ECAI 2024
U. Endriss et al. (Eds.)

© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms
of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA240594

On Explaining with Attention Matrices

Omar Naim[a,*] **and Nicholas Asher**[b,**]

[a]Université Toulouse 3, IRIT France
[b]CNRS, IRIT France

Abstract. This paper explores the much discussed, possible explanatory link between attention weights (AW) in transformer models and predicted output. Contrary to intuition and early research on attention, more recent prior research has provided formal arguments and empirical evidence that AW are not explanatorily relevant. We show that the formal arguments are incorrect. We introduce and effectively compute *efficient attention*, which isolates the effective components of attention matrices in tasks and models in which AW play an explanatory role. We show that efficient attention has a causal role (provides minimally necessary and sufficient conditions) for predicting model output in NLP tasks requiring contextual information, and we show, contrary to [7], that efficient attention matrices are probability distributions and are effectively calculable. Thus, they should play an important part in the explanation of attention based model behavior. We offer empirical experiments in support of our method illustrating various properties of efficient attention with various metrics on four datasets.

1 Introduction

Transformer and other attention based models optimize attention weights (AW), the distribution of weights in transformer models' implicit representations of the hidden states of tokens, as inputs to multilayer perceptrons (MLPs) in order to solve a variety of tasks. As such it stands to reason that AW should have an explanatory role [8, 20, 43, 10]. However, many have criticized using AW to explain model predictions [38, 24, 45, 17, 7, 5]. [7] formally proved that under certain assumptions an infinite number of AW could yield the same prediction; i.e., AW are *unidentifiable* from output–the relation from AW to predictions is not 1-1. Empirical testing by [24, 45, 42, 11] has confirmed non-identifiability of AW in at least some tasks and models, which [2, 9] among others cite as a drawback for explanability.

This paper makes three main contributions. First, we bring together in a novel way, AW, identifiability and explainability in transformer models, linking identifiability of AW with counterfactual and causal explanations and explanatory faithfulness [37, 23, 46].

Second, we show that [7]'s approach fails to make AW viable explanatory tools. While they identify a projection into a representation space that looks promising they are unable to show that the weights in that projection have the properties of a probability distribution. Moreover, they do not see a way to calculate the result of the projection effectively into a space with the requisite properties. This vitiates twice over the explanatory power of their solution. We introduce a new projection whose result we call , *efficient attention*. Efficient attention restores identifiability; and we show that its weights define a probability distribution. Finally, we show how to calculate this projection effectively.

Third, we present a series of experiments[1] on various datasets using the experimental set up of [45, 24]. The point of these experiments is to show empirically the effects of efficient AW. We show first that AW matrix makes the same predictions as its efficient AW projection. Second, we take AW and their adversarial AW from [45] with the same predictions, and we show empirically that they have the same efficient AW projection within the limits of integer precision. Finally, we show that intervening on model and replacing one efficient AW with another shifts the predictions, confirming empirically the identifiability of efficient AW and their potential role in counterfactual, faithful explanations, contrary to [24, 45, 9].

Section 2 starts with an analysis of explainability and identifiability. Section 3 provides a detailed look at the theory behind the claims of [7] about non-identifiability of AW and their solution. We introduce *efficient attention*, prove its identifiability and that it is a probability distribution. We also show how to calculate it. Section 5 empirically validates the identifiability of efficient attention contrary to the claims of [45] using their experimental set up and datasets.

2 Explainability and Interpretability for Transformers

Attention and transformers Generative models or transformers with decoders predict the next token y_{i+1} using the conditional probability $P(y_{i+1}|y_0, \ldots, y_i, C_{i+1})$ where y_0 is the seed vector, $\{y_1, \ldots, y_i\}$ are prior predicted tokens and C_{i+1} is a latent representation of y_i together with other information given to the model. Encoder transformer and other attention based models also use these latent representations of input tokens. Attention matrices or their AW determine these representations C_i from entries that are typically already vector encodings. More specifically, C_i is a weighted sum of C_i^h over h attention heads, where each C_i^h is:

$$C_i^h = \Sigma_{k=1}^{K} \alpha_{i,k}^h (W_V^h e_k) \tag{1}$$

$\alpha_{i,k}^h$ is the attention weight providing the import of token representation e_k to e_i and W_V^h is the value weight matrix of the head h.

A typical transformer has multiple AW that link with a residual stream from the previous layer to a multi-layer perceptron (MLP). Since early work on AW ([24, 45] did not even use full transformer models), we have learned a lot more about how transformer models

* Corresponding Author. Email: omar.naim.docs@gmail.com
** Corresponding Author. Email: nicholas.asher@irit.fr

[1] https://github.com/omyokun/On-explaining-with-attention-matrices/

exploit context input tokens and AW. The Transformer architecture has layers with each having different blocks in the following order: attention, residual learning [21], layer normalization [4] and MLP (Multi Layer Perceptron). [26, 19, 15, 16] show how to decompose attention blocks in a way that allows us to interpret token/token interactions in each block. AW rearranges the input embeddings in a transformer model to optimize the internal representation input to the MLP that must execute some task. The role of residual stream is to incorporate elements from previous layers to avoid overfitting at each layer; but this feeds into the next block's attention layer. The role of the layer normalization block is to significantly reduce training time. Finally the MLP block learn the function we want from the transformer.

All of this work points to the important role of attention in transformer model predictions. If attention did not play an explanatory role, by giving a better contextualization to facilitate the learning for the MLP, the Transformer performance would reduce to that of the MLP component, but this is manifestly not the case. [35]shows the importance of the attention layer in their theoretical analysis of the computational power of transformer architectures. Thus, the optimized representation of input data by the attention layer should have an important causal effect on the MLP's computations and the eventual output, in any task where the first input to the transformer model is not sufficient for an MLP to solve the task. We know from formal results about classifiers that Heavyside MLPs [13] benefit greatly from a representation space that minimizes the number of "jumps" the classification function must make in order to classify some data.[2] Moreover, the effects of AW at different layers of a transformer model may be different, as in circuit analysis [36], mechanistic interpretability [33, 32, 18].

We grant that not all tasks may need the contextual information of equation 1, at least not in all layers or in the same way. For instance, simple classification tasks like gender identification do not seem to require the contextual encodings of attention heads, at least not through all layers [25]. Nevertheless, it stands to reason that AW should play an important explanatory role. We will show, contrary to [24, 45, 7], that AWs are causally determinative of the output and so should in principle have explanatory value.

Explainability With this in mind, we turn to explainability with AW. Discussions of explainability for transformers often start from different interpretations of explainability and interpretability. [23] use them equivalently, but [37] distinguishes them using the concepts of faithfulness and plausibility. Plausibility has to do with how acceptable the purported explanation is to humans [23]. Faithfulness involves a causal connection, what is causally necessary and sufficient for the model to produce its prediction given a certain input; a faithful explanation is one that captures the causal relations in the explanations it generates of a model's predictions. That means a faithful explanation should ideally identify minimally sufficient and necessary conditions for the prediction.

A faithful explanation of the behavior $\hat{f}$ thus has to support counterfactual interventions of the form: A *and* B and *had* A *not been the case,* B *would not have been the case either* [23]. Using [29]'s definition of causation and given $\hat{f} : X \rightarrow Y$, we say that A faithfully explains why $f(x) = \pi$ for $x \in X$ iff we can establish (i) x has A ($A(x)$), and (ii) if $A(x)$ hadn't been the case, $\hat{f}(x) \neq \pi$.

To establish such counterfactuals we typically need a background theory T that could be formal (as in [40]), but it could also be based on observation. T requires a range of *cases* Φ, different set ups or variations on inputs that support counterfactual interventions, in which we remove our putative explanatory factors ϕ from x but keep the rest of our set up $\hat{f}$ and the other properties of x the same or as similar as possible. Counterfactuals require a distance metric over Φ [28] to make precise the notion of similarity. The appropriate metric depends on the nature of the cases and the task. There are many candidates for counterfactual theories of a transformer $\hat{f}$ [3]. One can set Φ to the feature space of the inputs to the model. Alternatively we can set Φ to a set of possible AW in $\hat{f}$ or $W =$ a set of possible parameter settings for the final layer of $\hat{f}$ as suggested in [45, 14], or any among a large number of other possibilities of parameter values for intermediate states in $\hat{f}$. Not all of these candidates may provide faithful explanations, however, a choice of cases may not turn out to support any nontrivial counterfactuals.

To have faithful explanations and non trivial counterfactuals, we need identifiability of the explanatory factors from the predictions. The non-identifiability of AW precludes counterfactual forms of explanation of the form *had the AW been different, the prediction would have differed*, and thus threatens the faithfulness of any explanation based on AW. However, there are a couple of caveats. In general, in light of finite integer precision and the approximation of exact real values and computations in transformers, we should not expect small variations in AW numerical values to practically affect prediction. We are interested in the possibility of faithful explanations using sets $\mathcal{A}$ of AW where for $x, y \in \mathcal{A}$, $\|x - y\| < \lambda$ for λ determined by the task, approximation and integer precision, the predictions are more or less the same and where for distinct $\mathcal{A}, \mathcal{A}'$ and for $x \in \mathcal{A}, y \in \mathcal{A}'$ with $\|x - y\| > \lambda$, we should get distinct predictions. [24, 45, 42] provide empirical evidence that at least in some NLP tasks, what appear to be numerically significant shifts in AW do not affect the predictions over a large number of instances in an NLP task.

If the variations in AW values are large and are not explicable in terms of approximations, then the explanatory plausibility carried by such a large set of AW is much less. Such sets, which are a mark of causal overdetermination [30], leave humans wanting more: what is the common element that makes all of these cases causally relevant? In some NLP tasks, for instance, text classification, this overdetermination might be innocuous or even indicative that the task does not distinguish between a broader class of AW numerical values; a single text input might provide many sufficient clues for a classification; a single movie review might be classified as favorable because of several passages, a high attention value on any one of which would suffice for a favorable classification. But in general, given a particular input, identifiable AW distributions (up to a certain arithmetic precision) for a particular prediction would carry a higher explanatory value in terms of plausibility than sets of heterogeneous AW values. There is a point then to searching for identifiability of AW within the limits of integer precision. Identifiable attention matrices can also furnish the basis for more sophisticated strategies for plausible explanations with attention as in [22].

There is another dimension of plausibility to consider. A plausible explanation using AW needs to link AW to input tokens. Recalling what attention layers do, they provide "mixtures" of input token effects on each token. These mixtures are difficult if not impossible to interpret when the values are negative, and they are most easily interpretable when they are probability distributions as motivated in a different context by [27]. That is, attention weights as probability distributions provide *a priori* for humans more plauible explanations.

[2] For a discussion, see https://nitishpuri.github.io/posts/books/a-visual-proof-that-neural-networks-can-compute-any-function/.

3 Mathematics for Attention and Identifiability

[7] showed a sufficient condition for non identifiability: for a given input sequence longer than the attention head dimension, a transformer model can generate an infinite set of attention weight distributions producing the same output prediction. This is not just an abstract possibility; in cases with relatively long inputs, non identifiability is real. [7] also proposed a means for removing the weight components that do not influence the model's prediction to make AW identifiable. But their approach doesn't guarantee that AW are probability distributions.

To analyze and unpack this problem, we first review and give a detailed proof of [7]'s necessary and sufficient condition for identifiability of AW. We then describe their solution of effective attention, which provides a projection of AW values that renders the function from AW values to predictions injective but at the price of making those values no longer a probability distribution. We then define a new and computable projection of AW into what we call *efficient AW* that solves the mathematical problem with the effective attention [7] while also restoring identifiability of attention weight distributions.

3.1 Identifiability on AW

As background to the work of [7], multi-headed attention output combine h independent attention heads with reduced head dimension $d_v = d/h$ in parallel fashion through a linear layer. The output is achieved by summing outputs over each single head and multiplying them by the matrix $H \in \mathbb{R}^{d_v * d}$ of the linear projection. We will keep the same notations as [7], then for ease of understanding, we omit layer and head indices in the upcoming proofs since they remain valid for each head and layer in transformer models.

$$Attention(Q, K, V)H = (A.V).H = \tag{2}$$

$$A.(E.W^V).H = A.T$$

where the attention matrix A is defined as:

$$A = softmax(\frac{QK'}{\sqrt{d_q}}) \in \mathbb{R}^{d_s * d_s}$$

and $Q \in \mathbb{R}^{d_s * d_q}$ is the query matrix, $K \in \mathbb{R}^{d_s * d_q}$ is the key matrix, $V = E.W^V \in \mathbb{R}^{d_s * d_v}$ is the value matrix (the bias is assumed to be zero), $W^V \in \mathbb{R}^{d * d_v}$ is the matrix that projects embeddings E to the value matrix V, $H \in \mathbb{R}^{d_v * d}$ is the matrix corresponding to the linear layer that reshapes the concatenation of contextualization matrices into the same dimension as our inputs vectors and $T = V.H$ is a matrix considered to simplify notations, with d_s corresponding to the input sequence length, d to embedding dimension and d_v to the attention head dimension.

Concerning the identifiability of AW, consider the function $f : A \mapsto AT$. We note that f is linear with respect to A, then f is injective $\Leftrightarrow Ker(f) = \{0\}$. Let " $'$ " stand for the transpose of a matrix. We now study the injectivity of f : for $A \in Ker(f), f(A) = 0 \Leftrightarrow AT = 0 \Leftrightarrow T'A' = 0$. The following proposition establishes necessary and sufficient conditions for injectivity:

Proposition 1. *f is injective iff $Ker(T') = \{0_{d_s}\}$*

For a full proof of Proposition 1, see Appendix A.

3.2 Effective Attention restores identifiability

[7] shows that a sufficient condition for f to fail to be 1-1 and for AW to be nonidentifiable is that: $d_s > d_v$. To restore identifiability, [7], proposed a solution, effective attention, which consists of the decomposition of each row of AW A into the component in Ker(T') and the component orthogonal to the null space. As all possible AW and effective AW are in finite dimensions $\mathbb{R}^n$, then we can write: $\mathbb{R}^{d_s} = Ker(T)'^{\perp} + Ker(T')$. Then $AT = (proj_{Ker(T')}(A) + proj_{Ker(T')^{\perp}}(A))T = proj_{Ker(T')^{\perp}}(A)T = (A - proj_{Ker(T')}(A))T$.
There are two problems with the solution proposed by [7]. First, elements of $Ker(T')$ might include negative weights. In addition even if these constraints hold, there is no guarantee that for $A \in Ker(T')$: $A \geq 0$ and $A1 = 1$. [7]'s solution does not respect probability constraints; and as a result, their effective attention will not lead to easily interpretable elements. We will tackle these two problems separately.

4 A new solution for Identifiability:Efficient Attention

In the transformer models typically considered in the literature, A is the result of a softmax, hence its rows are constrained by: $A \geq 0$, and $A1 = 1$, where $1 \in R^{ds}$ is the unit vector of all ones. The condition for nonidentifiability shows that there are infinite number of distributions that verify the same output, but it does not establish that there are several matrices that verify the softmax constraints. So we need to project attention weights into a different space. Proposition 2 gives details on that space.

Proposition 2. *$d_s > d_v + 1$ suffices for A to be non identifiable and for the null space of $[T, 1]$ to be non-zero.*[3]

A full proof of Proposition 2 is in Appendix A.

Using Proposition 2, we use instead of [7]'s $Ker(T')^{\perp}$, a projection into the null space $Ker([T, 1]')^{\perp}$, which induces the same prediction as original attention matrix ($AT = A^{\perp}T$) and an additional constraint ($A^{\perp}1 = 1$) which is important for $A^{\perp}$ to be a probability distribution. To solve the problem with [7]'s approach for explanations based on attention, we need to prove that the projection into $Ker([T, 1]')^{\perp}$ has the global properties of a probability space and that there is an effective means of calculating $Ker([T, 1]')^{\perp}$.

4.1 Global properties of $Ker([T, 1]')^{\perp}$

Our work relies on the decomposition of each row of attention matrix into two supplementary spaces $Ker([T, 1]')$ and its orthogonal. Once again for finite dimensions $\mathbb{R}^{d_s}$, $\mathbb{R}^{d_s} = Ker[T, 1]'^{\perp} + Ker[T, 1]'$, which means that any vector of $x \in \mathbb{R}^{d_s}$ can be written with a unique decomposition $x = x^{\sharp} + x^{\perp}$, where $x^{\sharp}$ is the projection of x into $Ker([T, 1]')$ and $x^{\perp}$ into its orthogonal.

Note that $AT = [A_1T, ..., A_{d_s}T]$ with A_i are rows of A. We do the same decomposition for each row of attention matrix A, then: $A = A^{\sharp} + A^{\perp}$. For any matrix A, we have $AT = (A^{\sharp} + A^{\perp})T = A^{\sharp}T + A^{\perp}T$ but each row $A_i^{\sharp} \in Ker[T, 1]'$, then $A^{\sharp}T = 0$, which means that $AT = A^{\perp}T$.

We also know that since A is a softmax, then $A.1 = 1$, therefore $A^{\perp}1 + A^{\sharp}1 = 1$ but each row $(A^{\sharp})_i \in Ker[T, 1]'$, then $A^{\sharp}1 = 0$, so $A1 = A^{\perp}1$. This means that we can identify each attention matrix

[3] $[T, 1]$ is the matrix formed by adding a column vector of ones to T

responsible for an output, by a unique projection that verifies the same output and has the sum of each row equal to one.

4.2 *Positivity of efficient attention weights*

While we have proved that the weights of efficient attention matrices have certain properties of probability distributions (they sum to one) but we have not shown that every weight is positive $w \geq 0$, which is essential for each row in an AW to be a probability distribution.
Let's prove that $A^{\perp} = proj_{Ker([T,1]')^{\perp}}(A))$ is a probability distribution, under the following conditions: for each row $i \in \{1,..,n\}$: $0 < A_i = A_i^{\perp} + A_i^{\sharp} \leq 1$, $A^{\perp}.1 = 1$, $A_i^{\sharp}.1 = 0$ and $< A_i^{\perp}, A_i^{\sharp} >= 0$. These conditions are ensured by the projection. So now we need to prove is that under those conditions, $A^{\perp} \geq 0$.
We prove by induction that under the conditions mentioned above $\forall n \geq 2$, $A^{\perp}$ can't have negative weights.
For $n = 2$, we suppose that $A^{\perp}$ has a negative weight, for simplification, let's suppose the negative weight is $a_{i,1}^{\perp} < 0$ with $i \in \{1, 2\}$. With constraints we have it means that: $a_{i,1}^{\perp} + a_{i,2}^{\perp} = 1$ and $a_{i,1}^{\sharp} + a_{i,2}^{\sharp} = 0$ then $a_{i,2}^{\perp} = 1 - a_{i,1}^{\perp}$ and $a_{i,2}^{\sharp} = -a_{i,1}^{\sharp}$
We know also that $a_{i,1}^{\perp} a_{i,1}^{\sharp} + a_{i,2}^{\perp} a_{i,2}^{\sharp} = 0$ then
$a_{i,1}^{\perp} a_{i,1}^{\sharp} + (1 - a_{i,1}^{\perp})(-a_{i,1}^{\sharp}) = 0$, so $a_{i,1}^{\sharp}(a_{i,1}^{\perp} - (1 - a_{i,1}^{\perp})) = 0$
We know that $a_{i,1}^{\sharp} + a_{i,1}^{\perp} > 0$ then $a_{i,1}^{\sharp} > -a_{i,1}^{\perp} > 0$
Which means that $a_{i,1}^{\perp} - (1 - a_{i,1}^{\perp})) = 0$, then $a_{i,1}^{\perp} = \frac{1}{2} > 0$ which is in contradiction with our supposition. The same reasoning remains true by taking $a_{i,2}^{\perp}$ instead of $a_{i,1}^{\perp}$.
We suppose that this property is true for $n \geq 2$, and to prove that it remains true for for $n + 1$, we have only to get back to the case of n variables by taking: $a_{i,n+1}^{\perp} = 1 - \sum_{j=1,..,n} a_{i,j}^{\perp}$ and $a_{i,n+1}^{\sharp} = -\sum_{j=1,..,n} a_{i,j}^{\sharp}$ and then use the induction assumption. □

4.3 *Computing the projection $Ker([T, 1]')^{\perp}$*

The final challenge is to determine mathematically $Ker([T, 1]')^{\perp}$. We do this now. We first prove a new simpler form for the projection and then providing a computable and analytic method to find that simpler form mathematically. We show how to find this mathematically using the image of $[T, 1]$. The image of a linear function f, $Im(f)$, is the set of values that the function f can take by applying the linear transformation to elements of its domain.

Definition 1. *For a linear function $f : V \rightarrow W$ that maps from a vector space V to a vector space W, we define $Im(f)$ or $f(V)$: $Im(f) = \{f(v) : v \in V\}$.*

Proposition 3. $Ker([T, 1]')^{\perp} = Im([T, 1])$

$x \in Ker([T, 1]') \Leftrightarrow [T, 1]'x = 0 \Leftrightarrow \forall y \in \mathbb{R}^{d+1} : \langle [T, 1]'x, y \rangle = 0 \Leftrightarrow \forall y \in \mathbb{R}^{d+1} : \langle x, [T, 1]y \rangle = 0.$ $\Leftrightarrow x \in Im([T, 1])^{\perp} \Leftrightarrow Ker([T, 1]') = Im([T, 1])^{\perp} \Leftrightarrow Ker([T, 1]')^{\perp} = Im([T, 1])^{\perp.\perp}$, since we are in finite dimensions then $Im([T, 1])^{\perp\perp} = Im([T, 1])$. As a result, $Ker([T, 1]')^{\perp} = Im([T, 1])$. □

We now show an analytical way to find $Im([T, 1])$.
Consider the standard basis of $\mathbb{R}^{d+1}$: $(e_1, ..., e_{d+1})$. Now $Im([T, 1]) = vect([T, 1]e_1, ...[T, 1]e_{d+1})$ gives us a generating family that we can transform into a basis using the Gauss Pivot algorithm and then get $proj_{Im([T,1])} = proj_{Ker([T,1]')^{\perp}}$ to finally have our new efficient matrix $A_{\text{efficient}} = proj_{Ker([T,1]')^{\perp}}(A)$. This procedure is general and holds for all attention matrices; it provides an effective computation of identifiable AW matrices that are identifiable.

Proposition 4. *When A is identifiable, $A_{\textit{efficient}} = A$*

This follows from the fact that when A is identifiable, $Ker([T, 1]'^{T}) = \{0\}$, and so $A = proj_{Ker([T,1]')}(A) + proj_{Ker([T,1]')^{\perp}}(A) = proj_{Ker([T,1]')^{\perp}}(A)) = A_{eff}$

To summarize, efficient attention restores identifiability of AW by removing the weight components that do not influence the generation of the contextualization vectors AT, and consequently the model predictions. We can thus extract the distribution that is responsible for the generation of contextualization vectors and which is unique. That is, we can have two different distribution of attention that generate the same prediction, but their efficient attention will be the same : $A_1.T = A_2.T$ with $A_1 \neq A_2$ but $proj_{Ker([T,1]')^{\perp}}(A_1)$ will be equal to $proj_{Ker([T,1]')^{\perp}}(A_2)$ and thus correspond to $A_{\textit{eff}}$. Our technique isolates the factors that can serve an explanatory role, eliminating the noise from AW.

5 Empirical investigations

Datasets	#test	$W(P_A, P_{A_{\textit{eff}}})$	$RS(P_A, P_{A_{\textit{eff}}})$
IMDB	4356	0.0016	0.02
AGNEWS	3798	0.0026	0.02
SST	1725	0.0039	0.03
20NEWS	357	0.0051	0.02

Table 1: Comparison between predictions generated by attention matrices P_A and those generated by their efficient attention projections $P_{A_{\textit{eff}}}$. #test gives the number of test samples; the metrics W and RS are defined below.

In this section, we support our approach with three empirical experiments. Our formal results are directly relevant to the observations of [24, 45], who respectively claim that « attention is not explanation » and that « attention is not explanation if you don't need it». They produce different attention matrices generating the same predictions, and claim shows AW cannot be explanatorily important. We use the same models and datasets as they do and show that efficient attention is identifiable and explanatorily relevant in the cases they consider. To make our case for efficient AW in the strongest terms, we used [45]'s architecture, which is made up of an encoder (LSTM/RNN/MLP), followed by an attention layer, then a decoder(LSTM/RNN/MLP). We note, however, that our experimental set up transfers easily to other transformer architectures.

First, we show that predictions generated by attention matrices and their efficient attention matrices are the same. Second, we show that [45]'s different attention matrices generating the same predictions have the same efficient attention matrices. Third, we show that two attention matrices with different efficient attention matrices generate different predictions. For practical reasons, the values will be very close rather than the same, given the limits of integer precision and approximation.

Following [24, 45], we conducted our experiments on the following datasets: IMDB Large Movie Reviews Corpus [31], in which the task is to predict positive or negative sentiment from movie reviews, AG News Corpus [47] to discriminate between world and business articles, Stanford Sentiment Treebank (SST) [39] to predict positive or negative sentiment from text, And 20 News Corpus to discriminate between belonging to baseball and hockey stories. We used [24]'s train-test split and dataset versions. Our analysis is done on the test set as for [24, 45]. Note that all datasets are in English.

T passes through the attention layer to generate the attention matrix A. Then $A.T$ is passed as input to the decoder to generate prediction P_A. Our manipulations are carried out in the step between the encoder and the decoder, where we calculate efficient attention A_{eff} from A and T, then its predictions $P_{A_{eff}}$. Note that even though our results have been demonstrated for transformers, they remain valid for attention-based systems, since the role of attention remains to generate contextualization vectors and that we can extract efficient attention responsible for their values from AW matrix A by projecting it onto the space constructed from hidden states matrix T in the same way.

We used a variety of metrics in our experiments. First, "Earth-Mover" distance or Wasserstein-1 : $W(P_A, P_{A_{eff}})$, is used to compare probability distributions [1, 34]. Since predictions as well as both original attention A and its efficient projection A_{eff} are distributions, this is a natural measure.

$$W(P_A, P_{A_{eff}}) = \inf_{\gamma \in \Pi(P_A, P_{A_{eff}})} \mathbb{E}_{(x,y)\sim\gamma}[||x - y||]$$

where $\Pi(P_A, P_{A_{eff}})$ is the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively P_A and $P_{A_{eff}}$. Intuitively, $\gamma(P_A, P_{A_{eff}})$ indicates how much "mass" must be transported from x to y in order to transform the distributions P_A into the distribution $P_{A_{eff}}$. We also use a Mean Wasserstein distance over the set of matrices for a given database in Tables 2 and 3, as well as the tables in the appendix.

A second measure we use is root-mean-square deviation, $RS(P_A, P_{A_{eff}})$, a measure of accuracy, to compare forecasting errors of different models on a particular dataset.

$$RS(P_A, P_{A_{eff}}) = \sqrt{\frac{\sum_{i=1}^{n}(P_{A_i} - P_{A_{i_{eff}}})^2}{n}}$$

We also define and use Pierson (r^2) and weighted L_2 distance in the Appendix B.

In all the experiments we are not talking about strict identity between outcomes and shifted attention matrices. Given the approximation errors due to calculations of projection, we will have approximate values. But the differences should be very small. If that is the case, as we noted in Section 2, this will suffice for identifiablity of AW. Similar remarks apply to the causal efficacy of AW.

For our first experiment, to see whether AW matrix A and its efficient projection A_{eff} have the same predictions, we saw that results were very close, indeed identical, up to approximation errors. To detail our experimental method for each dataset we examined, we extracted for each attention matrix, the efficient attention matrix that corresponds to it. We then compared the predictions generated by the efficient attention matrix and the original matrix using three distinct metrics. Table 1 shows that the predictions are the same with a few very small variations due to calculation errors, which confirms our assertion that $A \to P \Rightarrow A_{eff} \to P$. We replicated this finding in Table 6 in Appendix B for all the metrics we tried (also Table 4 for relative RS comparisons in the Appendix).

For our second experiment, where we compare original and adversarial matrices of [45], for all datasets, we apply the adversarial algorithm of [45], so that for each sample we have its attention matrix and its adversarial attention matrix, we extract the efficient matrix of each and compare them. As can be seen from the results, in cases where the adversarial algorithm generates predictions close to the values of the original predictions, the effective attention weights of the two are close. Note also that the algorithm proposed by [45] depends on the number of datasets, as can be seen very clearly in the table that the quality of predictions is degraded when the number of datasets is reduced, as in the case of 20News, and as a result there will obviously be a difference between the original efficient matrices and adversarial efficient matrices. We also note that our algorithm can also be seen as an alternative way of generating adversarial matrices and that is not depending on the length of dataset. Our results are summarized in Table 2, where once again the differences between the two efficient matrices is negligible. More extensive tests with other metrics can be found in Table 7 in the appendix.

Datasets	WA-eff	W-Pred	RS-Pred
IMDB	0.0047	0.0034	0.08
AGNEWS	0.0066	0.0021	0.05
SST	0.0288	0.0055	0.10
20NEWS	0.0236	0.0509	0.20

Table 2: WA-eff: $W_M(A_{eff}, A_{eff}^{adv})$), Mean Wasserstein distances between efficient attention matrices generated by the attention matrix and those generated by their adversarial attention matrices. W-Pred, RS-Pred:comparisons between their predictions $P_{A_{eff}}$ and $P_{A_{eff}^{adv}}$ using Wasserstein, $W(P_{A_{eff}}, P_{A_{eff}^{adv}})$, and RS, $RS(P_{A_{eff}}, P_{A_{eff}^{adv}})$.

To complete our study of adversarial AW and efficient AW, Table 5 in Appendix B gives a full comparison using Wasserstein distances for the predictions raw AW and efficient AW, predictions of raw AW and adversarial AW from [45], adversarial AW and the efficient projection of those adversarial weights, and finally the predictions of efficient AW and the projection of adversarial AW.

Third, we show that shifting efficient attention in a counterfactual intervention on the model causes a difference in prediction. We note first that efficient attention separates AW matrices more distinctly than the adversarial operation of [44]. To illustrate this,we generate for each sample its attention matrix A, its efficient attention projection A_{eff} and prediction $P_{A_{eff}}$, simultaneously, we take $1 - A$, we generate its efficient attention projection $(1 - A)_{eff}$ and prediction $P_{(1-A)_{eff}}$. We tested this procedure on all our datasets, and found that each time efficient attention matrices are different, predictions are different, which what we predicted mathematically. Table 3 shows the results with Wasserstein distances between predictions of distinct efficient AW is approximately 10x larger for individual AW and the mean Wasserstein difference on distinct efficient matrices is much larger than that for $W(P_{A_{eff}}, P_{A_{eff}^{adv}})$ for our datasets.

Datasets	$mean(W(A^1_{eff}, A^2_{eff}))$	$W(P_{A^1_{eff}}, P_{A^2_{eff}})$
IMDB	0.98	0.13
AGNEWS	0.94	0.06
SST	0.88	0.16
20NEWS	0.97	0.18

Table 3: Comparison to show that different efficient attention matrices generates different predictions.

Tables 4 and 5 below give an overview of RS and Wasserstein distances on predictions of various AW.

6 Discussion

Given that identifiability of AW as distributions maximizes their explanatory value in the sense that we explained in Section 2, our efficient AW provide the strongest explanation possible from attention weight distributions. Combining this with the experimental results in

Datasets	$RS(P_A, P_{A^{adv}})$	$RS(P_A, P_{A_{eff}})$	$RS(P_{A^{adv}}, P_{A^{adv}_{eff}})$	$RS(P_{A_{eff}}, P_{A^{adv}_{eff}})$
IMDB	0.03	0.02	0.07	0.08
AGNEWS	0.01	0.02	0.04	0.05
SST	0.08	0.03	0.05	0.10
20NEWS	0.20	0.02	0.03	0.20

Table 4: Comparison between RS values.

Datasets	$W(P_A, P_{A_{eff}})$	$W(P_A, P_{A^{adv}})$	$W(P_{A^{adv}}, P_{A^{adv}_{eff}})$	$W(P_{A_{eff}}, P_{A^{adv}_{eff}})$
IMDB	0.0016	0.0038	0.0020	0.0034
AGNEWS	0.0026	0.0020	0.0032	0.0021
SST	0.0039	0.0058	0.0036	0.0055
20NEWS	0.0051	0.0481	0.0067	0.0509

Table 5: Comparison of Wasserstein distances

prior literature that AW are sometimes not identifiable shows that using efficient attention is superior to using either raw attention or [7]'s solution for explanatory purposes in many datasets.

A question that arises is, how different is a model M' which differs from a model M only insofar as original AW in A have been replaced by efficient AW A_{eff} in M' ($M' = M[A_{eff}/A]$? Is M' for instance a surrogate model, an approximation of the original model?

Given our results, M' is not a surrogate model of M. To see this, note that transformer models are finite automata with (internal) states S and Λ labeled transitions $p \xrightarrow{\alpha} q$ for $\alpha \in \Lambda$ for $p, q \in S$. If 2 states are input output equivalent we write $p \equiv q$. Examples of states could be the output of a given attention layer.

Definition 2. *Two finite automata $\mathcal{S} = S \times \Lambda \times S)$ and $\mathcal{T} = T \times \Lambda \times T)$ are strongly equivalent iff there is a 1-1 and onto function $\beta : S \to T$ such that for every $s \in S$ and every $t \in T$, (i) if $s \xrightarrow{\alpha} s' \in \mathcal{S}$, then $\exists t \xrightarrow{\alpha} t' \in \mathcal{T}$ such that $(s' \equiv t')$; (ii) if $t \xrightarrow{\alpha} t' \in \mathcal{T}$, then $\exists s \xrightarrow{\alpha} s' \in \mathcal{S}$ such that $(t' \equiv s')$.*

Since A_{eff} and A have the same predictions in the sense that $A_{eff}.T = A.T$, we see that:

Proposition 5. *Let $M' = M[A_{eff}/A]$ be defined as above. Then M and M' are strongly equivalent.*

Any compelling argument against the usefulness of AW as explanatory devices should exploit our efficient AW. To this end, we now reconsider the criticisms of AW as explanatory vehicles for NLP tasks involving text classification. We look at two: the use of uniform weights, the use of adversarial weights using the adversarial technique of [45].

Uniformity: Note that the term *uniform* here does not mean that all weights have the same value, but that non-zero weights are the same. [45, 42, 41] have used uniform weights on contextual elements instead of attention matrix weights to test if attention matrices are important for the output. They found that uniform weights did not substantially alter the predictions at least for some tasks. This argument is far from conclusive because it does not look at the nature of the effective components in AW. [7]'s effective AW are more uniform, in general,than raw attention, as are our efficient AW, since they are included in the [7] effective attention space. Thus uniform weights at least for classification task matrices do not differ much from those of efficient attention. We have verified this experimentally; for uniform A, A_{eff} also has rather uniform values in our results; our matrices in the tables are uniform.

Finally, the uniform values must be close to original matrix values of [45]. [7] show experimentally that the values that are accentuated in raw attention remain those with the greatest value in effective attention but their value decreases to the point of changing only slightly from the other scores. The same holds for our efficient AW rendering the attention weight identifiable makes it close to uniform. Our experiment 1 (Table 1) also demonstrates that A and A_{eff} have close values. This may explain the good performance achieved by the uniform weights set by [45] in some datasets.

Adversarial AW. During our experiments, we noticed that the quality of the predictions generated by the adversarial algorithm proposed by [45] depends on the number of samples in each dataset, as can be seen in our metrics in (Table 2) and our appendix tables, especially for 20News and slightly for SST. On the other hand, our efficient attention matrix can also be seen as an adversarial algorithm that generates a different attention matrix with the same prediction as the original, but in this case is independent of the number of samples in each dataset.

Permutations of AW. [24] argue against attention as an explanatory vehicle by showing that permuting the AW only slightly varies the output. However, their attention is a vector not a matrix. Where AW are true matrices, our experiments have shown the values are relatively uniform, so permutation is not expected to produce large differences in results.

Finally we offer a detailed comparison with [7] that inspired our work. First, [7] investigate identifiability but they don't say why it is important. We say why identifiability is important for explainability contra [45]. This is crucial for our paper as we are looking at the explanatory potential of AW. [6] is not explicitly motivated by explanatory concerns but rather by mechanistic interpretability.

Second, [7] propose the orthogonal of Ker(T') to remove weight components that do not influence model predictions; we show that the orthogonal of Ker([T,1]') preserves identifiability and we show both mathematically and empirically that it removes components that do not affect model predictions.

The projection by [7] is not guaranteed to be a probability space and they provide examples where it could not be; we show that our projection guarantees that the image of an AW under a projection into the orthogonal of Ker([T,1]') is also an AW. [7]'s effective AW are not real AW, according to definitions in the literature on transformer architectures. We also provide and use an algorithmic method for calculating the orthogonal of Ker([T,1]').

We verify experimentally that our AW are distinct when results are distinct and that there is 1 efficient AW for distinct raw AW replicating [45]'s empirical set up; [7] does not. We compare efficient AW using the well known Wasserstein metric appropriate for probability distributions, [7] cannot. This experimental component provides a method for also using efficient AW in faithful and plausible explanations of model behavior. We argue [7] cannot do this, since their effective AW are not real AW but unknown objects.

Datasets	$r^2(P_A, P_{A_{eff}})$	$mse(P_A, P_{A_{eff}})$	$L_2^1(P(A), P(A_{eff})$	$L_2^2(P(A) - P(A_{eff})$
IMDB	0.99	0.0006	0.02	0.0003
AGNEWS	0.99	0.0006	0.01	0.0004
SST	0.98	0.0013	0.02	0.0008
20NEWS	0.99	0.0005	0.01	0.0012

Table 6: Different metrics comparing predictions generated by attention matrices and those generated by their efficient attention projections.

Datasets	$r^2(P_A, P_A^{adv})$	$r^2(P_{A_{eff}}, P_{A_{eff}^{adv}})$	$MSE(P_A, P_A^{adv})$	$L_2^1(P(A), P(A^{adv})$	$L_2^1(A_{eff}, A_{eff}^{adv})$
IMDB	0.99	0.95	0.0014	0.029	0.33
AGNEWS	0.99	0.98	0.0001	0.009	0.15
SST	0.94	0.91	0.0069	0.067	0.29
20NEWS	0.64	0.62	0.0413	0.167	0.70

Table 7: Different metrics comparing attention matrices, its efficient attention matrices, adversarial attention matrices, their efficient attention matrices and their predictions.

7 Conclusions

We have provided formal results, restoring the potential explanatory power of attention. We have shown how to render AW effectively identifiable, reducing noise in AW, while also ensuring their interpretability with *efficient attention*. Substituting efficient attention for original attention provides a strongly equivalent model.

Our novel concept of efficient attention removes an important obstacle to making AW a plausible basis for faithful and plausible explanations. As we pointed out, however, if a task for a transformer model does not require contextualization or much contextualization, it may be that AW are not causally involved as they are not needed to complete the task. Our results for efficient attention justify placing more sophisticated learning strategies on top of AW [2, 22] for explaining model behavior in tasks where contextualization is needed.

We have illustrated our approach in three empirical experiments over four databases. We compared predictions from an AW and its efficient projection, and we show that the predictions are very close, and for practical purposes identical given integer and approximation precision. Second, we took the adversarial AW provided by [45] A^{adv} and compared their efficient AW with the efficient AW of the original matrix. Third, we showed that given two different efficient matrices A_{eff} and $A^{\dagger}_{eff}$ that their projections differ.

Our mathematical results are general and apply not only to the simplified attention models of [45] and [24] but to full transformer models both of the encoder and full kind. We conducted our experiments on [45] architecture, as it offered a way of having two different attention matrices presenting the same predictions (adversarial technique). This enabled us to show that they shared the same efficient attention matrix in addition to our other experiments. Space precluded us from examining more modern transformer models empirically, but counterparts of our results here should hold.

Given that efficient AW are identifiable from predictions and that we can trace in a transformer model links from input tokens to AW, we can see how shifts in AW occur with shifts in input. This will also enable linking efficient AW to human annotations of “rationales” on input data to test plausibility of the explanations in more detail as in [12]. We feel this is an illuminating path for future research.

AW is only part of a transformer model. We have shown why this part is especially important in model performance, but other parts, e.g. the Residual stream, Layer Normalization or the MLP, will also play a part in a complete explanation of the model’s performance.

8 Appendices

A. Proof of Proposition 1

To prove the right to left direction, suppose that : $Ker(T') = \{0_{d_s}\}$ $A \in Ker(f)$, we consider : $A' = [a_1, ..., a_{d_s}]$ with a_i is the i column of A’ then : $T'A' = [T'a_1, ..., T'a_{d_s}] = [0, .., 0]$ $\Rightarrow \forall i \in 1, .., d_s : T'a_i = 0$ we know that $Ker(T') = \{0_{d_s}\}$ then $\forall i \in 1, .., d_s : a_i = 0 \Rightarrow A' = 0 \Rightarrow A = 0 \Rightarrow Ker(f) = \{0_{d_s * d_s}\} \Rightarrow$ f is injective (1)

To prove the left to right version we use the contrapositive, $Ker(T') \neq \{0_{d_s}\} \Rightarrow Ker(f) \neq \{0_{d_s * d_s}\}$

Assume that $Ker(T') \neq \{0_{d_s}\} \Rightarrow \exists a \in R^{ds}\_\{0\} : T'a = 0$ we take the matrix $A' = [a, .., a] \neq 0 : T'A' = [T'a, ..., T'a] = [0, ..., 0] \Rightarrow A \in Ker(f) \Rightarrow Ker(f) \neq \{0_{d_s * d_s}\}$ (2)

Finally $Ker(f) = \{0_{d_s * d_s}\} \Leftrightarrow Ker(T') = \{0_{d_s}\}$, then f is injective iff $Ker(T') = \{0_{d_s}\}$

Consider now $Ker(T')$: We know that $T \in R^{d_s * d}$; so $T' \in R^{d * d_s}$. Now apply the Rank Theorem : $dim(Ker(T')) + rg(T') = d_s$. Since we know that $rg(T') = rg(T)$ then $dim(Ker(T')) = d_s - rg(T)$. Finally if f_A is injective, $Ker(T') = \{0\}$ which means that $rg(T) = d_s$.

A. Proof of Proposition 2

As $[T, 1]' \in R^{(d+1) * d_s}$, we apply the rank theorem and so: $dim(Ker([T, 1]')) + rk([T, 1]') = d_s$.

We know $rk([T, 1]') = rk([T, 1])$, then $dim(Ker([T, 1]')) = d_s - rk([T, 1])$. So A is identifiable if $d_s = rk([T, 1])$, but here we search for a sufficient condition for non-identifiability.

$rk([T, 1]) \leq rk(T) + 1 \leq min(rk(E), rk(W^V), rk(H)) + 1 \leq min(d_s, d, d_v) + 1 \leq max(d_s, d_v) + 1 \leq max(d_s + 1, d_v + 1)$.

Therefore $dim(Ker([T, 1]')) \geq d_s - max(d_s + 1, d_v + 1) \Rightarrow dim(Ker([T, 1]')) \geq max(-1, ds - dv - 1)$ And we know that $dim(Ker([T, 1]')) \geq 0$ finally $dim(Ker([T, 1]')) \geq max(0, ds - dv - 1)$.

Then for $d_s > d_v + 1 : ker([T, 1]') \neq \{0\} \Rightarrow \exists B \in ker([T, 1]') \neq 0$ that verifies $f_{A+B} = f_A$ and $B.[T, 1] = 0$.

B. More metrics and tables

We used MSE, mean squared error, and $r^2(P_A, P_{A_{eff}})$ or Pierson distance, the coefficient of determination between P_A and $P_{A_{eff}}$:

$1 - \frac{\sum_{i=1}^{n}(P_{A_i} - P_{A_{eff}})^2}{\sum_{i=1}^{n}(P_{A_i} - mean(P_{A_i}))^2}$

We also used two mean, normalized L_2 distances, one for matrices and one for predictions:

$L_2^1(A_1, A_2)$: *mean*$(\frac{\|A_1 - A_2\|}{\|A_1\| + \|A_2\|})$ and $L_2^2(A_1, A_2)$: *mean*$(\frac{\|A_1 A_2\|}{n})$

The latter metrics were less robust and sensitive to the size of the database, especially with [45]’s adversarial method.

Acknowledgements

We gratefully acknowledge the support of the EU grant TUPLES as well as the SARER and Summ-RE ANR-20-CE23-0017 grants.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] I. Arous, L. Dolamic, J. Yang, A. Bhardwaj, G. Cuccu, and P. Cudré-Mauroux. Marta: Leveraging human rationales for explainable text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 5868–5876, 2021.

[3] N. Asher, L. De Lara, S. Paul, and C. Russell. Counterfactual models for fair and adequate explanations. *Machine Learning and Knowledge Extraction*, 4(2):316–349, 2022.

[4] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[5] A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, T. François, and P. Watrin. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, 2022.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] G. Brunner, Y. Liu, D. Pascual, O. Richter, M. Ciaramita, and R. Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJg1f6EFDB.

[8] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 2014.

[9] G. Chrysostomou and N. Aletras. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 477–488, 2021.

[10] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, 2019.

[11] A. de Santana Correia and E. L. Colombini. Attention, please! a survey of neural attention models in deep learning. *Artificial Intelligence Review*, 55(8):6037–6124, 2022.

[12] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

[13] G. F. D. Duff, D. Naylor, and D. A. McQuarrie. *Differential equations of applied mathematics*. John Wiley and sons, 1966.

[14] T. Fel, A. Picard, L. Bethune, T. Boissin, D. Vigouroux, J. Colin, R. Cadène, and T. Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721, 2023.

[15] J. Ferrando, G. I. Gállego, and M. R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, 2022.

[16] J. Ferrando, G. I. Gállego, I. Tsiamas, and M. R. Costa-jussà. Explaining how transformers use context to build predictions. *arXiv preprint arXiv:2305.12535*, 2023.

[17] A. Galassi, M. Lippi, and P. Torroni. Attention in natural language processing. *IEEE transactions on neural networks and learning systems*, 32(10):4291–4308, 2020.

[18] A. Geiger, H. Lu, T. Icard, and C. Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34: 9574–9586, 2021.

[19] M. Geva, R. Schuster, J. Berant, and O. Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.

[20] R. Ghaeini, X. Fern, and P. Tadepalli. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957, 2018.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] J. Heo, J. Park, H. Jeong, K. j. Kim, J. Lee, E. Yang, and S. J. Hwang. Cost-effective interactive attention learning with neural attention processes. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4228–4238, 2020.

[23] A. Jacovi and Y. Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, 2020.

[24] S. Jain and B. C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, pages 3543–3556, 2019.

[25] F. Jourdan, L. Risser, J.-M. Loubes, and N. Asher. Are fairness metric scores enough to assess discrimination biases in machine learning? *arXiv preprint arXiv:2306.05307*, 2023.

[26] G. Kobayashi, T. Kuribayashi, S. Yokoi, and K. Inui. Incorporating residual and normalization layers into analysis of masked language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.

[27] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

[28] D. Lewis. *Counterfactuals*. Basil Blackwell, Oxford, 1973.

[29] D. Lewis. A subjectivist's guide to objective chance. In *IFS: Conditionals, Belief, Decision, Chance and Time*, pages 267–297. Springer, 1981.

[30] L. E. Loeb. Causal theories and causal overdetermination. *The Journal of Philosophy*, 71(15):525–544, 1974.

[31] A. 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, 2011.

[32] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

[33] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[34] V. M. Panaretos and Y. Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

[35] J. Pérez, P. Barceló, and J. Marinkovic. Attention is turing-complete. *Journal of Machine Learning Research*, 22(75):1–35, 2021.

[36] T. Räuker, A. Ho, S. Casper, and D. Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483. IEEE, 2023.

[37] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[38] S. Serrano and N. A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, 2019.

[39] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[40] A. Tarski, A. Mostowski, and R. M. Robinson. *Undecidable theories*, volume 13. Elsevier, 1953.

[41] M. Tutek and J. Šnajder. Staying true to your word:(how) can attention become explanation? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 131–142, 2020.

[42] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*, 2019.

[43] J. Vig and Y. Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, 2019.

[44] J. Wei, J. Wei, Y. Tay, D. Tran, A. Webson, Y. Lu, X. Chen, H. Liu, D. Huang, D. Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

[45] S. Wiegreffe and Y. Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, 2019.

[46] K. Yin and G. Neubig. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, 2022.

[47] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.