An Evaluation Framework for Explainability Approaches in Seq2Seq Machine Translation Models

Anonymous EMNLP submission

Abstract

The study of the attribution of input features to the output of neural network models is an active area of research. While numerous explainability techniques have been proposed to interpret these models, the systematic and automated evaluation of these methods in the sequence-to-sequence models remains underexplored. This paper introduces a novel approach for evaluating explainability methods in transformer-based seq2seq models, building upon forward simulation of XAI methods. Our method transfers learned knowledge in the form of attribution maps from a larger model to a smaller one and quantifies the resulting impact on performance. We evaluate eight explainability methods using the Inseq library to extract attribution scores linking input and output sequences. This information is then injected into the attention mechanism of an encoder-decoder transformer for machine translation. Our results show that this framework serves both as an automatic evaluation tool for explainability techniques and as a knowledge distillation strategy that enhances model performance. Our experiments demonstrate that Attention attributions and Value Zeroing methods consistently improved results on three machine translation tasks and four composition settings. The codes will be available on Github [1].

1 Introduction

In recent years, natural language processing (NLP) generative models have advanced rapidly and found applications in a wide range of domains (Chang et al., 2024; Kalyan, 2024). These models are typically built on complex neural network architectures, which are often described as "black boxes" due to their opaque internal mechanisms (Dayhoff and DeLeo, 2001; Burkart and Huber, 2021). To address the challenges posed by these opaque models, the field of Explainable AI (XAI) aims to enhance the transparency and interpretability of machine learning systems. A key objective of XAI approaches is to assess and quantify the importance of input features or attributions on the final output of these models (Arya et al., 2019; Vieira and Digiampietri, 2022; Saeed and Omlin, 2023). Several XAI methods and algorithms have been developed and applied in NLP models to assess the attribution of input tokens in the final output for different tasks (Madsen et al., 2022b). However, determining which explainability method most accurately reflects the underlying relationships learned by the model remains an open challenge.

Since explanations are intended for human interpretation, the validation of XAI methods has been predominantly human-centered (Kim et al., 2024). Nevertheless, some approaches for the automatic evaluation of XAI methods in other domains have been proposed (Nauta et al., 2023). For example, in image classification tasks, techniques such as removing important features identified by XAI methods and retraining the model based on the remaining features (Hooker et al., 2019; Ribeiro et al., 2016a), as well as covering important features (Chang et al., 2018), have been explored. However, such approaches are less common in NLP (Madsen et al., 2022a), where human evaluation remains the dominant method (Madsen et al., 2022b; Leiter et al.). Furthermore, prior research has primarily focused on interpreting the attention mechanism (Moradi et al., 2021; Serrano and Smith, 2019) rather than comparing different explainability methods.

In the context of the evaluation of XAI methods, one proposed approach is the concept of simulatability (Doshi-Velez and Kim, 2017; Hase and Bansal, 2020), which measures how well explanations allow humans to predict the behavior of a model after receiving its explanation. However, human evaluation is costly and not easily scalable, motivating the development of automated evalua-

[1] https://github.com/

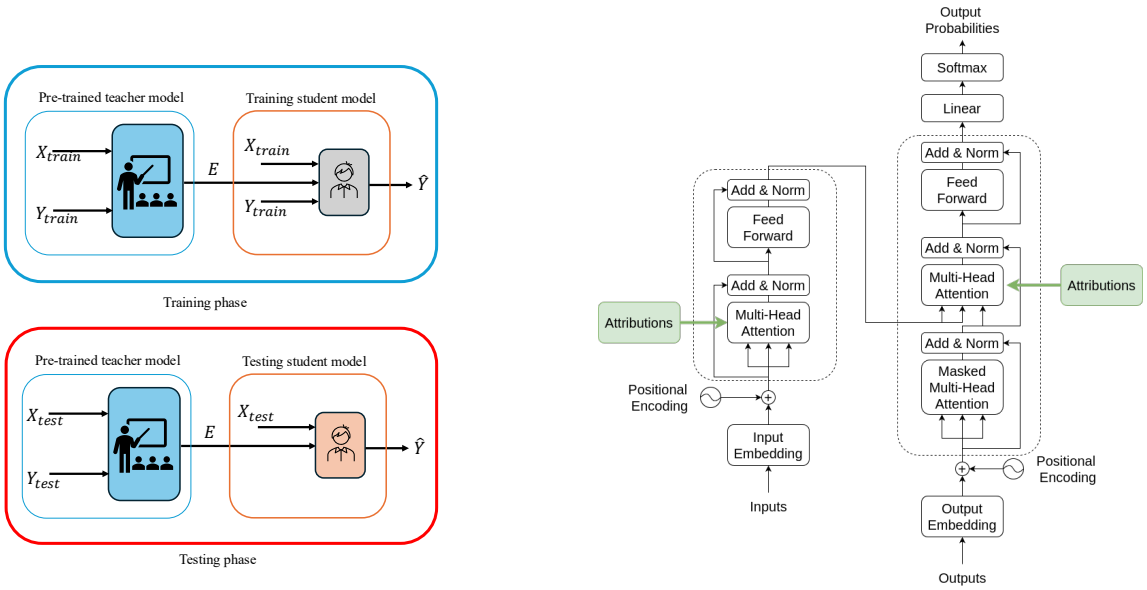


Figure 1: **(a)** Illustrates the overall design of our approach. The input sequence and the gold output (X, Y) are given to a teacher model, and their attributions E are obtained. Then, a new untrained model is trained using the same (X, Y, E) triples. In the testing phase, the model gets the $(X, E) \rightarrow \hat{Y}$. **(b)** Shows two places where we inject the attributions obtained from XAI methods.

tion pipelines for XAI methods. Building on this concept, we hypothesize that each explainability method encodes unique information to account for the feature attribution. We further conjecture that this information can be leveraged to train a downstream model that benefits from exposure to the explanations generated by the original model. In other words, the better the explanation, the higher the performance should be for a model exposed to this information. This approach offers a systematic framework to evaluate and compare different XAI methods.

These questions become particularly pertinent in the context of sequence-to-sequence (seq2seq) architectures (Sutskever, 2014), which employ an encoder-decoder framework and are central to neural machine translation, summarization, and dialogue systems. However, the many-to-many mappings and intricate encoding-decoding processes of seq2seq models make them more challenging to interpret than simpler classification models (Gurrapu et al., 2023). Understanding the causal relationships between source and target tokens is key to explaining the behavior of seq2seq models (Alvarez-Melis and Jaakkola, 2017), and several approaches aim to provide such insights.

In this work, we propose a new approach to evaluate XAI attribution methods in seq2seq models. In our approach, we train a new-untrained model by feeding source, target, and explanation attribution maps between the two sequences. In the test time, the model receives the source (i.e., input) along with the attribution explanation, and its performance in generating the correct output is measured. We apply this approach to three Machine Translation tasks, using Opus-MT (Tiedemann and Thottingal, 2020) models. To generate attribution explanations, we utilize the Inseq Python library (Sarti et al., 2023), which offers an interface to common XAI attribution methods for seq2seq models.

In summary, the main contributions of this work are as follows.

- We propose a novel framework for evaluating explainability methods in seq2seq models by integrating attribution mapping information directly into the encoder-decoder architecture.
- We conduct extensive experiments exploring multiple strategies for incorporating explanations within the Transformer architecture and systematically compare their effects on model performance across various MT language pairs.
- We provide empirical evidence that XAI attribution methods significantly influence the

performance of seq2seq models. Our findings demonstrate that the quality and type of explanations can enhance or degrade model output relative to baseline models without attribution guidance.

2 Related work

2.1 Explainable AI in Seq2seq Models

Seq2seq models, particularly those based on the transformer architecture (Vaswani, 2017), have revolutionized tasks such as machine translation and text summarization by capturing complex dependencies between input and output sequences (Stahlberg, 2020; Shakil et al., 2024). Nevertheless, their encoder-decoder structure introduces several challenges for explainability methods (Zhao et al., 2024). For instance, *Intermediate representations* pose a challenge as the transformation of inputs through multiple layers makes it difficult to directly correlate input features with outputs (Sutskever, 2014). *Attention mechanisms* are often used to explain decisions in transformer models, but their reliability as faithful explanations has been questioned (Jain and Wallace, 2019; Madsen et al., 2022a). Furthermore, *evaluation metrics* used in standard explainability methods may not fully capture the nuances of seq2seq models, necessitating specialized evaluation frameworks (Hase and Bansal, 2020; Nauta et al., 2023).

To address some of these challenges, recent research has focused on developing explainability methods designed specifically for seq2seq models (Burkart and Huber, 2021; Zhao et al., 2024). In this context, Lakhotia et al. (2021) introduced **FiD-Ex**, a framework that improves the faithfulness of explanations in seq2seq models by incorporating sentence markers and fine-tuning on structured datasets. Furthermore, Sarti et al. (2023) developed **Inseq**, a Python library that provides a comprehensive tool for analyzing and comparing different explainability methods for generative language models. The library offers a range of gradient-based, perturbation-based, and internal representation-based explainability techniques.

2.2 Evaluation of XAI methods

The evaluation of XAI methods has frequently drawn on established benchmarks introduced by works such Hooker et al. (2019), DeYoung et al. (2019), and Nguyen (2018). A common thread across these studies is the reliance on feature removal or manipulation strategies, where input features deemed important by a given XAI method are systematically ablated or masked to assess the impact on model performance. In our work, we use the XAI attribution as weights to strengthen (or diminish) the relation between input and output sequences.

Moreover, Tourni and Wijaya (2023) introduced an approach that leverages Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) explanations to weight intermediate features according to their relevance to the final output. However, the primary aim of their work was not to systematically compare different explainability methods, and their approach was limited to a single attribution technique. Li et al. (2020) introduced an approach by approximating the Alignment Error Rate metric through proxy models that utilize the most relevant source words identified by explanation methods.

In this work, we evaluate XAI attribution methods for seq2seq models in the context of machine translation tasks. Building on the concept of simulatability, our approach centers on injecting learned explanation mappings from a large trained model into a smaller one and training it from scratch. Then we can assess the impact of the explanation on translation performance. This framework enables us to systematically investigate how different attribution methods contribute to the transfer of mapping knowledge by affecting the model's performance. Specifically, we first use a pre-trained Opus-MT model (Tiedemann and Thottingal, 2020) to generate explanations for src-target pairs. We then train another model from scratch by providing the attribution weights to the encoder-decoder attention mechanism. By measuring performance changes, we evaluate how effectively these explanations influence model behavior. To systematically compare different XAI techniques, we use the **Inseq** library[2], applying the eight explainability methods described in the following subsection.

2.3 AI Explainability Methods

XAI attribution methods can be broadly categorized into three main types: gradient-based, internal-based, and perturbation-based approaches (Sarti et al., 2023). In this work, we use a multitude of XAI methods to generate input feature attribution scores.

[2] https://github.com/inseq-team/inseq

Saliency: Attribution scores with the saliency method (Simonyan et al., 2013) are calculated with the following formula:

$$arg\ max_I S_c(x) - \lambda||x||_2^2 \quad (1)$$

Where $S_c(x)$ is the score of class c computed by the last layer of a network for an input sentence x (Simonyan et al., 2013).

Input X Gradient: is calculated on the basis of the saliency method. The key difference is that the saliency map is multiplied with the input feature values x[3]:

$$x \times Saliency(x) \quad (2)$$

Layer Gradient × Activation: For this method, the same formula as for *Input* × *Gradient* is used, except that it is targeted at a specific layer, specifically the last encoder layer.

Integrated Gradients: integrates the gradient at dimension i of an input x:

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\delta F(x' + \alpha \times (x - x')}{\delta x_i} d\alpha \quad (3)$$

where F is a neural network and x the input (Sundararajan et al., 2017), i.e. the tokenized source sentence in our case.

Gradient SHAP: Approximates Shapley values through gradients. Specifically, the entire dataset is used as the background by 'approximating the model with a linear function between each background data sample and the current input to be explained, and we assume the input features are independent'[4].

DeepLIFT: is a method that generates input feature attribution scores based on a given reference (Shrikumar et al., 2019). The result expresses the difference between the reference and the output.

$$r_i^{(L)} = \begin{cases} S_i(x) - S_i(\overline{x}) & \text{if unit i is the target} \\ & \text{unit of interest} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

with

$$r_i^{(l)} = \sum_j \frac{z_{ji} - \overline{z}_{ji}}{\sum_{i'} z_{ji} - \sum_{i'} \overline{z}_{ji}} r_j^{l+1} \quad (5)$$

and $\overline{z}_{ji} = w_{ji}^{l+1,l} \overline{x}_i^{(l)}$ (Ancona et al., 2017).

Attention: In this method, the values of the attention heads of the teacher models are used directly:

$$Attention = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

(Vaswani, 2017). As the teacher and learner models have the same architecture in our experiments, the attention values are aggregated from the layers and heads of the teacher model.

Value Zeroing: is calculated by determining the distance between a changed output representation $\tilde{x}_i^{\neg j}$. $\tilde{x}_i^{\neg j}$ is calculated by removing token j from the input (Mohebbi et al., 2023):

$$C_{i,j} = cosine(\tilde{x}_i^{\neg j}, \tilde{x}_i) \quad (7)$$

All the above-mentioned methods are adapted to the sequence-to-sequence tasks. Therefore, attributions are calculated for every input token with respect to all output tokens.

3 Methodology

Inspired by forward simulation of XAI methods (Hase and Bansal, 2020), we design a pipeline to compare different explainability attribution mappings based on their impact on model's performance in downstream tasks. To analyze the forward simulation of various XAI methods, we use a teacher-student model (Fig. 1). In the first step, we use the **Inseq** library to extract input-output attributions using the eight explainability algorithms specified in subsection 2.3. At this stage, the teacher model receives a source-target language pair (X, Y) as input, and the output of **Inseq** is a set of attributions $(X, Y) \rightarrow E$ mapping the output to the input tokens.

All of these attributions, except for the Attention-based ones, are in the shape $e \in \mathbb{R}^{j \times k \times l}$, where j is the input sequence length, k is the output sequence length, and l is the hidden dimension of the model. Gradient-based methods get the weight of the gradient for each individual input feature in the vector space. We aggregate these values along the last dimension by averaging them, which results in a final shape of $e \in \mathbb{R}^{j \times k}$. With a slight difference,

[3] https://captum.ai/docs/attribution_algorithms#input-x-gradient

[4] https://github.com/shap/shap

the Attention attributions are extracted in the shape $e \in \mathbb{R}^{j \times k \times n \times h}$, where j and k are the same as before, n represents the number of layers (here, only on the encoder side, $n = 6$), and h is the number of attention heads (8 in this case). We then compute the average along both of the last two axes to obtain a final shape of $e \in \mathbb{R}^{j \times k}$. To handle negative values and normalize the attribution matrices, we apply the MinMaxScaler[5] as follows:

$$\mathbf{e}'_{i,j} = \frac{\mathbf{e}_{i,j} - \min_i(\mathbf{e}_{:,j})}{\max_i(\mathbf{e}_{:,j}) - \min_i(\mathbf{e}_{:,j})} \quad (8)$$

Then, the input to the student model is the triple of $(X, Y, \mathbf{E}')$. In the next step, we apply four different operations on the attention $A = QK^T$ product:

Add: Simply add attributions to attention weights:

$$\tilde{\mathbf{A}}^{(h)} = \mathbf{A}^{(h)} + \mathbf{E}' \quad (9)$$

Multiply: Elementwise multiplication with the attention weights:

$$\tilde{\mathbf{A}}^{(h)} = \mathbf{A}^{(h)} \odot \mathbf{E}' \quad (10)$$

Average: Take the average of attributions and attention weights:

$$\tilde{\mathbf{A}}^{(h)} = \frac{\mathbf{A}^{(h)} + \mathbf{E}'}{2} \quad (11)$$

Replace: Substitute the standard attention mechanism with attribution-guided attention:

$$\tilde{\text{Attention}}(Q, K, V, \mathbf{E}') = f\left(\text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \mathbf{E}'\right) V \quad (12)$$

Where $\mathbf{E}' = \{e'_1, e'_2, \ldots, e'_b\}$ represents the explainability attributions based on the batch size used for the model, and Q, K, and V are the query, key, and value matrices of the transformer model. f is one of the operators mentioned above in 1-3. The last operation replaces $\mathbf{E}'$ to completely substitute $\frac{QK^\top}{\sqrt{d_k}}$.

Now, we train the student Opus-MT model (Tiedemann and Thottingal, 2020) from scratch with two settings: 1) We apply one of the above-mentioned operators to all layers of the encoder block. 2) We apply the attributions to the cross-attention mechanism between the encoder and decoder blocks. The source and target language pairs remain the same as those used for obtaining the attributions from the teacher model.

[5] In our limited experiments with different normalizing functions, MinMaxScaler normalization yielded better results.

4 Experimental Results

4.1 Evaluation Datasets and Metrics

To evaluate the proposed pipeline, we train the Opus-MT model (Tiedemann and Thottingal, 2020) on three datasets. We select two datasets from the same language family: German→English (de-en) and French→English (fr-en). For the third dataset, we choose Arabic→English (ar-en) due to its encoding and linguistic differences from the target language. For de-en and fr-en, we use the WMT14 dataset (Bojar et al., 2014), and for ar-en, we use the UN Parallel Corpus (Ziemski et al., 2016).

We select $200,000$ sample pairs from each dataset and preprocess them to suit our experimental setup. Given the large number of seq2seq models we train from scratch, we impose constraints to efficiently manage the training process. Specifically, we limit both input and output sequences to a maximum of 128 tokens. Additionally, we discard samples with fewer than three tokens and filter out pairs where the input-to-output length ratio (or vice versa) exceeds 1.7. For the de-en and fr-en datasets, we further exclude samples with an excessively high normalized Levenshtein distance. Since the validation and test sets of the WMT datasets are relatively small, we select an additional 15,000 samples from their training sets (without overlap with our training data). The UN Parallel Corpus does not include separate validation and test sets, so we extract $15,000$ samples from the main dataset for this purpose.

Throughout our experiments, we use the implementation of BLEU score (Papineni et al., 2002) to evaluate the student models.

4.2 Experimental Settings

We train the Opus-MT models[6] for 20 epochs and apply an early stopping of three consecutive epochs without improvement in validation loss. The model follows an encoder-decoder architecture, with each containing six layers with eight attention heads. The model employs the Swish activation function, as proposed by Ramachandran et al. (2017) (Ramachandran et al., 2017), which has been shown to

[6] https://huggingface.co/Helsinki-NLP

enhance training dynamics and convergence compared to traditional activation functions like ReLU. For training the models, we utilized 20 Nvidia V100 GPUs.

4.3 Results and Discussion

In our analysis, we evaluate the proposed methods through three key comparisons: a) we assess eight XAI methods for their effectiveness in improving translation quality when their attributions are injected into the model. b) We compare the impact of injecting attributions into the encoder self-attention versus the cross-attention layer to understand their influence on information flow and source-target alignment. c) We examine the effect of applying the attributions to half of the attention heads, exploring whether selective attribution merge shows a different behavior of attribution maps. As a baseline, we report the results of training the student model without attribution injection on the three datasets. Table 1 presents the BLEU scores for training the model from scratch for each language pair.

	de-en	fr-en	ar-en
Baseline	22.85	28.79	16.85

Table 1: baseline BLEU score results

Comparison of XAI Methods – This analysis evaluates eight different explainability methods in terms of their impact on translation quality. Table 2 shows the result of the injection of the attribution score to the 8 attention heads of the encoder vs. cross-attention part of the Opus-MT model trained from scratch. Across all three language pairs, Attention and Value Zeroing tend to have the highest values among the attribution methods. Results suggest that these two mechanisms capture a strong signal relevant to the translation process. Gradient-based methods (IxG, LGxA, IG) generally yield lower scores compared to the aforementioned methods. DeepLIFT attributions, except for the addition operation, decrease the result for de-en and fr-en. French-English (fr-en) consistently exhibits higher attribution scores than German-English (de-en), while Arabic-English (ar-en) shows the highest scores overall across most methods. Value-zeroing and Attention attribution help to double the BLEU score of this language pair. This finding may indicate that the morphological and syntactic complexity of the source language influences the attributions and hence affects the result of the student model.

German-English (de-en) Attribution scores are generally the lowest among the three language pairs. This is likely due to the high word reordering requirements in German, which may lead to weaker local alignment between input tokens and model outputs (Avramidis et al., 2019; Macketanz et al., 2021). French-English (fr-en) Attribution scores are higher than de-en, suggesting that French and English have more direct word alignment, which leads to stronger feature attributions. This aligns with linguistic expectations and empirical evidence (Legrand et al., 2016), as French and English share more lexical and syntactic similarities. Arabic-English (ar-en) This pair exhibits the highest result of providing attribution mappings, particularly for Attention (46.69–51.53) and ValueZeroing (46.29–51.64). It is possible that Arabic's rich morphology and non-concatenative structure likely cause the model to rely more heavily on attention mechanisms, explaining the higher increase of the result after being exposed to the attribution maps across the board.

Overall, Attention and Value Zeroing tend to contribute to the highest scores among attribution methods across all the three language pairs. The results suggest these two mechanisms capture a strong signal relevant to the translation process. Gradient-based methods (IxG, LGxA, IG) generally yield lower scores.

Encoder Self-Attention vs. Cross-Attention – This analysis examines the impact of injecting attribution scores into encoder self-attention layers versus cross-attention layers. In contrast to the encoder self-attention, cross-attention bridges the source and target languages by guiding the decoder's focus on the encoder's output. This mechanism is more sensitive because it manages the alignment between the source input and the target output. Any modification here can directly influence how the source information is integrated into the target generation process. For this reason, the initial hypothesis was that injecting attributes—which describe the relation between the source and target—into the cross-attention layer might enhance the flow of relevant information. However, the experimental results tell a different story.

In most cases, injecting these attributes into cross-attention either blocks or corrupts the flow of information. For example, when we replace the

de-en Encoder	IxG	Saliency	LGxA	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
add	**23.10**	**27.36**	**23.10**	**27.68**	**23.17**	**23.26**	**31.58**	**33.12**
multiply	21.59	**27.98**	21.85	**27.75**	21.65	21.73	**35.08**	**35.01**
average	**23.18**	**26.65**	**23.18**	**26.51**	**22.90**	**22.99**	**30.47**	**32.27**
replace	21.78	**26.84**	21.75	**26.31**	21.68	21.68	**31.57**	**33.39**
de-en CrossAttention	**IxG**	**Saliency**	**LGxA**	**IG**	**GSHAP**	**DeepLIFT**	**Attention**	**ValueZeroing**
add	22.50	16.82	22.49	19.41	22.83	22.54	14.21	11.99
multiply	7.40	7.57	4.69	8.18	10.32	8.76	3.14	2.27
average	20.06	19.42	20.01	19.72	20.63	**22.87**	14.89	14.96
replace	0.25	0.08	0.21	0.04	0.25	0.38	4.69	0.25

fr-en Encoder	IxG	Saliency	LGxA	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
add	**29.04**	**36.99**	**29.04**	**35.52**	**30.14**	**29.04**	**44.16**	**46.97**
multiply	**29.29**	**38.54**	**29.30**	**35.94**	**29.66**	**28.88**	**49.31**	**49.14**
average	28.68	**36.31**	28.68	**34.16**	**29.55**	**28.84**	**42.62**	**45.43**
replace	28.15	**36.15**	28.15	**34.42**	**29.26**	28.26	**42.77**	**45.35**
fr-en CrossAttention	**IxG**	**Saliency**	**LGxA**	**IG**	**GSHAP**	**DeepLIFT**	**Attention**	**ValueZeroing**
add	24.31	26.50	24.32	23.78	26.53	24.59	24.50	22.25
multiply	14.69	3.66	14.68	6.29	7.49	15.86	5.82	1.62
average	22.12	23.50	28.76	28.76	24.40	20.55	25.75	26.12
replace	0.77	0.06	0.77	0.01	0.70	1.60	5.89	1.63

ar-en Encoder	IxG	Saliency	LGxA	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
add	**32.60**	**38.72**	**32.60**	**30.75**	**33.91**	**33.59**	**46.46**	**46.29**
multiply	**37.06**	**43.74**	**36.78**	**40.28**	**37.59**	**37.03**	**51.53**	**51.64**
average	**27.70**	**34.14**	**27.70**	**30.55**	**28.75**	**26.56**	**46.69**	**41.17**
replace	**36.76**	**43.4**	**36.69**	**40.17**	**37.75**	**36.81**	**49.77**	**51.48**
de-en CrossAttention	**IxG**	**Saliency**	**LGxA**	**IG**	**GSHAP**	**DeepLIFT**	**Attention**	**ValueZeroing**
add	**33.87**	**31.31**	**33.87**	**29.24**	**34.94**	**33.61**	**26.28**	**29.61**
multiply	**17.40**	5.81	**17.41**	12.52	**22.63**	**17.41**	6.17	1.72
average	**18.47**	10.32	**18.47**	12.94	14.38	**18.10**	11.05	13.08
replace	1.81	0.25	1.85	0.01	0.97	1.78	9.56	5.48

Table 2: BLEU score comparison of various attribution methods across different composition strategies (add, multiply, average, replace) to 8 heads applied to encoder and cross-attention modules in neural machine translation models. Results are reported for three language pairs—German–English (de–en), French–English (fr–en), and Arabic–English (ar–en)—with columns corresponding to attribution techniques (IxG, Saliency, LGxA, IG, GSHAP, DeepLIFT, Attention, and ValueZeroing). Scores that beat the baseline model for each setting are boldfaced and the highest BLEU score for each dataset are highlighted in green.

cross-attention weights entirely with the attribute values (using the 'replace' operator), the performance degrades drastically to the point where the model essentially fails to learn anything. This is likely due to reduced focus on the source input within the cross-attention mechanism.

For the operators addition (+) and average, scores better than multiplication ($\odot$) in the cross-attention context. Adding the attributions seems to augment the existing attention values in a beneficial way, whereas multiplying them often leads to an overly aggressive modification that harms the model's ability to propagate information from the encoder. The addition might act as a mild corrective signal that helps the decoder focus better, while multiplication can excessively amplify or diminish the weights, leading to a loss of critical alignment information.

Effect of Attention Head Reduction (8 Heads vs. 4 Heads) – In another setting, we applied the attribution methods to only 4 heads out of the 8 heads of the encoder attention. Figure 4 shows the result of this comparison. This analysis investigates how reducing the number of attention heads affects the performance of the model when integrating attribution scores. By selectively applying attributions to only four heads (every other head), we assess whether information flow can still be captured and

whether the model retains its translation quality. The changes in BLEU scores between 8-head and 4-head settings are relatively minor. Some methods and operators show slight improvements. The results suggest that a mix of normal attention mechanisms and attribution operators can help the model to learn the mapping better and show the robustness of the learned attribution mappings.

4.4 Attrbutions methods difference

While a linguistic and qualitative analysis of the differences between each attribution method is out of the scope of this work, we are interested in quantifying the differences between these attribution methods. The entropy of the attribution matrices can tell us the disparity of the mapping scores. We randomly selected 2000 samples for each method for this matter. We compared the entropy scores produced by all the methods using the Wilcoxon signed-rank test. Only for methods IxG and LGxA did we not find significantly different entropy scores between the two methods, $p > 0.5$. Moreover, from the visualization of the Figure 5 we see that the higher scoring attributes (i.e., Saliency, Attention, and Value Zeroing) have a lower average entropy.

5 Conclusion

In this work, we investigated integrating XAI-driven attributions into sequence-to-sequence NMT models to assess their impact on enhancing target sequence generation, using this improvement as a proxy to evaluate the quality of the learned mappings. Our extensive analysis across German–English, French–English, and Arabic–English language pairs included comparing eight XAI methods and various composition strategies (i.e., addition, multiplication, averaging, and replacement) for injecting attribution scores into the Transformer's attention mechanisms.

The effectiveness of attribution methods can be summarized as follows: Attention-based and Value Zeroing attribution techniques consistently yielded the greatest improvements in BLEU scores. In contrast, gradient-based methods (e.g., IxG, LGxA, DeepLift) resulted in lower performance gains and, in some cases, decreased the results. Statistical analysis of these attribution maps revealed that higher-scoring attributes tend to have lower entropy, indicating they convey more structured and organized information.

Injecting attribution scores into encoder self-attention layers generally reinforced intra-pairs relationships and improved translation quality. Conversely, modifications in the cross-attention layers do not benefit from the extra knowledge provided to this part of the Encoder-Decoder Transformers architecture. Finally, reducing the number of attention composition heads from eight to four demonstrated that selective merging can refine the attention mechanism, and, in some cases, further enhance the performance.

Limitations

There are some limitations to this work worth noting. First, we compared attribution information across explainability methods, the majority of which were gradient-based. This choice was primarily due to the computational cost associated with extracting attributions using other methods, particularly perturbation-based approaches such as LIME (Ribeiro et al., 2016b) and reAGent (Zhao and Shan, 2024), which are resource-intensive to obtain. Furthermore, some of these methods, provided by Inseq, generate (self-)attributions for the decoder side of the seq2seq models. However, at this stage, we limited our experiments to encoder self-attention and cross-attention between the decoder and encoder.

In this work, we restricted our experiments to a single evaluation metric—the BLEU score. Incorporating additional metrics that capture semantic similarity, such as METEOR and BERTScore, along with human evaluations assessing fluency, coherence, and relevance, could provide deeper insights and a more comprehensive evaluation of model performance. Future research should explore these alternative metrics to achieve a more nuanced assessment of generated sequences by the help of attributions.

Additionally, our focus was limited to machine translation tasks. Future work could extend this evaluation framework to other sequence-to-sequence models, including those applied in question answering and text summarization.

References

David Alvarez-Melis and Tommi S Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv preprint arXiv:1707.01943*.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2017. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*.

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. 2018. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Judith E Dayhoff and James M DeLeo. 2001. Artificial neural networks: opening the black box. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 91(S8):1615–1635.

Jay DeYoung, Sarthak Jain, Nazneen Rajani, Eric P. Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. In *Annual Meeting of the Association for Computational Linguistics*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, and Feras A Batarseh. 2023. Rationalization for explainable nlp: a survey. *Frontiers in Artificial Intelligence*, 6:1225093.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Katikapalli Subramanyam Kalyan. 2024. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048.

Jenia Kim, Henry Maathuis, and Danielle Sent. 2024. Human-centered evaluation of explainable ai applications: a systematic review. *Frontiers in Artificial Intelligence*, 7:1456486.

Kushal Lakhotia, Bhargavi Paranjape, Asish Ghoshal, Scott Yih, Yashar Mehdad, and Srini Iyer. 2021. FiD-ex: Improving sequence-to-sequence models for extractive rationale generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3712–3727, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joël Legrand, Michael Auli, and Ronan Collobert. 2016. Neural network-based word alignment through score aggregation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 66–73, Berlin, Germany. Association for Computational Linguistics.

C Leiter, P Lertvittayakumjorn, M Fomicheva, W Zhao, Y Gao, and S Eger. Towards explainable evaluation metrics for natural language generation (2022). *Preprint at https://arxiv. org/abs/2203.11131*.

Jierui Li, Lemao Liu, Huayang Li, Guanlin Li, Guoping Huang, and Shuming Shi. 2020. Evaluating explanation methods for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 365–375, Online. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for german to english and english to german. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2022a. Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1731–1751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022b. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802, Online. Association for Computational Linguistics.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1069–1078. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016a. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016b. "why should i trust you?": Explaining the predictions of any classifier.

Waddah Saeed and Christian Omlin. 2023. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Hassan Shakil, Ahmad Farooq, and Jugal Kalita. 2024. Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, page 128255.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2019. Learning important features through propagating activation differences.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

I Sutskever. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Isidora Chara Tourni and Derry Wijaya. 2023. Relevance-guided neural machine translation.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

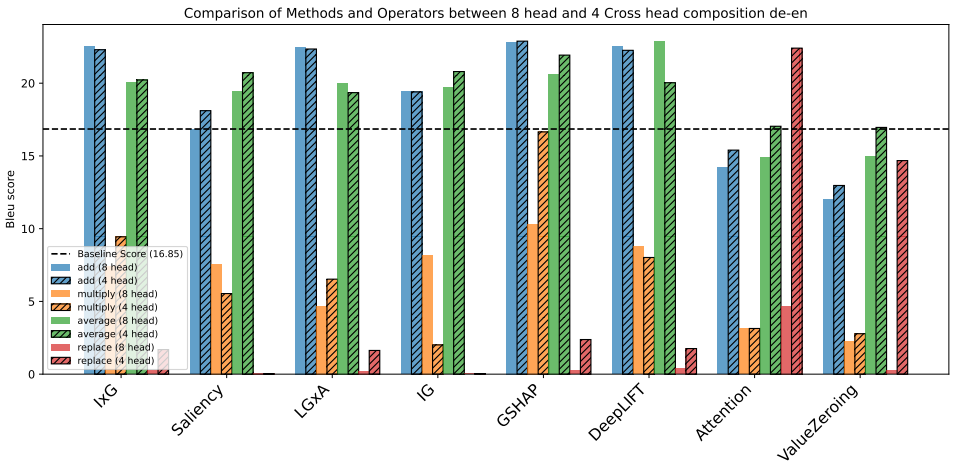
Carla Piazzon Vieira and Luciano Antonio Digiampietri. 2022. Machine learning post-hoc interpretability: A systematic mapping study. In *Proceedings of the XVIII Brazilian Symposium on Information Systems*, pages 1–8.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

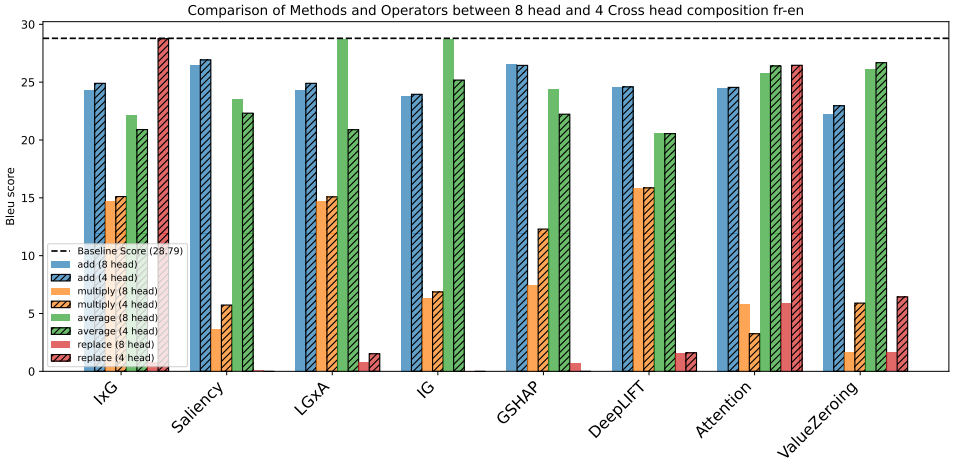
Zhixue Zhao and Boxuan Shan. 2024. Reagent: A model-agnostic feature attribution method for generative language models.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

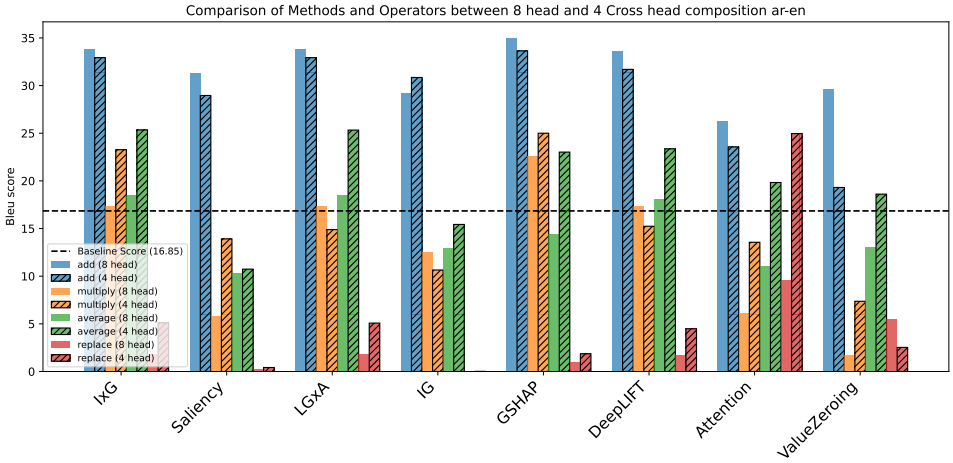
6 Appendix



(a) de-en

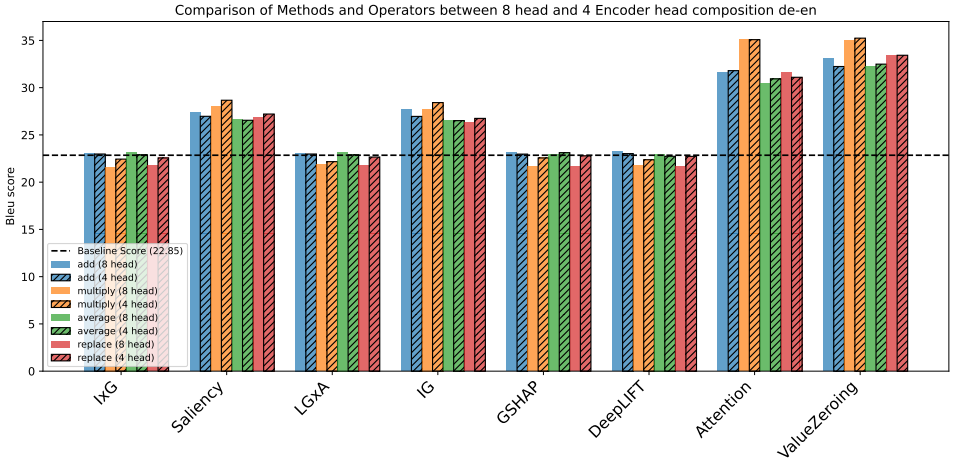


(b) fr-en

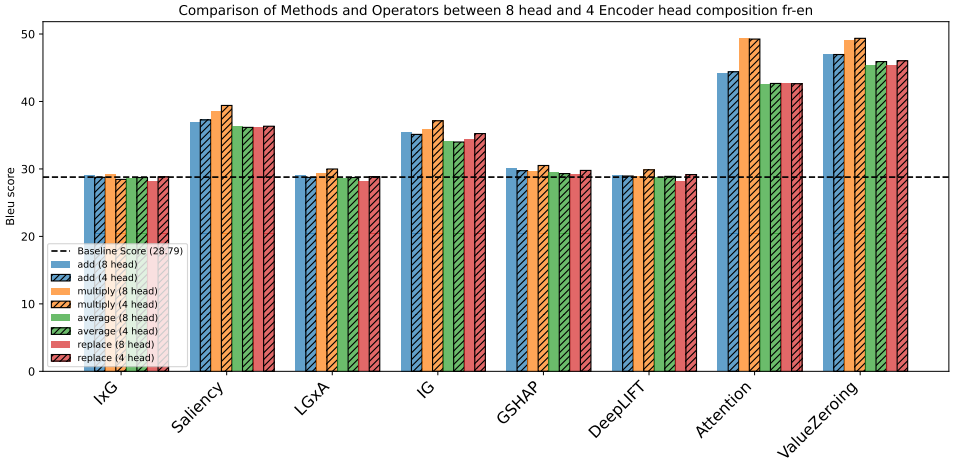


(c) ar-en

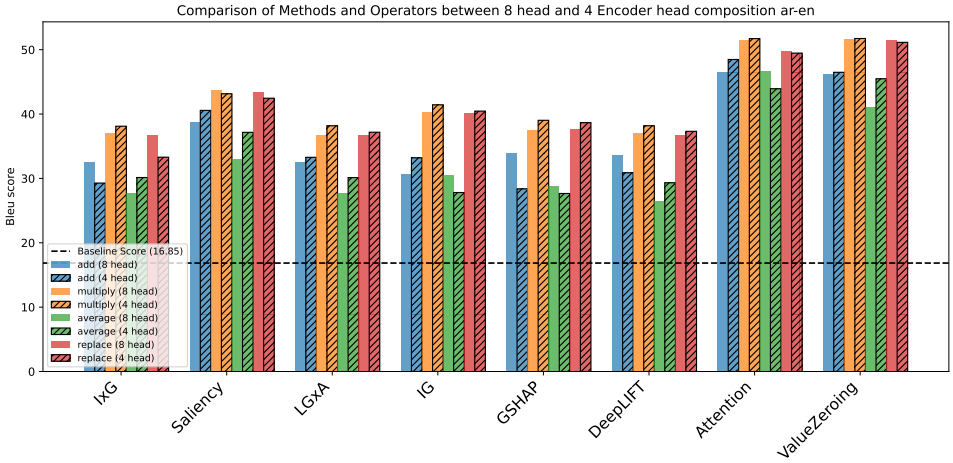
Figure 3: Result of attribution composition of cross-attention weights on the de-en (a), fr-en(b) and ar-en(c) datasets: comparing 4-Head (striped bars) and 8-Head (plain bars) configurations.



(a) de-en

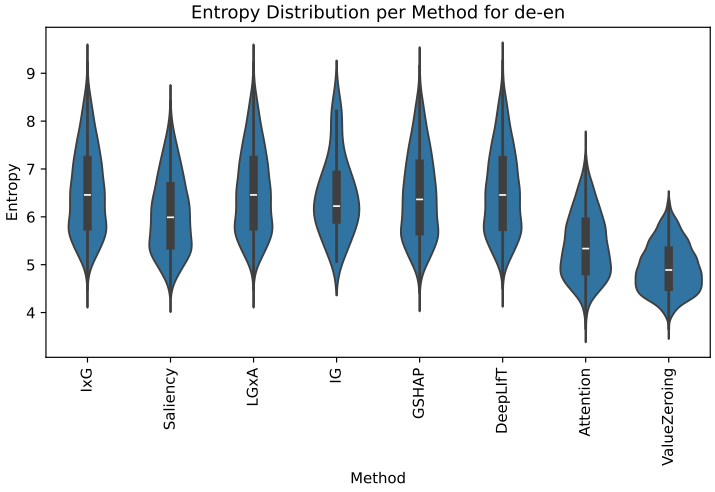


(b) fr-en

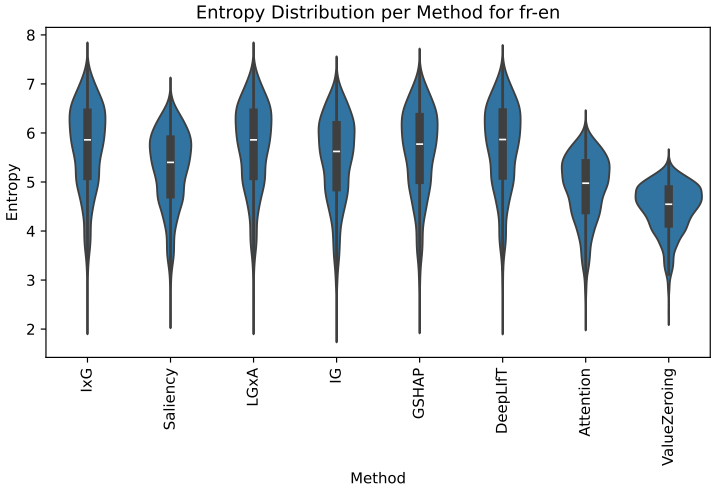


(c) ar-en

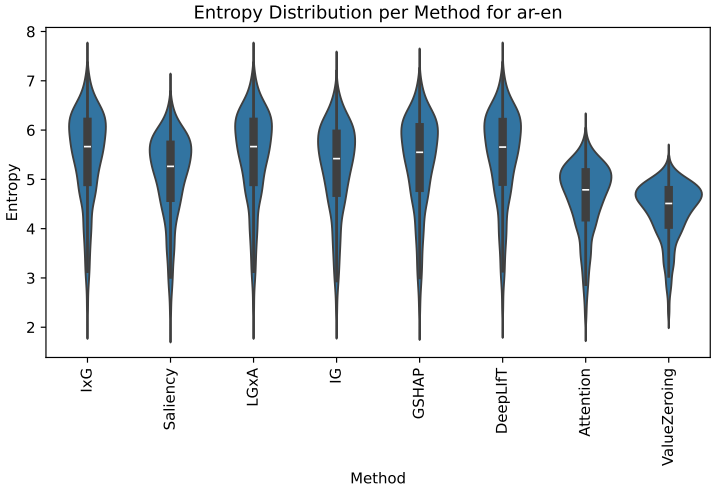
Figure 4: Result of Attribution composition of encoder self-attention weights on the de-en (a), fr-en(b) and ar-en(c) datasets: comparing 4-Head (striped sars) and 8-Head (plain bars) configurations.



(a) de-en



(b) fr-en



(c) ar-en

Figure 5: Violin plots showing the distribution of entropy values for attribution maps generated by different XAI methods. Each violin represents the spread of entropy across all samples for a given method. Lower entropy values indicate more focused attribution maps.