Evaluating Saliency Explanations in NLP by Crowdsourcing

Xiaotian Lu[*1] **Jiyi Li** [2] **Zhen Wan** [1] **Xiaofeng Lin** [1] **Koh Takeuchi** [1] **Hisashi Kashima**[1]

[1]Kyoto University   [2]University of Yamanashi

[1]{lu@ml.ist.,zhenwan@nlp.ist.,lxf@ml.ist.,takeuchi@,kashima@}i.kyoto-u.ac.jp

[2]jyli@yamanashi.ac.jp

Abstract

Deep learning models have performed well on many NLP tasks. However, their internal mechanisms are typically difficult for humans to understand. The development of methods to explain models has become a key issue in the reliability of deep learning models in many important applications. Various saliency explanation methods, which give each feature of input a score proportional to the contribution of output, have been proposed to determine the part of the input which a model values most. Despite a considerable body of work on the evaluation of saliency methods, whether the results of various evaluation metrics agree with human cognition remains an open question. In this study, we propose a new human-based method to evaluate saliency methods in NLP by crowdsourcing. We recruited 800 crowd workers and empirically evaluated seven saliency methods on two datasets with the proposed method. We analyzed the performance of saliency methods, compared our results with existing automated evaluation methods, and identified notable differences between NLP and computer vision (CV) fields when using saliency methods. The instance-level data of our crowdsourced experiments and the code to reproduce the explanations are available at `https://github.com/xtlu/lreccoling_evaluation`.

1 Introduction

Recently, the development of deep neural network (DNN) models for natural language processing, such as BERT (Devlin et al., 2019), and GPT-3 (Brown et al., 2020), has attracted considerable attention as a topic of active research. Given that DNNs can recognize important patterns in large and intricate datasets, DNN models exhibited significant achievements on various tasks for different applications. However, most deep models are black boxes, i.e., the relationship between their inputs and outputs is difficult to understand.

In most low-risk tasks, errors made by DNNs do not have severe impacts; for example, in a recommendation system designed to suggest fiction novels, the impact of recommendation errors is relatively low. However, in some situations, the DNN model performing an NLP task could cause psychological harm to users. For example, an online chat service could output unethical or offensive sentences. To avoid this, one of the most fundamental and effective solution is to understand the decision-making process of the relevant models. Therefore, as the complexity of DNN models increases, the need for research on the interpretation of the output of such models becomes increasingly pressing.

Various interpretation methods have emerged to attempt to understand DNNs (Liu et al., 2019a; Moradi et al., 2021; Carton et al., 2020), particularly for black-box predictions made by already-trained neural networks. One of the major approaches to this problem involves the estimation of the influence of a subset of input features on the predictions of a model. By understanding the important features, the model can be improved, its predictions can be trusted more, and undesirable behaviors can be isolated (Hooker et al., 2019). Saliency explanations that output a contribution score for each feature in an input have been widely adopted in DNNs. Important features in input can be found by ranking contributions scores. Thus, saliency methods can be regarded as a general solution for the issue of black-box models, and various saliency methods have been proposed such as VANILLA GRADIENTs (Baehrens et al., 2010; Erhan et al., 2009; Simonyan et al., 2013), INTEGRATED GRADIENTs (Sundararajan et al., 2017; Ancona et al., 2018), and Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016).

The explanations rendered by various saliency methods often diverge; for example, we apply saliency methods on 100 movie review samples and found that the overlap rates, i.e. consistency of

---

[*]Corresponding author.

	ALL ATTENTION	LAST ATTENTION	VANILLA GRADIENT	INPUT×GRAD	INTEGRATED GRADIENT	DEEPLIFT	LIME
ALL ATTENTION	100	50.4	57.5	31.2	34.8	31.8	28.7
LAST ATTENTION	50.4	100	48.5	32.0	33.9	32.0	29.0
VANILLA GRADIENT	57.5	48.5	100	37.65	40.3	37.9	28.8
INPUT×GRAD	31.2	32.0	37.65	100	37.0	87.2	28.3
INTEGRATED GRADIENT	34.8	33.9	40.3	37.0	100	37.5	29.7
DEEPLIFT	31.8	32.0	37.9	87.2	37.5	100	28.3
LIME	28.7	29.0	28.8	28.3	29.7	28.3	100

Table 1: **Overlap rate of seven different saliency methods based on top-40 words from 100 movie review samples.** For example, if we choose top-40 important words from ALL ATTENTION and LIME, there are $28.7\% \times 40 = 11.48$ identical words on average. Value of 100 implies perfect overlap and 0.0 denotes no overlap.

different saliency methods, were relatively low, as shown in Table 1. Therefore, evaluating saliency methods becomes an important problem.

Along these lines, many automated evaluation metrics have been proposed to verify different interpretable models (Madsen et al., 2022; Hooker et al., 2019; Pruthi et al., 2022; Zhou et al., 2021). However, such automated evaluation measures are not guaranteed to correctly evaluate the interpretability of saliency methods. There are essential differences between machine cognition and human cognition. For example, Su et al. (2019) found that changing a single pixel in an image can change the output of models, whereas such changes are generally completely imperceptible to humans. Therefore, the interpretability of saliency methods should ultimately be decided by humans and a high rating by an automated evaluation metric does not necessarily mean a high interpretability for humans. Jacovi et al. (2023) found that neighboring words could affect saliency explanations for human. de Langis and Kang (2023) leveraged human eye-tracking to study saliency explanations. Riordan (2020) discussed human-machine agreements on different models for a gradient-based saliency method. However, evaluating saliency methods from the perspective of human intelligence has not been fully investigated.

Crowdsourcing provides an efficient and economical way to solve various tasks by utilizing human intelligence. For example, Wang et al. (2019) proposed a mechanism of trust evaluation on edge computing using crowdsourcing. Zanca et al. (2020) exploited a crowdsourcing platform to evaluate human visual attention paths predicted by models. Intuitively, we consider that a better saliency method should assign higher scores to more important words, enabling humans to correctly infer properties (e.g., categories) of the original text by the given several top important words. The key idea of this study is to apply crowdsourcing to evaluating

saliency methods by whether crowd workers could distinguish labels of text based on the given top important words, as inspired by the measurements proposed by Lu et al. (2021) on image classification tasks. Based on the experimental results of the proposed evaluation method, we reach a list of findings on the saliency explanation methods in the NLP area.

The contributions of this study are summarized as follows. 1) We propose a novel human-based method to evaluate saliency methods in NLP tasks by crowdsourcing. We recruited 800 workers from a crowdsourcing platform to evaluate seven saliency methods on two datasets and verified the reliability of the data of annotations provided by crowd workers. 2) We found that INTEGRATED GRADIENT, a saliency method, performed best in our experiments and some saliency methods were not significantly better than random baseline. 3) We compared our results based on human evaluation with existing automated evaluation methods and found that there were differences between ours and existing methods. 4) We found a phenomenon called Flip, which makes saliency methods more unreliable in NLP than CV tasks. 5) We evaluated saliency methods on misclassified samples and found that saliency methods were also effective for misclassified samples.

2 Related Work

2.1 Saliency Methods

We first briefly review the saliency methods used in this study. Table 2 summarizes the saliency methods considered in this study.

Saliency methods typically assign scores for each feature in an input, where important features with greater influence on the output are expected to receive a higher score. Most of the saliency methods can be divided into two types, including gradient-based and perturbation-based (Slack et al., 2021). Gradient-based saliency methods cannot be

Method	ALL ATTENTION	LAST ATTENTION	VANILLA GRADIENT	INPUT×GRAD	INTEGRATED GRADIENT	DEEPLIFT	LIME
Type	Hidden States	Hidden States	Gradient-based	Gradient-based	Gradient-based	Gradient-based	Perturbation-based
Class-specific	✗	✗	✓	✓	✓	✓	✓

Table 2: **Saliency methods considered in this study.** "Class-specific" refers to whether a target class should be specified when using the method.

used for models that cannot be differentiable, such as random forest and perturbation-based saliency methods can be used in any model that maps an input to an output.

Gradient-based saliency methods assume that more important features have larger gradients. VANILLA GRADIENT (Baehrens et al., 2010; Erhan et al., 2009; Simonyan et al., 2013) is the most basic method, which measures the direct gradient of input features to quantify the influence of small changes in each input dimension on the output of the network. INPUT×GRAD (Shrikumar et al., 2016) multiplies the features in the input with the gradient with respect to the features. However, zero gradient does not necessarily mean zero contribution to the output. To solve this problem, INTEGRATED GRADIENT (Sundararajan et al., 2017; Ancona et al., 2018) is calculated by cumulating gradients of a straight-line path from given baselines to inputs. According to Shrikumar et al. (2017), **Deep Learning Important FeaTures (DEEPLIFT)** decomposes the contribution to the prediction output of a network into the contribution to each hidden neuron on a given input feature.

Perturbation-based saliency methods assume that perturbing more important features exhibits larger influences on the output. **LIME** (Ribeiro et al., 2016) samples points around inputs and trains a simpler interpretable surrogate model, such as a linear model, based on the samples. The weights of the simple model are used as the contribution scores of inputs.

Except for common saliency methods, hidden states, i.e. variables or outputs of hidden layers, are sometimes explanations. In CV tasks, the output of convolutional layers may highlight important parts of the input image (Zhou et al., 2016). In this study, we define ALL ATTENTION (Bahdanau et al., 2015) as the mean of all layers and all heads of self-attention layers (Vaswani et al., 2017) for a given word and LAST ATTENTION (Fernandes et al., 2022) as the mean of the last layer and all heads.

2.2 Saliency Method Evaluation

We briefly introduce previous work of evaluating saliency methods.

1) Relative Area Between-Curves (RACU): Madsen et al. (2022) utilized RemOve And Retrain (ROAR) (Hooker et al., 2019) in NLP, which removes the top important words and retrains the model. The performance of saliency methods was measured in terms of the **RACU**, which indicates changes in accuracy decline.

2) Effectiveness (Eff.): Pruthi et al. (2022) trained a student model with explanations of saliency methods of a teacher model and used **Eff.** to measure how saliency methods helped student models train better.

3) Simulability (Simu.): Similarly, Fernandes et al. (2022) uses **Simu.** to measure the simulation accuracy, i.e., what percentage of the student and teacher predictions matched over a dataset.

4) Plausibility (Psi.) and Faithfulness (Fait.): Ding and Koehn (2021) evaluated saliency methods from two perspectives, including **Psi.** and **Fait.** of model consistency on the PTB dataset (Marcinkiewicz, 1994). **Psi.** measures the similarity of explanations given by saliency methods and human annotations about the decision mechanism of a given model. **Fait.** measures the consistency of the explanations given by a saliency method for the same input and two different models.

5) Area Under the Precision-Recall curve (AUPRC): DeYoung et al. (2020) collected annotations of words that humans think are important for decision-making and used **AUPRC** to measure how well saliency methods agree with annotations marked by humans.

6) Sufficiency (Suff.) and Comprehensiveness (Comp.) were also proposed by DeYoung et al. (2020) inspired by Samek et al. (2016). **Suff.** measures how model output confidence drops when the model only takes the important part given by a saliency method and assumes that confidence should drop slightly if the saliency method performs well. **Comp.**, which is similar to ROAR but is not required to retrain the model, measures how

Figure 1: **Example of a real task for crowd workers.** The top-10 words of a negative review given by a saliency method are shown. A "." corresponds to a hidden word. All punctuation and special tokens of the original text are ignored. Workers were asked to choose whether the review was likely to be negative, positive, or indeterminate based on the words shown. In this example, because the word "Poor" was included, this review may be inferred to be negative. Workers may obtain additional information or make inferences based on the position of words, which is consistent with the position encoding of the model. "."s prevent words that are far away from each other in the original text from being visually close, and will not mislead workers.

model output confidence drops when the model takes the unimportant part, i.e., removing the important part from the original input, as the input and assumes that confidence should drop significantly if the saliency method performs well.

In this study, we evaluated saliency methods by humans directly. Crowd workers cannot see the full text while evaluating the important words given by saliency methods to make it easier for workers to make objective choices that are less dependent on subjective judgments of workers.

3 Crowd-based Evaluation Method

We evaluated saliency methods by showing the top important words given by the method to crowd workers. In the web interface, crowd workers are asked to perform a text classification task, i.e., assigning a label to a text from a set of labels based on the words provided. Figure 1 shows an example of a task. We denote by $\mathcal{L}_{ijkm} \in \{-1, 0, 1, 2, ..., C\}$ the annotation given by crowd worker m in the task t_{ijk} of showing top-k important words by the j-th saliency method for sample i, where $i \in \{1, 2, ..., N\}$, N is the number of samples, $k \in \{\mathbb{Z}^+\}^K$, K is the number of different top-k words settings, and C is the number of classes. $\mathcal{L}_{ijkm} = -1$ indicates that crowd worker m did not evaluate task t_{ijk}, $\mathcal{L}_{ijkm} = 0$ indicates that crowd worker m selected "I don't know" in the task t_{ijk} and $\mathcal{L}_{ijkm} = c \in \{1, 2, ..., C\}$ indicates that crowd worker m selected label c in the task t_{ijk}.

We denote by $\mathcal{A}_{ijk} \in \{0, 1\}$ the results of whether task t_{ijk} can be correctly recognized by crowd workers, where $\mathcal{A}_{ijk} = 1$ means recognizable and vice versa.

We aggregate annotations from multiple workers by majority voting because of the large amount of tasks and the very sparse annotations with $\frac{3900}{4000} = 97.5\%$ missing rate. For a sample i with ground truth label c, the aggregation can be represented by the following equation.

$$\mathcal{A}_{ijk} = \begin{cases} 1 & \text{if } \forall x, |\mathcal{L}_{ijkm} = c| > |\mathcal{L}_{ijkm} = x|, \\ 0 & \text{otherwise.} \end{cases}$$
(1)

where $x \neq -1$ and $x \neq c$.

Next, we measure the performance of saliency methods in terms of the aggregation results $\mathcal{A}_{ijk}$. We consider that if a saliency explanation is sufficiently "interpretable", the accuracy, i.e., probability of labels can be correctly recognized more accurately by humans based on the given top-k important words.

We denote by $p_{jk} = \frac{\sum_{i=1}^{N} \mathcal{A}_{ijk}}{N}$ the accuracy of the j-th saliency method for the top-k words. The performance of the j-th saliency method s_j can be considered as the weighted average of p_{jk} of each k. When k is relatively small, obtaining higher accuracy is more difficult than when k is large, so we assign smaller k larger weights. We use the mean accuracy of all saliency methods of the top-k words as the difficulty of the top-k words. The performance score can be represented as the following equation.

$$s_j = \sum_k w_k p_{jk} \text{ , where } w_k = \frac{\sum_j \sum_k p_{jk}}{K^2 \sum_j p_{jk}}.$$
(2)

The evaluation procedure consists of the following steps:

1) Prepare a pre-trained prediction model, a dataset, and several saliency methods to be evaluated.

2) Apply all saliency methods and a random baseline to each text sample.

3) Obtain saliency scores from each saliency method and the random baseline.

4) Set the settings of top-k important words.

5) Prepare task t_{ijk} of a given sample i, the j-th specified saliency method and show words k.

6) Get annotations $\mathcal{L}_{ijkm}$ by crowdsourcing.

7) Aggregate annotations $\mathcal{A}_{ijk}$ by majority voting.

8) Calculate performances s_j of each saliency method.

4 Experiments

We conducted experiments to answer the following three questions. 1) Which saliency methods performed best in human evaluations? 2) Do our results agree with the previous studies? 3) What is different between texts and images in the saliency methods? 4) What is the findings of evaluating misclassified cases?

4.1 Experimental settings

We used two datasets for the evaluation, including IMDB (Maas et al., 2011) and AGNEWS, with a max length of 256 tokens. IMDB is a binary text classification dataset that contains negative or positive movie reviews. AGNEWS is a subset of AG's corpus (Del Corso et al., 2005), which contains four classes of news items, including "World", "Sports", "Business" and "Science". We randomly selected 100 samples from both the IMDB and AGNEWS datasets of which 50 were negative movie reviews and 50 were positive in IMDB and 25 for each class in AGNEWS. We used BERT-Base model (Devlin et al., 2019) as the backbone and trained classifiers by adding a linear layer to the BERT-Base model on the two datasets. The implementation of the BERT-base model was based on PyTorch and from HuggingFace [1]. All hyperparameters were set by default. The top-k important words were set as 5, 10, 20, 30, 40 words, i.e., $k \in \{5, 10, 20, 30, 40\}$. We evaluated the seven saliency methods in Table 2 and a random baseline, i.e., assigning input features random contribution scores. The explanations of saliency methods were calculated based on Captum [2] and all settings are default.

In total, 4000 [3] tasks t_{ijk} of (sample i, the j-th saliency method, top-k words) were generated for each of two datasets. We used Lancers[4] crowd-sourcing platform for the evaluation. Each task was evaluated five times to collect reductant labels; the redundant labels are aggregated to ensure the reliability of the crowdsourcing results and eliminate the bias of individual workers.

Each crowd worker was required to evaluate 100 tasks for a reward of 330 JPY on Lancers. Each

---

[1] https://huggingface.co/
[2] https://captum.ai/
[3] 100 samples × 8 (7 Saliency methods+Random baseline) × 5 different word settings = 4000.
[4] https://www.lancers.jp/

worker could only see each sample once to ensure that workers would not remember the content of the text. Therefore, cases that a worker completed a task with showing more words of a sample, then saw the same sample with fewer words and completed the task by the memory of the previous task would not happen. 400 [5] crowd workers participated in the evaluations in total.

Workers completed the tasks in roughly 18 minutes on average. The hourly pay was about 1,100 JPY, which exceeds the minimum hourly wage, about 1,000 JPY, in Japan. Submissions of workers who completed the tasks within an extremely short period (less than 2 minutes) or showed obvious suspicions of inferior quality, such as selecting "I don't know" for all 100 tasks, were removed. No personal information or data was used in this study. All crowd workers received appropriate wages in line with local standards.

We randomly sampled 100 tasks out of 4,000 tasks in each dataset and let experts (members of our lab) evaluate these tasks to verify the reliability of the annotation data. The results show that the quality of the data collected was relatively reliable. The rate of agreement between experts and crowd workers, the rate of whether a task could be correctly recognized by both experts and crowd workers, was 86% in IMDB dataset and 93% in AGNEWS dataset.

4.2 Results and Discusisons

Q1. Which saliency methods performed best in humans evaluations? – INTEGRATED GRADIENT.

Because different saliency methods provided different results of top-k important words, we first investigated which saliency method yielded more reliable explanations for humans. It may be observed from Table 3 and 4 that INTEGRATED GRADIENT exhibited the best performance in the crowdsourced evaluation on both two datasets. This may be attributed to the fact that INTEGRATED GRADIENT satisfies two axioms (desirable characteristics), including Sensitivity and Implementation Invariance (Sundararajan et al., 2017), whereas VANILLA GRADIENT as well as INPUT×GRAD violates the Sensitivity axiom and DEEPLIFT violates the Implementation Invariance axiom.

ALL ATTENTION, LAST ATTENTION, VANILLA GRADIENT, and INTEGRATED GRA-

---

[5] 4000 tasks * 2 datasets * 5 times / 100 tasks = 400.

	5 Words	10 Words	20 Words	30 Words	40 Words	Score
RANDOM	25 (8)	42 (6)	41 (8)	55 (6)	54 (6)	42.9 (7)
ALL ATTENTION	66 (3)	74 (3)	81 (2)	**87 (1)**	84 (3)	78.5 (3)
LAST ATTENTION	53 (4)	60 (4)	73 (3)	72 (4)	78 (4)	66.9 (4)
VANILLA GRADIENT	**75 (1)**	**81 (1)**	73 (3)	80 (2)	85 (2)	79.5 (2)
INPUT×GRAD	46 (6)	46 (5)	60 (5)	56 (5)	52 (7)	52.1 (6)
INTEGRATED GRADIENT	74 (2)	**81 (1)**	**87 (1)**	80 (2)	**87 (1)**	**82.3 (1)**
DEEPLIFT	52 (5)	37 (7)	60 (5)	52 (7)	67 (5)	53.5 (5)
LIME	30 (7)	28 (8)	47 (7)	48 (8)	44 (8)	39.0 (8)

Table 3: **Evaluation results of different saliency methods on the IMDB dataset.** The numbers in brackets show the ranks of the saliency method. The bold numbers show the best results and the underlined numbers show the top-3 methods. The results of top-k words are the accuracy p_{jk} (%) and the last column "Score" are the overall performance s_j (%) of the j-th saliency method. Larger values indicate better performance.

	5 Words	10 Words	20 Words	30 Words	40 Words	Score
RANDOM	46 (7)	57 (8)	70 (5)	77 (7)	73 (8)	64.1 (7)
ALL ATTENTION	73 (2)	73 (2)	77 (2)	79 (3)	80 (2)	76.9 (2)
LAST ATTENTION	56 (6)	72 (3)	77 (2)	79 (3)	**81 (1)**	72.8 (4)
VANILLA GRADIENT	65 (3)	72 (3)	69 (6)	80 (2)	77 (4)	72.9 (3)
INPUT×GRAD	61 (4)	68 (5)	76 (4)	78 (6)	74 (7)	71.5 (5)
INTEGRATED GRADIENT	**79 (1)**	**81 (1)**	**87 (1)**	79 (3)	80 (2)	**82.0 (1)**
DEEPLIFT	61 (4)	61 (6)	69 (6)	72 (8)	77 (4)	68.1 (6)
LIME	33 (8)	61 (6)	69 (6)	**82 (1)**	75 (6)	62.9 (8)

Table 4: **Evaluation results of different saliency methods on the AGNEWS dataset.** Same format with Table 3.

Metrics	Ours↑	RACU ↑	Eff.↑	Simu.↑	Psi.↑	Fait.↑	AUPRC ↑	Suff.↓	Comp.↑
					Settings				
Model	Bb	Rob	Bb	Tf	Tf	Tf	Bb+LSTM	Bb	Bb
Dataset	IMDB	IMDB	IMDB	IMDB	PTB	PTB	IMDB	IMDB	IMDB
Human-included	✓	✗	✗	✗	✓	✗	✓	✗	✗
					Results				
Random	42.9(7)	-	91.5(7)	-	-	-	0.259(4)	0.29(5)	0.04(7)
ALL ATTENTION	78.5(3)	-	**94.55(1)**	**86.27(1)**	-	-	**0.417(1)**	0.11(2)	0.28(2)
LAST ATTENTION	66.9(4)	-	-	85.18(2)	-	-	-	-	-
VANILLA GRADIENT	79.5(2)	25.4(2)	92(4)	83.14(4)	0.551(2)	0.160(2)	0.385(2)	0.25(4)	0.11(4)
INPUT×GRAD	52.1(6)	16.9(3)	91.95(5)	83.27(3)	-	-	-	0.33(6)	0.06(5)
INTEGRATED GRADIENT	**82.3(1)**	**35.1(1)**	93.1(2)	82.99(5)	**0.734(1)**	**0.239(1)**	-	0.13(3)	0.17(3)
DEEPLIFT	53.5(5)	-	92.7(3)	-	-	-	-	0.39(7)	0.06(5)
LIME	39.0(8)	-	91.95(5)	-	-	-	0.280(3)	**0.06(1)**	**0.32(1)**

Table 5: **Results of our method and previous studies on evaluating saliency methods.** ↑ indicates larger values are better and ↓ indicates smaller values are better. The bold numbers show the best results and the underlined numbers show top-3 methods. "Ours" refers to the overall performance scores of our proposed evaluation method. All abbreviations of metrics are defined in Section 2.2. All data are the original data reported in the previous studies. Especially, "Eff." and "Simu." are the average of different experiment settings. Data of "Comp." and "Suff." were reported by Pruthi et al. (2022) because of containing more saliency methods. "Bb", "Rob" and "Tf" are abbreviations of BERT-base, RoBERTa (Liu et al., 2019b) and Transformer, respectively.

DIENT performed significantly better than the random baseline on both two datasets; however, INPUT×GRAD, DEEPLIFT and LIME did not. This most likely occurred because saliency methods cannot directly calculate the contribution scores for the words (tokens) in NLP tasks, but can calculate scores directly for pixels in CV tasks. Instead, the scores are calculated on token embeddings, and the calculation process may thus

become more complicated and fail to achieve the desired performance. ALL ATTENTION performed better than LAST ATTENTION on both datasets, which indicates that each self-attention layer included valuable information.

Next, we discuss the impact of different datasets on the results. The rankings of saliency methods roughly remained the same on the two datasets with some differences. The results for IMDB are

generally better than those for AGNEWS when few words were shown, although IMDB is a binary classification dataset and AGNEWS is a 4-classification dataset. These results show that humans only need 5 words to achieve relatively high accuracy, but showing more words may not increase the accuracy for AGNEWS and INTEGRATED GRADIENT. We discuss this further below.

Q2. Do our results agree with the previous studies? – Some agree and some disagree.

We compare the results of previous evaluation methods mentioned in Section 2.2 with ours in Table 5. Although models, datasets, and experiment settings in these works differ from each other somewhat, findings can be summarized as follows.

1) These results show that RACU, Psi. and Fait. completely agree with our results, but the results of other automated metrics were not all completely the same as our results. Psi. and AUPRC include human participation but are not directly human evaluations. The differences between our results and AUPRC may be attributed to the fact that the spans or sentences marked by human annotations were not very accurate or rankings of words were not provided. It is also possible that the reason that these three metrics agree with ours completely was that relatively few saliency methods were included in the experiments.

2) INTEGRATED GRADIENT performed best among gradient-based saliency methods, and VANILLA GRADIENT performed better than INPUT×GRAD, except for Simu. metric.

3) Comp. and Suff. show that LIME performed best, which was the opposite of our results. These findings show that saliency methods that perform well in terms of automated evaluation metrics do not necessarily perform well in terms of human evaluation.

Q3. What is the difference between the results of the saliency methods for texts and images? – The Flip phenomenon.

Most studies on saliency methods were first introduced in CV. Finally, we discuss a phenomenon called **Flip** which occurs frequently in NLP but is rarely considered in CV. A sample i is Flipped with the j-th saliency method when humans can correctly distinguish the label of a sample i when showing fewer words/pixels, but can not distinguish with showing more words/pixels, i.e. $\forall k_1 \exists k_2 > k_1, \mathcal{A}_{ijk_1} = 1$ and $\mathcal{A}_{ijk_2} = 0$. We show
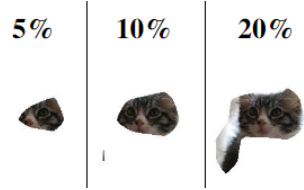

Figure 2: **An example (Lu et al., 2021) of top-**5% / 10% / 20% **important pixels given by vision saliency method GradCAM (Selvaraju et al., 2017).** Regardless of the saliency method, once humans can recognize that the images with fewer pixels show a cat, it is almost impossible for us to fail to recognize the same cat image with more pixels. However, there are two exceptions. The first is that the image is ambiguous. For example, in the cat/dog classification problem, an image may include both a cat and a dog. Another issue is rarer, that is, the position of important pixels can express the outline of an object's shape, such as by showing an outline of a cat's head against a gray sky (Gupta et al., 2022).
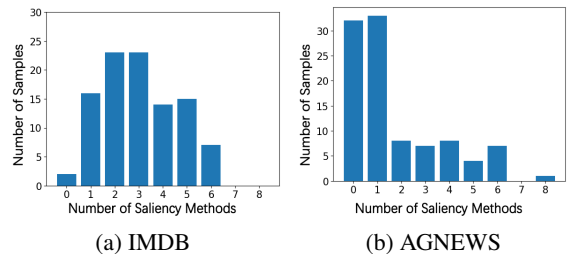

(a) IMDB   (b) AGNEWS

Figure 3: **Histograms of Flips in the two datasets.** The horizontal and vertical axes indicate the number of saliency methods where a sample Fliped with and the number of samples, respectively. For example, over 30 samples among 100 samples in the AGNEWS dataset did not Flip with all the 8 saliency methods (including the random baseline). Regardless of saliency methods, some samples are easier to Flip, i.e., contain more misleading words.

a text example in Table 6 and an image example in Figure 2, respectively.

We consider that the Flip phenomenon is saliency method-agnostic but attribute this to the properties of words themselves. Unlike pixels, each word has the potential to have strong expressive power to affect the meaning of the entire sample. Figure 3 indicates that some samples were inclined to be Flipped and Table 7 indicates that human recognition is not necessarily aided by showing more words. We noticed that Flip appeared more in the IMDB dataset, possibly because the sentiment of movie reviews was more easily influenced by individual words. Based on these findings, we suggest that future research on the application of saliency methods in NLP may consider the Flip phenomenon.

Text	# Words	Correct
....writing ....problem ....movie ....insults ....ruins	5	✓
...jason ....creative writing ....problem ....movie ....insults ....ruins .life ..this ....movie ..	10	✓
...jason ....creative writing .just ....crush ....problem ....saying ....aback .....movie.............course ....words .insults ....tell ....ruins . life.. this .the ....well ...movie .might	20	✗
...jason ....built ....finishes .creative writing .just .its .....fast ....crush ....problem .jason ....other ..saying ....aback .....movie ....course ...believe .....words .insults . jason ....tell ....ruins .life ..this .the short ...big ..well ...movie .might	30	✗
...jason ....built ....cannot ......finishes .creative writing .just.its .....fast .hand ......crush ....the problem .jason ....other way of saying ....movie ....aback .....movie ....course ...believe. ....words .insults .jason ..go ......tell ....ruins .life ..this.the short ...big fat .well .a .movie it might	40	✗

Table 6: **An example of the top important words of a negative movie review provided by INTEGRATED GRADIENT.** Humans correctly recognized that it is negative when showing 5 and 10 words, but failed when showing 20, 30 and 40 words. It is probably because the words "well" and "movie" affect human judgment.

	IMBD		AGNEWS	
	Flips	Aids	Flips	Aids
RANDOM	42	46	19	41
ALL ATTENTION	18	27	20	19
LAST ATTENTION	27	36	10	39
VANILLA GRADIENT	31	18	29	26
INPUT×GRAD	54	33	28	26
INTEGRATED GRADIENT	28	20	15	15
DEEPLIFT	50	34	23	33
LIME	54	42	23	54

Table 7: **Numbers of Flipped samples among 100 samples in two datasets.** "Aids" means that showing more words aids humans in correctly recognizing the label. That is $100-\#$ Flips$-\#$ Samples that can be correctly recognized in all the top-words settings.

Q4. What is the findings of evaluating misclassified cases? – Effective for explaining models.

In the previous experiments, we followed the settings that are usually used in the existing works (DeYoung et al., 2020; Samek et al., 2016), in which all chosen data samples were predicted by the model correctly, because saliency methods usually could not perform well on misclassified samples. However, misclassified cases are also important because analyzing misclassified cases is beneficial for understanding the model better. Therefore, we also conducted the experiments with misclassified cases by randomly choosing 100 misclassified samples from each dataset, and recruited 400 crowd workers to conduct the same experiments, as shown in Table 8 and 9. The findings of the experimental results are list as follows.

1) Comparing with Table 3 and 4, the performance of the random baseline has dropped significantly, which indicates the inherent difficulty of misclassified samples are relatively high and more ambiguous regardless of the model.

2) The performance of all methods also drops significantly, indicating that the correct top important words for correct prediction are not highlighted by the model. Therefore, saliency methods properly present the issue of misclassifications.

3) Showing more words does not necessarily improve the performance. This phenomenon is also present in the correct classified samples shown in Q3, but there are more ambiguous words in misclassified samples, which may even reduce the overall accuracy.

4) Especially, INTEGRATED GRADIENT obtained 50% accuracy and outperforms other saliency methods significantly when showing top 5 words on the AGNEWS dataset; however, the low performance of ALL ATTENTION indicates that the self-attention layers do not give high weights to the important words. Misclassification may tend not to occur when ALL ATTENTION performs well because attentions are based on the self-attention layers of the model. The low performance of VANILLA GRADIENT indicates that the model is not currently highlighting top important words, i.e. the model is not currently learning effectively on misclassified samples. Compared with VANILLA GRADIENT, the reason why INTEGRATED GRADIENT performs well when showing 5 words may be attributed to that INTEGRATED GRADIENT has the property of completeness, which may highlight potential top important words that can be learned by the model.

5 Conclusion

In this study, we proposed an approach to evaluate saliency explanations in natural language process-

	5 Words	10 Words	20 Words	30 Words	40 Words
RANDOM	19	45	43	49	38
ALL ATTENTION	15	25	41	35	38
LAST ATTENTION	12	32	35	40	37
VANILLA GRADIENT	33	39	37	31	50
INPUT×GRAD	28	20	45	35	43
INTEGRATED GRADIENT	21	44	31	38	30
DEEPLIFT	29	25	33	44	34
LIME	14	33	36	35	34

Table 8: **Evaluation results of misclassified cases in the IMDB dataset.** The number is the count of samples which could be recognized by humans correctly from 100 samples which were all misclassified by the model.

	5 Words	10 Words	20 Words	30 Words	40 Words
RANDOM	15	20	26	23	24
ALL ATTENTION	24	22	16	30	26
LAST ATTENTION	22	25	21	25	23
VANILLA GRADIENT	29	21	23	25	23
INPUT×GRAD	26	30	27	20	22
INTEGRATED GRADIENT	50	40	37	34	26
DEEPLIFT	28	31	21	22	21
LIME	18	22	23	25	24

Table 9: **Evaluation results of misclassified cases in the AGNEWS dataset.** The format is the same as Table 8.

ing by crowdsourcing. The proposed approach assesses the interpretability of saliency methods in terms of human evaluations. Based on this approach, We recruited a total of 800 crowd workers to conduct our experiments for evaluating seven different saliency methods on two datasets.

The scale of our experiments was limited by the cost of crowdsourcing because real crowdsourcing experiments are costly and adding more datasets or models will dramatically increase the cost. Although the number of datasets is relatively small and the results were only evaluated on the BERT-base model, the main purpose of this work is to claim that there are differences between human evaluation and automated evaluations and there are indeed differences in our experiments of one model and two datasets. As long as there are differences in one model, there are differences. Although more experiments will support our claim, we would like to caution further studies when using automated evaluation methods. Four takeaways can be summarized as follows based on the results.

1) INTEGRATED GRADIENT performed best in our experiments. ALL ATTENTION, LAST ATTENTION and VANILLA GRADIENT performed significantly better than the random baseline on both two datasets. DEEPLIFT and LIME performed close to or worse than the baseline, probably because the contribution scores were calculated on token embeddings but not directly on tokens (words), which

leads to additional complexity in the calculation.

2) There were differences between our results based on human evaluation and the results based on some of the existing automated evaluation metrics. Except for our results, INTEGRATED GRADIENT also performed best among Gradient-based saliency methods in most evaluation metrics.

3) A phenomenon called Flip occurred in NLP which is saliency method-agnostic and rarely applied in CV. We also found that showing more words to humans may not help them correctly recognize the labels of a given text. In future work, we will research how to avoid the Flip phenomenon of saliency methods in NLP.

4) The reasons for misclassification can be attributed to two perspectives. The first is that the inherent difficulty of misclassified samples is relatively high according to the performance of the random baseline. The other is that the model failed to highlight the correct top important words according to the poor performances of saliency methods.

Acknowledgements and Ethics Statement

No personal information or data was used in this study. All crowd workers received appropriate wages in line with local standards.

References

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Karin de Langis and Dongyeop Kang. 2023. A comparative study on textual saliency of styles from eye tracking, annotations, and language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 108–121, Singapore. Association for Computational Linguistics.

Gianna M Del Corso, Antonio Gulli, and Francesco Romani. 2005. Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web*, pages 97–106.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Shuoyang Ding and Philipp Koehn. 2021. Evaluating saliency methods for neural language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5034–5052, Online. Association for Computational Linguistics.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2009. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1.

Patrick Fernandes, Marcos Treviso, Danish Pruthi, André Martins, and Graham Neubig. 2022. Learning to scaffold: Optimizing model explanations for teaching. In *Advances in Neural Information Processing Systems*, volume 35, pages 36108–36122. Curran Associates, Inc.

Arushi Gupta, Nikunj Saunshi, Dingli Yu, Kaifeng Lyu, and Sanjeev Arora. 2022. New definitions and evaluations for saliency methods: Staying intrinsic, complete and sound. In *Advances in Neural Information Processing Systems*.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32:9737–9748.

Alon Jacovi, Hendrik Schuff, Heike Adel, Ngoc Thang Vu, and Yoav Goldberg. 2023. Neighboring words affect human interpretation of saliency explanations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11816–11833, Toronto, Canada. Association for Computational Linguistics.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019a. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaotian Lu, Arseny Tolmachev, Tatsuya Yamamoto, Koh Takeuchi, Seiji Okajima, Tomoyoshi Takebayashi, Koji Maruhashi, and Hisashi Kashima. 2021. Crowdsourcing evaluation of saliency-based xai methods. In *Joint European Conference on*

Machine Learning and Knowledge Discovery in Databases, pages 431–446. Springer.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

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

Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Using Large Corpora*, 273.

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

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating Explanations: How Much Do Explanations from the Teacher Aid Students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144.

Brian Riordan. 2020. Analyzing saliency in neural models for scoring content in science explanations. *BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

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

Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2021. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, 34:9391–9404.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tian Wang, Hao Luo, Xi Zheng, and Mande Xie. 2019. Crowdsourcing mechanism for trust evaluation in cpcs based on intelligent mobile edge computing. *ACM Trans. Intell. Syst. Technol.*, 10(6).

Dario Zanca, Stefano Melacci, and Marco Gori. 2020. Toward improving the evaluation of visual attention models: a crowdsourcing approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593.
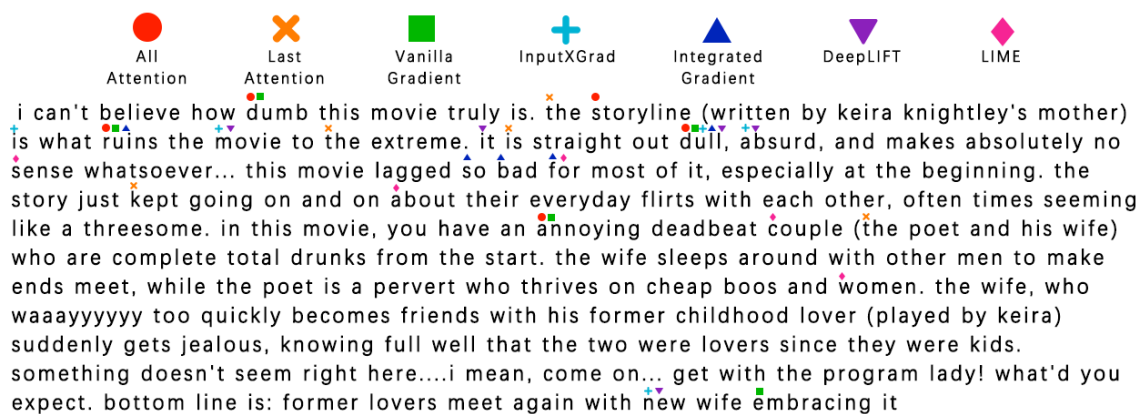
Figure 4: An example of the top five important words given by seven different saliency methods for a negative movie review in the IMDB dataset. A word with a mark (on the top of a word) represents as one of the top five important by a saliency method; marks with different colors and shapes represent different saliency methods. For example, the top five important words given by LIME consist of "sense", "woman", "for", "about", and "couple". The word "dull" is selected by five saliency methods. Even when some words are the same, different saliency methods are inconsistent with each other.

A   A demonstration of different saliency methods

Figure 4 shows the differing results (outputs) of different saliency methods for an example movie review. It could be found that LIME focused on some prepositions, such as "about" and "for" which are relatively meaningless for humans.

B   Task Instruction for Crowd Workers

The instruction for the AGNEWS dataset is as follows, which is also similar to the instruction for the IMDB dataset. The original instruction was in Japanese. Even though Japanese native speakers have disadvantages in reading English text, we found that the quality of annotations collected in Lancers is higher than Amazon Mechanical Turk[5] empirically.

[Read me first]

This task requires intermediate English reading ability.

We are collecting data on English news categories ("Worlds," "Sports," "Business," and "Science" categories) for use in research on artificial intelligence and machine learning.

Information that can identify an individual or account, such as the user id of Lancers, will not be disclosed to third parties other than the task requester.

If you have any other questions, please contact us by message.

[Task instruction]

This task displays 100 different English news. Classify each news item into one of the four categories "Worlds", "Sports", "Business" and "Science".

"Worlds" is news about world politics, crime, and conflict.

Example: "Guatemala to pay paramilitaries Guatemala's government agrees to pay millions of dollars to former paramilitaries, who some accuse of war crimes."

"Sports" is news about sports.

Example: "Iranian Will Not Meet Israeli in Olympics Due to Weight ATHENS (Reuters) - A diplomatic wrangle of Olympic proportions was avoided Sunday when Iranian world judo champion Arash Miresmaeili failed to make the weight for his clash with an Israeli opponent."

"Business" is news about finance, economy, and business.

---

[5]https://www.mturk.com/

Example: "SEC seeks to update stock sale rules BEIJING, Aug.18 (Xinhuanet) – The US Securities and Exchange Commission (SEC) may revise its rules on initial stock offerings to modernize 1930s-era guidelines, agency commissioners and staff said. ”

"Science" is news about science, technology, and technology products.

Example: "India's leading IT company TCS to list stock next week after share offer (AFP) AFP - Top Indian software exporter Tata Consultancy Services will list its share on the Bombay Stock Exchange next week following its successful maiden share offer, an official said. ”

However, in the task, only a few selected words from the original news are presented. For example, if the original news is "Israeli Attack in Gaza Kills Five Palestinians, Haaretz Says Aug. 18 (Bloomberg)".

What is presented is only a few discontinuous words, such as " . Attack .. Kills ... Says .. .. ". ("." indicates a hidden word.)

Judge the class of news only from the words presented.

In this example, since the words "Attack" and "Kills" are included, it can be inferred that this news is probably related to world conflicts so that "Worlds" is its class.

If you cannot judge the class from the displayed words alone, select "I don't know". Also, if there are any English words you don't understand, you can translate them into Japanese using a translation website (for example, Google Translate).

This task can only be completed once per person. Even if you have multiple Lancers accounts, you can only do this once. That is all for the instruction. If there is no problem, please press the "Start" button.