Natural Language Processing (2026), **32**, pp. 267–301
doi:10.1017/nlp.2026.10028

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Evaluating explainable AI attribution methods in neural machine translation via attention-guided knowledge distillation

Aria Nourbakhsh, Salima Lamsiyah, Adelaide Danilov and Christoph Schommer

University of Luxembourg, Luxembourg
Corresponding author: Aria Nourbakhsh; Email: aria.nourbakhsh@uni.lu

(Received 23 July 2025; revised 27 February 2026; accepted 19 April 2026)

Abstract

The study of the attribution of input features to the output of neural network models is an active area of research. While numerous Explainable AI (XAI) techniques have been proposed to interpret these models, the systematic and automated evaluation of these methods in sequence-to-sequence (seq2seq) models is less explored. This paper introduces a new approach for evaluating explainability methods in transformer-based seq2seq models, building upon the forward simulation of XAI methods. We use teacher-derived attribution maps as a structured side signal to guide a student model, and quantify the utility of different attribution methods through the student's ability to simulate targets. Using the Inseq library, we extract attribution scores over source–target sequence pairs and inject these scores into the attention mechanism of a student transformer model under four composition operators (addition, multiplication, averaging, and replacement). Across three language pairs (de–en, fr–en, ar–en) and attributions from Marian-MT and mBART models, Attention, Value Zeroing, and Layer Gradient × Activation consistently yield the largest gains in BLEU (and corresponding improvements in chrF) relative to baselines. In contrast, other gradient-based methods (Saliency, Integrated Gradients, DeepLIFT, Input × Gradient, GradientShap) lead to smaller and less consistent improvements. These results suggest that different attribution methods capture distinct signals and that attention-derived attributions better capture the alignment between source and target representations in seq2seq models. Finally, we introduce an Attributor transformer that, given a source–target pair, learns to reconstruct the teacher's attribution map. Our findings demonstrate that the more accurately the Attributor can reproduce attribution maps, the more useful an injection of those maps is for the downstream task. The source code can be found on GitHub.[a]

Keywords: Explainability; sequence-to-sequence model; attribution maps; transformer; knowledge distillation

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, especially transformer-based sequence-to-sequence (seq2seq) architectures (Sutskever, Vinyals, and Le 2014; Vaswani 2017), excel at capturing complex relationships between input and output sequences. However, they 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 this challenge, the field of Explainable AI (XAI) has attracted researchers seeking to improve transparency and interpretability and to explain models'

[a]https://github.com/ariana2011/seq2seq_xai_attributions/

behavior. One of the main objectives of XAI is to assess, quantify, or characterize the importance (or attribution) of input features in shaping the final outputs of these models (Arya *et al.* 2019; Vieira and Digiampietri 2022; Saeed and Omlin, 2023). Several XAI methods have been developed and applied specifically to NLP models to evaluate the contribution of input information to the model's output across various tasks (Madsen, Reddy, and Chandar 2022; Mohammadi *et al.* 2025). Despite these advances, identifying which explanation methods more accurately reflect model reasoning remains challenging, especially in seq2seq settings, which are characterized by intricate encoding-decoding dynamics and many-to-many mappings (Li *et al.* 2020; Gurrapu *et al.* 2023).

The calculation and extraction of the decision-making process of neural networks per input features, which are also applied to Neural Machine Translation (NMT) models (X. Li *et al.* 2019; He *et al.* 2019; Eksi *et al.* 2021; Fomicheva, Specia, and Aletras 2022; Perrella *et al.* 2024), are often grouped into three broad families of gradient-based, model-based, and perturbation-based methods (Dwivedi *et al.* 2023; Sarti *et al.* 2023; Fantozzi and Naldi 2024). Gradient-based approaches estimate the contribution of each input by computing derivatives of the model output with respect to the input or intermediate representations; examples include Saliency (Simonyan, Vedaldi, and Zisserman 2014) and Integrated Gradients (Sundararajan, Taly, and Yan 2017). Model-based approaches rely on components that already produce interpretable signals, such as the attention mechanism. Perturbation-based methods, instead, modify or remove parts of the input and measure the resulting change in model output; LIME (Ribeiro, Singh, and Guestrin 2016) and Value Zeroing (Mohebbi *et al.* 2023) fall into this category. However, the boundaries between these families are not always strict. Some techniques combine properties of multiple categories, such as GradientSHAP (Lundberg and Lee 2017), which blends gradient information with stochastic perturbations. Nevertheless, because these methods rest on different assumptions about how models encode and use information, they compute feature importance in different ways and can therefore produce divergent explanations for the same prediction, raising the question of which methods best reflect the model's behavior.

Despite the proliferation of explainability methods, comprehensive and scalable evaluation of XAI methods in NLP remains limited. Evaluation practices that rely on human validation, which, although insightful, are costly and difficult to scale (Leiter *et al.* 2022; Kim, Maathuis, and Sent 2024). Automated evaluation frameworks that are common in computer vision (Ribeiro *et al.* 2016; Chang *et al.* 2019; Hooker *et al.* 2019) are underrepresented in NLP and NMT, and existing work typically focuses on a small number of explanation methods. This signifies the need for systematic, model-based evaluation approaches capable of objectively comparing diverse explainability techniques in seq2seq settings. Prior evaluations have, for example, compared attribution maps with human-annotated word alignments (see Section 2.1). Yet, such alignments only approximate the underlying translation dynamics and may not represent the information flow within modern NMT systems.

In this work, we address these gaps by proposing an automated evaluation framework based on the simulatability of XAI methods (Doshi-Velez and Kim 2017; Hase and Bansal 2020), specifically designed to assess and compare multiple attribution methods within seq2seq models for NMT. Intuitively, if an attribution method captures a model's input–output dependencies, it should provide useful guidance for a student model to make better predictions. We operationalize this idea through a teacher–student setup: attribution maps are extracted from a pre-trained teacher NMT model and injected into the attention mechanism of a smaller, untrained student model. Concretely, we treat the attribution maps as attention priors within the encoder-decoder architecture and explore several ways to combine them with the student's own attention scores. The resulting student performance provides an automated, task-specific measure for evaluating different explanation methods. Within this framework, higher-quality explanations produce more informative attribution maps, which in turn allow the student to make more accurate predictions, thereby serving as a proxy for judging the effectiveness of the XAI attribution method. We apply this framework across three language pairs, using Marian-MT (Tiedemann and Thottingal 2020) and mBART (Liu *et al.* 2020) models' attributions.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

As part of our analysis, we deliberately compute attribution maps with respect to the gold reference translation, and in another part, we compute the attributions based on the teacher's generation. This defines an oracle setting in which the explanation is allowed to depend on the true target sequence and the resulting translation. Therefore, the model sees the source and attribution maps for all target tokens during encoding and at each autoregressive step. We use this oracle setup to address two questions. First, to what extent can attribution-guided attention priors help a student model reproduce the gold, human-generated target when the student model has access to the attributions? Second, given a fixed teacher-generated translation, which attribution methods produce maps that are most helpful for a student model to approximate that teacher, that is which explanations best capture the teacher's input–output behavior under our idealized conditions? In this way, the oracle setting serves as a controlled environment for comparing attribution methods, and we interpret the reported BLEU/chrF scores as relative indicators of explanation quality rather than as standard test-set performance.

To interpret the discrepancy in student performance under different attribution methods, we propose a hypothesis that their behavior can be explained by the closeness of their mappings to what a transformer is able to produce. To investigate that, we design a separate encoder-decoder transformer, named *Attributor*, and train it to reconstruct the teacher's attribution map based on the respective source-target pair. Our experiments confirm the claim and highlight that the Attributor's ability to reproduce scores of top-3 salient tokens per column of the attribution maps very strongly correlates with the student performance in the MT task utilizing those maps.

Beyond the primary goal of evaluating XAI attribution methods in NMT, this work also provides insight into the behavior of the attention mechanism itself. We show that attributions derived from the teacher model's attention tend to be more effective in guiding the student model. This is aligned with the fact that attention maps are the easiest for the Attributor to reproduce. We also observe an interesting and somewhat unintuitive pattern in how the student model responds to externally injected attribution signals. In particular, the intervention is more effective when applied to the encoder attention (see Section 3). A detailed discussion of these findings appears later in the paper.

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

- We propose an evaluation framework that uses knowledge distillation to systematically assess and compare explainability methods by integrating attribution explanations into seq2seq model architectures.
- We conduct extensive experiments exploring multiple strategies for incorporating explanations within the Transformer architecture and systematically compare their effects on model performance across various language pairs.
- We provide empirical evidence that XAI attribution methods 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.
- Finally, we investigate reasons why each attribution mapping yields different results when used within the student model for NMT tasks and show a strong correlation of those results with the ability of a transformer to approximate top-3 salient scores per target token of such mappings.

The paper is organized as follows: Section 2 reviews the background and related work. Section 3 describes the proposed approach. Sections 4 and 5 present and discuss the obtained results. In Section 6 we present the Attributor network. Section 7 concludes the paper and outlines the limitations of this work, highlighting directions for future research.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

2. Related work

In this section, we first summarize related work on evaluating and analyzing XAI attribution methods, with a particular focus on NMT. We then briefly review the seq2seq NMT architecture and outline the XAI methods used in this study.

2.1 Evaluation of explanations and their application in NMT

The XAI literature distinguishes several dimensions of what explanations can provide. A central distinction is between plausibility and faithfulness/fidelity: plausibility refers to how well an explanation aligns with human intuition, whereas faithfulness describes how accurately an explanation reflects the model's actual decision-making process (Arrieta *et al.* 2020; Jacovi and Goldberg 2020). Doshi-Velez and Kim (Doshi-Velez and Kim 2017, 2018) have proposed three approaches to evaluate XAI methods, one of which is a functionally grounded evaluation protocol. In this protocol, explanations are assessed by automatic task performance metrics rather than human judgments. This paradigm has been adopted in NLP for evaluating the faithfulness of saliency[b] methods. For example, Arras *et al.* (2016); Nguyen (2018); DeYoung *et al.* (2020); Atanasova *et al.* (2020); Nauta *et al.* (2023) proposed automatic, task-based metrics for classification models that quantify how well feature attributions capture model behavior.

Our focus is on the NMT task, in particular, transformer-based (Vaswani 2017) seq2seq models (Sutskever *et al.* 2014). Compared to standard classification, NMT introduces additional challenges for explanation because it decomposes prediction into a sequence of conditional next-token decisions, one per decoding timestep, and the importance of each source token depends not only on the current target token but also on the previous target prefix (Stahlberg 2020; Shakil, Farooq, and Kalita 2024). This temporal and source–target coupling complicates both the design and the evaluation of token-level attribution methods.

A large body of work has studied the attention mechanism as an interpretable component of NMT, using encoder–decoder attention weights to estimate word importance and to approximate word alignments (Ghader and Monz 2017; Raganato and Tiedemann 2018; Kobayashi *et al.* 2020; Ferrando and Costa-jussà 2021). These studies show that attention patterns correlate with, but do not faithfully reproduce, traditional word alignments, and that attention weights also reflect how the model balances source information against the evolving target prefix. On the other hand, other work has questioned the plausibility and faithfulness of attention as an explanation, arguing that attention weights should be interpreted cautiously based on the task and model, and in some cases, augmented with more explicit attribution mechanisms (Jain and Wallace 2019; Meister *et al.* 2021; Madsen *et al.* 2022).

Beyond attention, other explanation methods exist to compute per-token importance. The most common metrics to evaluate token importance are comprehensiveness (does removing the highlighted tokens reduce the model's confidence or translation quality?) and sufficiency (are the highlighted tokens alone sufficient to preserve model performance?) (DeYoung *et al.* 2020; Nauta *et al.* 2023). In NMT, such approaches have been explored to measure the drop in log-probability of the chosen next token after input perturbations or at the sequence level, where changes in BLEU scores after manipulating important source tokens, according to XAI methods, were used as a test bed for assessing whether attribution maps identify tokens that are necessary and/or sufficient for the model's predictions (He *et al.* 2019; Moradi, Kambhatla, and Sarkar 2021).

Word alignment (Brown *et al.* 1993) has also been widely used as a proxy for the plausibility of model attributions in NMT. X. Li *et al.* (2019) showed that attention-based alignments have clear limitations, motivating alternative alignment models and prediction-difference techniques to improve alignment quality. In parallel, Zenkel, Wuebker, and De Nero (2019) proposed

[b] In the broader literature, feature-importance methods are sometimes referred to collectively as "saliency methods." In this paper, however, later, we use Saliency to denote a specific gradient-based attribution method.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

augmenting NMT models with dedicated alignment layers, treating alignment as an auxiliary prediction task rather than a by-product of standard attention. Their approach improves alignment accuracy compared to classical tools such as GIZA++ (Och and Ney 2003) and FastAlign (Dyer, Chahuneau, and Smith 2013). Ding, Xu, and Koehn (2019) introduced saliency-driven, gradient-based methods that yield more interpretable alignment signals without modifying the underlying NMT architecture. Building on these ideas, Ferrando and Costa-jussà (2021) analyzed encoder–decoder attention in detail, highlighting systematic alignment errors and proposing techniques that explicitly quantify the relative contributions of source and target contexts. Ferrando *et al.* (2022) further developed ALTI+, an attention-rollout-based framework that traces contributions from both source and target contexts across layers in multilingual Transformer models. ALTI + has been used as an internal explanation metric for downstream diagnostic tasks; for instance, Dale *et al.* (2023) employ ALTI+-based scores to detect and mitigate hallucinations in NMT outputs. Related work by Voita, Sennrich, and Titov (2021) used layer-wise propagation (LPR) to analyze the intrinsic contributions of source and target contexts under different training regimes and dataset conditions, while Kobayashi *et al.* (2020) performed a norm-based analysis of attention and transformed representations to study internal alignment mechanisms within Transformers. Closer to the current work, Li *et al.* (2020) evaluated XAI methods in NMT by training surrogate models on the important words identified by the XAI methods and measuring the prediction success of each token *i* based on the top-k tokens identified as the most contributing tokens to the generation of that token. Other work has explored and compared XAI methods, such as gradient and perturbation methods, to detect word-level translation errors in NMT (Eksi *et al.* 2021; Fomicheva *et al.* 2022).

2.2 Simulatability of XAI methods

A complementary line of work evaluates explanations through simulatability. Simulatability is how well an explanation helps a user replicate a model's behavior. Doshi-Velez and Kim (2017); Hase *et al.* (2020) propose human-grounded protocols in which participants are asked to predict a model's output on a given input, first without and then with access to explanations. The underlying idea is that an explanation method is better if it improves human prediction accuracy of the model's decisions. These studies implement such user-in-the-loop evaluations on tabular and text classification tasks, using common XAI techniques to generate explanations that are shown to human subjects.

Closer to our setting, Pruthi *et al.* (2022) worked on a related idea without relying on human annotators. Building on the notion of simulatability, they transfer important information learned from one model to another. Token-level importance scores are used to guide a new model on several classification tasks, and the resulting change in performance is taken as evidence for the usefulness of the original attributions. In other words, an explanation is considered higher quality if it can be leveraged to train another model that better simulates or reconstructs the original model's behavior. Our work adopts a similar spirit of model-based simulatability, but in the more complex seq2seq NMT setting.

2.3 Injection of knowledge into the attention mechanism

There is also some work on injecting external or structured linguistic knowledge directly into the attention mechanism (Jiao *et al.* 2020; Bai *et al.* 2022; Zhao and Shan 2024). In NMT Bugliarello and Okazaki (2020) incorporate syntactic information by encoding, for each token, the distance to the syntactic 'parent' in a dependency tree, and using this signal to bias attention patterns. In their approach, knowledge injection is applied on the encoder side for the source sentence, enabling attention heads to exploit syntactic structure. Slobodkin, Choshen, and Abend (2022) augment

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

encoder attention with syntactic and semantic information in the form of alignment-like constraints over the input. Their architecture modifies the attention computation so that heads are explicitly informed by these external signals, rather than learning them implicitly. They report that enriching attention with semantic information benefits translation quality. Similar to the current work Nourbakhsh, Lamsiyah, and Schommer (2025) inject attributions in the form of hard alignments and compare them to linguistic information, such as POS and Dependency information. In this work, we deal with soft attributions and more diverse XAI methods. These works suggest that attention can serve as a locus for the injection of task-relevant prior knowledge.

Bringing together these strands of work, we propose an evaluation framework for XAI attribution methods to inject attribution matrix scores into the attention mechanism and compare the resulting effects on the translation task itself. Our design is inspired by Remove and Retrain (ROAR) retraining paradigms (Hooker *et al.* 2019), the simulatability of XAI methods (Hase *et al.* 2020), and prior work on knowledge injection into attention, and it contributes to this line of research by providing a comprehensive, functionally grounded comparison of several attribution methods in the seq2seq NMT model.

2.4 NMT with sequence-to-sequence models

Seq2seq models (Sutskever *et al.* 2014; Bahdanau, Cho, and Bengio 2015), originally introduced for machine translation (Cho *et al.* 2014), are conditional language models that learn to generate the target sentence token by token, conditioned on the source sentence and the previously generated target tokens. The original Transformer model, a precursor to encoder-based classifiers and decoder-based large language models, was an encoder–decoder (Vaswani 2017):

An NMT encoder–decoder model operates on two sequences:

$$\mathbf{x} = (x_1, x_2, \ldots, x_{T_x}), \quad \mathbf{y} = (y_1, y_2, \ldots, y_{T_y}).$$

Where $\mathbf{x}$ and $\mathbf{y}$ are the source and target sequences, an NMT model with a seq2seq architecture defines a conditional probability distribution over the target sequence given the source sequence and previous target tokens:

$$p(\mathbf{y} \mid \mathbf{x}; \theta) = \prod_{t=1}^{T_y} p(y_t \mid y_{<t}, \mathbf{x}; \theta),$$

where $y_{<t} = (y_1, \ldots, y_{t-1})$ and θ are all model parameters.

The encoder maps the source sequence to a sequence of continuous contextual representations based on stacked self-attention and feed-forward layers.

$$\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_{T_x}),$$

The decoder is another neural network that, at each time step t, takes as input the previously generated target tokens (through masked self-attention over $y_{<t}$) and attends to the encoder representations $\mathbf{H}$ (via cross-attention), producing a decoder representation $\mathbf{s}_t$ from which the next-token distribution $p(y_t \mid y_{<t}, \mathbf{x}; \theta)$ is computed.

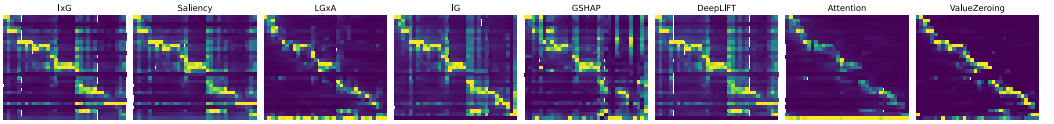
Given a training corpus

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(n)}, \mathbf{y}^{(n)} \right) \right\}_{n=1}^{N}$$

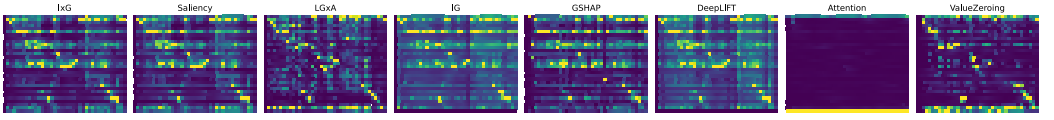
of source–target pairs, model parameters θ are typically learned by maximizing the conditional log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N} \sum_{t=1}^{T_y^{(n)}} \log p\left(y_t^{(n)} \mid y_{<t}^{(n)}, \mathbf{x}^{(n)}; \theta \right).$$

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press



(a) Marian-MT model.



(b) mBART model.

Figure 1. An example of attribution maps derived from different XAI methods. For the source sentence *'Dann gibt es noch Anbieter, die kaum Fahrraderfahrung, jedoch gute Fernostkontakte haben und so an günstige E-Bikes kommen.'* and the target *'Then there are suppliers with little or no experience in the bicycle industry but good contacts in the Far East, thus giving them access to low-cost e-bikes.'*. In the heatmaps, the rows correspond to source tokens, and the columns to target tokens. The heatmaps are generated from the normalized columns using the MinMax normalizer.

During training, teacher forcing is commonly used, meaning that the decoder receives the ground-truth previous token $y_{t-1}^{(n)}$ as input when predicting $y_t^{(n)}$.

2.5 AI explainability methods

XAI attribution methods can be broadly categorized into three main types: gradient-based, internal model-based, and perturbation-based approaches. Below, we briefly describe the attribution methods used in this work to extract attribution maps. To extract these attribution maps, we used the Inseq Python library (Sarti *et al*. 2023).

Saliency: The saliency method was originally introduced for image classification tasks using convolutional neural networks (CNNs) and is one of the earliest gradient-based explanation techniques (Simonyan *et al*. 2014). It treats a trained neural network as a locally linear function of its input around a given example. The primary objective is to identify which pixels in an input image are most influential for a particular class prediction. The 'saliency map' is defined as the gradient of the class score with respect to each input dimension, often visualized as the element-wise magnitude of this gradient. Intuitively, if a small change in a particular input dimension (e.g., a pixel intensity) leads to a large change in the class score, that dimension is considered important for the model's decision for class c.

Let $\mathbf{x} \in \mathbb{R}^d$ be an input and $S_c(\mathbf{x})$ the score for class c. The saliency map for class c at $\mathbf{x}_0$ is defined as:

$$M_c^{(i)}(\mathbf{x}_0) = \frac{\partial S_c(\mathbf{x}_0)}{\partial x_i}, \quad i = 1, \ldots, d. \tag{1}$$

In NLP, $\mathbf{x}$ corresponds to token embeddings; we aggregate per-dimension gradients to obtain a scalar attribution per token (see Section 3).

Input × Gradient (I×G): It is a simple extension of saliency that combines information about how sensitive a prediction is to a feature with how strongly that feature is present in the input (Denil, Demiraj, and De Freitas 2014). As in the saliency method, it considers the gradient of the class score with respect to the input, but instead of using the gradient alone, each input dimension is weighted by its own value. Intuitively, a feature should only be considered important if (i) small changes in that feature have a large effect on the score, and (ii) the feature is actually active in the current example. Raw gradients ignore the importance of the feature itself. I×G takes this

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

information into account by scaling the gradient by the input. In image models, this corresponds to weighting pixel-wise gradients by the pixel intensities. Let $\mathbf{x} \in \mathbb{R}^d$ be an input and $S_c(\mathbf{x})$ the score for class c. The Input × Gradient attribution for class c at $\mathbf{x}_0$ is defined as:

$$M_c^{(i)}(\mathbf{x}_0) = x_{0,i} \frac{\partial S_c(\mathbf{x}_0)}{\partial x_i}, \quad i = 1, \ldots, d. \tag{2}$$

Layer Gradient × Activation (LG×A): LG×A applies I×G to a chosen hidden layer. Let $\mathbf{h}(\mathbf{x}_0) \in \mathbb{R}^m$ be its activation vector; the attribution for unit j is:

$$M_c^{(j)}(\mathbf{x}_0) = h_j(\mathbf{x}_0) \frac{\partial S_c(\mathbf{x}_0)}{\partial h_j}, \quad j = 1, \ldots, m. \tag{3}$$

Integrated Gradients (IG): IG is motivated by two axioms: (i) Sensitivity, which requires that features responsible for a change in the model output relative to a baseline receive non-zero attribution, and (ii) Implementation invariance, which requires that attributions depend only on the input–output function $S(\mathbf{x})$, not on a particular network parameterization (Sundararajan *et al.* 2017). The authors define IG as:

$$\mathrm{IG}_i(\mathbf{x}_0; \mathbf{x}') = (x_{0,i} - x_i') \int_0^1 \frac{\partial S_c\left(\mathbf{x}' + \alpha(\mathbf{x}_0 - \mathbf{x}')\right)}{\partial x_i} d\alpha. \tag{4}$$

Given an input $\mathbf{x}$ and a baseline $\mathbf{x}'$ representing the absence of information (e.g., the zero vector), IG attributes feature i by integrating gradients along the straight-line path from $\mathbf{x}'$ to $\mathbf{x}$, capturing how the prediction changes as the input moves from the baseline to the actual example.

Gradient SHAP (GSHAP): GSHAP mixes IG with SHAP and estimates SHAP-style (Lundberg *et al.* 2017) attributions by averaging gradients over randomized reference points. For each of n samples, it adds small noise to the input, draws a random baseline from a set of baselines, and picks a random interpolation point along the straight line to compute the gradient[c].

Deep learning important features (DeepLIFT): DeepLIFT explains a prediction by comparing the network's response on an input $\mathbf{x}_0$ to a reference input $\mathbf{x}'$ (Shrikumar, Greenside, and Kundaje 2017). Instead of raw gradients, it propagates *differences from reference* layer by layer and assigns contribution scores to input features that sum to the output difference $S_c(\mathbf{x}_0) - S_c(\mathbf{x}')$, which helps mitigate gradient saturation and captures both positive and negative influences more reliably than standard gradient-based saliency methods.

Attention: In the Transformer model (Vaswani 2017), each token representation $x_t \in \mathbb{R}^{d_{\text{model}}}$ is linearly projected into a query, key, and value using learned parameter matrices $W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{model}} \times d_k}$. Scaled dot-product attention computes similarities between queries and keys, normalizes them with a softmax to obtain attention weights, and then uses these weights to form weighted sums of the values. In multi-head attention, H such projections $\{W_Q^{(h)}, W_K^{(h)}, W_V^{(h)}\}_{h=1}^H$ are used in parallel, and the concatenated head outputs are linearly projected back to the model dimension with a learned matrix $W_O \in \mathbb{R}^{(Hd_k) \times d_{\text{model}}}$.

Given queries $Q \in \mathbb{R}^{T_q \times d_k}$, keys $K \in \mathbb{R}^{T_k \times d_k}$, and values $V \in \mathbb{R}^{T_k \times d_v}$, scaled dot-product attention is

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \tag{5}$$

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

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

In (5), the score matrix $QK^\top$ can be interpreted as pairwise similarity scores between queries and keys; after softmax normalization, these become attention weights that determine how strongly each query attends to each key.

Value Zeroing (ValueZeroing): In the attention mechanism, the value vector V_j for token j carries contextual content that is mixed into the representation of other tokens via attention weights derived from the $QK^\top$ scores. ValueZeroing is an ablation technique that quantifies how much a context token j contributes to the representation of an output token at position i by recomputing the model's hidden representation at i after zeroing out the value vector of token j, while keeping all keys and queries fixed (Mohebbi *et al.* 2023).

Let $\tilde{\mathbf{x}}_i$ denote the original representation of the output token at position i, and let $\tilde{\mathbf{x}}_i^{\neg j}$ denote the representation obtained when the value vector of token j is replaced by the zero vector. The context-mixing score between output position i and token j is defined as the cosine distance between these two representations:

$$C_{i,j} = 1 - \cos\left(\tilde{\mathbf{x}}_i^{\neg j},\ \tilde{\mathbf{x}}_i\right). \tag{6}$$

Higher values of $C_{i,j}$ indicate that token j induces a larger change in the output representation at position i, and therefore has a stronger influence on that output.

In this subsection, we briefly summarize the XAI attribution methods used in our comparison. Our goal was to include representatives from all three major families of attribution methods while also respecting practical constraints on computation. In particular, many perturbation-based methods are computationally expensive, and generating their attribution maps at scale for our datasets would be infeasible within a reasonable runtime.

3. Methodology

Inspired by the forward simulation of XAI methods (Hase *et al.* 2020), we design a pipeline to compare different explainability attribution maps based on their impact on model performance in the NMT task. For this purpose, we use teacher–student knowledge distillation (Hinton 2015) (Figure 2). In the first step, we use the Inseq library[d] to extract input–output attribution maps using the eight explainability algorithms specified in Subsection 2.5. The teacher model receives source-target sample pairs as input, and the output of Inseq is a set of attributions $(\mathbf{x}, \mathbf{y}) \rightarrow E$ mapping the target to the source tokens . Then, the student models are trained under teacher forcing, receiving $(\mathbf{x}, \mathbf{y}, E)$ for training. During testing, the student model gets the source token and attributions to predict the target $(\mathbf{x}, E) \rightarrow \hat{\mathbf{y}}$.

Gradient-based attributions 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 getting the L2 norm $\|\mathbf{e}_i\|_2$ of the token vectors. L2 norm represents the magnitude of a vector and has a non-negative value.[e] The final result is in the shape of $e \in \mathbb{R}^{j \times k}$. For all the attribution methods, we get the scores from the first layer of the transformer model. However, for LG×A, where the attribution from the first layer is the same as I×G, we chose to obtain attributions from the encoder's last layer. Prior work has examined which task-relevant properties are encoded at different layers (Langedijk *et al.* 2024), but for our purposes, the encoder's final layer is the natural choice since its representations are the ones the decoder attends to when producing predictions.

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 of the transformers, and h is the number of attention

[d] https://inseq.org/en/latest/

[e] We did experiments with average pooling of the importance vectors; however, L2 norm representation always yielded better results.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

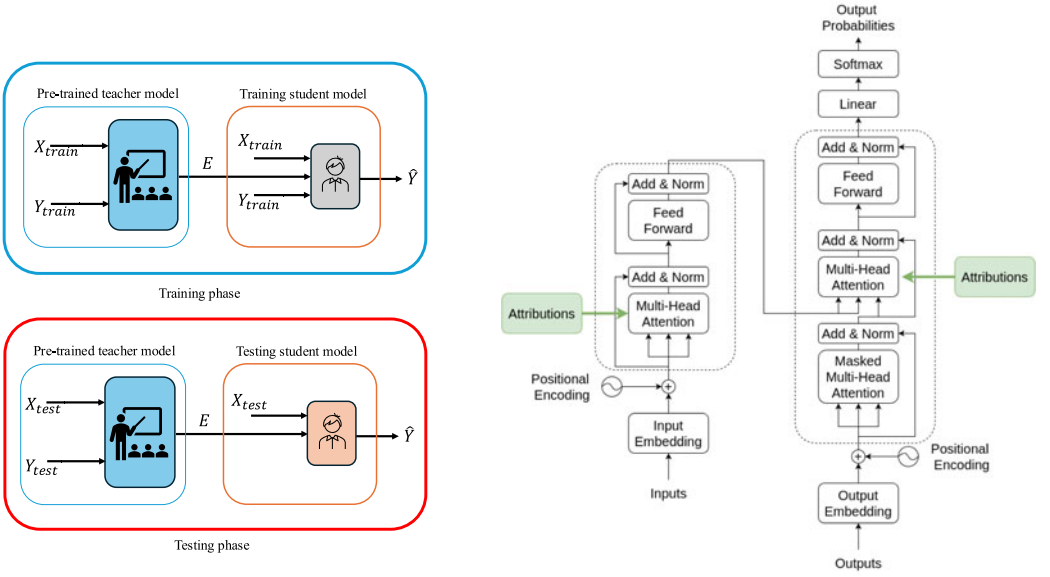


Figure 2. (a) Illustrates the overall design of our approach. The input sequence and the gold output $(\mathbf{x}, \mathbf{y})$ are given to a teacher model, and their attributions E are obtained. Then, a new untrained model is trained using the same $(\mathbf{x}, \mathbf{y}, E)$ triples. In the testing phase, the model gets the $(\mathbf{x}, E) \rightarrow \hat{\mathbf{y}}$. **(b)** Shows two places where we inject the attributions obtained from XAI methods.

heads. We then compute the average along both of the last two axes to obtain a final shape of $e \in \mathbb{R}^{j \times k}$. ValueZeroing yields the importance score for each layer, and therefore the scores are in the shape of $e \in \mathbb{R}^{j \times k \times n}$. Similarly, we get the average of the scores on the last dimension to reach $e \in \mathbb{R}^{j \times k}$. To normalize and handle negative values in the attribution matrices, we apply the MinMaxScaler[f] to the columns of the attribution maps as follows:

$$\mathbf{e}'_{i,j} = \frac{\mathbf{e}_{i,j} - \min\limits_{i}(\mathbf{e}_{:,j})}{\max\limits_{i}(\mathbf{e}_{:,j}) - \min\limits_{i}(\mathbf{e}_{:,j})} \tag{7}$$

Minmax transformation rescales values linearly to the $[0, 1]$ interval while preserving their rank and relative structure. This choice over the softmax function avoids the additional inductive bias introduced by a softmax transformation, which converts scores into a probability distribution whose shape depends nonlinearly on their scale and forces scores to compete with one another. In our setting, we were interested in comparing the pattern and relative magnitude of attributions across methods and tokens, rather than imposing a probabilistic interpretation. Minmax normalization, therefore, preserves the geometry of the original attribution map.[g]

Next, the student model receives as input the triple $(\mathbf{x}, \mathbf{y}, \mathbf{E}')$, where $\mathbf{E}'$ represents the (normalized) attribution map associated with the source–target pairs. Subsequently, we perform four distinct operations on the pre-softmax attention scores $\mathbf{A}^{(h)} = \frac{Q^{(h)}K^{(h)\top}}{\sqrt{d_k}}$ for each head h:

$$\tilde{\text{Attention}}(Q, K, V, \mathbf{E}') = \text{softmax}(f(\mathbf{A}, \mathbf{E}'))V, \tag{8}$$

[f] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[g] That said, the choice of the normalizer would affect the results; in our limited experiments, the minmax scaler scored better than the softmax; however, more investigations may be needed.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

where f is one of the following operations (applied element-wise):

Addition ($+$): add attributions to the attention scores:

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

Multiply ($\odot$): element-wise multiplication with the attention scores:

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

Average (μ): take the average of attributions and attention scores:

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

Replace (R): substitute the attention scores with attribution maps:

$$\tilde{\mathbf{A}}^{(h)} \leftarrow \mathbf{E}'. \tag{12}$$

The last operation replaces $\mathbf{A}^{(h)}$ with $\mathbf{E}'$, completely substituting $\frac{QK^\top}{\sqrt{d_k}}$. The point of applying these simple element-wise operators is to treat the normalized attribution matrices $\mathbf{E}'$ as soft importance weights over the similarity scores $A^{(h)} = QK^\top$ in each attention head. Acting directly on $QK^\top$ (i.e., at the level of the similarity matrix before softmax) rather than on the values V or the hidden states confines the intervention to the alignment structure between source and target tokens, which is precisely what attribution methods aim to characterize.

The four operators correspond to qualitatively different ways of using $\mathbf{E}'$ as a scaling factor for the similarity matrix. The multiplicative update $\tilde{\mathbf{A}}^{(h)} = \mathbf{A}^{(h)} \odot \mathbf{E}'$ implements a gating mechanism where attributions close to zero suppress specific query–key interactions, whereas attributions close to one leave them largely unchanged. Also, from another perspective, multiplication has a more dire effect if the attribution maps are incorrect. In contrast, the additive update $\tilde{\mathbf{A}}^{(h)} = \mathbf{A}^{(h)} + \mathbf{E}'$ behaves like a bias term on the similarity scores; when applied before softmax, it shifts probability mass toward positions preferred by the attribution map while preserving much of the relative structure induced by $QK^\top$. The averaging operator $\tilde{\mathbf{A}}^{(h)} = \frac{1}{2}(\mathbf{A}^{(h)} + \mathbf{E}')$ can be seen as a symmetric compromise between the model's own attention and the external explanation, which smooths extreme scores from both matrices. Finally, the replacement variant, which feeds $\mathbf{E}'$ directly into the attention module in place of $\frac{QK^\top}{\sqrt{d_k}}$, provides a raw testbed in which the attribution map is treated as the only alignment signal. This yields an approximate lower bound on how well a given XAI method can affect the translation task.

By comparing these operators within the same teacher–student framework, we can probe two questions at once: (i) whether they are strong enough to reliably gate or reroute information flow, and (ii) to quantify the influence of attribution maps on the MT task. In all cases, applying $\mathbf{E}'$ as element-wise weights on $QK^\top$ makes a source-target token pair suitable for forward-simulation style evaluation, since changes in translation quality can be traced back to the manipulation of the attention matrix by the XAI attribution maps.

4. Results

In this section, we first describe our experimental setup, including datasets, metrics, and implementation details. We then analyze the results along four dimensions: (1) the comparison of eight attribution methods based on their influence on translation quality when integrated into the model; (2) the impact of the attribution injection location, comparing encoder self-attention

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

and cross-attention modules; (3) the effect of selectively applying attributions to half of the attention heads (8 heads vs. 4 heads); (4) the ability of the student model to approximate the generation from the teacher model.

4.1 Experimental setup

To evaluate the proposed pipeline, we train the Marian-MT model (Tiedemann and Thottingal 2020)[h] on three datasets from scratch. We choose two datasets belonging to a more closely related 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, Junczys-Dowmunt, and Pouliquen 2016).

We select 220,000 sample pairs from each dataset and preprocess them to suit our experimental setup. Considering the numerous seq2seq models we train from scratch, we impose constraints to efficiently manage the training process. Specifically, we limit both the input and output sequences to at most 128 tokens. Additionally, we discard samples with fewer than ten tokens and filter out pairs where the input-to-output length ratio (or vice versa) exceeds 1.7 for de-en and fr-en. 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 validation and testing.

We use two teacher models from which we extract attribution maps: a) monolingual Marian-MT systems and b) multilingual mBART-large (Tang *et al.* 2020) models. We focus most of our analysis on Marian-MT, since it is substantially smaller than mBART and, therefore, more tractable for computing and training with attribution maps at scale, given its smaller vocabulary size. Marian-MT is a Transformer model with six encoder layers and six decoder layers, each layer containing eight attention heads. In contrast, mBART-large has 12 encoder and 12 decoder layers, each with 12 attention heads. As a multilingual model, mBART uses a much larger subword vocabulary than Marian-MT (on the order of 500k vs. roughly 50k token types), which further increases the computational cost of attribution extraction. The student model shares the same overall architecture as Marian-MT, with feed-forward dimensionality $d_{\text{ff}} = 2048$ and embedding dimensionality $d_{\text{model}} = 512$. We limit the maximum sequence length to $L_{\max} = 128$, while keeping the number of layers N_{layers} and attention heads H unchanged. Overall, Marian-MT has around 74 million parameters, and mBART has 610 million parameters.

We train the student Marian-MT models for 20 epochs and apply early stopping after three consecutive epochs without improvement in validation loss. The student model employs the Swish activation function, as proposed by Ramachandran, Zoph, and Le (2017), which has been shown to enhance training dynamics and convergence. We used 20 Nvidia V100 GPUs for our experiments.

Throughout our experiments, we report BLEU (Papineni *et al.* 2002), which measures n-gram overlap between system outputs and reference translations, and chrF (Popović 2015), a character-level n-gram F-score. This combination is particularly appropriate in our setting. We deliberately do not report semantic evaluation metrics such as COMET for two reasons. First, we conceptualize attribution maps as an auxiliary 'memory' that the student model can use to reconstruct the future target sequence Y from the teacher's representations. Our objective is thus fidelity to the reference at the token level, for which BLEU and chrF are sufficient. Second, most of our source–target segments are relatively short, and our datasets are modest in size and length. Thus, COMET can be noisy and add limited additional insight beyond n-gram overlap. All in all, consistent changes across both BLEU and chrF provide sufficient evidence that attribution-guided attention priors affect the underlying translation behavior.

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

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

4.2 Effectiveness of attribution methods (Encoder attention)

This analysis evaluates the impact of eight XAI attribution methods on translation quality, comparing models with injected attribution maps and the baseline. Tables 1–3 present the BLEU and chrF of this setting and their delta compared to the baseline, using Marian-MT attributions, while 4, 5, and 6 represent the results of mBART attribution maps. The baseline models are trained and evaluated on the same dataset and settings, but without integrating attribution *E*. The difference between the baseline and the models with attribution maps is an indicator that XAI attribution maps changed the results relative to the baseline.

Starting with attribution maps extracted from Marian-MT, across all three language pairs, injecting attribution maps into the encoder's attention mechanism consistently improved translation quality over the baseline models. The highest gains come from Attention, ValueZeroing, and LG×A. The highest BLEU gains ranged up to + 20.0 for de-en, +28.8 for fr-en, and + 27.9 for ar-en, with corresponding chrF gains up to + 13.3, +17.5, and + 14.9, respectively. Other gradient-based methods score quite similarly to each other, and among them, GSHAP scores lowest across all three language pairs. Importantly, BLEU and chrF changes were aligned, and the configurations that improved BLEU nearly always improved chrF by a similar margin and vice versa, indicating that the gains are not metric-specific but reflect genuine translation quality.

For mBART attributions, there are more nuances. ValueZeroing scores higher than all other attribution methods in all cases. Attention maps that used to achieve higher scores on Marian-MT now score lower, and even for de-en, they degrade results in three out of four operators. Among the gradient-based methods, LG×A scores highest for all the operators and all the language pairs. GSHAP yields the weakest results. Similarly, in mBART fr-en, Attention and ValueZeroing with the best operator reached 59.7 and 63.0 BLEU (+32.6 and + 35.9), while GSHAP-based injections remained close to the baseline or produced only small improvements.

The choice of operator used to combine attributions with the original attention weights had a strong and systematic effect. Across attribution methods, language pairs, and models, the element-wise product operator ($\odot$) consistently yielded the highest BLEU and chrF scores, while averaging (μ) was almost always the worst-performing operator, with + and *R* lying in between. For instance, in Marian-MT de-en with ValueZeroing, BLEU improved from 25.82 (baseline) to 41.8 (+16.0) with +, 40.2 (+14.4) with μ, 41.9 (+16.1) with *R*, and 45.8 (+20) with $\odot$. An analogous pattern appeared with Marian-MT fr-en and ar-en, where $\odot$ systematically dominated the other operators for all strong attribution sources. The same trend held with mBART: for fr-en, Attention with $\odot$ reached 59.7 BLEU (+32.6) above the other operators, and ValueZeroing with $\odot$ reached 63.0 BLEU (+35.9) versus 51.2, 45.2, and 48.9 BLEU for +, μ and *R* respectively. The magnitude of the difference between the operator for mBART Attention is higher than that for the others. We observed that mBART Attention assigns high values to the last source token (also visible in Figure 1), and that's why other operators can't contribute to the student model.

Finally, the patterns described above were consistent across de-en, fr-en, and ar-en for Marian-MT and across de-en and fr-en for mBART. Although the absolute baselines and magnitudes of improvement varied by language pair, the relative rankings of attribution sources and operators were nearly identical. This cross-lingual consistency suggests that certain XAI attribution methods provide more faithful target–source alignment signals than others, and that these signals can be leveraged as a reliable inductive bias for attention-guided knowledge distillation.

Overall ranking of attribution methods (best to worst).

For de-en, we obtain:

- **Marian-MT**: Attention ≈ ValueZeroing > LG×A > IG > DeepLIFT > Saliency > I×G > GSHAP
- **mBART**: ValueZeroing > Attention > LG×A > IG > DeepLIFT > Saliency > I×G > GSHAP

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

Table 1. BLEU and chrF scores for de-en Marian-MT attributions. Scores followed by Δ over the baseline

BLEU scores for de-en (Baseline: 25.82)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$32.1_{+6.3}$	$32.3_{+6.4}$	$35.5_{+9.6}$	$33.6_{+7.8}$	$27.5_{+1.7}$	$32.8_{+7.0}$	$40.5_{+14.7}$	$41.8_{+16.0}$
μ	$29.5_{+3.7}$	$29.7_{+3.9}$	$33.5_{+7.7}$	$32.0_{+6.2}$	$26.8_{+1.0}$	$30.5_{+4.7}$	$37.3_{+11.5}$	$40.2_{+14.4}$
⊙	$33.7_{+7.9}$	$33.8_{+8.0}$	$37.7_{+11.8}$	$35.4_{+9.6}$	$29.1_{+3.3}$	$34.2_{+8.3}$	$45.8_{+20.0}$	$45.8_{+19.9}$
R	$31.4_{+5.6}$	$31.7_{+5.9}$	$35.4_{+9.5}$	$33.0_{+7.1}$	$26.7_{+0.9}$	$32.3_{+6.5}$	$40.2_{+14.3}$	$41.9_{+16.1}$
chrF scores for de-en (Baseline: 49.27)								
+	$53.2_{+3.9}$	$53.3_{+4.0}$	$55.6_{+6.3}$	$54.2_{+4.9}$	$50.0_{+0.8}$	$53.7_{+4.4}$	$58.7_{+9.4}$	$60.0_{+10.7}$
μ	$51.3_{+2.0}$	$51.4_{+2.2}$	$54.1_{+4.8}$	$53.1_{+3.8}$	$49.7_{+0.4}$	$52.0_{+2.7}$	$56.4_{+7.1}$	$58.8_{+9.5}$
⊙	$54.3_{+5.0}$	$54.4_{+5.1}$	$57.2_{+7.9}$	$55.5_{+6.3}$	$51.1_{+1.8}$	$54.7_{+5.4}$	$62.6_{+13.3}$	$62.3_{+13.0}$
R	$52.7_{+3.5}$	$52.9_{+3.6}$	$55.5_{+6.3}$	$53.7_{+4.5}$	$49.6_{+0.3}$	$53.3_{+4.1}$	$58.5_{+9.3}$	$60.1_{+10.8}$

Table 2. BLEU and chrF scores for fr-en Marian-MT attributions. Scores followed by Δ over the baseline

BLEU scores for fr-en (Baseline: 27.01)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$35.5_{+8.5}$	$35.7_{+8.7}$	$39.1_{+12.1}$	$37.6_{+10.6}$	$31.5_{+4.5}$	$35.9_{+8.8}$	$45.1_{+18.1}$	$47.8_{+20.8}$
μ	$32.8_{+5.8}$	$33.0_{+6.0}$	$36.0_{+9.0}$	$34.8_{+7.8}$	$28.8_{+1.8}$	$33.4_{+6.4}$	$40.8_{+13.8}$	$45.6_{+18.6}$
⊙	$37.6_{+10.6}$	$37.9_{+10.9}$	$42.1_{+15.1}$	$40.1_{+13.1}$	$33.5_{+6.5}$	$37.7_{+10.7}$	$55.8_{+28.8}$	$54.1_{+27.1}$
R	$34.6_{+7.5}$	$34.6_{+7.6}$	$38.2_{+11.2}$	$37.2_{+10.1}$	$30.4_{+3.4}$	$35.1_{+8.1}$	$44.4_{+17.4}$	$47.1_{+20.1}$
chrF scores for fr-en (Baseline: 53.01)								
+	$57.6_{+4.6}$	$57.6_{+4.6}$	$60.0_{+7.0}$	$58.8_{+5.8}$	$55.0_{+2.0}$	$57.8_{+4.8}$	$63.4_{+10.4}$	$65.5_{+12.5}$
μ	$55.8_{+2.8}$	$55.9_{+2.9}$	$57.8_{+4.8}$	$57.0_{+4.0}$	$53.3_{+0.3}$	$56.1_{+3.1}$	$60.4_{+7.4}$	$63.9_{+10.9}$
⊙	$58.9_{+5.9}$	$59.1_{+6.1}$	$61.9_{+8.9}$	$60.4_{+7.4}$	$56.2_{+3.2}$	$58.9_{+5.9}$	$70.5_{+17.5}$	$69.1_{+16.1}$
R	$56.9_{+3.9}$	$56.9_{+3.9}$	$59.4_{+6.4}$	$58.5_{+5.5}$	$54.2_{+1.2}$	$57.3_{+4.2}$	$63.0_{+9.9}$	$65.1_{+12.1}$

Table 3. BLEU and chrF scores for ar-en Marian-MT attributions. Scores followed by Δ over the baseline

BLEU scores for ar-en (Baseline: 40.68)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$52.0_{+11.2}$	$51.7_{+10.9}$	$54.9_{+14.1}$	$52.1_{+11.3}$	$46.4_{+5.6}$	$52.7_{+11.9}$	$60.8_{+19.9}$	$62.9_{+22.1}$
μ	$47.3_{+6.5}$	$46.9_{+6.1}$	$51.3_{+10.5}$	$47.3_{+6.5}$	$43.7_{+2.9}$	$47.4_{+6.6}$	$55.7_{+14.9}$	$60.8_{+20.0}$
⊙	$54.1_{+13.2}$	$54.2_{+13.4}$	$57.5_{+16.7}$	$54.2_{+13.4}$	$48.1_{+7.3}$	$54.2_{+13.4}$	$66.4_{+25.6}$	$68.7_{+27.9}$
R	$51.1_{+10.3}$	$51.0_{+10.2}$	$54.4_{+13.6}$	$51.1_{+10.3}$	$45.1_{+4.3}$	$51.9_{+11.1}$	$60.2_{+19.4}$	$62.7_{+21.9}$
chrF scores for ar-en (Baseline: 66.07)								
+	$71.4_{+5.3}$	$71.2_{+5.1}$	$73.2_{+7.1}$	$71.4_{+5.4}$	$68.2_{+2.1}$	$71.8_{+5.7}$	$76.2_{+10.1}$	$77.6_{+11.6}$
μ	$68.7_{+2.6}$	$68.5_{+2.4}$	$71.0_{+4.9}$	$68.7_{+2.6}$	$66.8_{+0.7}$	$68.8_{+2.7}$	$73.2_{+7.1}$	$76.4_{+10.3}$
⊙	$72.5_{+6.4}$	$72.5_{+6.4}$	$74.5_{+8.4}$	$72.4_{+6.3}$	$68.9_{+2.9}$	$72.5_{+6.5}$	$79.5_{+13.4}$	$81.0_{+14.9}$
R	$70.6_{+4.6}$	$70.6_{+4.5}$	$72.7_{+6.6}$	$70.7_{+4.7}$	$67.3_{+1.3}$	$71.1_{+5.0}$	$75.8_{+9.7}$	$77.4_{+11.4}$

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

Table 4. BLEU and chrF scores for de-en mBART attributions. Scores followed by Δ over the baseline

BLEU scores for de-en (Baseline: 26.08)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$24.9_{-1.2}$	$25.9_{-0.2}$	$29.8_{+3.8}$	$26.0_{0.0}$	$23.2_{-2.8}$	$26.2_{+0.1}$	$25.0_{-1.1}$	$43.4_{+17.3}$
μ	$23.1_{-3.0}$	$23.3_{-2.8}$	$24.6_{-1.5}$	$23.5_{-2.6}$	$22.6_{-3.5}$	$22.9_{-3.2}$	$24.2_{-1.9}$	$40.4_{+14.3}$
⊙	$28.2_{+2.1}$	$28.4_{+2.3}$	$32.2_{+6.2}$	$28.8_{+2.7}$	$24.9_{-1.2}$	$28.4_{+2.4}$	$39.0_{+12.9}$	$51.4_{+25.3}$
R	$25.0_{-1.1}$	$25.4_{-0.7}$	$29.0_{+2.9}$	$26.2_{+0.1}$	$23.9_{-2.2}$	$25.4_{-0.7}$	$24.7_{-1.4}$	$41.4_{+15.3}$
chrF scores for de-en (Baseline: 48.4)								
+	$46.8_{-1.6}$	$47.6_{-0.8}$	$50.5_{+2.1}$	$47.3_{-1.1}$	$45.8_{-2.6}$	$47.8_{-0.6}$	$46.6_{-1.8}$	$60.5_{+12.1}$
μ	$45.3_{-3.1}$	$45.5_{-2.9}$	$46.2_{-2.2}$	$45.6_{-2.8}$	$45.2_{-3.2}$	$45.4_{-3.0}$	$46.3_{-2.1}$	$58.3_{+9.9}$
⊙	$49.2_{+0.8}$	$49.2_{+0.8}$	$52.2_{+3.8}$	$49.5_{+1.1}$	$46.9_{-1.5}$	$49.3_{+0.9}$	$56.7_{+8.3}$	$66.2_{+17.8}$
R	$47.0_{-1.4}$	$47.3_{-1.1}$	$49.9_{+1.5}$	$47.5_{-0.9}$	$46.4_{-2.0}$	$47.3_{-1.1}$	$46.3_{-2.1}$	$59.1_{+10.7}$

Table 5. BLEU and chrF scores for fr-en mBART attributions. Scores followed by Δ over the baseline

BLEU scores for fr-en (Baseline: 27.12)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$33.2_{+6.1}$	$33.9_{+6.8}$	$38.9_{+11.8}$	$34.8_{+7.7}$	$28.4_{+1.3}$	$34.0_{+6.9}$	$31.8_{+4.7}$	$51.2_{+24.1}$
μ	$29.5_{+2.4}$	$30.9_{+3.7}$	$35.2_{+8.1}$	$29.9_{+2.7}$	$26.8_{-0.3}$	$29.9_{+2.8}$	$30.8_{+3.7}$	$45.2_{+18.1}$
⊙	$37.5_{+10.4}$	$38.0_{+10.9}$	$42.9_{+15.8}$	$37.5_{+10.4}$	$32.7_{+5.6}$	$38.2_{+11.0}$	$59.7_{+32.6}$	$63.0_{+35.9}$
R	$32.2_{+5.1}$	$32.4_{+5.3}$	$37.4_{+10.2}$	$32.3_{+5.2}$	$28.0_{+0.9}$	$32.7_{+5.5}$	$31.3_{+4.2}$	$48.9_{+21.8}$
chrF scores for fr-en (Baseline: 53.01)								
+	$55.9_{+2.9}$	$56.3_{+3.3}$	$59.6_{+6.6}$	$56.9_{+3.9}$	$53.2_{+0.2}$	$56.3_{+3.3}$	$54.5_{+1.4}$	$67.7_{+14.7}$
μ	$53.6_{+0.6}$	$54.4_{+1.4}$	$57.1_{+4.1}$	$54.4_{+1.4}$	$52.3_{-0.7}$	$54.0_{+1.0}$	$54.0_{+1.0}$	$63.7_{+10.7}$
⊙	$58.4_{+5.4}$	$58.7_{+5.7}$	$62.1_{+9.0}$	$58.3_{+5.3}$	$55.3_{+2.3}$	$58.8_{+5.8}$	$73.2_{+20.2}$	$75.5_{+22.5}$
R	$55.1_{+2.1}$	$55.2_{+2.2}$	$58.4_{+5.4}$	$55.4_{+2.4}$	$52.8_{-0.2}$	$55.4_{+2.4}$	$53.9_{+0.9}$	$66.1_{+13.1}$

Table 6. BLEU and chrF scores for ar-en mBART attributions. Scores followed by Δ over the baseline

BLEU scores for ar-en (Baseline: 40.68)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$50.6_{+9.9}$	$51.2_{+10.5}$	$55.3_{+14.6}$	$50.5_{+9.8}$	$44.0_{+3.3}$	$50.3_{+9.6}$	$49.0_{+8.3}$	$63.6_{+22.9}$
μ	$45.1_{+4.4}$	$44.7_{+4.0}$	$49.7_{+9.1}$	$45.0_{+4.3}$	$42.1_{+1.5}$	$44.2_{+3.6}$	$45.7_{+5.0}$	$58.4_{+17.7}$
⊙	$53.7_{+13.0}$	$54.5_{+13.8}$	$59.5_{+18.9}$	$53.5_{+12.8}$	$48.3_{+7.6}$	$54.2_{+13.6}$	$73.4_{+32.7}$	$74.4_{+33.7}$
R	$48.3_{+7.6}$	$48.7_{+8.0}$	$52.8_{+12.1}$	$48.5_{+7.7}$	$42.0_{+1.3}$	$48.6_{+7.9}$	$47.6_{+6.9}$	$60.4_{+19.7}$
chrF scores for ar-en (Baseline: 65.41)								
+	$69.9_{+4.5}$	$70.3_{+4.9}$	$72.7_{+7.3}$	$69.1_{+3.7}$	$66.6_{+1.2}$	$69.8_{+4.4}$	$68.6_{+3.2}$	$77.7_{+12.3}$
μ	$67.0_{+1.6}$	$66.8_{+1.4}$	$69.5_{+4.1}$	$67.3_{+1.9}$	$66.0_{+0.6}$	$66.6_{+1.2}$	$67.0_{+1.5}$	$74.4_{+9.0}$
⊙	$71.4_{+6.0}$	$71.9_{+6.5}$	$75.1_{+9.7}$	$73.1_{+7.7}$	$68.1_{+2.7}$	$71.7_{+6.3}$	$83.8_{+18.4}$	$84.3_{+18.9}$
R	$68.2_{+2.8}$	$68.4_{+3.0}$	$70.9_{+5.5}$	$68.3_{+2.9}$	$65.0_{-0.4}$	$68.3_{+2.9}$	$67.1_{+1.7}$	$75.6_{+10.2}$

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

For fr-en, we obtain:

- **Marian-MT**: Attention > ValueZeroing > LG×A > IG > Saliency > DeepLIFT > I×G > GSHAP
- **mBART**: ValueZeroing > Attention > LG×A > DeepLIFT > Saliency > I×G ≈ IG > GSHAP

For ar-en, we obtain:

- **Marian-MT**: ValueZeroing > Attention > LG×A > DeepLIFT ≈ IG ≈ Saliency > I×G > GSHAP
- **mBART**: ValueZeroing > Attention > LG×A > Saliency >DeepLIFT > I×G > IG > GSHAP

4.3 Encoder self-attention versus cross-attention injection

In contrast to the encoder self-attention experiments, injecting attributions into cross-attention rarely improved translation quality and often degraded it (Tables 7–9). Across de-en, fr-en, and ar-en, most attribution–operator combinations reduced BLEU and chrF relative to the baseline. The only consistent but modest gains were observed for de-en and fr-en when using gradient-based attributions (IG, LG×A) with a multiplicative operator ($\odot$), yielding up to + 4.7 BLEU and + 1.0 chrF. ValueZeroing and teacher Attention, which were highly effective for encoder attention, provided little benefit in cross-attention and frequently harmed performance, while GSHAP was consistently detrimental. Replacement of cross-attention weights (R) was particularly destructive, often leading to large drops in BLEU and chrF. These patterns suggest that cross-attention is substantially more brittle than encoder attention, and its alignment structure can only tolerate very limited attribution guidance, and even then, the resulting gains are small and not always reflected consistently across evaluation settings. While the exact reason that attribution injection to the cross-attention does not work is obscure and hard to pinpoint, our primary hypothesis is that during autoregressive inference, due to a decoding strategy such as beam search, the fixed sequence of the attributions for generated targets confuses the model. In other words, the model can deviate from the ground-truth sentence during autoregressive generation, but attributions are added for the fixed sequence of tokens, which not only no longer matches the tokens generated by the model but also actively prevents it from being able to correct itself.[i]

4.4 Selective attribution injection (8 vs. 4 heads)

In another setting, as an ablation study, we applied the attribution methods to only 4 heads out of the 8 heads of the attention, only using Marian-MT attributions. We conduct these experiments only on the encoder side, as previous results showed that these attribution mappings can be useful at this part of the architecture. Tables 14–16 show the results of this comparison. This analysis investigates how reducing the number of attention heads affects the model's performance 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 and chrF scores between 8-head and 4-head settings are relatively minor. However, some methods and operators show slight improvements. The results suggest that combining standard attention mechanisms with attribution-based operators can be a 'best of both worlds' approach, as some heads learn the standard attention mechanisms while others utilize the attribution maps.

[i]To reduce energy consumption and training time, we omit training with mBART attributions when using 4-head encoder and cross-attention.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

Table 7. BLEU and chrF scores for de-en Cross-attention Marian-MT attributions (Baseline: 25.82). Scores followed by Δ over the baseline

BLEU scores for de-en (Baseline: 25.82)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$25.3_{-0.5}$	$25.4_{-0.4}$	$24.9_{-0.9}$	$24.8_{-1.0}$	$22.4_{-3.4}$	$25.6_{-0.3}$	$16.9_{-8.9}$	$23.9_{-1.9}$
μ	$24.8_{-1.0}$	$25.0_{-0.8}$	$26.0_{+0.2}$	$23.8_{-2.0}$	$22.3_{-3.5}$	$24.8_{-1.1}$	$20.0_{-5.8}$	$23.0_{-2.8}$
⊙	$26.1_{+0.3}$	$26.1_{+0.3}$	$28.6_{+2.8}$	$28.5_{+2.6}$	$21.8_{-4.1}$	$26.3_{+0.5}$	$25.6_{-0.2}$	$24.7_{-1.1}$
R	$24.2_{-1.6}$	$24.2_{-1.6}$	$22.6_{-3.2}$	$23.0_{-2.9}$	$18.2_{-7.6}$	$24.4_{-1.4}$	$15.7_{-10.1}$	$21.0_{-4.8}$
chrF scores for de-en (Baseline: 49.27)								
+	$47.5_{-1.8}$	$47.6_{-1.6}$	$48.7_{-0.6}$	$48.1_{-1.1}$	$46.0_{-3.3}$	$47.4_{-1.9}$	$44.1_{-5.1}$	$47.3_{-1.9}$
μ	$47.0_{-2.3}$	$47.3_{-1.9}$	$48.7_{-0.5}$	$47.8_{-1.4}$	$45.8_{-3.5}$	$47.0_{-2.2}$	$45.3_{-4.0}$	$46.6_{-2.6}$
⊙	$47.6_{-1.7}$	$47.4_{-1.9}$	$50.3_{+1.0}$	$49.6_{+0.4}$	$45.1_{-4.2}$	$47.2_{-2.1}$	$47.9_{-1.4}$	$47.9_{-1.4}$
R	$45.9_{-3.4}$	$45.9_{-3.4}$	$46.9_{-2.3}$	$45.9_{-3.4}$	$41.4_{-7.9}$	$44.1_{-5.2}$	$42.1_{-7.2}$	$45.3_{-4.0}$

Table 8. BLEU and chrF scores for fr-en Cross-attention Marian-MT attributions (Baseline: 27.01). Scores followed by Δ over the baseline

BLEU scores for fr-en (Baseline: 27.01)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$26.9_{-0.1}$	$26.9_{-0.1}$	$26.1_{-0.9}$	$28.2_{+1.2}$	$25.2_{-1.8}$	$25.3_{-1.7}$	$21.0_{-6.0}$	$27.6_{+0.6}$
μ	$26.4_{-0.6}$	$26.6_{-0.4}$	$26.9_{-0.1}$	$27.6_{+0.6}$	$24.7_{-2.3}$	$24.5_{-2.6}$	$22.2_{-4.8}$	$25.6_{-1.4}$
⊙	$27.7_{+0.6}$	$27.8_{+0.8}$	$29.1_{+2.1}$	$31.7_{+4.7}$	$25.2_{-1.8}$	$27.7_{+0.7}$	$19.9_{-7.2}$	$26.3_{-0.7}$
R	$25.4_{-1.6}$	$25.5_{-1.5}$	$23.3_{-3.7}$	$26.2_{-0.8}$	$21.5_{-5.5}$	$25.1_{-1.9}$	$16.6_{-10.4}$	$24.5_{-2.5}$
chrF scores for fr-en (Baseline: 53.01)								
+	$50.9_{-2.1}$	$50.8_{-2.2}$	$51.8_{-1.2}$	$52.1_{-0.9}$	$50.1_{-2.9}$	$49.4_{-3.6}$	$48.3_{-4.7}$	$53.0_{-0.0}$
μ	$50.4_{-2.6}$	$50.6_{-2.4}$	$51.8_{-1.2}$	$52.0_{-1.0}$	$49.7_{-3.4}$	$48.8_{-4.2}$	$48.2_{-4.8}$	$50.2_{-2.9}$
⊙	$51.3_{-1.7}$	$51.2_{-1.8}$	$52.8_{-0.2}$	$53.5_{+0.5}$	$49.9_{-3.1}$	$51.1_{-2.0}$	$50.1_{-2.9}$	$51.1_{-1.9}$
R	$48.7_{-4.3}$	$48.5_{-4.6}$	$49.9_{-3.1}$	$49.9_{-3.1}$	$46.2_{-6.8}$	$47.9_{-5.1}$	$43.8_{-9.2}$	$49.0_{-4.0}$

4.5 Faithfulness (predicting the model's output)

Up to this point, our experiments have relied on the assumption that attribution maps should help the student model better predict the gold (human) translation. Faithfulness, in contrast, relates to how precisely an attribution method captures the model's internal reasoning process that produces a particular output (Jacovi *et al.* 2020). Put differently, a faithful attribution provides a close approximation of the model's actual decision-making procedure that produced the prediction.

To investigate this notion in a teacher–student setting, we replace the human reference with the teacher's behavior as the supervision signal. Given a source input $\mathbf{x}$, we run the teacher model to produce a translation $\hat{\mathbf{y}}$ and compute the corresponding attribution map $\hat{\mathbf{E}}$ with respect to this generated output, yielding training triples $(\mathbf{x}, \hat{\mathbf{y}}, \hat{\mathbf{E}})$. The student model is then trained on $(\mathbf{x}, \hat{\mathbf{y}}, \hat{\mathbf{E}})$ exactly as in our earlier experiments, injecting $\hat{\mathbf{E}}$ into its attention mechanism. At test time, we

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

Table 9. BLEU and chrF scores for ar-en Cross-attention Marian-MT attributions (Baseline: 40.68). Scores followed by Δ over the baseline

BLEU scores for ar-en (Baseline: 40.68)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$36.2_{-4.7}$	$41.9_{+1.1}$	$35.9_{-4.9}$	$36.2_{-4.7}$	$36.0_{-4.9}$	$36.1_{-4.8}$	$36.1_{-4.7}$	$40.6_{-0.2}$
μ	$35.3_{-5.5}$	$41.7_{+0.9}$	$35.2_{-5.6}$	$35.3_{-5.5}$	$35.2_{-5.6}$	$35.3_{-5.5}$	$35.3_{-5.5}$	$40.5_{-0.3}$
⊙	$28.5_{-12.3}$	$42.2_{+1.4}$	$26.5_{-14.3}$	$29.6_{-11.2}$	$28.3_{-12.5}$	$29.3_{-11.5}$	$26.3_{-14.5}$	$42.9_{+2.1}$
R	$14.3_{-26.5}$	$40.0_{-0.8}$	$13.9_{-26.9}$	$14.5_{-26.3}$	$14.3_{-26.5}$	$14.2_{-26.6}$	$13.7_{-27.1}$	$39.1_{-1.7}$
chrF scores for ar-en (Baseline: 66.07)								
+	$62.7_{-3.4}$	$64.1_{-1.9}$	$62.7_{-3.3}$	$62.6_{-3.5}$	$62.7_{-3.4}$	$62.7_{-3.4}$	$62.8_{-3.3}$	$64.4_{-1.7}$
μ	$62.3_{-3.8}$	$64.1_{-2.0}$	$62.3_{-3.8}$	$62.3_{-3.8}$	$62.2_{-3.8}$	$62.3_{-3.7}$	$62.3_{-3.8}$	$64.1_{-2.0}$
⊙	$55.4_{-10.6}$	$64.3_{-1.7}$	$53.7_{-12.4}$	$56.1_{-9.9}$	$55.4_{-10.7}$	$56.0_{-10.1}$	$53.4_{-12.7}$	$65.3_{-0.8}$
R	$43.4_{-22.6}$	$62.6_{-3.5}$	$43.3_{-22.8}$	$43.5_{-22.6}$	$43.4_{-22.7}$	$43.4_{-22.7}$	$43.2_{-22.9}$	$63.0_{-3.1}$

Table 10. BLEU and chrF scores for de-en generated by the Marian-MT model. Scores followed by Δ over the baseline

BLEU scores for de-en (Baseline: 54.55)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$64.5_{+9.9}$	$64.6_{+10.1}$	$68.5_{+13.9}$	$65.5_{+11.0}$	$58.1_{+3.5}$	$65.1_{+10.5}$	$72.0_{+17.4}$	$73.7_{+19.1}$
μ	$58.4_{+3.8}$	$59.7_{+5.2}$	$65.2_{+10.7}$	$61.8_{+7.3}$	$55.1_{+0.5}$	$61.1_{+6.6}$	$68.8_{+14.3}$	$71.6_{+17.0}$
⊙	$66.5_{+11.9}$	$66.6_{+12.0}$	$70.0_{+15.5}$	$67.3_{+12.8}$	$59.7_{+5.1}$	$66.3_{+11.8}$	$75.4_{+20.9}$	$74.2_{+19.6}$
R	$63.4_{+8.8}$	$63.1_{+8.6}$	$67.9_{+13.3}$	$64.4_{+9.8}$	$56.0_{+1.4}$	$64.2_{+9.7}$	$72.0_{+17.4}$	$73.4_{+18.9}$
chrF scores for de-en (Baseline: 72.77)								
+	$78.3_{+5.5}$	$78.3_{+5.6}$	$80.8_{+8.0}$	$78.9_{+6.2}$	$74.4_{+1.6}$	$78.7_{+6.0}$	$82.8_{+10.0}$	$84.0_{+11.2}$
μ	$74.5_{+1.7}$	$75.4_{+2.6}$	$78.7_{+5.9}$	$76.6_{+3.9}$	$72.8_{+0.0}$	$76.2_{+3.4}$	$80.8_{+8.0}$	$82.6_{+9.8}$
⊙	$79.5_{+6.8}$	$79.6_{+6.9}$	$81.8_{+9.0}$	$80.1_{+7.3}$	$75.3_{+2.5}$	$79.5_{+6.7}$	$85.1_{+12.3}$	$84.0_{+11.2}$
R	$77.6_{+4.8}$	$77.4_{+4.7}$	$80.4_{+7.7}$	$78.2_{+5.4}$	$73.2_{+0.4}$	$78.1_{+5.3}$	$82.8_{+10.0}$	$83.8_{+11.0}$

provide the student model with $(\mathbf{x}, \hat{\mathbf{E}})$ and generate a student translation $\hat{\mathbf{y}}'$, that is $(\mathbf{x}, \hat{\mathbf{E}}) \mapsto \hat{\mathbf{y}}'$. We then compare $\hat{\mathbf{y}}'$ with $\hat{\mathbf{y}}$ to quantify how well the student model can reproduce the teacher's outputs when guided by a given attribution method. Under this setup, we hypothesize that attribution maps that better capture the teacher's input–output dependencies will provide more useful guidance and, in turn, enable the student to generate translations $\hat{\mathbf{y}}'$ that are closer to $\hat{\mathbf{y}}$. The resulting agreement between $\hat{\mathbf{y}}'$ and $\hat{\mathbf{y}}$ therefore serves as an empirical proxy for the utility of the underlying attribution maps for simulating the teacher model.

Here, we present the results of this setup for the Marian-MT model in Tables 10–12. The first observation is that the baseline scores are substantially higher for predicting the teacher's generation compared to the gold data (BLEU scores 54.55 vs 25.82 for de-en, 53.38 vs 27.01 for fr-en, 59.45 vs 40.68 for ar-en). As in the previous setting, injecting attribution maps into the encoder attention substantially improves performance over the baseline student for Attention and Value

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

Table 11. BLEU and chrF scores for fr-en generated by the Marian-MT model. Scores followed by Δ over the baseline

BLEU scores for fr-en (Baseline: 53.38)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$53.7_{+0.3}$	$53.6_{+0.2}$	$72.6_{+19.2}$	$70.7_{+17.3}$	$52.8_{-0.6}$	$53.6_{+0.2}$	$76.3_{+22.9}$	$80.6_{+27.2}$
μ	$52.8_{-0.6}$	$52.9_{-0.4}$	$63.6_{+10.2}$	$64.0_{+10.7}$	$52.4_{-1.0}$	$52.8_{-0.6}$	$69.5_{+16.1}$	$74.3_{+20.9}$
$\odot$	$62.2_{+8.8}$	$73.4_{+20.1}$	$76.5_{+23.1}$	$75.8_{+22.5}$	$58.7_{+5.3}$	$62.9_{+9.5}$	$82.9_{+29.5}$	$83.0_{+29.6}$
R	$54.1_{+0.7}$	$66.5_{+13.1}$	$72.8_{+19.4}$	$70.9_{+17.5}$	$52.2_{-1.2}$	$54.2_{+0.8}$	$77.4_{+24.0}$	$80.5_{+27.1}$
chrF scores for fr-en (Baseline: 74.05)								
+	$73.4_{-0.7}$	$73.3_{-0.7}$	$84.4_{+10.3}$	$83.2_{+9.1}$	$72.8_{-1.3}$	$73.2_{-0.8}$	$86.3_{+12.3}$	$89.0_{+15.0}$
μ	$73.1_{-1.0}$	$73.2_{-0.8}$	$79.5_{+5.5}$	$79.7_{+5.7}$	$72.9_{-1.2}$	$73.0_{-1.0}$	$82.5_{+8.5}$	$85.2_{+11.1}$
$\odot$	$78.1_{+4.1}$	$84.9_{+10.8}$	$86.7_{+12.6}$	$86.2_{+12.1}$	$76.1_{+2.0}$	$78.6_{+4.5}$	$90.3_{+16.2}$	$90.2_{+16.2}$
R	$73.5_{-0.5}$	$80.8_{+6.8}$	$84.5_{+10.5}$	$83.2_{+9.2}$	$72.4_{-1.7}$	$73.6_{-0.5}$	$87.0_{+12.9}$	$88.9_{+14.9}$

Table 12. BLEU and chrF scores for ar-en generated by the Marian-MT model. Scores followed by Δ over the baseline

BLEU scores for ar-en (Baseline: 59.45)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$69.1_{+9.6}$	$68.6_{+9.1}$	$73.1_{+13.6}$	$69.8_{+10.3}$	$63.7_{+4.2}$	$70.0_{+10.6}$	$76.8_{+17.4}$	$78.9_{+19.4}$
μ	$65.1_{+5.6}$	$64.6_{+5.1}$	$65.3_{+5.9}$	$63.8_{+4.3}$	$61.7_{+2.3}$	$65.1_{+5.7}$	$67.4_{+7.9}$	$75.1_{+15.7}$
$\odot$	$72.6_{+13.2}$	$72.8_{+13.3}$	$75.6_{+16.1}$	$72.4_{+13.0}$	$66.7_{+7.2}$	$72.9_{+13.5}$	$81.6_{+22.2}$	$82.8_{+23.3}$
R	$68.9_{+9.4}$	$68.5_{+9.0}$	$72.7_{+13.3}$	$68.9_{+9.4}$	$63.3_{+3.8}$	$69.5_{+10.1}$	$77.0_{+17.6}$	$78.7_{+19.2}$
chrF scores for ar-en (Baseline: 77.96)								
+	$82.5_{+4.5}$	$82.2_{+4.2}$	$84.8_{+6.8}$	$82.8_{+4.8}$	$79.5_{+1.6}$	$83.0_{+5.1}$	$86.7_{+8.7}$	$87.9_{+10.0}$
μ	$80.3_{+2.4}$	$80.1_{+2.1}$	$80.6_{+2.6}$	$79.6_{+1.7}$	$78.7_{+0.7}$	$80.4_{+2.4}$	$81.6_{+3.6}$	$85.8_{+7.8}$
$\odot$	$84.4_{+6.4}$	$84.5_{+6.5}$	$86.0_{+8.0}$	$84.2_{+6.2}$	$81.0_{+3.1}$	$84.6_{+6.6}$	$89.3_{+11.3}$	$90.0_{+12.0}$
R	$82.2_{+4.3}$	$82.0_{+4.0}$	$84.5_{+6.5}$	$82.2_{+4.2}$	$79.3_{+1.3}$	$82.6_{+4.6}$	$86.7_{+8.7}$	$87.7_{+9.7}$

Table 13. BLEU scores for random attribution matrices (left) and diagonal attribution matrices (right). Scores followed by Δ over the baseline

	Random matrix			Diagonal matrix		
Op.	de-en	fr-en	ar-en	de-en	fr-en	ar-en
$\odot$	$22.98_{-2.8}$	$24.85_{-2.1}$	$37.96_{-2.9}$	$21.64_{-4.2}$	$24.11_{-2.9}$	$37.34_{-3.3}$

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

Table 14. BLEU and chrF scores for de-en Marian-MT 4 head attribution injection. Scores followed by Δ over the baseline

BLEU scores for de-en (Baseline: 25.82)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$32.7_{+6.8}$	$32.4_{+6.6}$	$35.7_{+9.9}$	$33.7_{+7.8}$	$27.8_{+2.0}$	$32.9_{+7.1}$	$40.0_{+14.1}$	$41.8_{+16.0}$
μ	$29.9_{+4.0}$	$29.6_{+3.8}$	$33.5_{+7.7}$	$32.1_{+6.3}$	$26.2_{+0.4}$	$30.5_{+4.7}$	$36.1_{+10.2}$	$40.5_{+14.7}$
⊙	$34.4_{+8.5}$	$34.7_{+8.9}$	$38.3_{+12.5}$	$36.0_{+10.2}$	$29.9_{+4.0}$	$34.6_{+8.8}$	$47.2_{+21.4}$	$44.8_{+18.9}$
R	$32.4_{+6.6}$	$32.3_{+6.5}$	$35.9_{+10.1}$	$33.8_{+8.0}$	$27.9_{+2.1}$	$32.9_{+7.1}$	$40.2_{+14.4}$	$41.7_{+15.8}$
chrF scores for de-en (Baseline: 49.27)								
+	$53.6_{+4.4}$	$53.4_{+4.2}$	$55.8_{+6.6}$	$54.3_{+5.1}$	$50.3_{+1.0}$	$53.8_{+4.5}$	$58.3_{+9.1}$	$60.0_{+10.7}$
μ	$51.5_{+2.3}$	$51.3_{+2.0}$	$54.1_{+4.8}$	$53.1_{+3.8}$	$49.1_{-0.2}$	$52.0_{+2.8}$	$55.5_{+6.2}$	$58.9_{+9.6}$
⊙	$54.9_{+5.6}$	$55.1_{+5.8}$	$57.7_{+8.4}$	$56.0_{+6.7}$	$51.7_{+2.4}$	$55.0_{+5.7}$	$63.6_{+14.4}$	$62.0_{+12.7}$
R	$53.5_{+4.2}$	$53.4_{+4.1}$	$56.0_{+6.7}$	$54.4_{+5.1}$	$50.4_{+1.1}$	$53.8_{+4.6}$	$58.5_{+9.2}$	$59.9_{+10.6}$

Table 15. BLEU and chrF scores for fr-en Marian-MT 4 head attribution injection. Scores followed by Δ over the baseline

BLEU scores for fr-en (Baseline: 27.01)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$34.9_{+7.9}$	$34.6_{+7.6}$	$38.5_{+11.5}$	$37.5_{+10.5}$	$31.2_{+4.2}$	$35.1_{+8.1}$	$44.3_{+17.3}$	$47.5_{+20.5}$
μ	$32.1_{+5.1}$	$33.1_{+6.1}$	$34.9_{+7.9}$	$34.1_{+7.1}$	$28.5_{+1.5}$	$32.8_{+5.8}$	$39.9_{+12.9}$	$44.9_{+17.9}$
⊙	$38.2_{+11.2}$	$37.9_{+10.9}$	$42.3_{+15.3}$	$40.8_{+13.8}$	$34.1_{+7.1}$	$38.4_{+11.4}$	$56.1_{+29.1}$	$53.7_{+26.7}$
R	$35.3_{+8.3}$	$35.4_{+8.4}$	$38.9_{+11.9}$	$37.8_{+10.8}$	$31.1_{+4.1}$	$35.7_{+8.7}$	$43.5_{+16.5}$	$46.8_{+19.8}$
chrF scores for fr-en (Baseline: 53.01)								
+	$57.2_{+4.1}$	$58.6_{+5.6}$	$59.7_{+6.7}$	$58.8_{+5.8}$	$54.8_{+1.8}$	$57.4_{+4.4}$	$62.9_{+9.9}$	$65.3_{+12.3}$
μ	$55.4_{+2.4}$	$55.9_{+2.9}$	$57.2_{+4.2}$	$56.6_{+3.6}$	$53.1_{+0.1}$	$55.7_{+2.7}$	$59.8_{+6.7}$	$63.4_{+10.4}$
⊙	$59.4_{+6.4}$	$60.8_{+7.8}$	$62.1_{+9.1}$	$61.0_{+7.9}$	$56.6_{+3.6}$	$59.4_{+6.4}$	$70.8_{+17.8}$	$68.9_{+15.9}$
R	$57.4_{+4.4}$	$57.5_{+4.5}$	$59.9_{+6.9}$	$59.0_{+6.0}$	$54.8_{+1.8}$	$57.7_{+4.7}$	$62.3_{+9.3}$	$65.0_{+11.9}$

Zeroing. For de-en, the best configuration reaches 75.4 BLEU and 85.1 chrF (vs. 54.55/72.77 for the baseline), for fr-en 83.0 BLEU and 90.3 chrF (vs. 53.38/74.05), and for ar-en 82.8 BLEU and 90.0 chrF (vs. 59.45/77.96). In these cases, attribution-guided encoder attention markedly reduces the student–teacher gap, with absolute BLEU gains of + 20–30 points, comparable to the gains observed when training on human references, albeit from a higher baseline.

At a high level, the hierarchy of attribution maps remains similar to the human-reference setting. Across de-en, fr-en, and ar-en, the largest gains again come from Attention and ValueZeroing, followed by the gradient-based methods, with GSHAP consistently being the weakest. Aggregating across language pairs, the approximate ordering under the faithfulness objective is Attention ≈ ValueZeroing > LG×A > IG > Saliency > I×G ≈ DeepLIFT > GSHAP with slight nuances for ar-en where IG ranks lower. Because the baseline is already higher, some methods and operators fail to help the student model. For example, in Saliency for fr-en, addition

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

Table 16. BLEU and chrF scores for ar-en Marian-MT 4 head attribution injection. Scores followed by Δ over the baseline

BLEU scores for ar-en (Baseline: 40.68)								
Op.	I×G	Saliency	LG×A	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
+	$51.9_{+11.1}$	$51.6_{+10.8}$	$54.9_{+14.1}$	$51.8_{+11.0}$	$46.1_{+5.3}$	$52.7_{+11.9}$	$60.1_{+19.3}$	$63.0_{+22.2}$
μ	$47.5_{+6.7}$	$47.3_{+6.5}$	$50.3_{+9.5}$	$47.0_{+6.2}$	$43.8_{+3.0}$	$47.8_{+7.0}$	$54.6_{+13.8}$	$59.8_{+19.0}$
⊙	$54.9_{+14.1}$	$55.0_{+14.2}$	$58.1_{+17.3}$	$54.9_{+14.0}$	$49.0_{+8.2}$	$55.1_{+14.3}$	$66.9_{+26.1}$	$68.1_{+27.3}$
R	$51.9_{+11.1}$	$51.5_{+10.7}$	$55.1_{+14.3}$	$51.8_{+11.0}$	$46.2_{+5.4}$	$52.3_{+11.5}$	$60.7_{+19.9}$	$63.3_{+22.5}$
chrF scores for ar-en (Baseline: 66.07)								
+	$71.4_{+5.3}$	$71.2_{+5.1}$	$73.2_{+7.1}$	$71.3_{+5.3}$	$68.1_{+2.0}$	$71.8_{+5.8}$	$75.8_{+9.7}$	$77.8_{+11.7}$
μ	$68.9_{+2.8}$	$68.8_{+2.7}$	$70.4_{+4.4}$	$68.6_{+2.6}$	$66.9_{+0.9}$	$69.1_{+3.0}$	$72.6_{+6.5}$	$75.8_{+9.7}$
⊙	$73.0_{+6.9}$	$73.1_{+7.0}$	$74.9_{+8.8}$	$72.9_{+6.9}$	$69.7_{+3.6}$	$73.1_{+7.0}$	$79.9_{+13.9}$	$80.7_{+14.7}$
R	$71.4_{+5.3}$	$71.2_{+5.1}$	$73.3_{+7.2}$	$71.3_{+5.2}$	$68.1_{+2.0}$	$71.6_{+5.5}$	$76.2_{+10.1}$	$77.9_{+11.8}$

does not change the result from the baseline, whereas multiplication yields improvements. I×G and DeepLIFT can also degrade the results for fr-en with (μ) operator. From the operator's point of view, the element-wise product operator ($\odot$) yields the largest changes, simple averaging (μ) is almost always the weakest operator, and + and R lie in between. For example, in fr-en, the best $\odot$ configuration with ValueZeroing reaches 83.0 BLEU (+29.6) and 90.2 chrF (+16.2) relative to the baseline student, whereas the corresponding +, μ, and R configurations with ValueZeroing remain between 74.3–80.6 BLEU and 85.2–89.0 chrF. A similar pattern holds for de-en and ar-en. Table 17 (see Appendix) provides an example illustrating the output of encoder attention injection for Marian-MT, mBART and predicting Marian-MT generation.

5. Discussion

Because these attribution mappings come from trained models on larger datasets, they encode a learned relation between the source and target pairs, which otherwise cannot be learned by the student model. We conjectured that a better learned mapping can contribute to better results, and hence it can be a way of comparison between XAI attribution methods. In Subsection 4.2, we presented the main findings and contributions of this work. Expanding on our observations, we saw some consistent results for the injection of Marian-MT and mBART attributions into the encoder-attention, with some nuances for the mBART attributions. Particularly for the Marian-MT attributions, we saw improvements in almost all the attribution maps. As a sanity check, we conducted the same experiments with two settings: (1) Injecting random attribution maps taken from a uniform distribution and (2) using nearly diagonal matrices. Table 13 shows the results of these experiments:

The results of these experiments show that the network still learns, albeit with degraded performance. We conclude that non-sensical attribution maps can indeed reduce results, even when a pattern, such as a quasi-diagonal attribution matrix, is present, and that the increase is not accidental. There should be some meaningful alignment encoded by the XAI attribution maps.

While a linguistic and qualitative analysis of the differences between each attribution method, and whether there is a gold standard alignment in which these attributions approximate, is outside the scope of this work, we would like to examine some topological differences and similarities among the attribution matrices. For this reason, we treat each column of the matrices as a

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

probability distribution, from which we can calculate entropy as a measure of confidence in selecting the most important source tokens for the target tokens. Our manual inspection using heatmap visualization, such as Figure 1, showed that higher-scoring attribution methods exhibit more concentrated values around some tokens, and we observed more linear and diagonal patterns.

Figures 3–5 confirm that the three higher attribution maps, ValueZeroing, Attention, and LG×A, have lower entropies compared to other methods. For mBART, the entropy of gradient-based methods is higher than that of Marian-MT. One plausible explanation is that, in a deeper model with a longer computational graph, gradient signals become less sharp as they propagate toward earlier layers (Balduzzi *et al.* 2017). In particular, LG×A at the last layer of the encoder is more informative according to the results. For mBART, we observe that lower-scoring gradient-based attribution maps exhibit higher entropy than those of Marian-MT. Indeed, we can visually confirm that gradient-based methods have more chaotic representations. One point to note is that entropy does not necessarily completely correlate with the results. For example, for mBART, although the Entropy of Attention is lower, we observe that the last input token receives the most weight for all target tokens, which, in turn, leads to lower entropy. An observation that was confirmed by (Kobayashi *et al.* 2020), and in general, ValueZeroing scored higher for this model.

6. Attribution approximator

From the previous experiments, particularly those involving the attention mechanism (Attention attributions and ValueZeroing), we developed a hypothesis that the usefulness of an attribution method is largely determined by how geometrically close its maps are to those that a transformer can generate. This hypothesis motivates the introduction of a dedicated *Attributor* network. The Attributor is trained to approximate target–source attribution matrices corresponding to different explanation methods. Concretely, given a source sequence and its target sequence, the Attributor network learns to reconstruct the associated attribution matrix. We implement it as an encoder–decoder transformer so that the function class used to model attributions closely matches the inductive biases of the student translation model itself. Intuitively, if the student model can internally reproduce a particular attribution pattern, then it should also be better positioned to exploit that signal when the same attribution is injected into its attention layers, which in turn leads to more effective attribution-guided translation.

6.1 Attributor implementation

Our proposed *Attributor* network is a lightweight encoder–decoder transformer. The source sentence is tokenized and mapped to learned token embeddings, which are combined with learned positional embeddings and then processed by a 3-layer, 8-head self-attention encoder. The target sentence is embedded separately with its own learned positional embeddings and passed through 3 causal decoder layers. A standard triangular mask enforces autoregressive ordering to ensure that each target position attends only to previous target tokens. The resulting contextualized source and target embeddings are then fed into a modified cross-attention layer. For each target position of each attention head it returns a row attention score vector over source positions, which are essentially vectors of the per-head attention score matrices.

More formally, let H denote the number of heads and S the number of source positions. For each target token t, the cross-attention block produces per-head attention vectors $\vec{s}_{t,h} \in \mathbb{R}^S$:

$$\vec{s}_{t,h} = \frac{\vec{q}_{t,h} K_h^\top}{\sqrt{d_k}} + \text{mask}, \quad h = 1, \ldots, H, \tag{13}$$

where $\vec{q}_{t,h}, K_h \in \mathbb{R}^{d_k}$ are the query and key projections.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

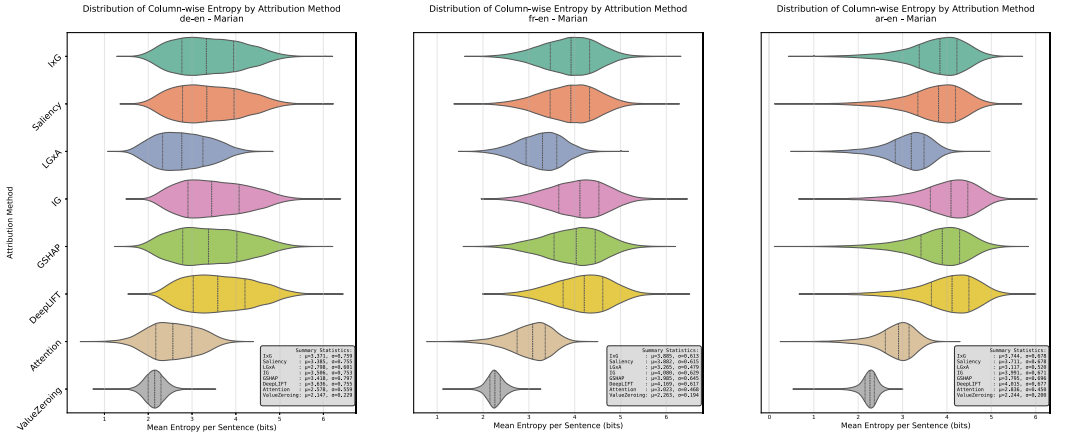


Figure 3. Column-wise entropy based on Marian-MT attributions.

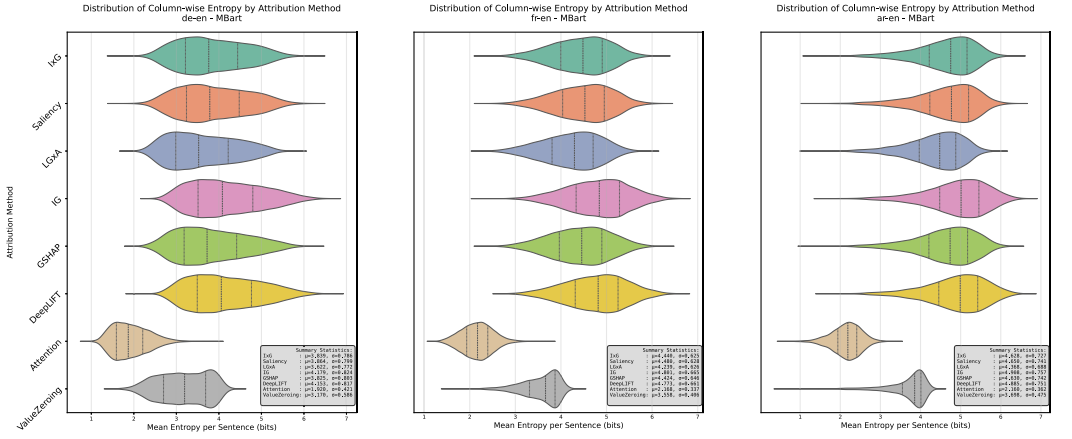


Figure 4. Column-wise entropy based on mBART attributions.

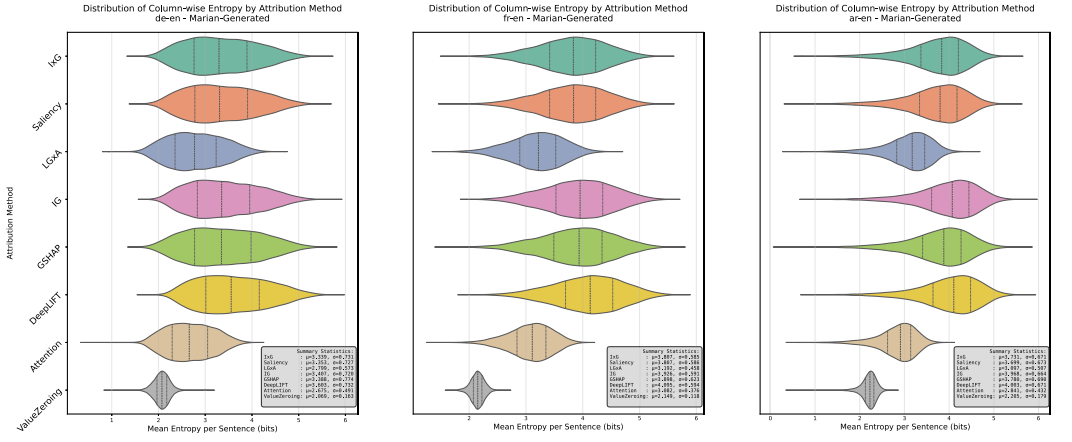


Figure 5. Column-wise entropy based on Marian-MT attributions (Marian-MT-generated targets).

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

To aggregate information across heads, we use a weighted mean of these score vectors across the head dimension. The weights are produced by a small MLP which maps each contextualized target token embedding $h_t \in \mathbb{R}^d$ to a probability vector over heads p_t:

$$\vec{p}_t = \text{softmax}\,(W_2\,\text{GELU}(W_1\vec{h}_t)) \in \mathbb{R}^H, \tag{14}$$

Finally, the combined logit vector over source positions for each target token is obtained as a weighted mean of the per-head attention scores vectors for that target token, after which it is passed to softmax to obtain a distribution $\hat{a}_t$ over source tokens:

$$\hat{\vec{a}}_t = \text{softmax}\left(\sum_{h=1}^{H} p_{t_h}\,\vec{s}_{t,h}\right) \in \mathbb{R}^S. \tag{15}$$

The attribution matrix $\hat{A}$ is essentially stacked $\hat{\vec{a}}_t$ for all t.

The objective function during the training is to minimize the average row-wise Kullback–Leibler divergence between the predicted attribution matrix $\hat{A}$ and the gold attribution matrix A, summed over non-padded target positions:

$$\mathcal{L}_{\text{KL}} = \frac{1}{T_{\text{valid}}} \sum_{t \in \mathcal{T}_{\text{valid}}} \text{KL}(A_t \,||\, \hat{A}_t), \tag{16}$$

where A_t and $\hat{A}_t$ denote the gold and predicted distributions over source tokens for target position t.

6.2 Attributor evaluation

We trained the Attributor on three datasets derived from our earlier experiments, covering all three language pairs. (a) A dataset constructed from Marian-MT attribution maps paired with gold (human) target translations. (b) A dataset constructed from mBART attribution maps paired with gold (human) target translations. (c) A faithfulness-oriented dataset in which Marian-MT attribution maps are paired with targets generated by the Marian-MT teacher model.

As a part of preprocessing, source and target sentences were tokenized according to the respective model's tokenizers, and attribution maps were normalized in the source dimension to represent a valid probability distribution (to accommodate the KL-divergence loss function). To assess the accuracy of the approximation of their attribution maps by the Attributor, we selected the following metrics:

- **KL-divergence**. Being the minimization target, it is the first choice to quantify the difference between the approximated and the gold attributions
- **Overlap@3** The overlap between sets of top-3 valued source token indices per target token for the approximated and gold attribution maps. As highlighted by the Nourbakhsh *et al.* (2025) we hypothesized that most of the useful signal attributions per target token comes from a small number of the highest-scoring source tokens. The overlap@3 is calculated as

$$\text{overlap@3}(t) = \frac{|\text{Top}_3(a_t) \cap \text{Top}_3(\hat{a}_t)|}{3} \tag{17}$$

- **Tau@3** Kendall's τ calculated on top-3 values over the source dimension of the gold attribution and values with the same indices but from the approximated attributions. This metric accompanies overlap@3 by quantifying the rank correlation.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

6.3 Attributor results

Obtained values for de-en, fr-en, and ar-en pairs are presented in the three settings mentioned above and can be seen in Figures 6–8.

To quantify the link between approximability and downstream gains, we correlate the BLEU scores obtained by Marian-MT when augmented with each attribution method with our three approximability metrics (mean KL divergence, Overlap@3 and Kendall's τ@3 between gold and predicted target–source maps). We compute BLEU separately for each injection operator (add, average, multiply, replace) and also consider the best BLEU per method across operators.

Across all three language pairs (ar-en, de-en, fr-en) and all operators, BLEU shows a very strong positive correlation with the 'top-3' alignment metrics. Pearson's (r) between BLEU and Overlap@3 lies in the range (r $\approx$ 0.88–0.97), and Kendall's (τ@3) achieves (r $\approx$ 0.74–0.95). In other words, between roughly 75% and 90% of the variance in BLEU across attribution methods can be explained by how well their target–source patterns can be reconstructed by the Attributor. Rank correlations show identical results: Spearman's (ρ) between BLEU and Overlap@3/(τ@3) is consistently high (typically $\rho \approx 0.65$–0.85), and for the fr-en pair with multiplication injection the ranking of methods by BLEU is *exactly* the same as the ranking by τ@3 ($\rho = 1.0$).

In contrast, KL divergence exhibits only weak and unstable association with BLEU. Pearson correlations between BLEU and mean KL are low–moderate and positive (r $\approx$ 0.27–0.56), while Spearman's (ρ) tends to be low negative and even changes sign for some operators ($\rho \approx -0.26$–0.11). This behavior, especially for Pearson, is largely driven by outliers such as ValueZeroing, which combines very high KL divergence with strong BLEU gains. These results suggest that global distribution similarity is a poor predictor of downstream effectiveness, whereas agreement on the locations and ordering of a few top-k source positions is highly predictive.

Taken together, these findings support our central claim: attribution methods whose target-source maps are closer to what an encoder-decoder transformer (Attributor) can reproduce are exactly the ones that yield the largest BLEU gains when injected into the student model. We first measured this closeness with full-distribution KL divergence, but the correlation analysis shows that the *best predictors of BLEU* are Overlap@3 and Kendall's τ@3 between the Attributor's predictions and the gold attribution maps, which correlate very strongly with BLEU, whereas KL shows only weak and unstable correlations. These results suggest that an attribution method tends to be useful precisely when a transformer can reliably recover the same few most-salient source tokens per target token as the teacher. In contrast, matching the full attention distribution beyond the top positions appears much less informative for predicting BLEU gains.

For the other two settings of generated targets, the manifested trend is the same. For mBART, the highest correlation is observed with the multiplication operator. It should be noted that the approximation accuracy is the highest for Attention from mBART: the last source token consistently gets the highest attribution scores and their reconstruction is easier. However, there are still residual scores for other tokens whose presence can guide the translation. The regression plot of the above approximation metrics for the three language pairs and the corresponding outcomes is provided in the figures 9–11 (see Appendix).

In short, the student benefits from what it can imitate: attribution methods for which target tokens a transformer can reliably reconstruct the same top-3 source tokens are precisely those that give student models the largest gains.

7. Conclusion and future work

In this work, we demonstrate that XAI attribution maps derived from a learned teacher model can be injected into a student model, and that the resulting changes in translation behavior provide a practical signal of the relative quality of the XAI attribution methods. We conducted an extensive

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

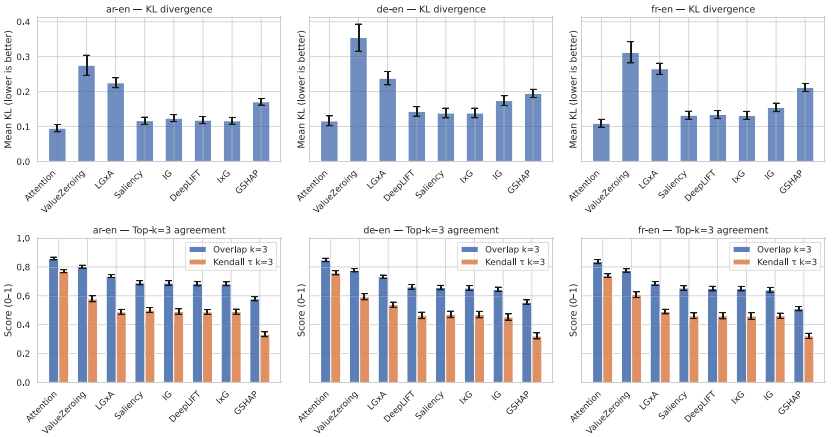


Figure 6. KL-divergence, Overlap@3, and τ@3 for the prediction of attribution on Marian-MT attribution and gold (human) data.

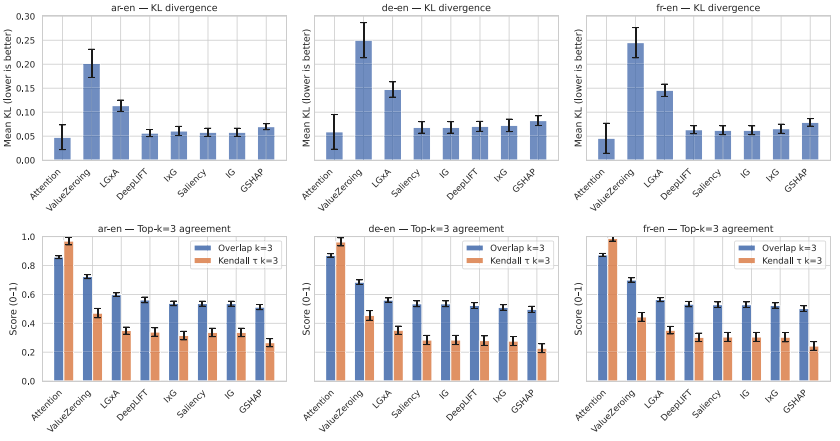


Figure 7. KL-divergence, Overlap@3, and τ@3 for the prediction of attribution on mBART attribution and gold (human) data.

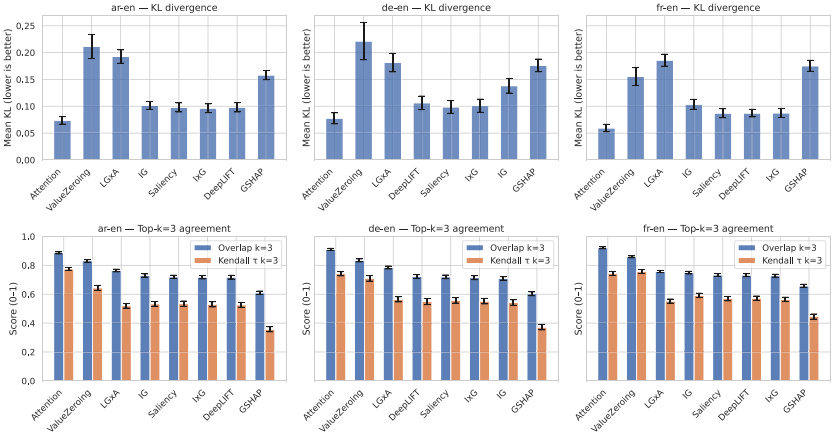


Figure 8. KL-divergence, Overlap@3, and τ@3 for the prediction of attribution on Marian-MT attribution generated target.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

evaluation across German–English, French–English, and Arabic–English language pairs, comparing eight XAI methods and multiple composition strategies for integrating attribution scores into Transformer attention (addition, multiplication, averaging, and replacement). We further examined where the scores are applied (encoder self-attention *vs.* cross-attention) and assessed robustness across two distinct teacher models, mBART and Marian-MT.

Our results indicate that Attention and ValueZeroing, as well as LG×A extracted from the final encoder layer, consistently produced the largest gains in BLEU and chrF. We also observed that injecting source–target attributions into the encoder self-attention can yield significant improvements, which is initially counterintuitive given that the encoder is designed to attend only within the source sequence. A plausible explanation is that the non-auto-regressive encoder attention setup makes future information accessible to the entire sequence.

We also showed that getting better results from the different attribution methods is not accidental: (1) Attribution maps with lower entropy tend to score higher. (2) We set up a transformer model, *Attributor*, which shows that the most beneficial attribution maps are the ones for whose target tokens the transformer model successfully recreates top-3 salient source tokens. These findings offer the NLP community a plausible explanation for the utility of attribution maps and highlight the importance of the attribution's most salient tokens.

There are some limitations to this work worth noting. First, this study provides an extensive exploration of attribution transfer between models. However, the overall pipeline is computationally expensive, both in deriving attributions at scale and in retraining student models. For this reason, we focused our comparison on attribution signals produced by a range of explainability methods, most of which are gradient-based. This choice was driven largely by practicality. Extracting attributions with perturbation-based approaches such as LIME (Ribeiro *et al.* 2016) and reAGent (Zhao, Wang, and Wang 2023) is substantially more resource-intensive and time-consuming. In addition, some methods available in Inseq produce (self-)attributions on the decoder side of seq2seq models. In this work, we restrict our experiments to encoder self-attention and encoder–decoder cross-attention, leaving decoder-side attribution transfer for future study.

More ablation studies are needed. In this work, we tried with cross-attention, 4 attention heads, and different operators. However, the effects of layer-wise injection, broader choices of head selection and head count, and alternative placement strategies remain underexplored. We also used a coarse attribution extraction procedure. We took attention-based attributions from all the layers and averaged attention scores, and for the gradient methods, our only alteration was the last encoder layer. More systematic investigation of layer-wise signals and aggregation strategies is needed. Moreover, it is important to note that, for the Attributor experiments, we used the top-3 tokens for overlap@3 and Kendall's τ; it may also be beneficial to extend the experiments to other top-k tokens to draw a line for the most salient tokens more precisely. Also, in this work, our focus was limited to machine translation tasks. Machine translation tasks are characterized by the alignments between source and target pairs. Future work could extend this evaluation framework to other generative models, including those applied in question answering and text summarization.

Funding statement. This work is supported by the Luxembourg National Research Fund (FNR) as part of the project C21 - Collaboration 21: IPBG2020/IS/14839977/C21.

Competing interests. The authors declare that the manuscript complies with the ethical standards of the journal and that there are no competing interests to disclose.

Acknowledgements. We thank Richard Albrecht for early assistance and helpful discussions during project initiation.

Artificial intelligence usage. We used Large Language Models, such as ChatGPT and Copilot, for code autocompletion, proofreading, and spellchecking of the manuscript.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

References

Arras L., **Horn F.**, **Montavon G.**, **Müller K.-R.** and **Wojciech S.** (2016). Explaining predictions of non-linear classifiers in NLP. In Phil B., Kyunghyun C., Shay C., Edward G., Karl M.H., Laura R., Jason W. and Scott W.-T.Y. (eds), *Proceedings of the 1st Workshop on Representation Learning for NLP*. Berlin, Germany: Association for Computational Linguistics, pp. 1–7. https://doi.org/10.18653/v1/W16-1601.

Arya V., **Bellamy R.K.E.**, **Chen P.-Y.**, **Dhurandhar A.**, **Hind M.**, **Hoffman S.C.**, **Houde S.**, **Liao Q.V.**, **Luss R.**, **Mojsilović A.**, **Mourad S.**, **Pedemonte P.**, **Raghavendra R.**, **Richards J.**, **Sattigeri P.**, **Shanmugam K.**, **Singh M.**, **Varshney K.R.**, **Wei D.** and **Zhang Y.** (2019). One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. arXiv preprint arXiv:1909.03012 [cs.AI]. https://arxiv.org/abs/1909.03012.

Atanasova P., **Simonsen J.G.**, **Lioma C.** and **Augenstein I.** (2020). A diagnostic study of explainability techniques for text classification. In Bonnie W., Trevor C., Yulan H. and Yang L. (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 3256–3274. https://doi.org/10.18653/v1/2020.emnlp-main.263.

Bahdanau D., **Cho K.** and **Bengio Y.** (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. San Diego, CA. arXiv:1409.0473. https://arxiv.org/abs/1409.0473.

Bai J., **Wang Y.**, **Sun H.**, **Wu R.**, **Yang T.**, **Tang P.**, **Cao D.**, **Zhang M.**, **Tong Y.**, **Yang Y.**, **Bai J.**, **Zhang R.**, **Sun H.** and **Shen W.** (2022). Enhancing self-attention with knowledge-assisted attention maps. In Marine C., Marie-Catherine de M. and Ivan V.M.R. (eds), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 107–115. https://doi.org/10.18653/v1/2022.naacl-main.8. https://aclanthology.org/2022.naacl-main.8/.

Balduzzi D., **Frean M.**, **Lewis J.P.**, **Leary L.**, **Kurt W.-D.M.** and **McWilliams B.** (2017). The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning*. PMLR, pp. 342–350.

Barredo Arrieta A., **Díaz-Rodríguez N.**, **Del Ser J.**, **Bennetot A.**, **Tabik S.**, **Barbado A.**, **Garcia S.**, **Gil-Lopez S.**, **Molina D.**, **Benjamins R.**, **Chatila R.** and **Herrera F.** (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58**, 82–115.

Bojar O., **Buck C.**, **Federmann C.**, **Haddow B.**, **Koehn P.**, **Leveling J.**, **Monz C.**, **Pecina P.**, **Post M.**, **Saint-Amand H.**, **Soricut R.**, **Specia L.** and **Tamchyna A.** (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 12–58. https://doi.org/10.3115/v1/W14-3302.

Brown P.F., **Pietra V.J.D.**, **Pietra S.A.D.** and **Mercer R.L.** (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* **19**(2), 263–311.

Bugliarello E. and **Okazaki N.** (2020). Enhancing machine translation with dependency-aware self-attention. In Dan J., Joyce C., Natalie S. and Joel T. (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 1618–1627. https://doi.org/10.18653/v1/2020.acl-main.147.

Burkart N. and **Huber M.F.** (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* **70**, 245–317.

Chang C.-H., **Creager E.**, **Goldenberg A.** and **Duvenaud D.** (2019). Explaining image classifiers by counterfactual generation. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*. New Orleans, LA. arXiv:1807.08024. https://arxiv.org/abs/1807.08024.

Chang Y., **Wang X.**, **Wang J.**, **Wu Y.**, **Yang L.**, **Zhu K.**, **Chen H.**, **Yi X.**, **Wang C.**, **Wang Y.**, **Ye W.**, **Zhang Y.**, **Chang Y.**, **Yu P.S.**, **Yang Q.** and **Xie X.** (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* **15**(3), 1–45.

Cho K., **van Merriënboer B.**, **Gulcehre C.**, **Bahdanau D.**, **Bougares F.**, **Schwenk H.** and **Bengio Y.** (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro M., Bo P. and Walter D. (eds), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734. https://doi.org/10.3115/v1/D14-1179.

Dale D., **Voita E.**, **Barrault L.** and **Costa-Jussà M.R.** (2023). Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In Anna R., Jordan B.-G. and Naoaki O. (eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 36–50. https://doi.org/10.18653/v1/2023.acl-long.3.

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

Denil M., **Demiraj A.** and **de Freitas N.** (2014). Extraction of salient sentences from labelled documents. Technical Report, University of Oxford. arXiv:1412.6815. https://arxiv.org/abs/1412.6815.

De Young J., **Jain S.**, **Rajani N.F.**, **Lehman E.**, **Xiong C.**, **Socher R.** and **Wallace B.C.** (2020). ERASER: A benchmark to evaluate rationalized NLP models. In Dan J., Joyce C., Natalie S. and Joel T. (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 4443–4458. https://doi.org/10.18653/v1/2020.acl-main.408.

Ding S., **Xu H.** and **Koehn P.** (2019). Saliency-driven word alignment interpretation for neural machine translation. In Ondrej B., Rajen C., Christian F., Mark F., Yvette G., Barry H., Matthias H., et al. (eds), *Proceedings of the Fourth Conference*

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

on Machine Translation (Volume 1: Research Papers). Florence, Italy: Association for Computational Linguistics, pp. 1–12. https://doi.org/10.18653/v1/W19-5201.

Doshi-Velez F. and **Kim B.** (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 [stat.ML]. https://arxiv.org/abs/1702.08608.

Doshi-Velez F. and **Kim B.** (2018). Considerations for evaluation and generalization in interpretable machine learning. In Hugo J.E., Sergio E., Isabelle G., Xavier B., Yagmur G., Umut G. and Marcel van G. (eds), *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Cham: Springer International Publishing, pp. 3–17. https://doi.org/10.1007/978-3-319-98131-4_1.

Dwivedi R., **Dave D.**, **Naik H.**, **Singhal S.**, **Omer R.**, **Patel P.**, **Qian B.**, **Wen Z.**, **Shah T.**, **Morgan G.** and **Ranjan R.** (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys* **55**(9), 1–33.

Dyer C., **Chahuneau V.** and **Smith N.A.** (2013). A simple, fast, and effective reparameterization of IBM Model 2. In Vanderwende L., Daumé III H. and Kirchhoff K. (eds), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 644–648. https://aclanthology.org/N13-1073/.

Eksi M., **Gelbing E.**, **Stieber J.** and **Vu C.V.** (2021). Explaining errors in machine translation with absolute gradient ensembles. In Yang G., Steffen E., Wei Z., Piyawat L. and Marina F. (eds), *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 238–249. https://doi.org/10.18653/v1/2021.eval4nlp-1.23.

Fantozzi P. and **Naldi M.** (2024). The explainability of transformers: current status and directions. *Computers* **13**(4), 92.

Ferrando J. and **Costa-Jussà M.R.** (2021). Attention weights in transformer NMT fail aligning words between sequences but largely explain model predictions. In Marie-Francine M., Xuanjing H., Lucia S. and Scott W.-T. Y. (eds), *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 434–443. https://doi.org/10.18653/v1/2021.findings-emnlp.39.

Ferrando J., **Gállego G.I.**, **Alastruey B.**, **Escolano C.** and **Costa-Jussà M.R.** (2022). Towards opening the black box of neural machine translation: source and target interpretations of the transformer. In Yoav G., Zornitsa K. and Yue Z. (eds), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. United Arab Emirates: Abu Dhabi, Association for Computational Linguistics, pp. 8756–8769. https://doi.org/10.18653/v1/2022.emnlp-main.599.

Fomicheva M., **Specia L.** and **Aletras N.** (2022). Translation error detection as rationale extraction. In Smaranda M., Preslav N. and Aline V. (eds), *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 4148–4159. https://doi.org/10.18653/v1/2022.findings-acl.327.

Ghader H. and **Monz C.** (2017). What does attention in neural machine translation pay attention to? In Greg K. and Taro W. (eds), *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 30–39. https://aclanthology.org/I17-1004/.

Gurrapu S., **Kulkarni A.**, **Huang L.**, **Lourentzou I.** and **Batarseh F.A.** (2023). Rationalization for explainable NLP: A survey. *Frontiers in Artificial Intelligence* **6**, 1225093.

Hase P. and **Bansal M.** (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In Dan J., Joyce C., Natalie S. and Joel T. (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 5540–5552. https://doi.org/10.18653/v1/2020.acl-main.491.

He S., **Tu Z.**, **Wang X.**, **Wang L.**, **Lyu M.** and **Shi S.** (2019). Towards understanding neural machine translation with word importance. In Kentaro I., Jing J., Vincent N. and Xiaojun W. (eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 953–962. https://doi.org/10.18653/v1/D19-1088.

Hinton G., **Vinyals O.** and **Dean J.** (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 [stat.ML]. https://arxiv.org/abs/1503.02531.

Hooker S., **Erhan D.**, **Kindermans P.-J.** and **Kim B.** (2019). A benchmark for interpretability methods in deep neural networks. In Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E. and Garnett R. (eds), *Advances in Neural Information Processing Systems*, vol. 32. Red Hook, NY: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

Jacovi A. and **Goldberg Y.** (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan J., Joyce C., Natalie S. and Joel T. (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386.

Jain S. and **Wallace B.C.** (2019). Attention is not Explanation. In Jill B., Christy D. and Thamar S. (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3543–3556. https://doi.org/10.18653/v1/N19-1357.

Jiao X., **Yichun Y.**, **Lifeng S.**, **Xin J.**, **Xiao C.**, **Linlin L.**, **Wang F.** and **Liu Q.** (2020). TinyBERT: Distilling BERT for natural language understanding. In Trevor C., Yulan H. and Yang L. (eds), *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 4163–4174. https://doi.org/10.18653/v1/2020.findings-emnlp.372.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

Kalyan K.S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal* **6**, 100048.

Kim J., **Maathuis H.** and **Sent D.** (2024). Human-centered evaluation of explainable AI applications: A systematic review. *Frontiers in Artificial Intelligence* **7**, 1456486.

Kobayashi G., **Kuribayashi T.**, **Yokoi S.** and **Inui K.** (2020). Attention is not only a weight: Analyzing transformers with vector norms. In Bonnie W., Trevor C., Yulan H. and Yang L. (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 7057–7075. https://doi.org/10.18653/v1/2020.emnlp-main.574.

Langedijk A., **Mohebbi H.**, **Sarti G.**, **Zuidema W.** and **Jumelet J.** (2024). DecoderLens: Layerwise interpretation of encoder-decoder transformers. In Kevin D., Helena G. and Steven B. (eds), *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics, pp. 4764–4780. https://doi.org/10.18653/v1/2024.findings-naacl.296.

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

Li J., **Liu L.**, **Li H.**, **Li Gu.**, **Huang G.** and **Shi S.** (2020). Evaluating explanation methods for neural machine translation. In Dan J., Joyce C., Natalie S. and Joel T. (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 365–375. https://doi.org/10.18653/v1/2020.acl-main.35.

Li X., **Li G.**, **Liu L.**, **Meng M.** and **Shi S.** (2019). On the word alignment from neural machine translation. In Anna K., David T. and Lluís M. (eds), *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1293–1303. https://doi.org/10.18653/v1/P19-1124.

Liu Y., **Gu J.**, **Goyal N.**, **Li X.**, **Edunov S.**, **Ghazvininejad M.**, **Lewis M.** and **Zettlemoyer L.** (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* **8**, 726–742.

Lundberg S.M. and **Lee S.-I.** (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, Vol. **30**. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Madsen A., **Meade N.**, **Adlakha V.** and **Reddy S.** (2022). Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining. In Yoav G., Zornitsa K. and Yue Z. (eds), *Findings of the Association for Computational Linguistics: EMNLP 2022*. United Arab Emirates: Abu Dhabi, Association for Computational Linguistics, pp. 1731–1751. https://doi.org/10.18653/v1/2022.findings-emnlp.125.

Madsen A., **Reddy S.** and **Chandar S.** (2022). Post-hoc interpretability for neural NLP: A survey. *ACM Computing Surveys* **55**(8), 1–42.

Meister C., **Lazov S.**, **Augenstein I.** and **Cotterell R.** (2021). Is sparse attention more interpretable? In Chengqing Z., Fei X., Wenjie L. and Roberto N. (eds), *Proceedings of the 57th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, pp. 122–129. https://doi.org/10.18653/v1/2021.acl-short.17.

Mohammadi H., **Bagheri A.**, **Giachanou A.** and **Oberski D.L.** (2025). Explainability in practice: A survey of explainable NLP across various domains. arXiv preprint arXiv:2502.00837 [cs.CL]. https://arxiv.org/abs/2502.00837.

Mohebbi H., **Zuidema W.**, **Chrupała G.** and **Alishahi A.** (2023). Quantifying context mixing in transformers. In Andreas V. and Isabelle A. (eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 3378–3400 . https://doi.org/10.18653/v1/2023.eacl-main.245.

Moradi P., **Kambhatla N.** and **Sarkar A.** (2021). Measuring and improving faithfulness of attention in neural machine translation. In Paola M., Jorg T. and Reut T. (eds), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, pp. 2791–2802. https://doi.org/10.18653/v1/2021.eacl-main.243.

Nauta M., **Trienes J.**, **Pathak S.**, **Nguyen E.**, **Peters M.**, **Schmitt Y.**, **Schlötterer J.**, **van Keulen M.** and **Seifert C.** (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys* **55**(13s), 1–42.

Nguyen D. (2018). Comparing automatic and human evaluation of local explanations for text classification. In Marilyn W., Heng J. and Amanda S. (eds), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1069–1078. https://doi.org/10.18653/v1/N18-1097.

Nourbakhsh A., **Lamsiyah S.** and **Schommer C.** (2025). Quantifying the overlap: Attribution maps and linguistic heuristics in encoder-decoder machine translation models. In Angelova G., Kunilovskaya M., Escribe M. and Mitkov R. (eds), *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing: Natural Language Processing in the Generative AI Era*. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, pp. 822–831. https://aclanthology.org/2025.ranlp-1.94/.

Och F.J. and **Ney H.** (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics* **29**(1), 19–51.

Papineni K., **Roukos S.**, **Ward T.** and **Zhu W.-J.** (2002). Bleu: A method for automatic evaluation of machine translation. In Isabelle P., Charniak E. and Lin D. (eds), *Proceedings of the 40th Annual Meeting of the Association for*

Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. https://doi.org/10.3115/1073083.1073135.

Perrella S., **Proietti L.**, **Cabot P.-L.H.**, **Barba E.** and **Navigli R.** (2024). Beyond correlation: Interpretable evaluation of machine translation metrics. In Yaser A.-O., Mohit B. and Yun-Nung C. (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 20689–20714. https://doi.org/10.18653/v1/2024.emnlp-main.1152.

Popović M. (2015). ChrF: Character n-gram F-score for automatic MT evaluation. In Ondrej B., Rajan C., Christian F., Barry H., Chris H., Matthias H., Varvara L. and Pavel P. (eds), *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. https://doi.org/10.18653/v1/W15-3049.

Pruthi D., **Bansal R.**, **Dhingra B.**, **Soares L.B.**, **Collins M.**, **Lipton Z.C.**, **Neubig G.** and **Cohen W.W.** (2022). Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics* **10**, 359–375.

Raganato A. and **Tiedemann J.** (2018). An analysis of encoder representations in transformer-based machine translation. In Tal L., Grzegorz C. and Afra A. (eds), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 287–297. https://doi.org/10.18653/v1/W18-5431.

Ramachandran P., **Zoph B.** and **Le Q.V.** (2017). Searching for activation functions. arXiv preprint arXiv:1710.05941 [cs.NE]. https://arxiv.org/abs/1710.05941.

Ribeiro M.T., **Singh S.** and **Guestrin C.** (2016). '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*. New York, NY: Association for Computing Machinery, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.

Saeed W. and **Omlin C.** (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems* **263**, 110273. https://doi.org/10.1016/j.knosys.2023.110273.

Sarti G., **Feldhus N.**, **Sickert L.** and **van der Wal O.** (2023). Inseq: An interpretability toolkit for sequence generation models. In Danushka B., Ruihong H. and Alan R. (eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Toronto, Canada: Association for Computational Linguistics, pp. 421–435. https://doi.org/10.18653/v1/2023.acl-demo.40.

Shakil H., **Farooq A.** and **Kalita J.** (2024). Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing* **603**, 128255.

Shrikumar A., **Greenside P.** and **Kundaje A.** (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning – Volume 70*. Sydney, NSW, Australia: JMLR.org, pp. 3145–3153.

Simonyan K., **Vedaldi A.** and **Zisserman A.** (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at the 2nd International Conference on Learning Representations (ICLR 2014)*. Banff, Canada. arXiv:1312.6034. https://arxiv.org/abs/1312.6034.

Slobodkin A., **Choshen L.** and **Abend O.** (2022). Semantics-aware attention improves neural machine translation. In Vivi N., Ellie P., Mohammad T.P., Jose C.-C. and Alessandro R. (eds), *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*. Seattle, Washington: Association for Computational Linguistics, pp. 28–43. https://doi.org/10.18653/v1/2022.starsem-1.3.

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

Sundararajan M., **Taly A.** and **Yan Q.** (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, pp. 3319–3328.

Sutskever I., **Vinyals O.** and **Le Q.V.** (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems– Volume 2, NIPS'14*. Montreal, Canada: MIT Press, pp. 3104–3112.

Tang Y., **Tran C.**, **Li X.**, **Chen P.-J.**, **Goyal N.**, **Chaudhary V.**, **Gu J.** and **Fan A.** (2020). Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401 [cs.CL]. https://arxiv.org/abs/2008.00401.

Tiedemann J. and **Thottingal S.** (2020). OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, pp. 479–480. https://aclanthology.org/2020.eamt-1.61

Vaswani A., **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A.N.**, **Kaiser Ł.** and **Polosukhin I.** (2017). Attention is all you need. In Guyon I., von Luxburg U., Bengio S., Wallach H., Fergus R., Vishwanathan S. and Garnett R. (eds), *Advances in Neural Information Processing Systems*, vol. 30. Red Hook, NY: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vieira C.P. and **Digiampietri L.A.** (2022). Machine learning post-hoc interpretability: A systematic mapping study. In *Proceedings of the XVIII Brazilian Symposium on Information Systems*. New York, NY: Association for Computing Machinery, article no. 1, pp. 1–8. https://doi.org/10.1145/3535511.3535512.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

Voita E., **Sennrich R.** and **Titov I.** (2021). Analyzing the source and target contributions to predictions in neural machine translation. In Chengqing Z., Fei X., Wenjie L. and Roberto N. (eds), *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)*. Association for Computational Linguistics, pp. 1126–1140. https://doi.org/10.18653/v1/2021.acl-long.91.

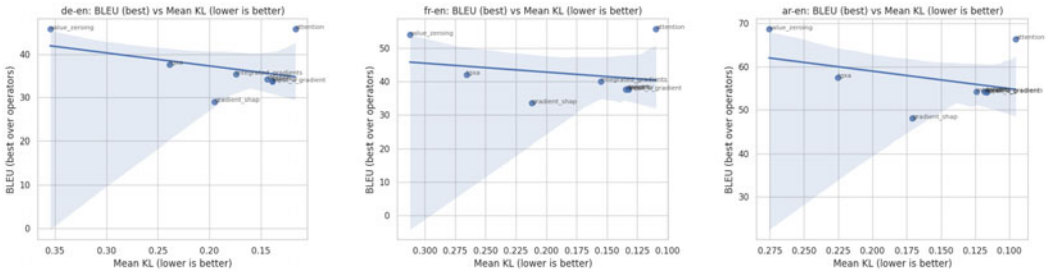
Zenkel T., **Wuebker J.** and **DeNero J.** (2019). Adding interpretable attention to neural translation models improves word alignment. arXiv preprint arXiv:1901.11359 [cs.CL]. https://arxiv.org/abs/1901.11359.

Zhao Z. and **Shan B.** (2024). ReAGent: A model-agnostic feature attribution method for generative language models. arXiv preprint arXiv:2402.00794 [cs.CL]. https://arxiv.org/abs/2402.00794.

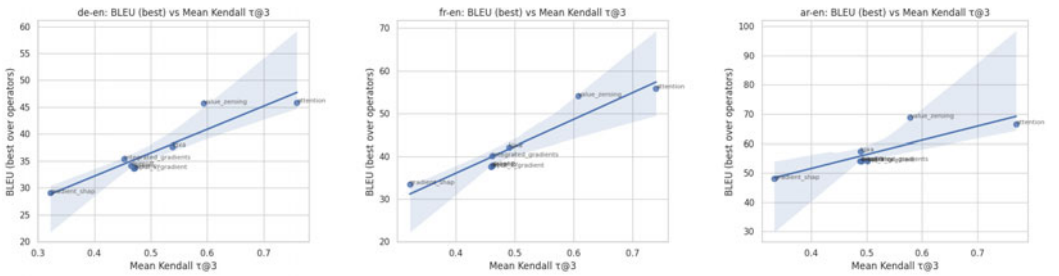
Zhao Z., **Wang Y.** and **Wang Y.** (2023). Knowledge-aware Bayesian co-attention for multimodal emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5.

Ziemski M., **Junczys-Dowmunt M.** and **Pouliquen B.** (2016). The United Nations parallel corpus v1.0. In Nicoletta C., Khalid C., Thierry D., Sara G., Marko G., Bente M., Joseph M., et al. (eds), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3530–3534. https://aclanthology.org/L16-1561/

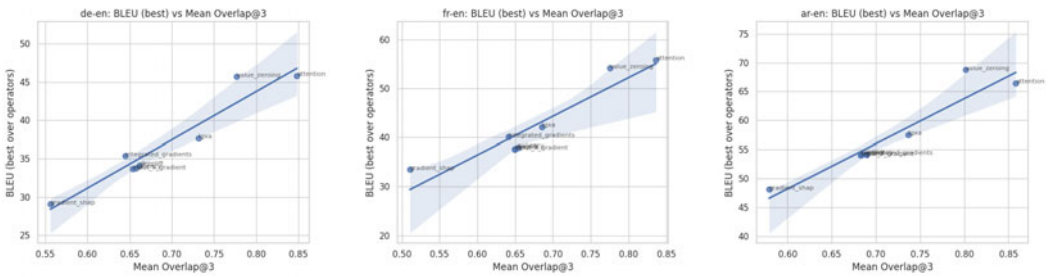
Appendix



(a) Mean KL (de-en, fr-en, ar-en).



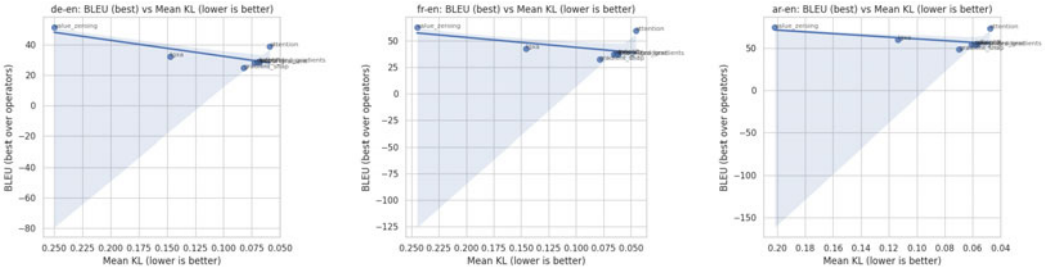
(b) Mean Kendall τ@3 (de-en, fr-en, ar-en).



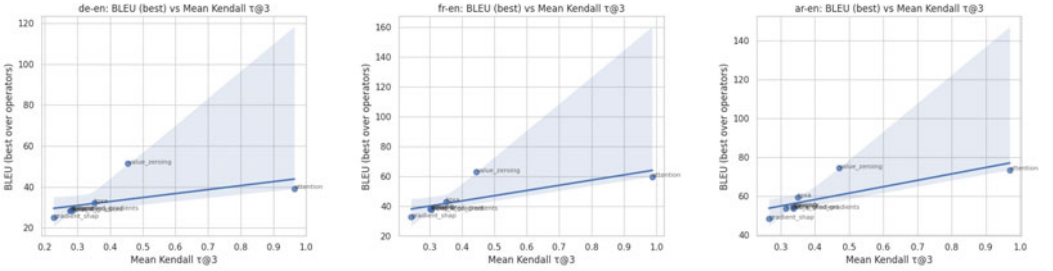
(c) Mean Overlap@3 (de-en, fr-en, ar-en).

Figure 9. Regression plots of the Marian-MT attributions with gold (human) target sentence.

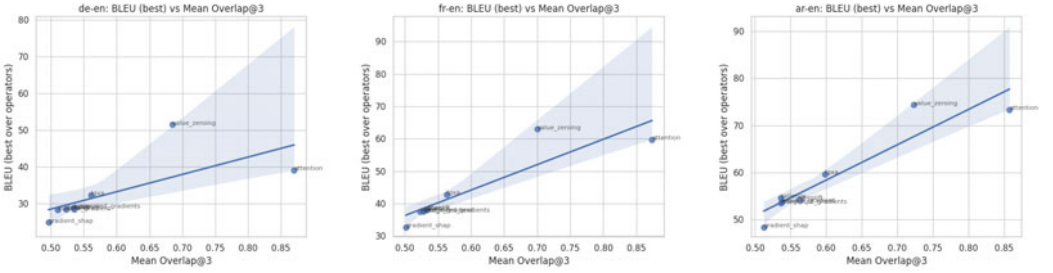
https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press



(a) Mean KL (de-en, fr-en, ar-en).



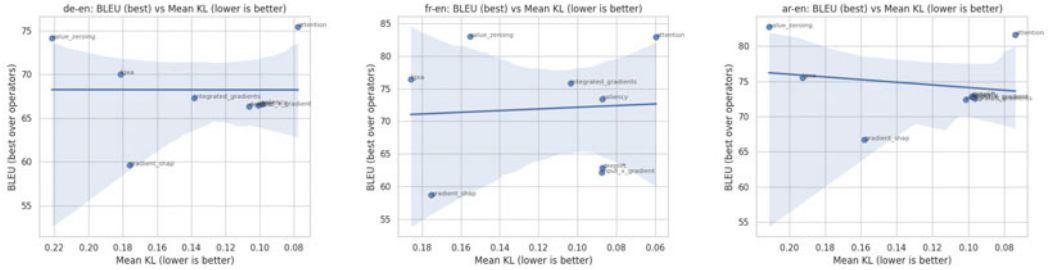
(b) Mean Kendall τ@3 (de-en, fr-en, ar-en).



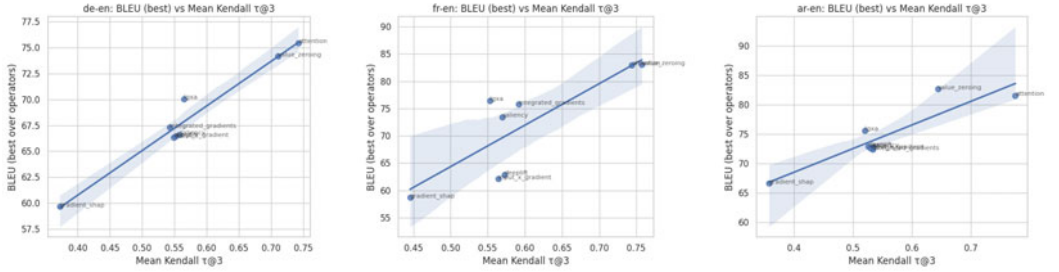
(c) Mean Overlap@3 (de-en, fr-en, ar-en).

Figure 10. Regression plots of the mBART attributions with gold (human) target sentence.

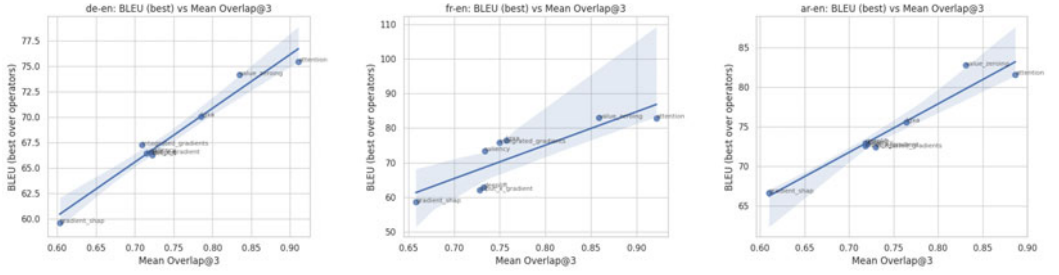
https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press



(a) Mean KL (de-en, fr-en, ar-en).



(b) Mean Kendall τ@3 (de-en, fr-en, ar-en).



(c) Mean Overlap@3 (de-en, fr-en, ar-en).

Figure 11. Regression plot of the Marian-MT attributions with generated target sentences.

https://doi.org/10.1017/nlp.2026.10028 Published online by Cambridge University Press

Table 17. Sample translations and metrics for Marian-MT and mBART under different attribution/XAI variants for a sample de-en.

Model	Text	BLEU	chrF
Marian-MT			
Source:	Ich dachte, dass es damit beendet wäre.	–	–
Reference:	I thought that would be the end of it.	–	–
Baseline:	I thought it was finished.	9.91	29.19
I×G	I thought it would have been finished with it.	12.55	43.76
Saliency	I thought it would have been finished with it.	12.55	43.76
LG×A	I thought that it had been finished with it.	19.64	47.02
IG	I thought it would have been over with it.	12.55	43.29
GSHAP	I thought it would be finished with it.	14.78	47.72
DeepLIFT	I thought it would have been finished with it.	12.55	43.76
Attention	I thought that would have an end to it.	32.47	61.15
ValueZeroing	I thought this might be the end of it.	58.14	65.49
mBART			
Reference:	I thought that would be the end of it.	–	–
I×G	I thought it'd be done to it.	12.86	33.19
Saliency	I thought it'd be done to it.	12.86	33.19
LG×A	I thought it'd be done with it.	12.86	33.28
IG	I thought it'd be done to it.	12.86	33.19
GSHAP	I thought it was the only way it was.	10.55	29.46
DeepLIFT	I thought it would end to it for it.	13.13	41.71
Attention	I thought it would be the end of it.	70.71	80.71
ValueZeroing	I thought it would be the end over it.	35.49	68.34
Marian-MT (Generated)			
Reference:	I thought it was over.	–	–
I×G	I thought it would be.	32.47	57.06
Saliency	I thought it was over.	100	100
LG×A	I thought it was over.	100	100
IG	I thought it would end.	32.47	56.33
GSHAP	I thought it would be.	32.47	57.06
DeepLIFT	I thought it would be.	32.47	57.06
Attention	I thought it was over.	100	100
ValueZeroing	I thought it was over.	100	100

Cite this article: Nourbakhsh A, Lamsiyah S, Danilov A and Schommer C (2026). Evaluating explainable AI attribution methods in neural machine translation via attention-guided knowledge distillation. *Natural Language Processing* **32**, 267–301. https://doi.org/10.1017/nlp.2026.10028