidden state $\hat{h}_{\text{rand}}$. Parameters are FF keys (top-left), FF values (top-right), attention values (bottom-left), and attention outputs (bottom-right).

viable choice for interpreting these vectors. Figure 7 shows that just like §4.2, correspondence is better when hidden states are not randomized, suggesting there parameter interpretations have an impact on the final prediction.

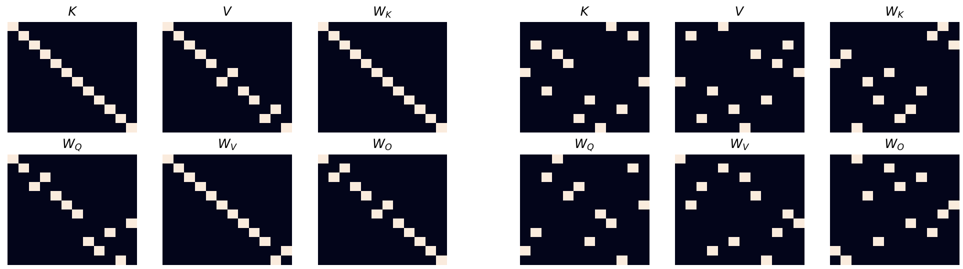
B.3 Parameter Alignment Plots for Additional Model Pairs

Alignment in embedding space of layers of pairs of BERT models trained with different random seeds for additional model pairs.

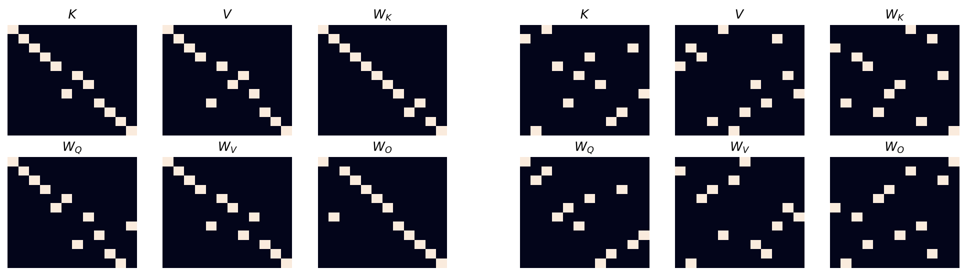
Seed 1 VS Seed 2



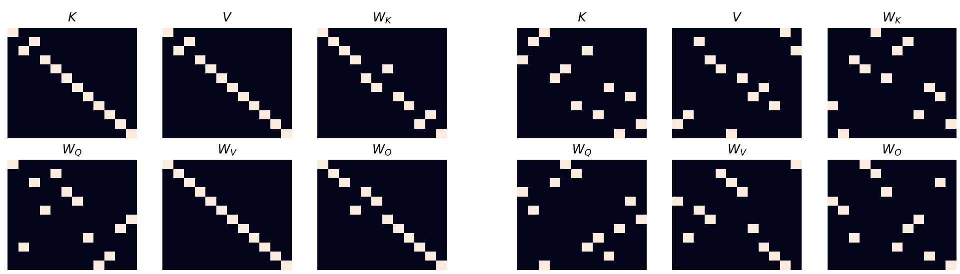
SEED 2 VS SEED 3



SEED 3 VS SEED 4



SEED 4 VS SEED 5



C EXAMPLE CASES

C.1 VALUE-OUTPUT MATRICES

Below we show value-output pairs from different heads of GPT-2 Medium. For each head, we show the 50 pairs with largest value in the $e \times e$ transition matrix. There are 384 attention heads in GPT-2 medium from which we manually choose a subset. Throughout the section some lists were marked with asterisks indicating the way this particular list was created:

- * - pairs of the form (x, x) were excluded from the list

C.1.1 LOW LEVEL LANGUAGE MODELING

Layer 21 Head 7*

```
('FN', 'NF'),
(' Ramos', 'Ram'),
(' Hughes', 'Hug'),
('GR', 'gran'),
('NF', 'FN'),
('CL', 'CLA'),
(' McCain', 'McC'),
(' Marshall', 'Marsh'),
('Hug', ' Hughes'),
(' Tanner', 'Tan'),
('NH', 'nih'),
('NR', 'NRS'),
('Bow', ' Bowman'),
('Marsh', ' Marshall'),
(' Jacobs', 'Jac'),
(' Hayes', 'Hay'),
('Hay', ' Hayes'),
(' McCorm', 'McC'),
('NR', 'NI'),
(' Dawson', ' sidx'),
('Tan', ' Tanner'),
('GR', 'gra'),
('jac', 'JA'),
('zo', 'zos'),
('NF', 'NI'),
(' McCull', 'McC'),
('Jac', ' Jacobs'),
(' Beet', ' Beetle'),
('FG', 'GF'),
('ja', 'jas'),
(' Wilkinson', 'Wil'),
('Ram', ' Ramos'),
('GR', 'GRE'),
('FN', ' NF'),
('McC', ' McCorm'),
(' Scarborough', 'Scar'),
('Ba', ' Baal'),
('FG', 'FP'),
('FN', 'FH'),
('Gar', ' Garfield'),
('jac', 'jas'),
('nut', 'nuts'),
(' Wis', 'WI'),
(' Vaughan', ' Vaughn'),
('PF', 'FP'),
('RN', 'RNA'),
('jac', ' Jacobs'),
('FN', 'FM'),
('Kn', ' Knox'),
('nic', 'NI')
```

Layer 19 Head 13 (guessing the first letter/consonant of the word)

```
('senal', ' R'),        # arsenal
('senal', 'R'),
('vernment', ' G'),     # government
(' Madness', ' M'),
(' Mayhem', ' M'),
('nesday', ' W'),       # wednesday
('vernment', 'G'),
(' Madness', 'M'),
('lace', ' N'),         # necklace
('nesday', 'W'),
('senal', 'Rs'),
('vernment', ' g'),
('farious', ' N'),      # nefarious
('eneg', ' C'),
('senal', ' r'),
('ruary', ' F'),        # february
('senal', 'RIC'),
('ondo', ' R'),
(' Mandela', ' N'),     # nelson
(' Mayhem', 'M'),
('senal', ' RD'),
('estine', ' C'),
('vernment', 'Gs'),
('senal', 'RF'),
('esis', ' N'),
('Reviewed', ' N'),
('arette', ' C'),       # cigarette
('rome', ' N'),
('theless', ' N'),      # nonetheless
('lace', 'N'),
('DEN', ' H'),
(' versa', ' V'),
('bably', ' P'),        # probably
('vernment', 'GF'),
('vernment', 'g'),
('vernment', 'GP'),
('ornia', ' C'),        # california
('ilipp', ' F'),
('umbered', ' N'),
('arettes', ' C'),
('senal', 'RS'),
('onsense', ' N'),
('senal', 'RD'),
('senal', 'RAL'),
('uci', ' F'),
('ondo', 'R'),
('senal', ' RI'),
('iday', ' H'),         # holiday
('senal', ' Rx'),
('odor', ' F')
```

Layer 20 Head 9

```
(' behalf', 'On'),
(' behalf', ' On'),
(' behalf', ' on'),
(' periods', 'during'),
(' bounds', 'within'),
(' envelope', ' inside'),
('door', 'outside'),
(' envelope', 'inside'),
(' regime', ' Under'),
(' periods', ' during'),
('lihood', ' LIKE'),
(' occasions', ' on'),
(' regime', 'Under'),
('door', 'inside'),
('period', 'during'),
('lihood', 'Like'),
(' periods', ' During'),
(' envelope', 'Inside'),
(' sake', 'for'),
(' doors', ' inside'),
(' regime', ' under'),
(' behalf', ' ON'),
(' purposes', 'for'),
(' occasions', 'On'),
(' doors', 'inside'),
(' basis', ' on'),
(' regimes', ' Under'),
('doors', 'outside'),
(' Osc', 'inside'),
(' periods', 'During'),
('door', ' inside'),
(' regime', ' UNDER'),
(' regimes', ' under'),
(' regimes', 'Under'),
('doors', 'inside'),
('zx', 'inside'),
(' period', 'during'),
('ascript', 'inside'),
('door', 'Inside'),
(' occasions', ' On'),
('ysc', 'BuyableInstoreAndOnline')
  ,
(' envelope', ' Inside'),
(' pauses', 'during'),
(' regime', 'under'),
(' occasion', ' on'),
(' doors', 'outside'),
(' banner', ' UNDER'),
(' envelope', 'within'),
('abouts', ' here'),
(' duration', 'during')
```

Layer 22 Head 5 (named entities, mostly made of two parts)

```
('enegger', ' Schwartz'),
('shire', ' Lincoln'),
('xual', 'Weiss'),
('nery', ' Nun'),
(' Qiao', ' Huang'),
('schild', ' Schwarz'),
('oslov', ' Czech'),
(' Rica', ' Costa'),
(' Qiao', ' Qiao'),
('xual', ' RW'),
(' Nadu', ' Tamil'),
(' Nadu', 'Tam'),
('shire', ' Baldwin'),
('swick', ' Hoff'),
('xual', ' Weiss'),
(' Takeru', ' Yamato'),
('xual', ' Grassley'),
('swick', ' Schwartz'),
('enegger', ' Schiff'),
('enegger', 'Weiss'),
('xual', 'RW'),
('shire', ' Nottingham'),
('shire', ' Barrett'),
('arest', ' Buch'),
(' Fei', ' Fei'),
('miah', 'Jere'),
('swick', ' Owl'),
('ufact', ' Swanson'),
('akuya', ' Tanaka'),
(' Sachs', ' Feinstein'),
('enegger', ' Wagner'),
('otle', 'Roberts'),
('shire', ' Neville'),
('oslov', ' Prague'),
('sburg', ' Hammond'),
(' ILCS', ' Dunham'),
(' Malfoy', ' Draco'),
('yip', 'Billy'),
('iversal', ' Monroe'),
('iversal', 'Murray'),
('Yang', 'Yang'),
('akuya', ' Krishna'),
('schild', ' Schwartz'),
('tz', ' Rabb'),
('shire', 'gow'),
('enegger', ' Feldman'),
('cair', ' Chou'),
('enegger', ' Duffy'),
('enegger', 'Sch'),
(' Jensen', ' Jensen')
```

Layer 22 Head 13

```
(' Additionally', ' the'),
(' Unfortunately', ' the'),
(' Nevertheless', ' the'),
(' Sadly', ' the'),
(' However', ' the'),
(' Furthermore', ' the'),
(' Additionally', ','),
(' During', ' the'),
(' Moreover', ' the'),
(' Whilst', ' the'),
(' Since', ' the'),
(' Unfortunately', ','),
(' Additionally', '-'),
(' Perhaps', ' the'),
(' Sadly', ','),
(' Throughout', ' the'),
(' Nevertheless', ','),
(' While', ' the'),
(' However', ','),
(' Although', ' the'),
(' There', ' the'),
(' Furthermore', ','),
```

```
(' Eventually', ' the'),
(' Meanwhile', ' the'),
(' Hopefully', ' the'),
(' Nevertheless', '-'),
(' During', ','),
(' Regardless', ' the'),
(' However', '-'),
(' Whilst', ','),
(' Additionally', ' and'),
(' Moreover', ','),
(' Unfortunately', '-'),
(' They', ' the'),
(' Sadly', '-'),
(' Whereas', ' the'),
(' Additionally', ' a'),
(' Furthermore', '-'),
(' Unlike', ' the'),
(' Typically', ' the'),
(' Since', ','),
(' Normally', ' the'),
(' Perhaps', ','),
(' During', '-'),
(' Throughout', ','),
(' While', ','),
(' Nevertheless', ' a'),
(' Interestingly', ' the'),
(' Unfortunately', ' and'),
(' Unfortunately', ' a')
```

C.1.2 GENDER

Layer 18 Head 1

```
(' Marie', 'women'),
(' Marie', ' actresses'),
(' Anne', 'women'),
(' Anne', 'Women'),
(' Marie', 'woman'),
(' Marie', 'Women'),
(' Anne', 'woman'),
(' Marie', 'Woman'),
(' Anne', ' actresses'),
(' Marie', ' heroine'),
('Jane', 'Women'),
(' Anne', ' heroine'),
('Jane', 'women'),
(' actresses', 'Women'),
(' Anne', 'Woman'),
(' Esther', 'Women'),
(' Esther', 'women'),
(' Marie', 'girls'),
(' Anne', 'Mrs'),
(' Marie', ' actress'),
(' actresses', 'women'),
('Jane', 'Woman'),
(' Marie', ' girls'),
('Jane', ' actresses'),
('Anne', 'Woman'),
(' Marie', 'Girls'),
('Anne', 'women'),
(' Anne', 'Girls'),
(' actresses', 'Woman'),
(' Marie', ' Women'),
(' Anne', ' Women'),
(' Anne', ' girls'),
(' Anne', 'girl'),
('Anne', 'Women'),
('Women', 'Woman'),
(' Anne', 'girls'),
('Anne', ' actresses'),
(' Michelle', 'women'),
(' Marie', ' Actress'),
(' Marie', 'girl'),
(' Anne', ' Feminist'),
(' Marie', ' women'),
(' Devi', 'Women'),
(' Elizabeth', 'Women'),
(' Anne', ' actress'),
('Anne', 'Mrs'),
('Answer', 'answered'),
('Anne', 'woman'),
('maid', 'Woman'),
('Marie', 'women')
```

C.1.3 GEOGRAPHY

Layer 16 Head 6*

```
(' Mumbai', ' Chennai'),
(' Mumbai', 'India'),
(' Chennai', ' Mumbai'),
(' Tasmania', ' Queensland'),
(' Rahul', 'India'),
(' Gujar', 'India'),
(' Bangalore', ' Chennai'),
('Scotland', 'England'),
(' Kerala', ' Chennai'),
(' Mumbai', ' Delhi'),
('Scotland', 'Britain'),
(' Mumbai', ' Bangalore'),
('India', 'Pakistan'),
('Ireland', 'Scotland'),
(' Bangalore', ' Mumbai'),
(' Chennai', ' Bangalore'),
(' Gujar', ' Aadhaar'),
(' Maharashtra', ' Mumbai'),
(' Gujarat', ' Maharashtra'),
(' Gujar', ' Gujarat'),
('Australia', 'Australian'),
(' Gujarat', 'India'),
(' Gujar', ' Rahul'),
(' Mumbai', ' Maharashtra'),
('England', 'Britain'),
(' Chennai', 'India'),
(' Bombay', ' Mumbai'),
(' Kerala', ' Tamil'),
(' Mumbai', ' Hindi'),
(' Tasman', ' Tasmania'),
('India', ' Mumbai'),
(' Gujar', ' Hindi'),
(' Gujar', ' Maharashtra'),
('Austral', ' Australians'),
(' Kerala', ' Maharashtra'),
(' Bangalore', 'India'),
(' Kerala', 'India'),
(' Bombay', 'India'),
('Austral', 'Australia'),
```

```
('India', ' Aadhaar'),
(' Mumbai', ' Sharma'),
('Austral', 'Australian'),
(' Kerala', ' Mumbai'),
('England', 'Scotland'),
(' Gujar', ' Mumbai'),
(' Mumbai', ' Rahul'),
(' Tasman', ' Queensland'),
(' Chennai', ' Tamil'),
(' Maharashtra', ' Gujarat'),
(' Modi', 'India')
```

Layer 18 Head 9

```
(' Winnipeg', ' Winnipeg'),
(' Edmonton', ' Winnipeg'),
(' Winnipeg', ' Ottawa'),
(' Calgary', ' Winnipeg'),
(' Ottawa', ' Winnipeg'),
(' Winnipeg', ' Calgary'),
(' Winnipeg', 'CBC'),
(' Winnipeg', 'Canada'),
(' Canberra', ' Canberra'),
(' RCMP', ' Winnipeg'),
(' Ottawa', 'CBC'),
(' Winnipeg', 'Canadian'),
('Toronto', ' Winnipeg'),
(' Winnipeg', ' Canadians'),
(' Edmonton', ' Ottawa'),
(' Winnipeg', ' RCMP'),
(' Winnipeg', ' Edmonton'),
(' Ottawa', 'Canadian'),
('Canadian', ' Winnipeg'),
('Toronto', ' Calgary'),
(' Winnipeg', ' Quebec'),
(' Winnipeg', ' Canad'),
('Toronto', 'Canadian'),
(' Edmonton', ' Edmonton'),
(' Ottawa', ' Calgary'),
(' Leafs', ' Winnipeg'),
(' Edmonton', ' Calgary'),
(' Ottawa', 'Canada'),
(' Calgary', 'Canadian'),
('Toronto', 'Canada'),
(' Calgary', ' Calgary'),
('Ott', ' Winnipeg'),
(' Winnipeg', ' Saskatchewan'),
(' Winnipeg', ' Canadian'),
(' Ottawa', ' Ottawa'),
(' Calgary', ' Ottawa'),
(' Winnipeg', ' Manitoba'),
(' Canadians', ' Winnipeg'),
(' Winnipeg', ' Canada'),
(' RCMP', ' Calgary'),
('Toronto', ' Manitoba'),
('Toronto', ' Ottawa'),
('CBC', ' Winnipeg'),
('Canadian', 'Canada'),
(' Edmonton', 'Canadian'),
(' RCMP', ' Ottawa'),
(' Winnipeg', 'ipeg'),
('Toronto', 'Toronto'),
('Canadian', ' Calgary'),
(' Ottawa', ' Canadians')
```

Layer 16 Head 2*

```
(' Australians', 'Austral'),
('Austral', 'Australia'),
('Austral', ' Canberra'),
(' Canberra', 'Austral'),
(' Edmonton', ' Winnipeg'),
('Austral', 'Australian'),
(' Edmonton', ' Alberta'),
(' Australians', 'Australia'),
('Austral', ' Australians'),
('ovych', 'Ukraine'),
(' Canad', ' Quebec'),
(' Australians', 'Australian'),
(' Manitoba', ' Winnipeg'),
(' Winnipeg', ' Manitoba'),
('Canada', 'Canadian'),
(' Bulgar', 'Moscow'),
(' Edmonton', ' Manitoba'),
('Austral', 'berra'),
('Australian', 'Austral'),
('ovych', ' Ukrainians'),
(' Canadians', 'Canada'),
(' Australians', ' Canberra'),
('Canadian', 'Canada'),
('ovych', ' Yanukovych'),
(' Trudeau', 'Canada'),
(' Bulgar', ' Dmitry'),
('Austral', ' Australia'),
(' Canad', ' Mulcair'),
(' Canberra', 'berra'),
('oglu', 'Turkish'),
('Canada', 'udeau'),
(' Oilers', ' Edmonton'),
(' Canberra', 'Australia'),
(' Edmonton', 'Canada'),
(' Calgary', ' Edmonton'),
(' Calgary', ' Alberta'),
(' Trudeau', 'udeau'),
(' Edmonton', ' Calgary'),
(' Trudeau', 'Canadian'),
(' Canberra', 'Australian'),
(' Canucks', ' Vancouver'),
('Australian', 'Australia'),
(' Fraser', ' Vancouver'),
(' Edmonton', 'Canadian'),
('elaide', 'Austral'),
(' Braz', 'Tex'),
(' RCMP', 'Canada'),
('sov', 'Moscow'),
(' Bulgar', 'Russia'),
('Canada', ' Canadians')
```

Layer 21 Head 12*

```
(' Indones', ' Indonesian'),
(' Nguyen', ' Vietnamese'),
(' Jakarta', ' Indonesian'),
(' Indonesia', ' Indonesian'),
('oglu', 'Turkish'),
(' Indones', ' Indonesia'),
(' Indones', ' Jakarta'),
(' Koreans', ' Korean'),
```

```
(’oglu’, ’ Turkish’),
(’ Taiwanese’, ’ Taiwan’),
(’ Nguyen’, ’ Thai’),
(’Brazil’, ’ Brazilian’),
(’ Indonesia’, ’ Indones’),
(’ Taiwanese’, ’Tai’),
(’oglu’, ’ Istanbul’),
(’ Indonesian’, ’ Indones’),
(’ Jakarta’, ’ Indones’),
(’ Nguyen’, ’ Laos’),
(’ Sloven’, ’ Slovenia’),
(’ Korean’, ’ Koreans’),
(’ Nguyen’, ’ Cambod’),
(’zzi’, ’Italy’),
(’Tai’, ’ Taiwanese’),
(’ Jakarta’, ’ Indonesia’),
(’ Indonesian’, ’ Indonesia’),
(’ Bulgaria’, ’ Bulgarian’),
(’ Icelandic’, ’ Iceland’),
(’ Koreans’, ’ Korea’),
(’ Brazilian’, ’Brazil’),
(’ Bulgar’, ’ Bulgarian’),
(’ Malays’, ’ Malaysian’),
(’oglu’, ’ Ankara’),
(’ Bulgarian’, ’ Bulgaria’),
(’ Indones’, ’ Malays’),
(’ Tai’, ’ Taiwanese’),
(’oglu’, ’Turkey’),
(’ Janeiro’, ’Brazil’),
(’zzi’, ’Italian’),
(’ Malays’, ’ Kuala’),
(’ Fuk’, ’Japanese’),
(’ Indonesian’, ’ Jakarta’),
(’ Taiwan’, ’ Taiwanese’),
(’oglu’, ’ Erdogan’),
(’ Nguyen’, ’ Viet’),
(’ Filipino’, ’ Philippine’),
(’ Indonesia’, ’ Jakarta’),
(’ Jong’, ’ Koreans’),
(’ Duterte’, ’ Filipino’),
(’ Azerbai’, ’ Azerbaijan’),
(’ Bulgarian’, ’ Bulgar’)
```

C.1.4 British Spelling

Layer 19 Head 4

```
(’ Whilst’, ’ realise’),
(’ Whilst’, ’ Whilst’),
(’ Whilst’, ’ realised’),
(’ Whilst’, ’ organise’),
(’ Whilst’, ’ recognise’),
(’ Whilst’, ’ civilisation’),
(’ Whilst’, ’ organisation’),
(’ Whilst’, ’ whilst’),
(’ Whilst’, ’ organising’),
(’ Whilst’, ’ organised’),
(’ Whilst’, ’ organis’),
(’ Whilst’, ’ util’),
(’ Whilst’, ’ apologise’),
(’ Whilst’, ’ emphas’),
(’ Whilst’, ’ analyse’),
(’ Whilst’, ’ organisations’),
(’ Whilst’, ’ recognised’),
(’ Whilst’, ’ flavours’),
(’ Whilst’, ’ colour’),
(’ Whilst’, ’colour’),
(’ Whilst’, ’ Nasa’),
(’ Whilst’, ’ Nato’),
(’ Whilst’, ’ analys’),
(’ Whilst’, ’ flavour’),
(’ Whilst’, ’ colourful’),
(’ Whilst’, ’ colours’),
(’ organising’, ’ realise’),
(’ Whilst’, ’ behavioural’),
(’ Whilst’, ’ coloured’),
(’ Whilst’, ’ learnt’),
(’ Whilst’, ’ favourable’),
(’ Whilst’, ’isation’),
(’ Whilst’, ’ programmes’),
(’ organis’, ’ realise’),
(’ Whilst’, ’ authorised’),
(’ Whilst’, ’ practise’),
(’ Whilst’, ’ criticised’),
(’ Whilst’, ’ organisers’),
(’ organising’, ’ organise’),
(’ Whilst’, ’ analysed’),
(’ Whilst’, ’ programme’),
(’ Whilst’, ’ behaviours’),
(’ Whilst’, ’ humour’),
(’ Whilst’, ’isations’),
(’ Whilst’, ’ tyres’),
(’ Whilst’, ’ aluminium’),
(’ organised’, ’ realise’),
(’ Whilst’, ’ favour’),
(’ Whilst’, ’ ageing’),
(’ organis’, ’ organise’)
```

C.1.5 Related Words

Layer 13 Head 8*

```
(’ mirac’, ’ miraculous’),
(’ mirac’, ’ miracle’),
(’ nuanced’, ’ nuance’),
(’Better’, ’ smarter’),
(’ equitable’, ’ healthier’),
(’ liberating’, ’ liberated’),
(’ unaffected’, ’ untouched’),
(’ equitable’, ’ unbiased’),
(’ inconsistent’, ’failed’),
(’ emanc’, ’ liberated’),
(’ equitable’, ’ humane’),
(’ liberated’, ’ liberating’),
(’ incompatible’, ’failed’),
(’ mirac’, ’ miracles’),
(’ consensual’, ’ peacefully’),
(’ uncond’, ’ unconditional’),
(’ unexpected’, ’ unexpectedly’),
(’ unconditional’, ’ untouched’),
(’Better’, ’ healthier’),
(’ unexpectedly’, ’ unexpected’),
(’ graceful’, ’ peacefully’),
(’ emanc’, ’ emancipation’),
(’ effortlessly’, ’ seamlessly’),
(’ honorable’, ’ peacefully’),
(’ unconditional’, ’ uncond’),
(’ rubbish’, ’ excuses’),
```

```
(' emanc', ' liberating'),
(' equitable', ' peacefully'),
(' Feather', ' gracious'),
(' emancipation', ' liberated'),
(' nuanced', ' nuances'),
('icable', ' avoids'),
(' liberated', ' freeing'),
(' liberating', ' freeing'),
(' inconsistent', ' lousy'),
(' lousy', 'failed'),
(' unconditional', ' unaffected'),
(' equitable', 'ivable'),
(' equitable', 'Honest'),
('erning', ' principled'),
(' survival', 'surv'),
('ocre', ' lackluster'),
(' equitable', ' liberating'),
('Bah', 'Instead'),
(' incompatible', ' inappropriate
  '),
(' emancipation', ' emanc'),
(' unchanged', ' unaffected'),
(' peacefully', ' peaceful'),
(' equitable', ' safer'),
(' unconditional', ' uninterrupted
  ')
```

Layer 12 Head 14*

```
(' perished', ' died'),
(' perished', ' dies'),
(' testify', ' testifying'),
(' intervened', ' interven'),
(' advises', ' advising'),
(' disbanded', ' disband'),
('lost', ' perished'),
(' died', ' perished'),
(' applauded', ' applaud'),
(' dictates', ' dictate'),
(' prev', ' prevailed'),
(' advise', ' advising'),
('shed', 'thood'),
('Reviewed', 'orsi'),
(' dies', ' perished'),
('published', ' publishes'),
(' prevailed', ' prevail'),
(' died', ' dies'),
(' testified', ' testifying'),
(' testifying', ' testify'),
(' dictates', ' governs'),
(' complicit', ' complicity'),
(' dictated', ' dictate'),
('enough', 'CHO'),
(' skelet', 'independence'),
(' Recomm', ' prescribe'),
('essential', ' perished'),
('noticed', 'CHO'),
('avorable', ' approving'),
(' perish', ' perished'),
(' overseeing', ' oversee'),
(' skelet', 'shed'),
('EY', 'chart'),
(' presiding', ' overseeing'),
(' fundament', 'pees'),
(' sanction', 'appro'),
(' prevail', ' prevailed'),
(' governs', ' regulates'),
('tails', 'shed'),
(' Period', 'chart'),
('lihood', 'hower'),
(' prev', ' prevail'),
(' aids', 'helps'),
(' dictated', ' dict'),
(' dictated', ' dictates'),
(' Dise', 'itta'),
('REC', 'CHO'),
('exclusive', 'ORTS'),
(' Helpful', 'helps'),
('bart', 'ciples')
```

Layer 14 Head 1*

```
(' misunderstand', ' incorrectly')
  ,
(' Proper', ' properly'),
(' inaccur', ' incorrectly'),
(' misunderstand', ' wrongly'),
(' misinterpret', ' incorrectly'),
(' incorrect', ' incorrectly'),
(' mistakes', ' incorrectly'),
(' misunderstanding', '
  incorrectly'),
(' proper', ' properly'),
('fail', ' incorrectly'),
(' faulty', ' incorrectly'),
(' misrepresent', ' incorrectly'),
(' failing', ' fails'),
(' inaccurate', ' incorrectly'),
(' errors', ' incorrectly'),
(' harmful', ' Worse'),
(' misunderstand', ' wrong'),
(' misunderstand', ' improperly'),
('wrong', ' incorrectly'),
(' harmful', ' incorrectly'),
(' mistake', ' incorrectly'),
(' mis', ' incorrectly'),
('fail', ' fails'),
(' detrimental', ' Worse'),
(' rightful', ' properly'),
(' misunderstand', '
  inappropriately'),
(' harmful', ' unnecessarily'),
(' neglect', ' unnecessarily'),
(' correctly', ' properly'),
(' Worst', ' Worse'),
(' failure', ' fails'),
(' satisfactory', ' adequately'),
(' defective', ' incorrectly'),
(' misunderstand', ' mistakenly'),
(' harming', ' Worse'),
(' mishand', ' incorrectly'),
('adequ', ' adequately'),
(' misuse', ' incorrectly'),
('Failure', ' fails'),
(' hurts', ' Worse'),
(' misunderstand', 'wrong'),
(' mistakenly', ' incorrectly'),
(' failures', ' fails'),
(' adequate', ' adequately'),
(' properly', ' correctly'),
```

```
(’ hurting’, ’ Worse’),
(’ Proper’, ’ correctly’),
(’ fail’, ’ fails’),
(’ mistaken’, ’ incorrectly’),
(’ harming’, ’ adversely’)
```

Layer 14 Head 13*

```
(’ editors’, ’ editorial’),
(’ broadcasters’, ’ broadcasting’),
(’ broadcasting’, ’ broadcasts’),
(’ broadcast’, ’ broadcasts’),
(’ Broadcasting’, ’ broadcasters’),
(’ editors’, ’ Editorial’),
(’ broadcasters’, ’ broadcast’),
(’ Broadcasting’, ’ broadcast’),
(’ lectures’, ’ lecture’),
(’ Broadcast’, ’ broadcasting’),
(’ broadcasters’, ’ broadcaster’),
(’ broadcasters’, ’ broadcasts’),
(’ Publishers’, ’ publishing’),
(’ broadcasting’, ’ broadcast’),
(’ broadcasters’, ’ Broadcasting’),
(’ Publishers’, ’ Publishing’),
(’ lecture’, ’ lectures’),
(’ Editors’, ’ editorial’),
(’ broadcast’, ’ broadcasting’),
(’ Broadcasting’, ’ broadcasts’),
(’ broadcasting’, ’ broadcasters’),
(’ journalism’, ’ journalistic’),
(’reports’, ’Journal’),
(’ Broadcast’, ’ Broadcasting’),
(’ Publishers’, ’Publisher’),
(’azeera’, ’ Broadcasting’),
(’Reporting’, ’Journal’),
(’ journalistic’, ’ journalism’),
(’ Broadcasting’, ’ broadcaster’),
(’ broadcasting’, ’ broadcaster’),
(’ broadcaster’, ’ broadcasting’),
(’ editors’, ’ publication’),
(’ journalism’, ’journal’),
(’ Journalists’, ’Journal’),
(’ documentary’, ’ documentaries’),
(’ filming’, ’ filmed’),
(’ publishers’, ’ publishing’),
(’ journalism’, ’Journal’),
(’ Broadcast’, ’ broadcasts’),
(’ broadcast’, ’ broadcasters’),
(’ articles’, ’Journal’),
(’ reporting’, ’reports’),
(’ manuscripts’, ’ manuscript’),
(’ publish’, ’ publishing’),
(’azeera’, ’ broadcasters’),
(’ Publishers’, ’ publication’),
(’ Publishers’, ’ publications’),
(’ newspapers’, ’ Newsp’),
(’ Broadcast’, ’ broadcasters’),
(’ Readers’, ’Journal’)
```

C.2 QUERY-KEY MATRICES

Layer 22 Head 1

```
(’ usual’, ’ usual’),
(’ occasional’, ’ occasional’),
(’ aforementioned’, ’ aforementioned’),
(’ general’, ’ usual’),
(’ usual’, ’ slightest’),
(’agn’, ’ealous’),
(’ traditional’, ’ usual’),
(’ free’, ’amina’),
(’ major’, ’ major’),
(’ frequent’, ’ occasional’),
(’ generous’, ’ generous’),
(’ free’, ’lam’),
(’ regular’, ’ usual’),
(’ standard’, ’ usual’),
(’ main’, ’ usual’),
(’ complete’, ’ Finished’),
(’ main’, ’liest’),
(’ traditional’, ’ traditional’),
(’ latest’, ’ aforementioned’),
(’ current’, ’ aforementioned’),
(’ normal’, ’ usual’),
(’ dominant’, ’ dominant’),
(’ free’, ’ministic’),
(’ brief’, ’ brief’),
(’ biggest’, ’liest’),
(’usual’, ’ usual’),
(’ rash’, ’ rash’),
(’ regular’, ’ occasional’),
(’ specialized’, ’ specialized’),
(’ free’, ’iosis’),
(’ free’, ’hero’),
(’ specialty’, ’ specialty’),
(’ general’, ’iosis’),
(’ nearby’, ’ nearby’),
(’ best’, ’liest’),
(’ officially’, ’ formal’),
(’ immediate’, ’mediate’),
(’ special’, ’ ultimate’),
(’ free’, ’otropic’),
(’ rigorous’, ’ comparative’),
(’ actual’, ’ slightest’),
(’ complete’, ’ comparative’),
(’ typical’, ’ usual’),
(’ modern’, ’ modern’),
(’ best’, ’ smartest’),
(’ free’, ’ free’),
(’ highest’, ’ widest’),
(’ specialist’, ’ specialist’),
(’ appropriate’, ’ slightest’),
(’ usual’, ’liest’)
```

Layer 0 Head 9

```
(’59’, ’27’),
(’212’, ’39’),
(’212’, ’38’),
(’217’, ’39’),
(’37’, ’27’),
(’59’, ’26’),
(’54’, ’88’),
(’156’, ’39’),
```

```
('212', '79'),
('59', '28'),
('57', '27'),
('212', '57'),
('156', '29'),
('36', '27'),
('217', '79'),
('59', '38'),
('63', '27'),
('72', '39'),
('57', '26'),
('57', '34'),
('59', '34'),
('156', '27'),
('91', '27'),
('156', '38'),
('63', '26'),
('59', '25'),
('138', '27'),
('217', '38'),
('72', '27'),
('54', '27'),
('36', '29'),
('72', '26'),
('307', '39'),
('37', '26'),
('217', '57'),
('37', '29'),
('54', '38'),
('59', '29'),
('37', '28'),
('307', '38'),
('57', '29'),
('63', '29'),
('71', '27'),
('138', '78'),
('59', '88'),
('89', '27'),
('561', '79'),
('212', '29'),
('183', '27'),
('54', '29')
```

Layer 17 Head 6*

```
(' legally', ' legal'),
(' legal', ' sentencing'),
(' legal', ' arbitration'),
(' boycot', ' boycott'),
(' legal', ' criminal'),
(' legal', ' Judicial'),
(' legal', ' rulings'),
(' judicial', ' sentencing'),
(' marketing', ' advertising'),
(' legal', ' confidential'),
(' protesting', ' protest'),
(' recruited', ' recruit'),
(' recruited', ' recruits'),
(' judicial', ' criminal'),
(' legal', ' exemptions'),
(' demographics', ' demographic'),
(' boycott', ' boycot'),
(' sentencing', ' criminal'),
(' recruitment', ' recruits'),
(' recruitment', ' recruit'),
(' Constitutional', ' sentencing')
  ,
(' Legal', ' sentencing'),
(' constitutional', ' sentencing')
  ,
(' legal', ' subpoena'),
(' injury', ' injuries'),
(' FOIA', ' confidential'),
(' legal', ' licenses'),
(' donation', ' donations'),
(' disclosure', ' confidential'),
(' negotiation', ' negotiating'),
(' Judicial', ' legal'),
(' legally', ' criminal'),
(' legally', ' confidential'),
(' legal', ' jur'),
(' legal', ' enforcement'),
(' legal', ' lawyers'),
(' legally', ' enforcement'),
(' recruitment', ' recruiting'),
(' recruiting', ' recruit'),
(' criminal', ' sentencing'),
(' legal', ' attorneys'),
(' negotiations', ' negotiating'),
(' legally', ' arbitration'),
(' recruited', ' recruiting'),
(' legally', ' exemptions'),
(' legal', ' judicial'),
(' voting', ' Vote'),
(' negotiated', ' negotiating'),
(' legislative', ' veto'),
(' funding', ' funded')
```

Layer 17 Head 7

```
('tar', 'idia'),
(' [...]', '..."'),
(' lecture', ' lectures'),
(' Congress', ' senate'),
(' staff', ' staffers'),
(' Scholarship', ' collegiate'),
(' executive', ' overseeing'),
(' Scholarship', ' academic'),
(' academ', ' academic'),
('."', '..."'),
(' [', '..."'),
('";', '..."'),
(' Memorial', 'priv'),
(' festival', 'conference'),
('crew', ' supervisors'),
(' certification', ' grading'),
(' scholarship', ' academic'),
(' rumored', ' Academic'),
(' Congress', ' delegated'),
(' staff', ' technicians'),
('Plex', ' CONS'),
(' congress', ' senate'),
(' university', ' tenure'),
(' Congress', ' appointed'),
(' Congress', ' duly'),
(' investigative', ' investig'),
(' legislative', ' senate'),
('ademic', ' academic'),
('bench', ' academic'),
(' scholarship', ' tenure'),
```

```
(’ campus’, ’ campuses’),
(’ staff’, ’ Facilities’),
(’ Editorial’, ’mn’),
(’ clinic’, ’ laboratory’),
(’ crew’, ’ crews’),
(’ Scholarship’, ’ academ’),
(’ staff’, ’ staffer’),
(’icken’, ’oles’),
(’?"’, ’..."’),
(’ Executive’, ’ overseeing’),
(’ academic’, ’ academ’),
(’ Congress’, ’atra’),
(’aroo’, ’anny’),
(’ academic’, ’ academia’),
(’ Congress’, ’ Amendments’),
(’ academic’, ’ academics’),
(’student’, ’ academic’),
(’ committee’, ’ convened’),
(’",’, ’..."’),
(’ove’, ’idia’)
```

Layer 16 Head 13

```
(’ sugg’, ’ hindsight’),
(’ sugg’, ’ anecdotal’),
(’ unsuccessfully’, ’ hindsight’),
(’didn’, ’ hindsight’),
(’orously’, ’staking’),
(’illions’, ’uries’),
(’until’, ’era’),
(’ lobbied’, ’ hindsight’),
(’ incorrectly’, ’ incorrect’),
(’ hesitate’, ’ hindsight’),
(’ECA’, ’ hindsight’),
(’ regret’, ’ regrets’),
(’inventoryQuantity’, ’imore’),
(’consider’, ’ anecdotal’),
(’ errone’, ’ incorrect’),
(’ someday’, ’ eventual’),
(’illions’, ’Murray’),
(’ recently’, ’recent’),
(’ Learned’, ’ hindsight’),
(’before’, ’ hindsight’),
(’ lately’, ’ealous’),
(’upon’, ’rity’),
(’ja’, ’ hindsight’),
(’ regretted’, ’ regrets’),
(’ unsuccessfully’, ’udging’),
(’ lately’, ’dated’),
(’ sugg’, ’ anecd’),
(’ inform’, ’imore’),
(’ lately’, ’recent’),
(’ anecd’, ’ anecdotal’),
(’orously’, ’ hindsight’),
(’ postwar’, ’ Era’),
(’ lately’, ’ recent’),
(’ skept’, ’ cynicism’),
(’ sugg’, ’informed’),
(’ unsuccessfully’, ’ealous’),
(’ebin’, ’ hindsight’),
(’ underest’, ’ overest’),
(’ Jinn’, ’ hindsight’),
(’ someday’, ’2019’),
(’ recently’, ’turned’),
(’ sugg’, ’ retrospect’),
(’ unsuccessfully’, ’didn’),
(’ unsuccessfully’, ’gged’),
(’ mistakenly’, ’ incorrect’),
(’assment’, ’)</’),
(’ja’, ’didn’),
(’illions’, ’ hindsight’),
(’ sugg’, ’ testimony’),
(’jri’, ’ hindsight’)
```

Layer 12 Head 9

```
(’ PST’, ’ usual’),
(’etimes’, ’ foreseeable’),
(’uld’, ’uld’),
(’ Der’, ’ Mankind’),
(’ statewide’, ’ yearly’),
(’ guarantees’, ’ guarantees’),
(’ Flynn’, ’ Logged’),
(’borne’, ’ foreseeable’),
(’ contiguous’, ’ contiguous’),
(’ exceptions’, ’ exceptions’),
(’ redist’, ’ costly’),
(’ downstream’, ’ day’),
(’ ours’, ’ modern’),
(’ foreseeable’, ’ foreseeable’),
(’ Posted’, ’ Posted’),
(’ anecdotal’, ’ anecdotal’),
(’ moot’, ’ costly’),
(’ successor’, ’ successor’),
(’ any’, ’ ANY’),
(’ generational’, ’ modern’),
(’ temporarily’, ’ costly’),
(’ overall’, ’ overall’),
(’ effective’, ’ incentiv’),
(’ future’, ’ tomorrow’),
(’ ANY’, ’ lifetime’),
(’ dispatch’, ’ dispatch’),
(’ legally’, ’ WARRANT’),
(’ guarantees’, ’ incentiv’),
(’ listed’, ’ deductible’),
(’ CST’, ’ foreseeable’),
(’ anywhere’, ’ any’),
(’ guaranteed’, ’ incentiv’),
(’ successors’, ’ successor’),
(’ weekends’, ’ day’),
(’iquid’, ’ expensive’),
(’ Trib’, ’ foreseeable’),
(’ phased’, ’ modern’),
(’ constitutionally’, ’
   foreseeable’),
(’ any’, ’ anybody’),
(’ anywhere’, ’ ANY’),
(’ veto’, ’ precedent’),
(’ veto’, ’ recourse’),
(’ hopefully’, ’ hopefully’),
(’ potentially’, ’ potentially’),
(’ ANY’, ’ ANY’),
(’ substantive’, ’ noteworthy’),
(’morrow’, ’ day’),
(’ancial’, ’ expensive’),
(’listed’, ’ breastfeeding’),
(’ holiday’, ’ holidays’)
```

Layer 11 Head 10

```
(’ Journalism’, ’ acron’),
```

```
(’ democracies’, ’ governments’),
(’/-’, ’verty’),
(’ legislatures’, ’ governments’),
(’ocracy’, ’ hegemony’),
(’osi’, ’ RAND’),
(’ Organizations’, ’ organisations
  ’),
(’ellectual’, ’ institutional’),
(’ Journalists’, ’ acron’),
(’eworks’, ’ sponsors’),
(’ Inqu’, ’ reviewer’),
(’ocracy’, ’ diversity’),
(’ careers’, ’ Contributions’),
(’gency’, ’\\-’),
(’ellectual’, ’ exceptions’),
(’ Profession’, ’ specializing’),
(’online’, ’ Online’),
(’ Publications’, ’ authorised’),
(’Online’, ’ Online’),
(’ sidx’, ’ Lazarus’),
(’eworks’, ’ Networks’),
(’ Groups’, ’ organisations’),
(’ Governments’, ’ governments’),
(’ democracies’, ’ nowadays’),
(’ psychiat’, ’ Mechdragon’),
(’ educ’, ’ Contributions’),
(’ Ratings’, ’ organisations’),
(’vernment’, ’spons’),
(’..."’, ’),"’),
(’ Caucas’, ’ commodity’),
(’ dictators’, ’ governments’),
(’istration’, ’ sponsor’),
(’iquette’, ’ acron’),
(’ Announce’, ’ answ’),
(’ Journalism’, ’ empowering’),
(’Media’, ’ bureaucr’),
(’ Discrimination’, ’
  organizations’),
(’ Journalism’, ’Online’),
(’FAQ’, ’sites’),
(’ antitrust’, ’ Governments’),
(’..."’, ’..."’),
(’Questions’, ’ acron’),
(’rities’, ’ organisations’),
(’ Editorial’, ’ institutional’),
(’ tabl’, ’ acron’),
(’ antitrust’, ’ governments’),
(’ Journalism’, ’ Everyday’),
(’icter’, ’ Lieberman’),
(’ defect’, ’SPONSORED’),
(’ Journalists’, ’ organisations’)
```

Layer 22 Head 5 (names and parts of names seem to attend to each other here)

```
(’ Smith’, ’ovich’),
(’ Jones’, ’ovich’),
(’ Jones’, ’Jones’),
(’ Smith’, ’Williams’),
(’ Rogers’, ’opoulos’),
(’Jones’, ’ovich’),
(’ Jones’, ’inez’),
(’ug’, ’ Ezek’),
(’ Moore’, ’ovich’),
(’orn’, ’roit’),
(’van’, ’actionDate’),
(’ Jones’, ’inelli’),
(’ Edwards’, ’opoulos’),
(’ Jones’, ’ Lyons’),
(’Williams’, ’opoulos’),
(’Moore’, ’ovich’),
(’ Rodriguez’, ’hoff’),
(’ North’, ’ suburbs’),
(’ Smith’, ’chio’),
(’Smith’, ’ovich’),
(’ Smith’, ’opoulos’),
(’Mc’, ’opoulos’),
(’Johnson’, ’utt’),
(’ Jones’, ’opoulos’),
(’Ross’, ’Downloadha’),
(’pet’, ’ilage’),
(’ Everett’, ’ Prairie’),
(’ Cass’, ’isma’),
(’ Jones’, ’zynski’),
(’Jones’, ’Jones’),
(’ McCl’, ’elman’),
(’ Smith’, ’Jones’),
(’ Simmons’, ’opoulos’),
(’ Smith’, ’brown’),
(’ Mc’, ’opoulos’),
(’ Jones’, ’utt’),
(’ Richards’, ’Davis’),
(’ Johnson’, ’utt’),
(’ Ross’, ’bred’),
(’ McG’, ’opoulos’),
(’ Stevens’, ’stadt’),
(’ra’, ’abouts’),
(’ Johnson’, ’hoff’),
(’ North’, ’ Peninsula’),
(’ Smith’, ’Smith’),
(’Jones’, ’inez’),
(’ Hernandez’, ’hoff’),
(’ Lucas’, ’Nor’),
(’ Agu’, ’hoff’),
(’Jones’, ’utt’)
```

Layer 19 Head 12

```
(’ 2015’, ’ADVERTISEMENT’),
(’ 2014’, ’2014’),
(’ 2015’, ’2014’),
(’ 2015’, ’Present’),
(’ 2013’, ’2014’),
(’ 2017’, ’ADVERTISEMENT’),
(’ 2016’, ’ADVERTISEMENT’),
(’itor’, ’ Banner’),
(’2015’, ’ Bulletin’),
(’2012’, ’ Bulletin’),
(’2014’, ’ Bulletin’),
(’ Airl’, ’Stream’),
(’2016’, ’ Bulletin’),
(’ 2016’, ’2014’),
(’2017’, ’ Bulletin’),
(’ 2013’, ’ 2014’),
(’ 2012’, ’2014’),
(’ stadiums’, ’ventions’),
(’ 2015’, ’ Bulletin’),
(’2013’, ’ Bulletin’),
(’ 2017’, ’2014’),
(’ 2011’, ’ 2011’),
```

```
(' 2014', ' 2014'),
(' 2011', ' 2009'),
(' mile', 'eming'),
(' 2013', 'ADVERTISEMENT'),
(' 2014', '2015'),
(' 2014', 'Present'),
(' 2011', '2014'),
(' 2011', '2009'),
(' 2015', ' 2014'),
(' 2013', ' Bulletin'),
(' 2015', '2015'),
(' 2011', ' 2003'),
(' 2011', ' 2010'),
(' 2017', 'Documents'),
('2017', 'iaries'),
(' 2013', '2015'),
('2017', 'Trend'),
(' 2011', '2011'),
(' 2016', 'Present'),
(' 2011', ' 2014'),
(' years', 'years'),
('Plug', 'Stream'),
(' 2014', 'ADVERTISEMENT'),
('2015', 'Present'),
(' 2018', 'thora'),
(' 2017', 'thora'),
(' 2012', ' 2011'),
(' 2012', ' 2014')
```

C.3 Feedforward Keys and Values

Key-value pairs, (k_i, v_i), where at least 15% of the top-k vocabulary items overlap, with $k = 100$. We follow our forerunner's convention of calling the index of the value in the layer "dimension" (Dim).

Layer 0 Dim 116

#annels	#Els
#netflix	#osi
telev	#mpeg
#tv	#vous
#avi	#iane
#flix	transmitter
Television	Sinclair
#outube	Streaming
#channel	#channel
Vid	mosqu
#Channel	broadcaster
documentaries	airs
#videos	Broadcasting
Hulu	broadcasts
channels	streams
#levision	channels
DVDs	broadcasters
broadcasts	broadcasting
#azeera	#RAFT
MPEG	#oded
televised	htt
aired	transmissions
broadcasters	playback
Streaming	Instruction
viewership	nic
#TV	Sirius
Kodi	viewership
ITV	radio
#ovies	#achers
channel	channel

Layer 3 Dim 2711

purposes	purposes
sake	sake
purpose	reasons
reasons	purpose
convenience	ages
reason	reason
Seasons	#ummies
#Plex	#going
Reasons	foreseeable
#ummies	Reasons
#asons	#reason
#lation	#pur
#alsh	Developers
#agos	#akers
#ACY	transl
STATS	Reason
#itas	consideration
ages	#purpose
#purpose	beginners
#=[	awhile
#gencies	Pur
Millennium	#benefit
Brewers	#atel
Festival	#tun
EVENT	pur
#payment	Ages
#=-	preservation
#printf	Metatron
beginners	um
Expo	#KEN

Layer 4 Dim 621

#ovie	headlined
newspapers	pestic
television	dime
editorial	describ
#journal	Afric
broadcasters	broadcasts
#Journal	#('
publication	#umbnails
Newsweek	#adish
Zeit	#uggest
columnist	splash
Editorial	#ZX
newsletter	objectionable
cartoon	#article
#eport	Bucc
telev	#London
radio	reprint
headlined	#azine
#ribune	Giov
BBC	#ender
reprint	headline
sitcom	#oops
reprinted	#articles
broadcast	snipp
tabloid	Ajax
documentaries	marqu

journalist	#("
TV	#otos
headline	mast
news	#idem

Layer 7 Dim 72

sessions	session
dinners	sessions
#cation	#cation
session	#iesta
dinner	Booth
#eteria	screenings
Dinner	booked
#Session	#rogram
rehears	vacation
baths	baths
Lunch	#pleasant
#hops	meetings
visits	#Session
Session	greet
#session	#athon
meetings	Sessions
chatting	boarding
lunch	rituals
chats	booking
festivities	Grape
boarding	#miah
#workshop	#session
#rooms	Pars
#tests	simulated
seated	Dispatch
visit	Extras
appointments	toile
#vu	Evening
#rations	showers
#luaj	abroad

Layer 10 Dim 8

Miy	Tai
#imaru	#jin
Gong	Jin
Jinn	Makoto
Xia	#etsu
Makoto	Shin
Kuro	Hai
Shin	Fuj
#Tai	Dai
Yamato	Miy
Tai	#iku
Ichigo	Yun
#Shin	Ryu
#atsu	Shu
Haku	Hua
Chun	Suzuki
#ku	Yang
Qing	Xia
Tsuk	#Shin
Hua	#iru
Jiang	Yu
Nanto	#yu
manga	Chang
Yosh	Nan
yen	Qian
Osaka	#hao
Qian	Fuk
#uku	Chun
#iku	Yong
Yue	#Tai

Layer 11 Dim 2

progressing	toward
#Progress	towards
#progress	Pace
#osponsors	progression
#oppable	#inness
advancement	onward
progress	canon
Progress	#progress
#senal	pace
#venge	#peed
queue	advancement
#pun	advancing
progression	progressing
#wagon	ladder
advancing	path
#cknowled	honoring
#Goal	ranks
momentum	standings
#zag	goal
#hop	#grand
pursuits	momentum
#encing	#ometer
#Improve	timetable
STEP	nearing
#chini	quest
standings	spiral
#eway	trajectory
#chie	progress
#ibling	accelerating
Esports	escal

Layer 15 Dim 4057

EDITION	copies
versions	Version
copies	#edition
version	#Version
Version	version
edition	#download
editions	download
reprint	versions
#edition	#Download
EDIT	copy
Edition	#release
reproduce	#version
originals	release
#edited	#copy
VERS	VERS
#Versions	#pub
#Publisher	Download
reprodu	#released
#uploads	editions
playthrough	edition
Printed	reprint
reproduction	Release
#Reviewed	#Available
copy	#published

#Version	#Published
paperback	EDITION
preview	print
surv	#Quantity
#Download	#available
circulate	RELEASE

Layer 16 Dim 41

#duino	alarm
#Battery	alarms
Morse	signal
alarms	circuit
GPIO	GPIO
LEDs	timers
batteries	voltage
#toggle	signals
signal	circuitry
circuitry	electrical
#PsyNetMessage	circuits
alarm	LEDs
autop	standby
signalling	signalling
#volt	signaling
volt	lights
signals	Idle
voltage	triggers
LED	batteries
electrom	Morse
timers	LED
malfunction	#LED
amplifier	button
radios	Signal
wiring	timer
#Alert	wiring
signaling	buzz
#Clock	disconnect
arming	Arduino
Arduino	triggered

Layer 17 Dim 23

responsibility	responsibility
Responsibility	respons
responsibilities	responsibilities
#ipolar	Responsibility
#responsible	oversee
duties	#respons
#respons	duties
superv	supervision
supervision	superv
#abwe	stewards
Adin	chore
respons	oversight
oversee	oversees
entrusted	responsible
overseeing	#responsible
helicop	handling
presided	handles
overseen	overseeing
#dyl	chores
responsible	manage
#ADRA	managing
reins	duty
#accompan	Respons
chores	charge
oversees	reins
supervised	handle
blame	oversaw
oversaw	CONTROL
#archment	RESP
RESP	tasks

Layer 19 Dim 29

subconscious	thoughts
thoughts	thought
#brain	Thoughts
#Brain	minds
memories	mind
OCD	thinking
flashbacks	#thought
brainstorm	imagination
Anxiety	Thinking
#mind	Thought
fantas	imagin
amygdala	thinker
impuls	#thinking
Thinking	#mind
#Memory	memories
Thoughts	#think
dreams	imagining
#ocamp	impulses
#Psych	fantasies
#mares	think
mentally	urges
#mental	desires
mind	dreams
#thinking	delusions
#Mind	subconscious
#dream	emotions
psyche	imag
prefrontal	#dream
PTSD	conscience
Memories	visions

Layer 20 Dim 65

exercises	volleyball
#Sport	tennis
#athlon	sports
Exercise	sport
#ournaments	#basketball
volleyball	Tennis
Recre	soccer
Mahjong	golf
#basketball	playground
exercise	Golf
bowling	athletics
skating	#athlon
spar	athletic
skiing	rugby
gymn	amusement
#sports	gymn
drills	sled
#Training	#Sport
tournaments	cricket
sled	Soccer
Volunte	amuse
skate	Activities

golf	recreational
#Pract	Ski
dunk	activities
#hower	basketball
athletics	#games
sport	skating
Solitaire	hockey
#BALL	#sports

Layer 21 Dim 86

IDs	number
identifiers	#number
surname	#Number
surn	Number
identifier	NUM
initials	numbers
#Registered	Numbers
NAME	#Numbers
#names	address
pseudonym	#address
#codes	#Num
nomine	#NUM
names	addresses
username	Address
#IDs	identifier
ID	#Address
registration	#num
#76561	ID
#soDeliveryDate	numbering
#ADRA	IDs
CLSID	#ID
numbering	identifiers
#ername	identification
#address	numer
addresses	digits
codes	#numbered
#Names	numerical
regist	Ident
name	numeric
Names	Identification

Layer 21 Dim 400

#July	Oct
July	Feb
#February	Sept
#January	Dec
#Feb	Jan
November	Nov
#October	Aug
January	#Oct
Feb	May
October	#Nov
#September	Apr
September	March
#June	April
#Sept	#Sept
February	June
#November	#Aug
#April	October
April	#Feb
June	July
#December	December
August	Sep
#March	November
Sept	#Jan
December	#May
Aug	August
March	Jul
#August	Jun
#Aug	September
#wcs	January
Apr	February

Layer 23 Dim 166

#k	#k
#ks	#K
#kish	#ks
#K	#KS
#kat	k
#kus	#kt
#KS	K
#ked	#kr
#kr	#kl
#kB	#kish
#kan	#kos
#kw	#king
#ket	#ked
#king	#kie
#kb	#KB
#kos	#kk
#kHz	#kowski
#kk	#KR
#kick	#KING
#kers	#KT
#kowski	#KK
#KB	#KC
#krit	#kw
#KING	#kb
#kt	#Ka
#ksh	#krit
#kie	#KN
#ky	#kar
#KY	#kh
#ku	#ket

Layer 23 Dim 907

hands	hand
hand	#Hand
#hands	Hand
#hand	#hand
fingers	hands
#feet	Hands
fingertips	fist
claws	#hands
paw	finger
paws	handed
metab	thumb
palms	fingers
fingert	foot
#Hand	#handed
fists	paw
wrists	handing
levers	#finger
thumbs	#hander
tentacles	fingertips
feet	claw

```
limb         fingert
slider       #Foot
#handed      Stick
#dimension   arm
jaws         #Accessory
skelet       #fing
lapt         Foot
ankles       index
weap         toe
foot         #auntlet
```

C.4 KNOWLEDGE LOOKUP

Given a few seed embeddings of vocabulary items we find related FF values by taking a product of the average embeddings with FF values.

Seed vectors:
`["python", "java", "javascript"]`
Layer 14 Dim 1215 (ranked 3rd)

```
filesystem
debugging
Windows
HTTP
configure
Python
debug
config
Linux
Java
configuration
cache
Unix
lib
runtime
kernel
plugins
virtual
FreeBSD
hash
plugin
header
file
server
PHP
GNU
headers
Apache
initialization
Mozilla
```

Seed vectors: `["cm", "kg", "inches"]`
Layer 20 Dim 2917 (ranked 1st)

```
percent
years
hours
minutes
million
seconds
inches
months
miles
weeks
pounds
#%
kilometers
ounces
kilograms
grams
kilometres
metres
centimeters
thousand
days
km
yards
Years
meters
#million
acres
kg
#years
inch
```

Seed vectors: `["horse", "dog", "lion"]`
Layer 21 Dim 3262 (ranked 2nd)

```
animal
animals
Animal
dogs
horse
wildlife
Animals
birds
horses
dog
mammal
bird
mammals
predator
beasts
Wildlife
species
#Animal
#animal
Dogs
fish
rabbits
deer
elephants
wolves
pets
veterinary
canine
beast
predators
reptiles
rodent
primates
hunting
livestock
creature
rabbit
rept
elephant
creatures
human
```

```
hunters
hunter
shark
Rept
cattle
wolf
Humane
tiger
lizard
```

D Sentiment Analysis Fine-Tuning Vector Examples

This section contains abusive language

Classification Head Parameters

Below we show the finetuning vector of the classifier weight. "POSITIVE" designates the vector corresponding to the label "POSITIVE", and similarly for "NEGATIVE".

```
POSITIVE      NEGATIVE
-----------   ------------
#yssey        bullshit
#knit         lame
#etts         crap
passions      incompetent
#etooth       inco
#iscover      bland
pioneers      incompetence
#emaker       idiots
Pione         crappy
#raft         shitty
#uala         idiot
prosper       pointless
#izons        retarded
#encers       worse
#joy          garbage
cherish       CGI
loves         FUCK
#accompan     Nope
strengthens   useless
#nect         shit
comr          mediocre
honoured      poorly
insepar       stupid
embraces      inept
battled       lousy
#Together     fuck
intrig        sloppy
#jong         Worse
friendships   Worst
#anta         meaningless
```

In the following sub-sections, we sample 4 difference vectors per each parameter group (FF keys, FF values; attention query, key, value, and output subheads), and each one of the fine-tuned layers (layers 9-11). We present the ones that seemed to contain relevant patterns upon manual inspection. We also report the number of "good" vectors among the four sampled vectors for each layer and parameter group.

FF Keys

Layer 9

4 out of 4

diff	-diff
amazing	seiz
movies	coerc
wonderful	Citiz
love	#cffff
movie	#GBT
cinematic	targ
enjoyable	looph
wonderfully	Procedures
beautifully	#iannopoulos
enjoy	#Leaks
films	#ilon
comedy	grievance
fantastic	#merce
awesome	Payments
#Enjoy	#RNA
cinem	Registrar
film	Regulatory
loving	immobil
enjoyment	#bestos
masterpiece	#SpaceEngineers

diff	-diff
reperto	wrong
congratulations	unreasonable
Citation	horribly
thanks	inept
Recording	worst
rejo	egregious
Profile	#wrong
Tradition	unfair
canopy	worse
#ilion	atro
extracts	stupid
descendant	egreg
#cele	bad
enthusiasts	terribly
:-)	ineffective
#photo	nonsensical
awaits	awful
believer	#worst
#IDA	incompetence
welcomes	#icably

diff	-diff
movie	seiz
fucking	Strongh
really	#etooth
movies	#20439
damn	#Secure
funny	Regulation
shit	Quarterly
kinda	concess
REALLY	Recep
Movie	#aligned
stupid	targ
#movie	mosqu
goddamn	#verning
crap	FreeBSD
shitty	PsyNet
film	Facilities
crappy	#Lago
damned	#Register
#Movie	#"}],"
cheesy	Regist

diff	-diff
incompetence	#knit
bullshit	#Together
crap	Together
useless	versatile
pointless	#Discover
incompetent	richness
idiots	#iscover
incompet	forefront
garbage	inspiring
meaningless	pioneering
stupid	#accompan
crappy	unparalleled
shitty	#Explore
nonexistent	powerfully
worthless	#"},{"
Worse	#love
lame	admired
worse	#uala
inco	innovative
ineffective	enjoyed

Layer 10

4 out of 4

diff	-diff
quotas	wonderfully
#RNA	wonderful
cessation	beautifully
subsidy	amazing
#SpaceEngineers	fantastic
placebo	incredible
exemptions	amazingly
treadmill	great
Labs	unforgettable
receipt	beautiful
moratorium	brilliantly
designation	hilarious
ineligible	love
reimbursement	marvelous
roundup	vividly
Articles	terrific
PubMed	memorable
waivers	#Enjoy
Citiz	loving
landfill	fascinating

diff	-diff
isEnabled	wonderfully
guiActiveUnfocu...	beautifully
#igate	cinem
waivers	cinematic
expires	wonderful
expire	amazing
reimb	Absolutely
expired	storytelling
#rollment	fantastic
#Desktop	Definitely
prepaid	unforgettable
#verning	comedy
#andum	movie
reimbursement	comedic
Advisory	hilarious
permitted	#movie
#pta	#Amazing
issuance	scenes
Priebus	Amazing
#iannopoulos	enjoyable

diff	-diff
horror	#deals
whim	#iband
subconscious	[&
unrealistic	#heid
imagination	#APD
viewers	withdrew
enjoyment	#Shares
nostalgia	mathemat
absolute	[+]
sentimental	#Tracker
unreal	#zb
Kubrick	testified
awe	#ymes
inspiration	mosqu
subtle	#Commerce
cinematic	administr
perfection	feder
comedic	repaired
fantasy	#pac
mindless	#Community

diff	-diff
#Leaks	loving
quotas	love
#RNA	loved
subsidy	lovers
#?’"	wonderful
Penalty	lover
#iannopoulos	nostalgic
#>]	alot
discredited	beautiful
#conduct	amazing
#pta	great
waivers	passionate
Authorization	admire
#admin	passion
HHS	lovely
arbitrarily	loves
#arantine	unforgettable
#ERC	proud
memorandum	inspiration
#Federal	#love

Layer 11

4 out of 4

diff	-diff
inco	cherish
pointless	#knit
Nope	#terday
bullshit	#accompan
crap	prosper
useless	versatile
nonsense	friendships
futile	#uala
anyways	Lithuan
anyway	cherished
meaningless	redes
clueless	inspires
lame	Proud
wasting	friendship
bogus	exceptional
vomit	#beaut
nonsensical	#ngth
retarded	pioneering
idiots	pioneers
shit	nurt

diff	-diff
#SpaceEngineers	love
nuisance	definitely
#erous	always
#aband	wonderful
Brist	loved
racket	wonderfully
Penalty	cherish
bystand	loves
#iannopoulos	truly
Citiz	enjoy
Codec	really
courier	#olkien
#>]	beautifully
#termination	#love
incapac	great
#interstitial	LOVE
fugitive	never
breaching	adore
targ	loving
thug	amazing

diff	-diff
#accompan	bad
Pione	crap
celebrate	inefficient
#Discover	stupid
#knit	worse
pioneering	mistake
recogn	incompetence
reunited	mistakes
comr	incompetent
thriving	miser
#iscover	garbage
commemorate	retarded
Remem	#bad
ecstatic	poor
forefront	ineffective
enthusi	retard
renewed	Poor
colle	bullshit
Inspired	inept
#uala	errors

diff	-diff
#knit	bullshit
passions	crap
#accompan	idiots
#ossom	goddamn
#Explore	stupid
welcomes	shitty
pioneering	shit
forefront	garbage
embraces	fuck
pioneers	incompetence
intertw	crappy
#izons	bogus
#iscover	useless
unparalleled	idiot
evolving	#shit
Together	pointless
vibrant	stupidity
prosper	fucking
strengthens	nonsense
#Together	FUCK

FF Values

Layer 9

0 out of 4

Layer 10

0 out of 4

Layer 11

0 out of 4

W_Q Subheads

Layer 9

3 out of 4

```
diff          -diff
------------  ------------
#ARGET        kinda
#idal         alot
#--+          amazing
Prev          interesting
#enger        wonderful
#iannopoulos  definitely
#report       unbelievable
#RELATED      really
issuance      amazingly
#earcher      pretty
Previous      nice
Legislation   absolutely
#astical      VERY
#iper         wonderfully
#>[           incredible
#</           hilarious
Vendor        funny
#">           fantastic
#phrine       quite
#wcsstore     defin
```

```
diff          -diff
------------  -----------
bullshit      strengthens
bogus         Also
faux          #helps
spurious      adjusts
nonsense      #ignt
nonsensical   evolves
inept         helps
crap          grew
junk          grows
shitty        #cliffe
fake          recognizes
incompetence  #assadors
crappy        regulates
phony         flourished
sloppy        improves
dummy         welcomes
mediocre      embraces
lame          gathers
outrage       greets
inco          prepares
```

```
diff        -diff
----------  ------------
alot        Provision
kinda       coerc
amazing     Marketable
definitely  contingency
pretty      #Dispatch
tho         seiz
hilarious   #verning
VERY        #iannopoulos
really      #Reporting
lol         #unicip
wonderful   Fiscal
thats       issuance
dont        provision
pics        #Mobil
doesnt      #etooth
underrated  policymakers
funny       credential
REALLY      Penalty
#love       #activation
alright     #Officials
```

Layer 10

4 out of 4

diff	-diff
crap	#Register
shit	Browse
bullshit	#etooth
stupid	#ounces
shitty	#verning
horrible	#raft
awful	#egu
fucking	#Lago
comedic	Payments
crappy	#orsi
cheesy	Coinbase
comedy	#ourse
fuck	#iann
mediocre	#"}],"
terrible	#onductor
movie	#obil
bad	#rollment
gimmick	#ivot
filler	#Secure
inept	#ETF

diff	-diff
love	Worse
unforgettable	Nope
beautiful	#Instead
loved	Instead
#love	#Unless
loving	incompetence
amazing	incapable
#joy	Unless
inspiring	#failed
passion	incompet
adventure	incompetent
loves	ineffective
excitement	#Fuck
joy	#Wr
LOVE	inept
together	spurious
memories	#Failure
wonderful	worthless
enjoyment	obfusc
themes	inadequate

diff	-diff
#knit	crap
#"},{"	bullshit
#"}],"	stupid
#estones	inept
#Learn	shit
#ounces	idiots
#egu	shitty
#Growing	crappy
#ributes	incompetence
#externalAction...	fuck
#encers	pointless
Browse	nonsense
jointly	nonsensical
Growing	stupidity
#ossom	gimmick
honoured	inco
#accompan	lame
#agos	incompetent
#raft	mediocre
#iership	bland

diff	-diff
crap	#egu
bullshit	#etooth
shit	#verning
:(	#ounces
lol	#accompan
stupid	coh
filler	#assadors
shitty	#pherd
fucking	#acio
pointless	#uchs
idiots	strengthens
anyways	#reprene
nonsense	Scotia
anyway	#rocal
crappy	reciprocal
stupidity	Newly
fuck	fost
#shit	#ospons
anymore	#onductor
Nope	governs

Layer 11

3 out of 4

```
diff           -diff
-------------  ------------
#utterstock    amazing
#ARGET         movie
#cffff         alot
#etooth        scenes
#Federal       comedy
POLITICO       movies
#Register      cinematic
#Registration  greatness
#rollment      wonderful
#ETF           storytelling
#ulia          film
Payments       tho
#IRC           masterpiece
Regulatory     films
Alternatively  Kubrick
#RN            realism
#pta           comedic
Regulation     cinem
#GBT           #movie
#":""},{"      genre
```

```
diff           -diff
-------------  ------------------
#also          meaningless
#knit          incompetence
helps          inco
strengthens    pointless
:)             incompetent
broaden        Worse
#ossom         inept
incorporates   nonsensical
#Learn         coward
incorporate    unint
#"},{"         obfusc
enjoy          excuses
enjoyed        panicked
complementary  useless
#etts          bullshit
enhances       stupid
integrates     incompet
#ospons        incomprehensibl...
differs        stupidity
#arger         lifeless
```

```
diff           -diff
-------------  ---------------
amazing        #iannopoulos
beautifully    expired
love           ABE
wonderful      Yiannopoulos
wonderfully    liability
unforgettable  #SpaceEngineers
beautiful      #isance
loving         Politico
#love          waivers
#beaut         #utterstock
enjoyable      excise
#Beaut         #Stack
inspiring      phantom
fantastic      PubMed
defin          #ilk
incredible     impunity
memorable      ineligible
greatness      Coulter
amazingly      issuance
timeless       IDs
```

W_K Subheads

Layer 9

3 out of 4

```
diff     -diff
-------  ----------
enclave  horrible
#.       pretty
#;       alot
#omial   MUCH
apiece   VERY
#assian  nothing
#.</     #much
#ulent   terrible
#,[      crappy
#eria    strange
#ourse   everything
exerc    very
#\/      shitty
#Wire    nice
#arium   many
#icle    wonderful
#.[      genuinely
#/$      beautiful
#API     much
#ium     really
```

```
diff           -diff
-------------  -----------
Then           any
Instead        #ady
Unfortunately  #imate
Why            #cussion
Sometimes      #ze
Secondly       appreci
#Then          #raq
But            currently
Luckily        #kers
Anyway         #apixel
And            active
Suddenly       significant
Thankfully     #ade
Eventually     #imal
Somehow        specific
Fortunately    #ability
Meanwhile      anyone
What           #ker
Obviously      #unction
Because        reap
```

```
diff         -diff
-----------  ---------
bullshit     #avorite
anyway       #ilyn
crap         #xtap
anyways      #insula
unless       #cedented
nonsense     #aternal
#falls       #lyak
fuck         #rieve
#.           #uana
fallacy      #accompan
#tics        #ashtra
#punk        #icer
damned       #andum
#fuck        Mehran
stupidity    #andise
shit         #racuse
commercials  #assadors
because      #Chel
despite      rall
movies       #abella
```

Layer 10

2 out of 4

diff	-diff
#,	Nope
work	Instead
#icle	Thankfully
#.	Surely
outdoors	#Instead
inspiring	Fortunately
exped	Worse
ahead	Luckily
together	#Thankfully
touches	Unless
out	Apparently
personalized	Perhaps
#joy	#Unless
#unction	#Fortunately
warm	Sorry
exceptional	Secondly
experience	#Luckily
lasting	#Rather
integ	Hence
#astic	Neither

diff	-diff
#sup	#etting
Amazing	#liness
#airs	#ktop
awesome	#ulkan
Bless	#enthal
Loving	#enance
my	#yre
#OTHER	#eeds
#BW	omission
#perfect	#reys
#-)	#lihood
amazing	#esian
#adult	#holes
perfect	syndrome
welcome	grievance
Rated	offenders
#Amazing	#wig
#anch	#hole
FANT	#creen
#anche	#pmwiki

Layer 11

2 out of 4

diff	-diff
shots	#Kind
shit	suscept
bullshit	Fathers
stuff	#Footnote
tits	concess
crap	#accompan
boobs	Strait
creepy	#orig
noises	#ESE
spectacle	#ufact
boring	Founder
things	#iere
everything	#HC
noise	#Prev
#anim	#alias
ugly	participated
garbage	#Have
stupidity	#coe
visuals	#Father
selfies	strugg

diff	-diff
#ly	#say
storytelling	actionGroup
sounding	prefers
spectacle	#ittees
#ness	#reon
#hearted	presumably
cinematic	waivers
#est	#aucuses
portrayal	#Phase
quality	#racuse
paced	#arge
combination	#hers
juxtap	#sup
representation	#later
mixture	expired
#!!!!!	stricter
filmmaking	#onds
enough	#RELATED
thing	#rollment
rendition	#orders

W_V SUBHEADS

Layer 9

4 out of 4

```
diff          -diff
-----------   ----------
#":""},{"     honestly
#etooth       definitely
#ogenesis     hilarious
#verning      alot
broker        amazing
#ounces       funn
threatens     cinem
#astical      Cinem
foothold      comedic
intruder      Absolutely
#vernment     comedy
#activation   absolutely
#Oracle       amazingly
fugitive      satire
visitor       underrated
#assian       really
barrier       fantastic
#":[          enjoyable
#vier         REALLY
#oak          wonderful
```

```
diff       -diff
--------   -------------
crap       jointly
shit       #verning
bullshit   #pora
fucking    #rocal
idiots     #raft
fuck       #etooth
goddamn    #estead
stupid     #ilitation
FUCK       #ourse
#fuck      migr
shitty     #ourses
damn       #iership
#shit      Pione
lol        #iscover
fuckin     pioneering
nonsense   #egu
crappy     #ivities
kinda      neighbourhood
Fuck       pioneer
idiot      nurt
```

```
diff           -diff
------------   --------------
crap           Pione
bullshit       pioneers
shit           complementary
vomit          pioneering
nonsense       #knit
stupid         #raits
idiots         Browse
fucking        #iscover
#shit          strengthened
idiot          #rocal
fuck           prosper
gimmick        Communities
stupidity      neighbourhoods
goddamn        #Learn
shitty         strengthens
incompetence   #iscovery
lame           #ributes
FUCK           strengthen
inco           #izons
blah           Mutual
```

```
diff        -diff
---------   --------------
anime       #rade
kinda       #jamin
stuff       #ounces
shit        #pherd
lol         Unable
tho         #pta
realism     Roche
damn        Payments
:)          Gupta
fucking     #odan
alot        #uez
movie       #adr
funny       #ideon
anyways     #Secure
enjoyable   #raught
crap        Bei
comedy      sovere
genre       unsuccessfully
anyway      #moil
fun         #Register
```

Layer 10

4 out of 4

diff	-diff
#knit	crap
welcomes	bullshit
Together	idiots
Growing	stupid
#Explore	shitty
pioneering	incompetence
complementary	pointless
milestone	goddamn
pioneer	retarded
#Together	lame
strengthens	Worse
#ossom	crappy
pioneers	incompet
#Learn	shit
jointly	stupidity
#Growing	fucking
embraces	Nope
#"},{"	FUCK
sharing	incompetent
#Discover	pathetic

diff	-diff
#"}],"	crap
#verning	stupid
#etooth	shit
#"},{"	fucking
Browse	fuck
#Register	shitty
#Lago	bullshit
#raft	crappy
#egu	idiots
jointly	horrible
#iership	stupidity
strengthens	kinda
Scotia	goddamn
#ounces	awful
#uania	mediocre
#iann	pathetic
workspace	#fuck
seiz	damn
Payments	FUCK
#Learn	damned

diff	-diff
bullshit	inspiring
incompetence	unforgettable
Worse	#knit
idiots	#love
crap	passions
dummy	cherish
incompetent	richness
Nope	timeless
stupid	loves
retarded	passionate
lame	beautifully
nonexistent	overcoming
wasting	unique
#Fuck	highs
bogus	nurture
worse	unparalleled
nonsense	vibrant
ineligible	#beaut
pointless	intertw
inco	insepar

diff	-diff
bullshit	Pione
crap	pioneers
stupid	pioneering
nonsense	complementary
incompetence	#knit
idiots	#Learn
shit	#accompan
stupidity	pioneer
pointless	invaluable
inco	#ossom
retarded	#Together
idiot	Browse
vomit	versatile
lame	welcomes
meaningless	#"},{"
goddamn	admired
nonsensical	jointly
garbage	Sharing
#shit	Together
useless	#Discover

Layer 11

4 out of 4

```
diff          -diff
------------  ------------
Provision     alot
issuance      amazing
Securities    kinda
#ogenesis     fucking
Holdings      awesome
Regulatory    funny
indefinitely  damn
Advisory      REALLY
designation   hilarious
unilaterally  tho
Province      unbelievable
Regulation    fuckin
#Lago         wonderful
issued        doesnt
Recep         definitely
Advis         thats
#verning      yeah
broker        fantastic
#Mobil        badass
Policy        dont
```

```
diff          -diff
------------  ---------
crap          #rocal
fucking       #verning
bullshit      #etooth
fuck          #uania
goddamn       caches
shit          Browse
#fuck         #"},{"
stupidity     #imentary
pathetic      exerc
spoiler       #Lago
stupid        #"}],"
inept         #cium
blah          #enges
FUCK          #ysis
awful         quarterly
shitty        #iscover
trope         Scotia
Godd          #resso
inco          #appings
incompetence  jointly
```

```
diff            -diff
--------------  ------------
pioneers        bullshit
pioneering      crap
Browse          shit
Pione           idiots
complementary   stupid
#knit           vomit
prosper         incompetence
#raits          nonsense
#Trend          gimmick
#ributes        stupidity
#Learn          idiot
strengthen      shitty
strengthened    fucking
#ossom          lame
pioneer         crappy
#iscover        goddamn
#Growing        pointless
prosperity      inco
neighbourhoods  #shit
#owship         Nope
```

```
diff          -diff
------------  -------------
Worse         #knit
bullshit      pioneers
Nope          pioneering
crap          inspiring
incompetence  #iscover
idiots        complementary
incompetent   pioneer
stupid        #ossom
incompet      passionate
pointless     passions
inco          journeys
Stupid        unique
meaningless   embraces
nonsense      admired
lame          forefront
idiot         richness
worse         invaluable
#Fuck         prosper
whining       vibrant
nonsensical   enriched
```

W_O Subheads

Layer 9

0 out of 4

Layer 10

0 out of 4

Layer 11

0 out of 4