# Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection

**Weijia Xu**
Microsoft Research, Redmond, USA
weijiaxu@microsoft.com

**Sweta Agrawal**
University of Maryland, USA
sweagraw@cs.umd.edu

**Eleftheria Briakou**
University of Maryland, USA
ebriakou@cs.umd.edu

**Marianna J. Martindale**
University of Maryland, USA
mmartind@umd.edu

**Marine Carpuat**
University of Maryland, USA
marine@cs.umd.edu

## Abstract

Neural sequence generation models are known to ''hallucinate'', by producing outputs that are unrelated to the source text. These hallucinations are potentially harmful, yet it remains unclear in what conditions they arise and how to mitigate their impact. In this work, we first identify internal model symptoms of hallucinations by analyzing the relative token contributions to the generation in contrastive hallucinated vs. non-hallucinated outputs generated via source perturbations. We then show that these symptoms are reliable indicators of natural hallucinations, by using them to design a lightweight hallucination detector which outperforms both model-free baselines and strong classifiers based on quality estimation or large pre-trained models on manually annotated English-Chinese and German-English translation test beds.

## 1 Introduction

While neural language generation models can generate high quality text in many settings, they also fail in counter-intuitive ways, for instance by ''hallucinating'' (Wiseman et al., 2017; Lee et al., 2018; Falke et al., 2019). In the most severe case, known as ''detached hallucinations'' (Raunak et al., 2021), the output is completely detached from the source, which not only reveals fundamental limitations of current models, but also risks misleading users and undermining trust (Bender et al., 2021; Martindale and Carpuat, 2018). Yet, we lack a systematic understanding of the conditions where hallucinations arise, as hallucinations occur infrequently among translations of naturally occurring text. As a workaround, prior work has largely focused on black-box detection methods which train neural classifiers on synthetic data constructed by heuristics (Falke et al., 2019; Zhou et al., 2021), and on studying hallucinations given artificially perturbed inputs (Lee et al., 2018; Shi et al., 2022).

In this paper, we address the problem by first identifying the internal model symptoms that characterize hallucinations given artificial inputs and then testing the discovered symptoms on translations of natural texts. Specifically, we study hallucinations in Neural Machine Translation (NMT) using two types of interpretability techniques: saliency analysis and perturbations. We use saliency analysis (Bach et al., 2015; Voita et al., 2021) to compare the relative contributions of various tokens to the hallucinated vs. non-hallucinated outputs generated by diverse adversarial perturbations in the inputs (Table 1) inspired by Lee et al. (2018) and Raunak et al. (2021). Results surprisingly show that source contribution patterns are stronger indicators of hallucinations than the relative contributions of the source and target, as had been previously hypothesized (Voita et al., 2021). We discover two distinctive source contribution patterns, including 1) concentrated contribution from a small subset of source tokens, and 2) the staticity of the source contribution distribution along the generation steps (§ 3).

We further show that the symptoms identified generalize to hallucinations on natural inputs by using them to design a lightweight hallucination classifier (§ 4) that we evaluate on manually annotated hallucinations from English-Chinese and German-English NMT (Table 1). Our study shows that our introspection-based detection model largely outperforms model-free baselines and the classifier based on quality estimation scores. Furthermore, it is more accurate and robust to domain shift than black-box detectors based on large pre-trained models (§ 5).



Transactions of the Association for Computational Linguistics, vol. 11, pp. 546–564, 2023. https://doi.org/10.1162/tacl_a_00563
Action Editor: Ivan Titov. Submission batch: 7/2022; Revision batch: 12/2022; Published 6/2023.

| | |
|---|---|
| **Counterfactual hallucination from perturbation** | |
| *Source* | Republicans Abroad are not running a similar election, nor will they have delegates at the convention. Recent elections have emphasized the value of each vote. |
| *Good* NMT | 国外的共和党没有举行类似的选举，也没有代表参加大会。最近的选举强调了每次投票的价值。 |
| *Perturbed Source* | Repulicans Abroad ar not runing a simila election, nor will they have delegates at the convention. Recent elections have emphasized the value o each vote. |
| *Hallucination* | 大耳朵评论管理人员有权保留或删除其管辖评论中的任意内容。<br>*Gloss: The big ear comments that administrators have the right to retain or delete any content in the comments under their jurisdiction.* |
| **Natural hallucination** | |
| *Source* | DAS GRUNDRECHT JEDES EINZELNEN AUF FREIE WAHL DES BERUFS, DER AUSBILDUNGSSTÄTTE SOWIE DES AUSBILDUNGS - UND BESCHÄFTIGUNGSORTS MUSS GEWAHRT BLEIBEN.<br>*Gloss: The fundamental right of every individual to freely choose their profession, their training institution and their employment place must remain guaranteed.* |
| *Hallucination* | THE PRIVACY OF ANY OTHER CLAIM, EXTRAINING STANDARDS, EXTRAINING OR EMPLOYMENT OR EMPLOYMENT WILL BE LIABLE. |

Table 1: Contrasting counterfactual English-Chinese hallucinations derived from source perturbations (top) with a natural hallucination produced by a German-English NMT model (bottom).

Before presenting these two studies, we review current findings about the conditions in which hallucinations arise and formulate three hypotheses capturing potential hallucination symptoms.

## 2 Hallucinations: Definition and Hypotheses

The term ''hallucinations'' has varying definitions in MT and natural language generation. We adopt the most widely used one, which refers to output text that is unfaithful to the input (Maynez et al., 2020; Zhou et al., 2021; Xiao and Wang, 2021; Ji et al., 2022), while others include fluency criteria as part of the definition (Wang and Sennrich, 2020; Martindale et al., 2019). Different from previous work that aims to detect partial hallucinations at the token level (Zhou et al., 2021), we focus on **detached hallucinations** where a major part of the output is unfaithful to the input, as these represent severe errors, as illustrated in Table 1.

Prior work on understanding the conditions that lead to hallucinations has focused on training conditions and data noise (Ji et al., 2022). For MT, Raunak et al. (2021) show that hallucinations under perturbed inputs are caused by training samples in the long tail that tend to be memorized by Transformer models, while natural hallucinations given unperturbed inputs can be linked to corpus-level noise. Briakou and Carpuat (2021) show that models trained on samples where the source and target side diverge semantically output degenerated text more frequently. Wang and Sennrich (2020) establish a link between MT hallucinations under domain shift and exposure bias by showing that Minimum Risk Training, a training objective which addresses exposure bias, can reduce the frequency of hallucinations. However, these insights do not yet provide practical strategies for handling MT hallucinations.

A complementary approach to diagnosing hallucinations is to identify their symptoms via model introspection at inference time. However, there lacks a systematic study of hallucinations from the model's internal perspective. Previous work is either limited to an interpretation method that is tied to an outdated model architecture (Lee et al., 2018) or to pseudo-hallucinations (Voita et al., 2021). In this paper, we propose to shed light on the decoding behavior of hallucinations on both artificially perturbed and natural inputs through model introspection based on Layerwise Relevance Propagation (LRP) (Bach et al., 2015), which is applicable to a wide range of neural model architectures. We focus on MT tasks with the widely used Transformer model (Vaswani et al., 2017), and examine existing and new hypotheses for how

hallucinations are produced. These hypotheses share the intuition that anomalous patterns of contributions from source tokens are indicative of hallucinations, but operationalize it differently.

The **Low Source Contribution Hypothesis** introduced by Voita et al. (2021) states that hallucinations occur when NMT overly relies on the target context over the source. They test the hypothesis by inspecting the relative source and target contributions to NMT predictions on Transformer models using LRP. However, their study is limited to pseudo-hallucinations produced by force decoding with random target prefixes. This work will test this hypothesis on actual hallucinations generated by NMT models.

The **Local Source Contribution Hypothesis** introduced by Lee et al. (2018) states that hallucinations occur when NMT model overly relies on a small subset of source tokens across all generation steps. They test it by visualizing the dot-product attention in RNN models, but it is unclear whether these findings generalize to other model architectures. In addition, they only study hallucinations caused by random token insertion. This work will test this hypothesis on hallucinations under various types of source perturbations as well as on natural inputs, and will rely on LRP to quantify token contributions more precisely than with attention.

Inspired by the previous observation on attention matrices that an NMT model attends repeatedly to the same source tokens throughout inference when it hallucinates (Lee et al., 2018; Berard et al., 2019b) or generates a low-quality translation (Rikters and Fishel, 2017), we formalize this observation as the **Static Source Contribution Hypothesis**—the distribution of source contributions remains static along inference steps when an NMT model hallucinates. While prior work (Lee et al., 2018; Berard et al., 2019b; Rikters and Fishel, 2017) focuses on the static attention to the EOS or full-stop tokens, this hypothesis is agnostic about which source tokens contribute. Unlike the Low Source Contribution Hypothesis, this hypothesis exclusively relies on the source and does not make any assumption about relative source versus target contributions. Unlike the Local Source Contribution Hypothesis, this hypothesis is agnostic to the proportion of source tokens contributing to a translation.

In this work, we evaluate in a controlled fashion how well each hypothesis explains detached hallucinations, first on artificially perturbed samples that let us contrast hallucinated vs. non-hallucinated outputs in controlled settings (§ 3), and second on natural source inputs that let us test the generalizability of these hypotheses when they are used to automatically detect hallucinations in more realistic settings (§ 5).[1]

## 3 Study of Hallucinations under Perturbations via Model Introspection

Hallucinations are typically rare and difficult to identify in natural datasets. To test the aforementioned hypotheses at scale, we first exploit the fact that source perturbations exacerbate NMT hallucinations (Lee et al., 2018; Raunak et al., 2021). We construct a perturbation-based counterfactual hallucination dataset on English→Chinese by automatically identifying hallucinated NMT translations given perturbed source inputs and contrast them with the NMT translations of the original source (§ 3.1). This dataset lets us directly test the three hypotheses by computing the relative token contributions to the model's predictions using LRP (§ 3.2), and conduct a controlled comparison of patterns on the original and hallucinated samples (§ 3.4).

### 3.1 Perturbation-based Hallucination Data

To construct the dataset, we randomly select $50k$ seed sentence pairs to perturb from the NMT training corpora, and then we apply the following perturbations on the source sentences:[2]

- We randomly misspell words by deleting characters with a probability of 0.1, as Karpukhin et al. (2019) show that a few misspellings can lead to egregious errors in the output.
- We randomly title-case words with a probability of 0.1, as Berard et al. (2019a) find that this often leads to severe output errors.
- We insert a random token at the beginning of the source sentence, as Lee et al. (2018) and Raunak et al. (2021) find it a reliable trigger of hallucinations. The inserted token is chosen from 100 most frequent, 100 least frequent, mid-frequency tokens (randomly

[1] Code and data are released at `https://github.com/weijia-xu/hallucinations-in-nmt`.

[2] For better contrastive analysis, we select samples with source length of $n = 30$ and clip the output length by $T = 15$.

sampled 100 tokens from the remaining tokens), and punctuations.

Inspired by Lee et al. (2018), we then identify hallucinations using heuristics that compare the translations from the original and perturbed sources. We select samples whose original NMT translations $y'$ are of reasonable quality compared to the reference $y$ (i.e., $bleu(y, y') > 0.3$). The translation of a perturbed source sentence $\tilde{y}$ is identified as a hallucination if it is very different from the translation of the original source (i.e., $bleu(y', \tilde{y}) < 0.03$) and is not a copy of the perturbed source $\tilde{x}$ (i.e., $bleu(\tilde{x}, \tilde{y}) < 0.5$).[3] This results in 623, 270, and 1307 contrastive pairs of the original (non-hallucinated) and hallucinated translations under misspelling, title-casing, and insertion perturbations, respectively.

We further divide the contrastive pairs into degenerated and non-degenerated hallucinations. Degenerated hallucinations are ''bland, incoherent, or get stuck in repetitive loops'' (Holtzman et al., 2020), i.e., hallucinated translations that contain 3 more repetitive $n$-grams than the source are identified as degenerated hallucinations, while the non-degenerated group contains relatively fluent but hallucinated translations.

### 3.2 Measuring Relative Token Contributions

We test the three source contribution hypotheses described in § 2 on the resulting dataset by contrasting the contributions of relevant tokens to the generation of a hallucinated versus a non-hallucinated translation using LRP (Bach et al., 2015). LRP decomposes the prediction of a neural model computed over an input instance into relevance scores for input dimensions. Specifically, LRP decomposes a neural model into several layers of computation and measures the relative influence score $R_i^{(l)}$ for input neuron $i$ at layer $l$. Different from other interpretation methods that measure the absolute influence of each input dimension (Alvarez-Melis and Jaakkola, 2017; Ma et al., 2018; He et al., 2019), LRP adopts the principal that the relative influence $R_i^{(l)}$ from all neurons at each layer should sum up to a constant:

$$\sum_i R_i^{(1)} = \sum_i R_i^{(2)} = \ldots = \sum_i R_i^{(L)} = C \quad (1)$$

[3]The BLEU thresholds are selected based on manual inspection of the translation outputs.

To back-propagate the influence scores from the last layer to the first layer (i.e., the input layer), we need to decompose the relevance score $R_j^{(l+1)}$ of a neuron $j$ at layer $l+1$ into messages $R_{i\leftarrow j}^{(l,l+1)}$ sent from the neuron $j$ at layer $l+1$ to each input neuron $i$ at layer $l$ under the following rules:

$$R_{i\leftarrow j}^{(l,l+1)} = v_{ij} R_j^{(l+1)}, \sum_i v_{ij} = 1 \quad (2)$$

There exist several versions of LRP, including LRP-$\varepsilon$, LRP-$\alpha\beta$, and LRP-$\gamma$, which compute $v_{ij}$ differently (Bach et al., 2015; Binder et al., 2016; Montavon et al., 2019). Following Voita et al. (2021), we use LRP-$\alpha\beta$ (Bach et al., 2015; Binder et al., 2016), which defines $v_{ij}$ such that the relevance scores are positive at each step. Consider first the simplest case of linear layers with non-linear activation functions:

$$u_j^{(l+1)} = g(z_j),\ z_j = \sum_i z_{ij} + b_j,\ z_{ij} = w_{ij} u_i^{(l)} \quad (3)$$

where $u_i^{(l)}$ is the $i$-th neuron at layer $l$, $w_{ij}$ is the weight connecting the neurons $u_i^{(l)}$ and $u_j^{(l+1)}$, $b_j$ is a bias term, and $g$ is a non-linear activation function. The $\alpha\beta$ rule considers the positive and negative contributions separately:

$$z_{ij}^{+} = \max(z_{ij}, 0),\ b_j^{+} = \max(b_j, 0)$$

$$z_{ij}^{-} = \min(z_{ij}, 0),\ b_j^{-} = \min(b_j, 0)$$

and defines $v_{ij}$ by the following equation:

$$v_{ij} = \alpha \cdot \frac{z_{ij}^{+}}{\sum_i z_{ij}^{+} + b_j^{+}} + \beta \cdot \frac{z_{ij}^{-}}{\sum_i z_{ij}^{-} + b_j^{-}} \quad (4)$$

Following Voita et al. (2021), we use $\alpha = 1$, $\beta = 0$. This rule is directly applicable to linear, convolutional, maxpooling, and feed-forward layers. To back-propagate relevance scores through attention layers in the Transformer encoder-decoder model (Vaswani et al., 2017), we follow the propagation rules in Voita et al. (2021), where the weighting $v_{ij}$ is obtained by performing a first order Taylor expansion of each neuron $u_j^{(l+1)}$.

In the context of NMT, LRP ensures that, at each generation step $t$, the sum of contributions $R_t(x_i)$ and $R_t(y_j)$ from source tokens $x_i$ and target prefix tokens $y_j$ remains equal:

$$\forall t, \sum_i R_t(x_i) + \sum_{j<t} R_t(y_j) = 1 \quad (5)$$

We further define normalized source contribution $\bar{R}(x_i)$ at source position $i$ averaged over all generation steps $t$ as:

$$\bar{R}(x_i) = \frac{1}{T}\sum_{t}^{T}\frac{n \cdot R_t(x_i)}{\sum_i^n R_t(x_i)} \quad (6)$$

where $n$ is the length of each source sequence and $T$ is the length of the output sequence.

We then test the aforementioned hypotheses based on the distribution of relative token contributions and compare it with the attention matrix.

### 3.3 NMT Setup

We build strong Transformer models on two high-resource language pairs: English→Chinese (En-Zh) and German→English (De-En). They produce acceptable translation outputs overall, thus making hallucinations particularly misleading.

**Data** For En-Zh, we use the $18M$ training samples from WMT18 (Bojar et al., 2018) and *newsdev2017* as the validation set. For De-En, we use all training corpora from WMT21 (Akhbardeh et al., 2021) except for ParaCrawl, which yields $5M$ sentence pairs after cleaning as in Chen et al. (2021).[4] We use *newstest2019* for validation. We tokenize English and German sentences using the Moses scripts (Koehn et al., 2007) and Chinese sentences using the Jieba segmenter.[5] For En-Zh, we train separate BPE models for English and Chinese using $32k$ merging operations for each language. For De-En, we train a joint BPE model using $32k$ merging operations.

**Models** All models are based on the *base* Transformer (Vaswani et al., 2017). We apply label smoothing of 0.1. We train all models using the Adam optimizer (Kingma and Ba, 2015) with initial learning rate of 4.0 and batch sizes of 4,000 tokens for maximum $800k$ steps. We decode with beam search with a beam size of 4. The resulting NMT models achieve close or higher BLEU scores than comparable published results.[6]

[4] https://github.com/browsermt/students/tree/master/train-student/clean.

[5] https://github.com/fxsjy/jieba.

[6] The En-Zh model achieves 33.5 BLEU on *newstest2017*, which is close to the 34.5 achieved by the most comparable model in Xu and Carpuat (2018). The De-En model achieves 35.0 BLEU on *newstest2019*, which is higher than the strong baseline (29.6 BLEU) from Germann (2020).

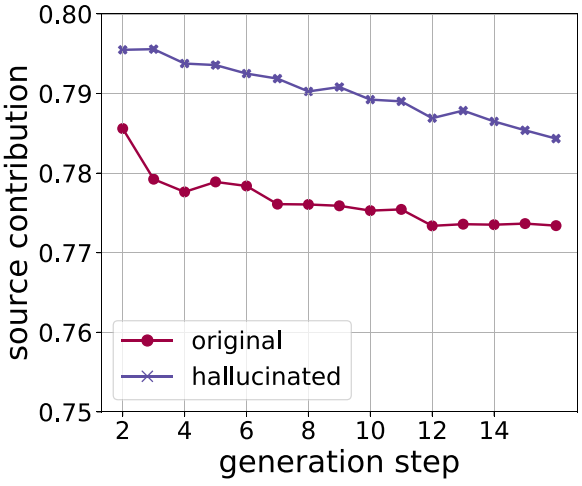



Figure 1: Relative source contributions $\sum_i R_t(x_i)$ at varying generation step $t$ averaged over the original or hallucinated samples under a mixture of the misspelling, title-casing, and insertion perturbations.

### 3.4 Findings

We test the aforementioned hypotheses on the perturbation-based counterfactual hallucination dataset constructed on English→Chinese.

First, we test the **Low Source Contribution Hypothesis** by computing the relative source contributions $\sum_{i=1}^{n} R_t(x_i)$ at each generation step $t$, where $n$ is the length of each source sequence.[7] We plot the average contributions over a set of samples in Figure 1. It shows that hallucinations under source perturbations have only slightly higher source contributions ($\Delta \approx 0.1$) than the original samples. This departs from previous observations on pseudo-hallucinations (Voita et al., 2021), where the relative source contributions were lower on pseudo-hallucinations than on reference translations, perhaps because actual model outputs differ from pseudo-hallucinations created by inserting random target prefixes. We show that the hypothesis does not hold on actual hallucinations generated by the model itself.

To explain this phenomenon, we further examine the source contribution from the end-of-sequence (EOS) token. Previous work hypothesizes that a translation is likely to be a hallucination when the attention distribution is concentrated on the source EOS token, which carries little information about the source (Berard et al., 2019b; Raunak et al., 2021). However, this hypothesis

[7] Since LRP ensures that the sum of source and target contributions at each generation step is a constant, we only visualize the relative source contributions.

| | Contrib Ratio | | Staticity | |
|---|---|---|---|---|
| | D | N | D | N |
| Attention | $-1.03^{\dagger}$ | $+0.51^{\dagger}$ | $1.92^{\dagger}$ | $-1.10^{\dagger}$ |
| LRP | $\mathbf{-1.05}^{\dagger}$ | $\mathbf{-1.13}^{\dagger}$ | $\mathbf{3.16}^{\dagger}$ | $\mathbf{2.16}^{\dagger}$ |

Table 2: Standardized mean difference in High-Contribution Ratio (*Contrib Ratio*) and Source Contribution Staticity (*Staticity*) (computed on attention and LRP-based contribution matrices) between pairs of hallucinated and original samples. We show the score differences on degenerated (*D*) and non-degenerated (*N*) hallucinations separately. † indicates that the difference is statistically significant with $p < 0.05$.

has only been supported by qualitative analysis on individual samples. Our quantitative results on the perturbation-based hallucination dataset do not support it, and align instead with the recent finding that the proportion of attention paid to the EOS token is not indicative of hallucinations (Guerreiro et al., 2022). Specifically, our results show that the proportion of source contribution from the EOS token is slightly higher on the original samples (11.2%) than that on the hallucinated samples (10.8%). We will show in the next part that the source contribution is more concentrated on the beginning than the end of the source sentence when the model hallucinates.

Second, we test the **Local Source Contribution Hypothesis** by computing the **High-Contribution Ratio** $r(\lambda_0)$—the ratio of source tokens with normalized contribution $\bar{R}(x_i)$ larger than a threshold $\lambda_0$:

$$r(\lambda_0) = \sum_{i=1}^{n} \mathbb{I}(\bar{R}(x_i) > \lambda_0)/n \qquad (7)$$

The ratio will be lower on hallucinated samples than on original samples if the hypothesis holds. We compute the standardized mean difference in High-Contribution Ratio between the hallucinated and original samples (Table 2).[8] The negative score differences in LRP-based scores support the hypothesis, which is consistent with the findings of Lee et al. (2018) based on attention weights. However, the attention-based score patterns are not consistent on degenerated and non-degenerated samples.

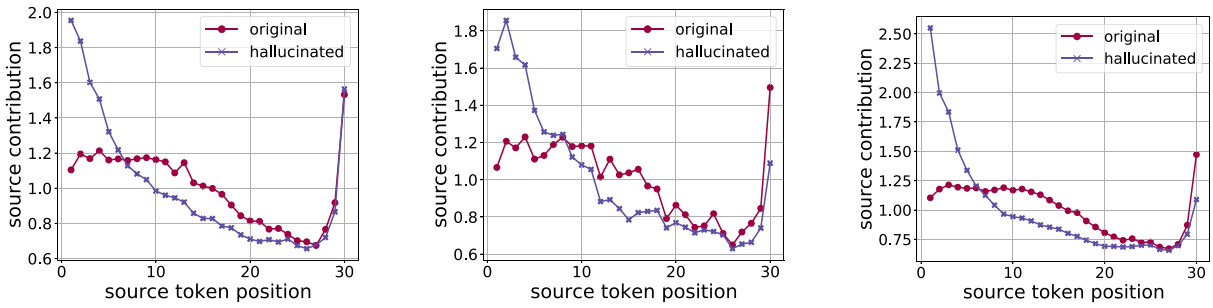
[8] $\lambda_0$ is set to yield the largest score difference for each measurement type.

Furthermore, we investigate whether there is any positional bias for the local source contribution. We visualize the normalized source contribution $\bar{R}(x_i)$ averaged over all samples with a source length of 30 in Figure 2. The source contribution of the hallucinated samples is disproportionately high at the beginning of a source sequence. By contrast, on the original samples, the normalized contribution is higher at the end of the source sequence, which could be a way for the model to decide when to finish generation. The positional bias exists not only on hallucinations under insertions at the beginning of the source, but also on hallucinations under misspelling and title-casing perturbations that are applied at random positions.

Third, we examine the **Static Source Contribution Hypothesis** hypothesis by first visualizing the source contributions $R_t(x_i)$ at varying source and generation positions on individual pairs of original and hallucinated samples. The heatmaps of source contributions for the example from Table 1 are shown in Figure 3. On the original outputs, the source contribution distribution in each column changes dynamically when moving horizontally along target generation steps. By contrast, when the model hallucinates, the source contribution distribution remains roughly static.

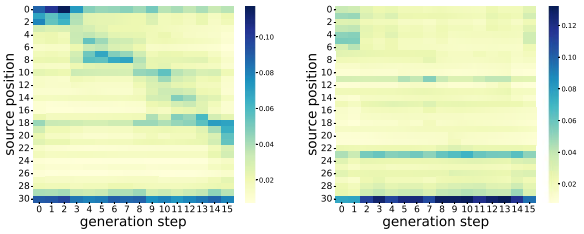
To quantify this pattern, we introduce **Source Contribution Staticity**, which measures how the source contribution distribution shifts over generation steps. Specifically, given a window size $k$, we first divide the target sequence into several non-overlapping segments, each containing $k$ tokens. Then, we compute the average vector over the contribution vectors $\boldsymbol{R_t} = [R_t(x_0) \dots R_t(x_n)]$ at steps $t$ within each segment. Finally, we measure the cosine similarity between the average contribution vectors of adjacent segments and average over the cosine similarity scores at all positions as the final score $s_k$ of window size $k$. Figure 4 illustrates this process for a window size of 2.

Table 2 shows the standardized mean difference in Source Contribution Staticity between the hallucinated and original samples in the degenerated and non-degenerated groups, taking the maximum staticity score among window sizes $k \in [1, 3]$ for each sample. The positive differences in LRP-based scores supports the Static Source Contribution Hypothesis—the source contribution distribution is more static on




(a) Misspelling (b) Title-Casing (c) Insertion

Figure 2: Normalized source contribution $\bar{R}(x_i)$ (Eq. 6) at each source token position averaged over the original or hallucinated samples under (a) misspelling, (b) title-casing, and (c) insertion perturbations.




(a) Original (b) Hallucinated

Figure 3: Heatmaps of relative contributions of source tokens ($y$-axis) at each generation step ($x$-axis) computed on the example of the original translation and the counterfactual hallucination from the perturbed source in Table 1. The source contribution distribution remains static across almost all generation steps on the hallucinated sample, unlike on the original sample.

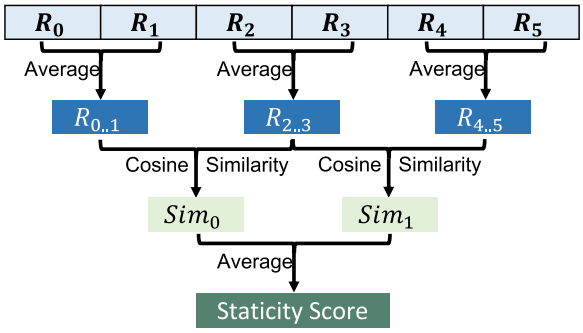



Figure 4: Computing the Source Contribution Staticity of window size $k = 2$ given the source contribution vectors $\boldsymbol{R_t} = [R_t(x_0) \dots R_t(x_n)]$ at generation step $t$.

the hallucinated samples than that on the original samples. Furthermore, LRP distinguishes hallucinations from non-hallucinations better than attention, especially on non-degenerated samples where the translation outputs contain no repetitive loops.

In summary, we find that, when generating a hallucination under source perturbations, the NMT model tends to rely on a small proportion of the source tokens, especially the tokens at the beginning of the source sentence. In addition, the distribution of the source contributions is more static on hallucinated translations than that on non-hallucinated translations. We turn to applying these insights on natural hallucinations next.

## 4 A Classifier to Detect Natural Hallucinations

Based on these findings, we design features for a lightweight hallucination detector trained on samples automatically constructed by perturbations.

**Classifier** We build a small multi-layer perceptron (MLP) with a single hidden layer and the following input features:

- **Normalized Source Contribution** of the first $K_1$ source tokens and the last $K_1$ source tokens: $\bar{R}(x_i)|i = 1, \dots, K_1, n - K_1 + 1, \dots, n$ (where $n$ is the length of the source sequence and $K_1$ is a hyper-parameter), as we showed in the Local Source Contribution Hypothesis that the contributions of the beginning and end tokens distribute differently between hallucinated and non-hallucinated samples.

- **Source Contribution Staticity** $s_k$ given the source contributions $R_t(x_i)$ and a window size $k$ as defined in § 3.4. We include the similarity scores of window sizes $k = \{1, 2, \dots, K_2\}$ as input features, where $K_2$ is a hyper-parameter.

This yields small classifiers with input dimension of 9. For each language pair, we train 20 classifiers

with different random seeds and select the model with the highest validation F1 score.

**Data Generation** We construct the training and validation data using the same approach to constructing the perturbation-based hallucination dataset (§ 3.1), but with longer seed pairs—we randomly select seed sentence pairs with source length between 20 and 60 from the training corpora. We split the synthetic data randomly into the training (around $1k$ samples) and validation (around 200 samples) sets with roughly equal number of positive and negative samples.

## 5 Detecting Natural Hallucinations

While the hallucination classifier is trained on hallucinations from perturbations, we collect more realistic data to evaluate it against a wide range of relevant models.

### 5.1 Natural Hallucination Evaluation Set

We build a test bed for detached hallucination detection for different language pairs and translation directions (En-Zh and De-En), and release the data together with the underlying NMT models (described in § 3.3).

Since hallucinations are rare, we collect samples from large pools of out-of-domain data for our models to obtain enough positive examples of hallucinations for a meaningful test set. We use TED talk transcripts from the IWSLT15 training set (Cettolo et al., 2015) for En-Zh, and the JRC-Acquis corpus (Steinberger et al., 2006) of legislation from the European Union for De-En. To increase the chance of finding hallucinations, we select around 200, 50, and 50 translation outputs with low BLEU, low COMET (Rei et al., 2020a), or low LASER similarity (Artetxe and Schwenk, 2019) scores, respectively. We further combine them with 50 randomly selected samples.

Three bilingual annotators assess the faithfulness of the NMT output given each input. While we ultimately need a binary annotation of outputs as hallucinated or not, annotators were asked to choose one of five labels to improve consistency:

- *Detached hallucination*: a translation with large segments that are unrelated to the source.
- *Faithful translation*: a translation that is faithful to the source.
- *Incomplete translation*: a translation that is partially correct but misses part(s) of the source.
- *Locally unfaithful*: a translation that contains a few unfaithful phrases but is otherwise faithful.
- *Incomprehensible but aligned*: a translation that is incomprehensible even though most phrases can be aligned to the source.

| | En-Zh | De-En |
|---|---|---|
| Detached hallucination | 111 | 189 |
| Non hallucination, including: | | |
| *Faithful translation* | 154 | 153 |
| *Incomplete translation* | 80 | 17 |
| *Locally unfaithful* | 58 | 31 |
| *Incomprehensible but aligned* | 5 | 33 |
| **Total** | 408 | 423 |

Table 3: Human annotation label distribution on the En-Zh and De-En natural hallucination test sets (with random tie breaking on fine-grained labels; there are no ties on binary labels post-aggregation).

All labels except for the ‘‘detached hallucination’’ are aggregated into the ‘‘non-hallucination’’ category. The inter-annotator agreement on aggregated labels is substantial, with a Fleiss's Kappa (Fleiss, 1971) score of $FK = 0.77$ for De-En and $FK = 0.64$ for En-Zh. Disagreements are resolved by majority voting for De-En, and by adjudication by a bilingual speaker for En-Zh. This yields $27\%$ of detached hallucinations on En-Zh and $45\%$ on De-En. The non-hallucinated NMT outputs span all the fine-grained categories above, as can be seen in Table 3. Hallucinations are over-represented compared to what one might expect in the wild, but this is necessary to provide enough positive examples of hallucinations for evaluation.

### 5.2 Experimental Conditions

#### 5.2.1 Introspection-based Classifiers

We implement the **LRP-based classifier** described in § 4. To lower the cost of computing source contributions, we clip the source length at 40, and only consider the influence back-propagated through the most recent 10 target tokens—prior work shows that nearby context is more influential than distant context (Khandelwal et al., 2018).

We tune the hyper-parameters $K_1$ and $K_2$ within the space $K_1 \in \{1, 3, 5, 7, 9\}$, $K_2 \in \{4, 8, 12, 16\}$ based on the average F1 accuracy on the validation set over three runs. We compare it with an **attention-based classifier**, which uses the same features, but computes token contributions using attention weights averaged over all attention heads.

#### 5.2.2 Model-free Baselines

We use three simple baselines to characterize the task. The **random classifier** that predicts hallucination with a probability of 0.5. The **degeneration** detector marks as hallucinations degenerated outputs that contain $K$ more repetitive $n$-grams than the source, where $K$ is a hyper-parameter tuned on the perturbation-based hallucination data. The **NMT probability scores** are used as a coarse model signal to detect hallucinations based on the heuristic that the model is less confident when producing a hallucination. The output is classified as a hallucination if the probability score is lower than a threshold tuned on the perturbation-based hallucination data.

#### 5.2.3 Quality Estimation Classifier

We also compare the introspection-based classifiers with a baseline classifier based on the state-of-the-art quality estimation model—**COMET-QE** (Rei et al., 2020b). Given a source sentence and its NMT translation, we compute the COMET-QE score and classify the translation as a hallucination if the score is below a threshold tuned on the perturbation-based validation set.

#### 5.2.4 Large Pre-trained Classifiers

We further compare the introspection-based classifiers with classifiers that rely on large pre-trained multilingual models, to compare the discriminative power of the source contribution patterns from the NMT model itself to extrinsic semantically driven discrimination criteria.

We use the cosine distance between the **LASER** representations (Artetxe and Schwenk, 2019; Heffernan et al., 2022) of the source and the NMT translation. It classifies a translation as a hallucination if the distance score is higher than a threshold tuned on the perturbation-based validation set.

Inspired by local hallucination (Zhou et al., 2021) and cross-lingual semantic divergence (Briakou and Carpuat, 2020) detection methods, we build an **XLM-R classifier** by fine-tuning the XLM-R model (Conneau et al., 2020) on synthetic hallucination samples. We randomly select $50K$ seed pairs of source and reference sentences with source lengths between 20 and 60 from the parallel corpus and use the following perturbations to construct examples of detached hallucinations:

- Map a source sentence to a random target from the parallel corpus to simulate natural, detached hallucinations.
- Repeat a random dependency subtree in the reference many times to simulate degenerated hallucinations.
- Drop a random clause from the source sentence to simulate natural, detached hallucinations.

We then collect diverse non-hallucinated samples:

- Original seed pairs provide faithful translations.
- Randomly drop a dependency subtree from a reference to simulate incomplete translations.
- Randomly substitute a phrase in the reference keeping the same part-of-speech to simulate translations with locally unfaithful phrases.

The final training and validation sets contain around $300k$ and 700 samples, respectively. We fine-tune the pre-trained model with a batch size of 32. We use the Adam optimizer (Kingma and Ba, 2015) with decoupled weight decay (Loshchilov and Hutter, 2019) and an initial learning rate of $2 \times 10^{-5}$. We fine-tune all models for 5 epochs and select the checkpoint with the highest F1 score on the validation set.

### 5.3 Findings

As shown in Table 4, we compare all classifiers against the baselines by the Precision, Recall, and F1 scores. Since false positives and false negatives might have a different impact in practice (e.g., does the detector flag examples for review by humans, or entirely automatically? what is MT used for?), we also report the Area Under the Receiver Operating Characteristic Curve (AUC), which characterizes the discriminative power of each method at varying threshold settings.

| | Params | De-En | | | | En-Zh | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | P | R | F1 | AUC | P | R | F1 | AUC |
| *Model-free Baselines* | | | | | | | | | |
| Random | 0 | 44.0 | 49.9 | 46.8 | 50.2 | 27.6 | 49.8 | 35.5 | 48.0 |
| Degeneration | 1 | 49.1 | 59.3 | 53.7 | – | 63.2 | 71.2 | 66.9 | – |
| NMT Score | 1 | 33.3 | 3.4 | 6.2 | 37.7 | 35.4 | **91.9** | 51.1 | 49.8 |
| *Quality Estimation Classifier* | | | | | | | | | |
| COMET-QE | $363M$ | 72.2 | 71.4 | **71.8** | 82.4 | 32.4 | **99.1** | 48.9 | 89.4 |
| *Large Pre-trained Classifiers* | | | | | | | | | |
| LASER | $45M$ | 81.6 | 54.0 | 65.0 | **89.5** | 54.6 | 64.0 | 58.9 | 75.3 |
| XLM-R | $125M$ | 91.3 | 21.0 | 33.8 | 45.6 | **94.9** | **83.2** | **88.6** | **93.3** |
| *Introspection-based Classifiers* | | | | | | | | | |
| Attention-based | $< 400$ | 54.3 | **89.0** | 67.4 | 70.1 | 36.0 | 71.0 | 47.7 | 68.6 |
| LRP-based | $< 400$ | 87.3 | **76.2** | **81.2** | **91.4** | 87.5 | **85.6** | **86.4** | **96.5** |
| *Ensemble Classifier* | | | | | | | | | |
| LRP + LASER | $45M$ | **100.0** | 45.7 | 62.7 | – | **94.5** | 59.5 | 72.9 | – |
| LRP + XLM-R | $125M$ | 95.3 | 21.5 | 35.1 | – | **97.6** | 72.4 | 83.1 | – |

Table 4: Precision (P), Recall (R), F1, and Area Under the Receiver Operating Characteristic Curve (AUC) scores of each classifier on English-Chinese (En-Zh) and German-English (De-En) NMT outputs (means of three runs). We boldface the highest scores based on independent Student $t$-test with Bonferroni correction ($p < 0.05$). The *Params* column indicates the total number of parameters used for each method (in addition to the NMT parameters).

**Main Results** The LRP-based, XLM-R, and the LASER classifiers are the best hallucination detectors, reaching AUC scores around 90 for either or both language pairs, which is considered outstanding discrimination ability (Hosmer Jr et al., 2013).

The LRP-based classifier is the best and most robust hallucination detector overall. It achieves higher F1 and AUC scores than LASER on both language pairs. Additionally, it outperforms XLM-R by +47 F1 and +46 AUC on De-En, while achieving competitive performance on En-Zh. This shows that the source contribution patterns identified on hallucinations under perturbations (§ 3) generalize as symptoms of natural hallucinations even under domain shift, as the domain gap between training and evaluation data is bigger on De-En than En-Zh. It also confirms that LRP provides a better signal to characterize token contributions than attention, improving F1 by 14–39 points and AUC by 21–28 points. These high scores represent large improvements of 41–54 points on AUC and 20–75 points on F1 over the model-free baselines.

**Model-free Baselines** These baselines shed light on the nature of the hallucinations in the dataset. The degeneration baseline is the best among them, with 53.7 F1 on De-En and 66.9 F1 on En-Zh, indicating that the Chinese hallucinations are more frequently degenerated than the English hallucinations from German. However, ignoring the remaining hallucinations is problematic, since they might be more fluent and thus more likely to mislead readers. The NMT score is a poor predictor, scoring worse than the random baseline on De-En, in line with previous reports that NMT scores do not capture faithfulness well during inference (Wang et al., 2020). Manual inspection shows that the NMT score can be low when the output is faithful but contains rare words, and it can be high for a hallucinated output that contains mostly frequent words.

**Quality Estimation Classifier** The COMET-QE classifier achieves higher AUC and F1 scores than the model-free classifiers, except for En-Zh, where the degeneration baseline obtains higher F1

than the COMET-QE classifier. However, compared with the LRP-based classifier, COMET-QE lags behind by 9-38 points on F1 and 7-9 points on AUC. This is consistent with previous findings that quality estimation models trained on data with insufficient negative samples (e.g., COMET-QE) are inadequate for detecting critical MT errors such as hallucinations (Takahashi et al., 2021; Sudoh et al., 2021; Guerreiro et al., 2022).

**Pre-trained Classifiers** The performance of pre-trained classifiers varies greatly across language pairs. LASER achieves a competitive AUC score to the LRP-based classifier on De-En but lags behind on En-Zh, perhaps because the LASER model is susceptible to the many rare tokens in the En-Zh evaluation data (from TED subtitles). XLM-R obtains better performance on En-Zh, approaching that of the LRP-based classifier, but lags behind greatly on De-En. This suggests that the XLM-R classifier suffers from domain shift, which is bigger on De-En (News→Law) than En-Zh (News→TED). Fine-tuning the model on the synthetic training data generalizes more poorly across domains. By contrast, the introspection-based classifiers are more robust.

**Ensemble Classifiers** The LASER and XLM-R classifiers emerge as the top classifiers apart from the LRP-based one, but they make different errors than LRP—the confusion matrix comparing their predictions shows that the LASER and LRP classifiers agree on 68–78% of samples, while the XLM-R and LRP classifiers agree on 64–88% of samples. Thus an ensemble of LRP + LASER or LRP + XLM-R (which detects hallucinations when the two classifiers both do so) yields a very high precision (at the expense of recall).

**LRP Ablations** The LRP-based classifier benefits the most from Source Contribution Staticity features (Table 5). Removing them hurts AUC by 15–17 points and F1 by 28–31 points, confirming that the Static Source Contribution Hypothesis holds on natural hallucinations. Ablating the Normalized Source Contribution features also causes a significant drop in F1 on De-En, while its impact on En-Zh is not significant.

**Error Analysis** Incomprehensible but aligned translations suffer from the highest false positive rate for the LRP classifier, followed by incomplete translations. Additionally, the classifier can fail to detect hallucinations caused by the mistranslation of a large span of the source with rare or previously unseen tokens, rather than by pathological behavior at inference time as shown by the example in Table 6.

| | De-En | | En-Zh | |
|---|---|---|---|---|
| | F1 | AUC | F1 | AUC |
| All features | **81.2** | **91.4** | **86.4** | **96.5** |
| - Src Contrib | 74.4 | **92.7** | **85.3** | **96.1** |
| - Staticity | 50.7 | 76.6 | 58.3 | 80.0 |

Table 5: Ablating the Normalized Source Contribution (*Src Contrib*) and Source Contribution Staticity (*Staticity*) features used in the LRP-based classifier. We boldface the highest scores based on independent student's $t$-test with Bonferroni Correction($p < 0.05$).

| |
|---|
| **Source:** C) DASS DIE WAREN IN DEM ZUSTAND IN DIE GEMEINSCHAFT VERSANDT WORDEN SIND, IN DEM SIE ZUR AUSSTELLUNG GESANDT WURDEN; |
| **Correct Translation:** C) THAT THE GOODS WERE SHIPPED TO THE COMMUNITY IN THE CONDITION IN WHICH THEY ARE SENT FOR EXHIBITION; |
| **Output:** C) THAT THE WOULD BE CONSIDERED IN THE COMMUNITY, IN which YOU WILL BE EXCLUSIVE; |

Table 6: Example of a detached hallucination produced by the De-En NMT being classified as non-hallucination by the LRP-based classifier.

**Toward Practical Detectors** Detecting hallucinations in the wild is challenging since they tend to be rare and their frequency may vary greatly across test cases. We provide a first step in this direction by stress testing the top classifiers in an in-domain scenario where hallucinations are expected to be rare. Specifically, we randomly select $10k$ English sentences from the *News Crawl: articles from 2021* from WMT21 (Akhbardeh et al., 2021) and use the En-Zh NMT model to translate them into Chinese. We measure the *Precision@20* for hallucination detection by manually examining the top-20 highest scoring hallucination

predictions for each method. The LASER, XLM-R, and LRP-based classifiers evaluated above (without fine-tuning in this setting) achieve 35%, 45%, and 45% Precision@20, respectively (compared to 0% for the random baseline). More interestingly, after tuning the threshold on the predicted probabilities (which is originally set to 0.5) so that each classifier predicts hallucination 1% of the time, the LRP + LASER ensemble detects 9 hallucinations with a much higher precision of 89%, and the LRP + XLM-R ensemble detects 12 hallucinations with a precision of 83%. These ensemble detectors thus have the potential to provide useful signals for detecting hallucinations even when they are needles in a haystack.

### 5.4 Limitations

Our findings should be interpreted with several limitations in mind. First, we exclusively study detached hallucinations in MT. Thus, we do not elucidate the internal model symptoms that lead to partial hallucinations (Zhou et al., 2021), although the methodology in this work could be used to shed light on this question. Second, we work with NMT models trained using the parallel data from WMT without exploiting monolingual data or comparable corpora retrieved from collections of monolingual texts (e.g., WikiMatrix [Schwenk et al., 2021]). It remains to be seen whether hallucination symptoms generalize to NMT models trained with more heterogeneous supervision. Finally, we primarily test the hallucination classifiers in roughly balanced test sets, while hallucinations are expected to be rare in practice. We conducted a small stress test which shows the promise of our LRP +LASER classifier in more realistic conditions. However, further work is needed to systematically evaluate how these classifiers can be used for hallucination detection in the wild.

## 6 Related Work

Hallucinations occur in all applications of neural models to language generation, including abstractive summarization (Falke et al., 2019; Maynez et al., 2020), dialogue generation (Dušek et al., 2018), data-to-text generation (Wiseman et al., 2017), and machine translation (Lee et al., 2018). Most existing detection approaches view the generation model as a black-box, by 1) training hallucination classifiers on synthetic data constructed by heuristics (Zhou et al., 2021; Santhanam et al., 2021), or 2) using external models to measure the faithfulness of the outputs, such as question answering or natural language inference models (Falke et al., 2019; Durmus et al., 2020). These approaches ignore the signals from the generation model itself and could be highly biased by the heuristics used for synthetic data construction, or the biases in the external semantic models trained for other purposes. Concurrent to this work, Guerreiro et al. (2022) explore glass-box detection methods based on model confidence scores or attention patterns (e.g., the proportion of attention paid to the EOS token and the proportion of source tokens with attention weights higher than a threshold). They evaluate these methods based on hallucination recall, and find that model confidence is a better indicator of hallucinations than attention patterns. In this paper, we investigated varying types of glass-box patterns based on the relative token contributions instead of attention, and find that these patterns yield more accurate hallucination detectors than model confidence.

Detecting hallucinations in MT has not yet been directly addressed by the MT quality estimation literature. Most quality estimation work has focused on predicting a direct assessment of translation quality, which does not distinguish adequacy and fluency errors (Guzmán et al., 2019; Specia et al., 2020). More recent task formulations target critical adequacy errors (Specia et al., 2021), but do not separate hallucinations from other error types, despite arguments that hallucinations should be considered separately from other MT errors (Shi et al., 2022). The critical error detection task at WMT 2022 introduces an Additions error category, which refers to hallucinations where the translation content is only partially supported by the source (Zerva et al., 2022). Additions includes both detached hallucinations (as in this work) and partial hallucinations. Methods for addressing all these tasks fall in two categories: 1) black-box methods based on the source and output alone (Specia et al., 2009; Kim et al., 2017; Ranasinghe et al., 2020), and 2) glass-box methods based on features extracted from the NMT model itself (Rikters and Fishel, 2017; Yankovskaya et al., 2018; Fomicheva et al., 2020). Black-box methods typically use resource-heavy deep neural networks trained on large amounts of annotated data. Our work is inspired by the glass-box methods that

rely on model probabilities, uncertainty quantification, and the entropy of the attention distribution, but shows that relative token contributions computed through LRP provide sharper features to characterize hallucinations.

This paper combines interpretability techniques to identify the symptoms of hallucinations. We adopt a saliency method to measure the importance of each input unit through a back-propagation pass (Simonyan et al., 2014; Bach et al., 2015; Li et al., 2016a; Ding et al., 2019). While other saliency-based methods measure an abstract quantity reflecting the importance of each input feature by the partial derivative of the prediction with regard to each input unit (Simonyan et al., 2014), LRP (Bach et al., 2015) measures the proportional contribution of each input unit. This makes it well-suited to compare model behavior across samples. Furthermore, LRP does not require neural activations to be differentiable and smooth, and can be applied to a wide range of architectures, including RNN (Ding et al., 2017) and Transformer (Voita et al., 2021). We apply this technique to analyze counterfactual hallucination samples inspired by perturbation methods (Li et al., 2016b; Feng et al., 2018; Ebrahimi et al., 2018), but crucially show that the insights generalize to natural hallucinations.

## 7 Conclusion

We contribute a thorough empirical study of the notorious but poorly understood hallucination phenomenon in NMT, which shows that internal model symptoms exhibited during inference are strong indicators of hallucinations. Using counterfactual hallucinations triggered by perturbations, we show that distinctive source contribution patterns alone indicate hallucinations better than the relative contributions of the source and target. We further show that our findings can be used for detecting natural hallucinations much more accurately than model-free baselines and quality estimation models. Our detector also outperforms black-box classifiers based on pre-trained models. We release human-annotated test beds of natural English-Chinese and German-English hallucinations to enable further research. This work opens a path toward detecting hallucinations in the wild and improving models to minimize hallucinations in MT and other generation tasks.

## Acknowledgments

We thank our TACL action editor, the anonymous reviewers, and the UMD CLIP lab for their feedback. Thanks also to Yuxin Xiong for helping examine German outputs. This research is supported in part by an Amazon Machine Learning Research Award and by the National Science Foundation under Award No. 1750695. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1042

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610. https://doi.org/10.1162/tacl_a_00288

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, 10(7):e0130140. https://doi.org/10.1371/journal.pone.0130140, PubMed: 26161953

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. 2019a. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 168–176, Hong Kong. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5617

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019b. Naver labs Europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5361

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer. https://doi.org/10.1007/978-3-319-44781-0_8

Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation*, pages 272–307. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6401

Eleftheria Briakou and Marine Carpuat. 2020. Detecting fine-grained cross-lingual semantic divergences without supervision by learning to rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.121

Eleftheria Briakou and Marine Carpuat. 2021. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7236–7249, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.562

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 evaluation campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 2–14, Da Nang, Vietnam.

Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli Barone, Jonas Waldendorf, Alexandra Birch, and Kenneth Heafield. 2021. The University of Edinburgh's English-German and English-Hausa submissions to the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 104–109, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451,

Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5201

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1106

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.454

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328, Tilburg University, The Netherlands. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6539

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1213

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1407

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378. https://doi.org/10.1037/h0031619

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555. https://doi.org/10.1162/tacl_a_00330

Ulrich Germann. 2020. The University of Edinburgh's submission to the German-to-English and English-to-German tracks in the WMT 2020 news translation and zero-shot translation robustness tasks. In *Proceedings of the Fifth Conference on Machine Translation*, pages 197–201, Online. Association for Computational Linguistics.

Nuno M. Guerreiro, Elena Voita, and André F. T. Martins. 2022. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. https://doi.org/10.48550/arXiv.2208.05309

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and*

*the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1632

Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael Lyu, and Shuming Shi. 2019. Towards understanding neural machine translation with word importance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 953–962, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1088

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. https://doi.org/10.48550/arXiv.2205.12654

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

David W. Hosmer Jr., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*, volume 398. John Wiley & Sons. https://doi.org/10.1002/9781118548387

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. https://doi.org/10.48550/arXiv.2202.03629

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5506

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1027

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4763

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3th International Conference on Learning Representations*, San Diego, CA, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. In *NeurIPS Interpretability and Robustness for Audio, Speech and Language Workshop*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1082

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220. https://doi.org/10.48550/arXiv.1612.08220

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xutai Ma, Ke Li, and Philipp Koehn. 2018. An analysis of source context dependency in neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, page 189.

Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in mea suring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 13–25, Boston, MA. Association for Machine Translation in the Americas.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 233–243, Dublin, Ireland. European Association for Machine Translation.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: An overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 193–209. https://doi.org/10.1007/978-3-030-28954-6_10

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.445

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.92

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Matīs Rikters and Mark Fishel. 2017. Confidence through attention. In *Proceedings of the 16th Machine Translation Summit (MT Summit 2017)*. Nagoya, Japan.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *CoRR*, abs/2110.05456. https://doi.org/10.48550/arXiv.2110.05456

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.115

Ruikang Shi, Alvin Grissom II, and Duc Minh Trinh. 2022. Rare but severe neural machine

translation errors induced by minimal deletion: An empirical study on Chinese and English. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5175–5180, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Katsuhito Sudoh, Kosuke Takahashi, and Satoshi Nakamura. 2021. Is this translation error critical?: Classification-based human and automatic machine translation evaluation focusing on critical errors. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 46–55, Online. Association for Computational Linguistics.

Kosuke Takahashi, Yoichi Ishibashi, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Multilingual machine translation evaluation metrics fine-tuned on pseudo-negative examples for WMT 2021 metrics task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1049–1052, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, Long Beach, CA, USA. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.91

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.326

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.278

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document

generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1239

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.236

Weijia Xu and Marine Carpuat. 2018. The University of Maryland's Chinese-English neural machine translation systems at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 535–540, Belgium, Brussels. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6431

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2018. Quality estimation with force-decoded attention and cross-lingual embeddings. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 816–821, Belgium, Brussels. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6466

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.120