Saliency Map Verbalization: Comparing Feature Importance Representations from Model-free and Instruction-based Methods

Nils Feldhus[1] **Leonhard Hennig**[1] **Maximilian Dustin Nasert**[1,2]
Christopher Ebert[1,2] **Robert Schwarzenberg** **Sebastian Möller**[1,2]
[1] German Research Center for Artificial Intelligence (DFKI)
[2] Technische Universität Berlin
{firstname.lastname}@dfki.de

Abstract

Saliency maps can explain a neural model's predictions by identifying important input features. They are difficult to interpret for laypeople, especially for instances with many features. In order to make them more accessible, we formalize the underexplored task of translating saliency maps into natural language and compare methods that address two key challenges of this approach – what and how to verbalize. In both automatic and human evaluation setups, using token-level attributions from text classification tasks, we compare two novel methods (search-based and instruction-based verbalizations) against conventional feature importance representations (heatmap visualizations and extractive rationales), measuring simulatability, faithfulness, helpfulness and ease of understanding. Instructing `GPT-3.5` to generate saliency map verbalizations yields plausible explanations which include associations, abstractive summarization and commonsense reasoning, achieving by far the highest human ratings, but they are not faithfully capturing numeric information and are inconsistent in their interpretation of the task. In comparison, our search-based, model-free verbalization approach efficiently completes templated verbalizations, is faithful by design, but falls short in helpfulness and simulatability. Our results suggest that saliency map verbalization makes feature attribution explanations more comprehensible and less cognitively challenging to humans than conventional representations. [1]

1 Introduction

Feature attribution methods, or (input) saliency methods, such as attention- or gradient-based attribution, are the most prominent class of methods for generating explanations of NLP model behavior (Wallace et al., 2020; Madsen et al., 2022) and can be used to produce word-level importance scores without human supervision (Wallace et al., 2019; Sarti et al., 2023). A major limitation of saliency maps is that they require expert knowledge to interpret (Alvarez-Melis et al., 2019; Colin et al., 2022). Furthermore, Schuff et al. (2022) revealed visual perception and belief biases which may influence the recipient's interpretation.

Natural language explanations (NLEs), on the other hand, exceed other explainability methods in plausibility (Lei et al., 2016; Wiegreffe and Pinter, 2019; Jacovi and Goldberg, 2020), accessibility (Ehsan and Riedl, 2020), and flexibility (Brahman et al., 2021; Chen et al., 2023), i.e. they can be adapted to both different target tasks and different audiences. Most previous approaches in generating NLEs depend on datasets of human-annotated text highlights (Zaidan et al., 2007; Lei et al., 2016; Wiegreffe and Marasović, 2021) or carefully constructed gold rationales for supervised training (Camburu et al., 2020; Wiegreffe et al., 2022), which are costly to obtain and task-specific. Alignment of model rationales with very few human-acceptable gold rationales may raise issues of trust (Jacovi et al., 2021) and the models trained on them may suffer from hallucinations (Maynez et al., 2020).

In this work, we revisit and formalize the task of verbalizing saliency maps, i.e. translating the output of feature attribution methods into natural language (Forrest et al., 2018; Mariotti et al., 2020; Slack et al., 2022). Verbalizations can describe relations between words and phrases and their associated saliency scores. Contrary to conventional heatmap visualizations, we can adjust the comprehensiveness of an explanation more precisely and infuse it with additional semantics such as word meanings, concepts, and context about the task.

We find that verbalization also comes with a few caveats: Similar to human explainers, who communicate only the most relevant explanations to avoid cognitive overload of the recipient (Hilton,

[1]Code and data at `https://github.com/DFKI-NLP/SMV`.

Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2023), pages 30–46
June 13, 2023

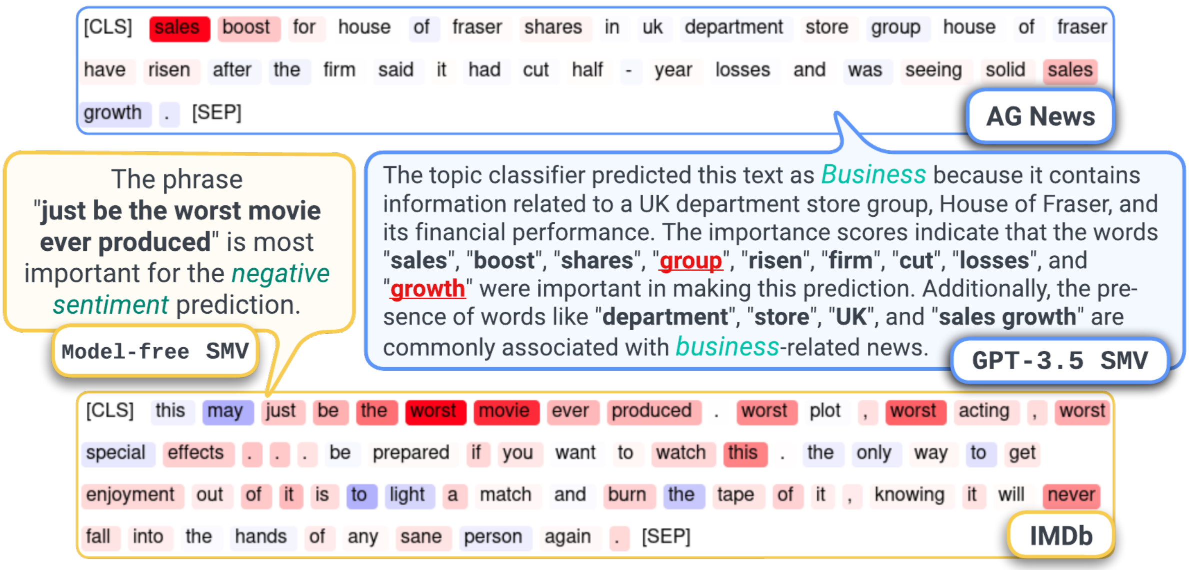

Figure 1: Heatmap visualizations generated by the Integrated Gradients feature attribution method explaining the predictions of a BERT model: Correct classifications of an instance from AG News (top) as *Business* and an instance from IMDb (bottom) as *Negative sentiment*. Tokens with red backgrounds have higher importance scores, while blue backgrounds indicate the contrast case. Two verbalizations (SMVs) are depicted in the center of the figure: The left (yellow) is produced by our model-free approach, while the right (blue) is produced by GPT-3.5. The predicted labels are highlighted in cyan and italic. The model-generated verbalization conveys semantic information such as associations with the target label (*Business*) and reasoning that is disconnected from the underlying model. GPT-3.5 wrongly deems two of the least attributed tokens salient ("group" and "growth", highlighted in red).

2017; Miller, 2019), verbalization methods need to address the problem of deciding "what" to say, i.e. selecting the most informative and useful aspects of the saliency maps and communicating them in a concise manner. We therefore compare different methods for verbalizing saliency maps: Supervised rationales, prompting LLMs, and model/training-free templates.

We address the problem of saliency map verbalization (**SMV**) with the following contributions:

- We formalize the underexplored task of SMV and establish desiderata, i.e. simulatability, explainer-faithfulness, plausibility, and conciseness (§2.1);
- We conduct a comparative study on various representations of feature attribution in two text classification setups, measuring the effects of verbalizations methods on both automated (explainer-faithfulness) and human evaluation metrics (simulatability, helpfulness, ease of understanding) (§3, §5).
- We propose a novel, model-free, template-based SMV approach, and design instructions for GPT-3.5-generated SMVs (§4) (examples from our two setups are depicted in Fig. 1);
- We show that model-free SMVs perform slightly better than heatmaps and extractive rationales on ease of understanding and are faithful by design, while instruction-based SMVs achieve the highest average simulation accuracy and are preferred in subjective ratings (§6);
- We publish a large dataset of model-free and GPT-generated SMVs alongside extractive rationales and results from both evaluations, and open-source code to produce all kinds of SMVs.

2 Verbalizing saliency maps

2.1 Formalization

The setup of the saliency map verbalization task consists of an underlying (to-be-explained) **model** m whose prediction $\hat{y} \subset Y$ on source tokens $W = w_1 \ \ldots \ w_n$ we want to explain (against the set of possible outcomes Y).

m is equipped with a feature explanation method (or short: **explainer**) e which produces a **saliency map** $S = s_1 \ldots s_n$:

$$e(W, m) = S \tag{1}$$

Here, we call token w_i salient *towards* outcome y if its associated saliency score $s_i > 0$ and salient *against* y for $s_i < 0$. e can have many sources, e.g. gradient-based methods such as Integrated Gradients (Sundararajan et al., 2017) which we employ in our experiments (§5), or even human experts assigning relevance scores.

A **verbalized saliency map** S_{V} is produced by some verbalizer v that receives the output of e:

$$v(W, S) = S_{\text{V}} \tag{2}$$

v can be any function that discretizes attribution scores and constructs a natural language representation S_{V}. This is connected to the concept of hard selection in DeYoung et al. (2020) and heuristics for discretizing rationales (Jain et al., 2020). In the taxonomy of Wiegreffe and Marasović (2021), verbalized saliency maps can be categorized as free-text rationales with varying degrees of structure imposed through templates. Moreover, verbalized explanations are procedural and deterministic by nature, i.e. they function as instructions that one can directly follow (Tan, 2022) to understand a model's decision, similar to compositional explanations (Hancock et al., 2018; Yao et al., 2021).

2.2 Desiderata

In the following, we outline the common evaluation paradigms for explanations (faithfulness, simulatability, plausibility) and how we adapt them to saliency map verbalizations.

Faithfulness Saliency maps express that "certain parts of the input are more important to the model reasoning than others" (*linearity assumption* in Jacovi and Goldberg (2020)). For verbalizations, explainer e and verbalizer v are two separate processes, so the saliency map S can be seen as static. Therefore, the faithfulness of e to the model m is extrinsic to the verbalization. Instead, it is essential to faithfully translate S into natural language, which we coin **explainer-faithfulness**. The verbalizer breaks faithfulness, e.g. if words are referenced as salient in S_{V} that are made up (do not appear in W) or if the polarity of any s_i is falsely interpreted.

Simulatability Another type of faithfulness is the model assumption which requires two models to "make the same predictions [iff] they use the same reasoning process" (Jacovi and Goldberg, 2020). By extension this means a model has to be simulatable (Doshi-Velez and Kim, 2017; Hase and Bansal, 2020), i.e. a human or another model should be able to predict a model's behaviour on unseen examples while exposed only to the explanation and not the model's prediction.

Plausibility The plausibility of explanations is commonly measured by correlation with ground-truth explanations (DeYoung et al., 2020; Jacovi and Goldberg, 2020), since gold rationales are influenced by human priors on what a model should do.

Conciseness In addition to these paradigms, verbosity is also an important aspect. A full translation into natural language is nonsensical, however, because all relations between the continuous-valued saliency scores and the associated tokens would normally overload human cognitive abilities. We want S_{V} to be concise, yet still contain the key information, similar to sufficiency and comprehensiveness measures from DeYoung et al. (2020). Thus, we define a **coverage** measure to indicate how much information is retained going from S to S_{V}, i.e. how much of the total attribution in $S = s_1 \ldots s_n$ is referenced by the tokens mentioned in $S_{\text{V}} = v_1 \ldots v_m$:

$$\text{Coverage}(S_{\text{V}}) = \frac{\sum |v_i|}{||S||} \tag{3}$$

The goal here is not to achieve a coverage of 1 with all of S, but depending on the use case, S_{V} should mention the most influential tokens, so a trivial solution for $k = 5$ would be to include the top k tokens with the highest attribution in S.

3 Study setup

3.1 Human Evaluation

Inspired by previous crowd studies in explainability (Chandrasekaran et al., 2018; Strout et al., 2019; Hase and Bansal, 2020; Sen et al., 2020; González et al., 2021; Arora et al., 2022; Joshi et al., 2023), we propose to measure **simulatability** as well as ratings for helpfulness and ease of understanding (**plausibility**). We evaluate the quality of different verbalization methods in a study involving 10 human participants. All participants have a computational linguistics background, with at least a Bachelor's degree, limited to no prior exposure to explainability methods, and are proficient in English (non-native speakers). After an introduction to the goal of the study and a brief tutorial, annotators are to complete the tasks described below. For each task, we present text instances along with their explanations, using a simple Excel interface.[2]

[2] See Appendix C, Figure 7

Task A: Simulation In the first task, participants are asked to simulate the model, i.e. predict the model's outcome, based only on one type of explanation plus the input text ("What does the model predict?"). They are given the possible class labels and were given an example for each dataset in the tutorial before starting the session. If the explanation does not provide any sensible clues about the predicted label, they still have to select a label, but may indicate this in the following question B1.

Task B: Rating In the second task, participants have to provide a rating on a seven-point Likert scale about (B1) "how helpful they found the explanation for guessing the model prediction" and (B2) "how easy they found the explanation to understand". A higher rating indicates a higher quality of the explanation.

Task C: Questionnaire Finally, participants are asked to complete a post-annotation questionnaire to obtain overall judgements for each verbalization method. They are prompted for Likert scale ratings about time consumption, coherence, consistency and qualitative aspects of each verbalization method, as listed in Table 1.

3.2 Automated Evaluation

We expect hallucinations (synthesized, factually incorrect text due to learned patterns and statistical cues) from GPT-type models and thus devise the following tests measuring **explainer-faithfulness** and **conciseness**:

1. Have the referred words been accurately cited from the input text?
2. How often do the referred words represent the top k most important tokens? (Eq. 3)

We obtain the results by simple counting and automated set intersection.

4 Methods

To complement heatmap visualizations and extractive rationales, we propose and analyze two additional verbalization methods: Model-free (§4.1, Fig. 2) and instruction-based (§4.2, Fig. 3) saliency map verbalization.

4.1 Model-free verbalization

For our model-free approach we employ hand-crafted templates for surface realization, different binary filter algorithms as search methods (§4.1.1) and scoring metrics (§4.1.2) to select tokens for filling the templates. This approach does not require architectural changes to the underlying model or modifications to an existing saliency method. The most similar approach to our selection heuristics, to our knowledge, are the discretization strategies in Jain et al. (2020, §5.2).

In the following, we will present two distinct candidate generation methods that can both be combined with one of two scoring metrics. A final candidate selection (§4.1.3) will collect the results from both searches, concatenate them to possibly larger spans and filter the top scoring candidates once more while maximizing coverage (Eq. 3). These salient subsets are then used to complete hand-crafted templates (App. E). We argue that this is more human-interpretable than simple top k single token selection, at the cost of a lower coverage. Our methodology allows to set parameters in accordance to how faithful the verbalization should be to the underlying explainer.

4.1.1 Explanation search

To acquire potentially salient snippets from a given text, we perform a binary selection on a window of attributions from the input of size c and then compare the sum of our selection to one of our scoring methods, performing basic statistical analysis on the window and the input.

Convolution Search Inspired by the convolutions of neural networks, we compare tokens that are located close to each other but are not necessarily direct neighbors. Coherence between pairs of tokens is solely determined by looking at their attributions with the following binary filters. In short, the following method firstly generates template-vectors that we then permute and keep as our binary filters. After computing all valid and sensible permutations, we can start calculating possibly salient or coherent snippets of our input. We choose $b \in \mathbb{N}$ vectors with a length of $c \in \mathbb{N}$. We describe these b vectors v_i as follows:

$$v_i = [1_{1,i}, 0_{1,c-i}], \quad v_i \in \mathbb{Z}^{1,c}. \tag{4}$$

e.g., for $i = 3$, $c = 5$, $v_i = (1\ 1\ 1\ 0\ 0)$

We only keep those v_i where $\sum v_i \notin \{0, 1, c\}$ in order to perform sensible permutations. For each v_i, we define a filter $\boldsymbol{f}_{i,j}$, where each distinct entry in $\boldsymbol{f}_i$ is a unique permutation of v_i. Let A be our attribution input, with $A \in \mathbb{R}^{1,k}$, where k is the

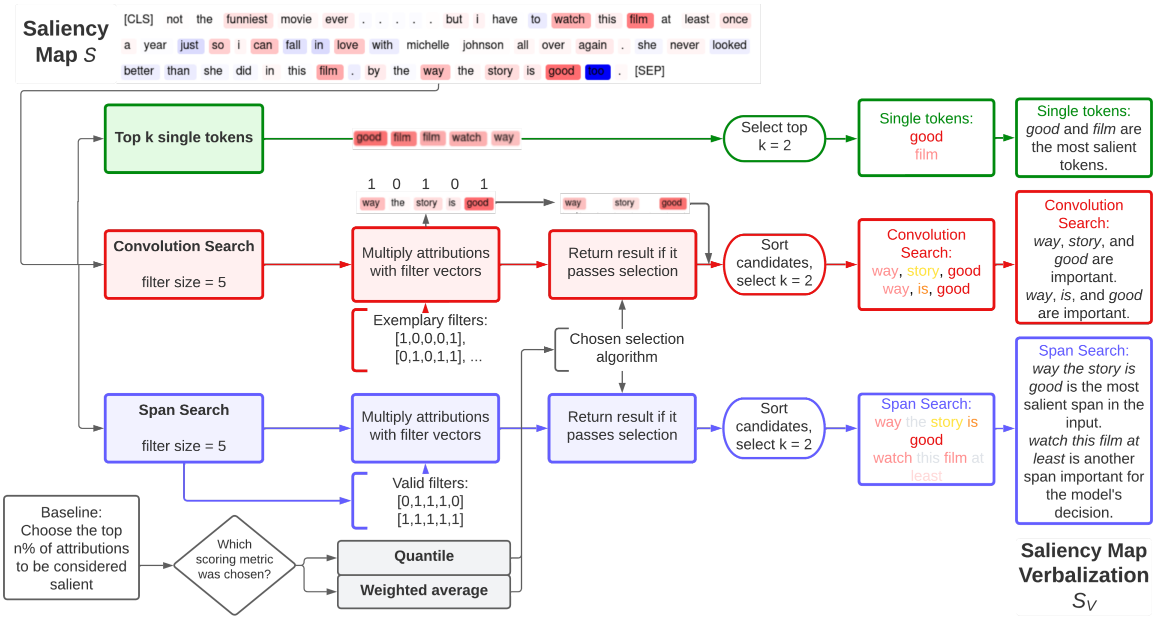

Figure 2: Model-free saliency map verbalizations ($\text{SMV}_{\text{Templ}}$) as generated from three different search methods (Top k single tokens, Convolution Search, Span Search) and two scoring metrics (Quantile, Weighted Average).

length of our input $k > c$, then we multiply a subset of our input with every binary filter

$$\boldsymbol{r}_{i,j,l} = \boldsymbol{f}_{i,j} \cdot A_l^{l+c}, \quad l \in L, L = \{l \in \mathbb{Z} | 1 \leq l \leq k - c\}. \tag{5}$$

From this, we receive result vectors containing possibly coherent attributions and tokens.

Span Search Instead of looking for token pairs in a local neighborhood, we can also look for contiguous spans of tokens by adapting our proposed convolutional search.

We generate b vectors of length of c with c being odd. We describe these b vectors as follows: Choose $i \in \mathbb{N}$ with i being odd, which ensures symmetry of our filters.[3]

$$v_i = [0_{1,\lfloor\frac{c-i}{2}\rfloor}, 1_{1,i}, 0_{1,\lfloor\frac{c-i}{2}\rfloor}], \quad v_i \in \mathbb{Z}^{1,c} \tag{6}$$

We calculate attribution vectors $\boldsymbol{r}_{i,l}$ as such:

$$\boldsymbol{r}_{i,l} = v_i \cdot A_l^{l+c}, \quad l \in L, L = \{l \in \mathbb{Z} | 1 \leq l \leq k - c\} \tag{7}$$

4.1.2 Candidate scoring metrics

We score and filter the snippets $\mathbf{r}$ so that we can present the most salient samples. As a threshold, we calculate the average of the $n\%$ most salient tokens of the given input sample A. This simple method does not filter for saliency, but it reduces the likelihood of presenting non-salient sample snippets. We call this our baseline β.

Weighted average The weighted average sums up the attribution values of r and divides the resulting scalar by the length of r, calculating the "saliency per word" of r. Then the result gets compared to β. Is the result larger than β, r is considered salient and will be a candidate for the verbalization.

Quantile The quantile method relies on the standard deviation within our current sample A. Given a quantile $n, n \in \mathbb{R}_0^+$, we calculate the corresponding standard deviation value σ and compare it to the average of the values of our snippet. If the score is greater than σ and β, it will be marked for verbalization.

4.1.3 Summarized explanation

On top of the two search methods in §4.1.1, we construct a summarized explanation to be used in our human evaluation (§3.1) by considering the k single tokens with the highest attribution scores. After generating k candidates from each search method, we concatenate neighboring token indices to (possibly) longer sequences and recalculate their coverage. We compute the q-th quantile of the remaining candidates according to their coverage to

[3] In contrast to our proposed Convolution Search, we don't need permutations of v_i to generate filters $\mathbf{f}$, so we directly use v_i. Thus, the result vector $\mathbf{r}$ has only two indices.

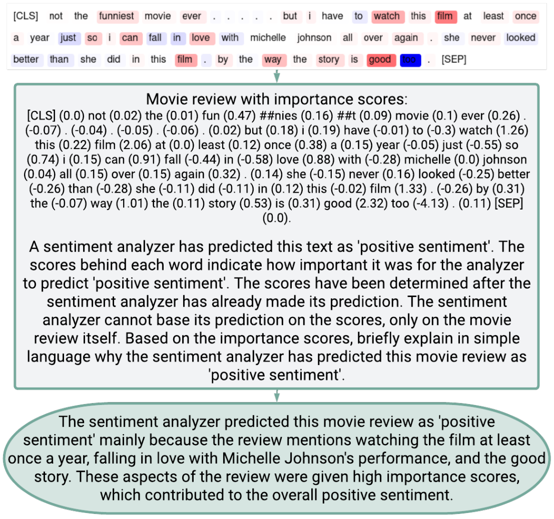

Figure 3: Instruction-based verbalizations SMV_{GPT} using GPT-3.5 of a *negative sentiment* instance from IMDb that was wrongly classified by BERT.

select the final input(s) to our templates. If no candidate is within the q-th quantile, the top-scoring span will be chosen.

4.2 Instruction-based Verbalizations

In light of very recent advances in instructing large language models to perform increasingly complex tasks (Wei et al., 2022), we additionally construct "rationale-augmented" verbalizations (Fig. 3) next to template-based and search-based ones. The instruction contains an overview of the saliency map verbalization task and the associated caveats, e.g. "The classifier cannot base its prediction on the scores, only on the input text itself.". Our most consistently accurate result was achieved by then representing S as bracketed scores rounded to two digits put behind each word, e.g. "definitely (0.75) a (0.14) girl (-0.31) movie (0.15)".

In practice, we manually engineered task-agnostic instruction templates to work with GPT-3.5 (March '23) aka ChatGPT.[4] To our knowledge, there are no datasets with gold verbalizations available and we do not want to enforce any specific format of the explanation, so we use the API in a zero-shot setting. We post-process all outputs by removing all occurrences of the predicted label and semantically very similar words (App. G).

[4]We describe the task-specific instructions in App. F and document the edits to mitigate label leakage in App. G.

Explanations...	**Templ**	**GPT**
were concise & not time-consuming.	4.00	2.38
were not too complex.	3.63	3.88
were not inconsistent/contradictory.	-	3
helped me detect wrong predictions.	2.63	3
with more diverse sentences are useful.	4.25*	-
with numeric scores are useful.	2.63*	2.38
with associations/context are useful.	4.00*	4.50
summarizing the input are useful.	-	4.75

Table 1: Questionnaire asking participants about their overall impressions on both types of verbalizations. All aspects were rated based on a 5-point Likert scale (1: "strongly disagree"; 5: "strongly agree"). Starred values: $\text{SMV}_{\text{Templ}}$ do not have this property, so we asked if the participants *would have liked them* to have it.

5 Data

We choose datasets that cover a selection of English-language text classification tasks. In particular, we select IMDb (Maas et al., 2011) for sentiment analysis, and AG News (Zhang et al., 2015) for topic classification.

We retrieve predictions from BERT models on the test partitions of IMDb and AG News made available through TextAttack (Morris et al., 2020) and their Integrated Gradients (Sundararajan et al., 2017) explanations with 25 samples exactly as they appear in Thermostat (Feldhus et al., 2021).

We then take subsets (IMDb: $n = 80$, AG News: $n = 120$) of each dataset according to multiple heuristics (App. D) that make the tasks more manageable for annotators. Each annotator was shown 340 explanations consisting of equal amounts of each type of representation or rationale. We randomize the order in which they are presented to the annotators. Every instance was evaluated by seven different annotators.

6 Results

Human evaluation Tab. 2 shows that both kinds of SMVs are generally **easier to understand** (B2) than heatmaps or extractive rationales. In a post-annotation questionnaire, we asked 8 out of 10 participants 14 questions about both types of SMVs. Tab. 1 lists the results. While template-based explanations are preferred in being less time-consuming, we can see that GPT-generated verbalizations outperform them in all other aspects. Unsurprisingly, associations and summarizations are the preferred characteristics of verbalizations.

Downstream tasks According to Jacovi et al. (2023a), a feature attribution explanation aggre-

		A: Simulation Accuracy				B1: Helpfulness				B2: Ease of understanding			
		HM Vis	Rat Extr	SMV Templ	SMV GPT	HM Vis	Rat Extr	SMV Templ	SMV GPT	HM Vis	Rat Extr	SMV Templ	SMV GPT
IMDb	All	90.75	85.94	87.5	**94.06**	4.73	4.19	4.46	**5.80**	4.35	4.00	4.67	**5.88**
	$\mathrm{Cov}(S_{\mathrm{VT}})^{\nearrow}$	94.38	89.45	92.19	**96.09**	4.98	4.50	4.91	**5.94**	4.47	4.34	4.99	**5.99**
IAA	$y \neq \hat{y}$	74.49	58.43	63.90	**84.65**	3.67	3.09	3.21	**5.01**	3.48	2.92	3.61	**5.25**
$\kappa = 0.731$	$\hat{y} \neq y_{\mathrm{sim}}$	n.a. (0.00)				3.40	3.10	2.85	**3.94**	3.48	3.18	3.35	**4.33**
AG News	All	**79.83**	-	79.50	77.60	5.26	-	4.65	**5.63**	5.02	-	4.90	**5.77**
	$\mathrm{Cov}(S_{\mathrm{VT}})^{\nearrow}$	**85.31**	-	84.57	81.13	5.41	-	4.98	**5.80**	5.18	-	5.13	**5.89**
IAA	$y \neq \hat{y}$	**70.17**	-	69.37	64.53	5.02	-	4.52	**5.36**	4.84	-	4.84	**5.61**
$\kappa = 0.721$	$\hat{y} \neq y_{\mathrm{sim}}$	n.a. (0.00)				4.14	-	3.34	**4.40**	4.08	-	3.89	**5.10**

Table 2: Results of the human evaluation. Task **A**: Simulation accuracy (annotators guessing the label predicted by the underlying BERT correctly). Task B: Average rating of annotators (1 "bad" - 7 "good") for helpfulness (**B1**) and ease of understanding (**B2**). HM-Vis = Heatmap visualization. Rat-Extr = Extractive rationalizer of Treviso and Martins (2020). SMV-Templ = Template-based saliency map verbalization. SMV-GPT = GPT-3.5-based saliency map verbalization. **All**: Overall result. $\mathbf{Cov(S_{VT})}^{\nearrow}$: Coverage above average. $y \neq \hat{y}$: Explained BERT model made a false prediction. $\hat{y} \neq y_{\mathbf{sim}}$: False human simulation. Inter-annotator agreement in Fleiss κ below the dataset names.

gates counterfactual contexts. This becomes apparent in our overall results on the AG News dataset where more than one potential alternative (multi-class classification with $|C| = 4$) outcome exists. Annotators' simulation accuracy drops from as high as 94 % (IMDb) to 78 %. $\mathrm{SMV}_{\mathrm{GPT}}$ beats all other representations across all three measures in IMDb, but surprisingly underperforms in AG News.

Coverage of the verbalization Fig. 4 and App. A show that $\mathrm{SMV}_{\mathrm{GPT}}$ focuses less on the actual most important tokens that might not be intuitive for recipients, such as function words. The subset of instances with higher-than-average coverage according to $\mathrm{SMV}_{\mathrm{Templ}}$ ($\mathrm{Cov}(S_{\mathrm{VT}})^{\nearrow}$) is substantially easier to simulate (IMDb) and elicits the highest ratings and accuracies from annotators. We utilize this as a proxy for (low) complexity of S, because usually only a single or few tokens that are very salient make these explanations easy to decipher in most representations.

Therefore, we conducted an automated simulatability evaluation on all SMV types, documented in Appendix B, confirming the suspicions about the faithfulness of GPT verbalizations.

Model predictions Lastly, we investigate the subsets of wrong model predictions: The drop in simulation accuracy and ratings when we filter the instances where the model predicts something different from the true label ($y \neq \hat{y}$) is more severe for IMDb throughout all types of explanations. In AG News, the simulatability and the ease of understanding turn out to be higher for SMVs. Our consistently worse results in this subset reveal the belief bias (González et al., 2021), i.e. explanations have a hard time convincing humans about a model behavior when they already have prior assumptions about the true label of an instance. For instances where the human simulation mismatched with the predicted label ($\hat{y} \neq y_{\mathrm{sim}}$), the drop in scores is even harsher: Only $\mathrm{SMV}_{\mathrm{GPT}}$ still achieves ratings that are slightly above average.

6.1 Evaluating instruction-based verbalizations

While there are no invented words in the human evaluation subset, our automated mapping between explanation and input text still detected cases where words are **auto-corrected** and not accurately copied, especially fixing capitalization and small typos. We also found examples in which words or spans are replaced with a synonym, e.g. "not reliable" → "unreliable", but most strikingly, in an IMDb example, "good premise" was replaced with "bad premise" which entirely changed the meaning and the polarity of the sentiment.

In Tab. 3, we manually count what **type of task-related information and semantics** $\mathrm{SMV}_{\mathrm{GPT}}$ provides on top of the translation of the importance scores. We can see that the "negative sentiment" in IMDb is often a confounder for the correct interpretation of the negative saliency scores. Without explicit instructions, GPT still questioned some of the wrong prediction the underlying BERT has made, particularly for IMDb. In terms of linguistic aspects of the verbalizations, associations are frequently included, while summarizations of the

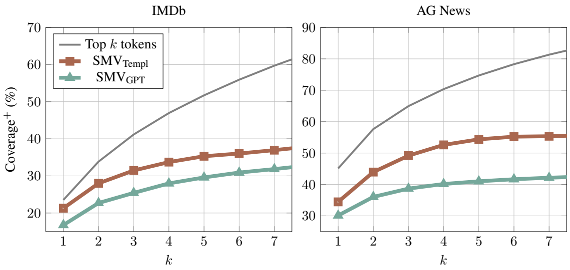

Figure 4: Coverage$^+$@k of SMV$_{\text{Templ}}$ and SMV$_{\text{GPT}}$. Top k tokens is the upper bound for explainer-faithfulness.

	IMDb	AG News
Saliency-related *"because of the high importance scores of words such as 'oil', 'supply', [...]"*	100.00	99.17
Correct interpretation of neg. saliency *"[...] predicted this movie review as 'negative sentiment' because of the high negative importance scores [...]"*	72.50	100.00
Suspecting a wrong prediction *"[...] it is unclear why the classifier predicted this article as 'Business'."*	55.00 FP: 0.00	23.21 FP: 0.83
Associations *"These words are associated with positive emotions and experiences."*	47.50	90.00
Summarizations *"[...] the reviewer enjoyed these aspects of the movie."*	10.00	27.50

Table 3: Occurrences of semantics and accuracies of task comprehension (both in %) in GPT-3.5-generated verbalizations for both datasets. FP = False positives.

input or the decision are rare.

6.2 Discussion

By choosing parameters that prefer longer spans to be selected, we show that SMV$_{\text{Templ}}$ can be more plausible to humans than single token selection. We acknowledge that SMV$_{\text{Templ}}$ are repetitive and, while the results show that they can guarantee a minimum degree of understandability (Ehsan et al., 2019), sufficiency and conciseness, they will not be satisfying enough for lay recipients on their own.

For SMV$_{\text{GPT}}$, the choice of instruction can greatly impact the faithfulness to the explainer. Plausible explanations driven by world knowledge and semantics allow laypeople to contextualize the prediction w.r.t. the input text, but reliable and generalizable methods for auditing these rationales for faithfulness have yet to be discovered.

7 Related Work

To our knowledge, the only previous saliency map **verbalization** approach is by Forrest et al. (2018) who used LIME explanations and a template-based NLG pipeline on a credit dataset. While they mostly included numerical values in explanations, we focus on most important features and free-text rationales, because humans are more interested in reasoning than in numerical values (Reiter, 2019). Ampomah et al. (2022) created a dataset of tables summarizing the performance metrics of a text classifier and trained a neural module to automatically generate accompanying texts. The HCI community highlighted the advantages of verbalization as a complementary medium to visual explanations (Sevastjanova et al., 2018; Hohman et al., 2019; Szymanski et al., 2021; Chromik, 2021). Zhang and Lim (2022) advocated for adding concepts and associations to make explanations more understandable, particularly in contrastive setups.

Hsu and Tan (2021) introduced the task of **decision-focused summarization**. While there are overlaps in the selection of important subsets of the input, the textual nature of the output and the employment of saliency methods, our work is concerned with summarizing the token-level information provided by a saliency map from an arbitrary source for a single instance. Okeson et al. (2021) found in their study that global feature attributions obtained by ranking features by different summary statistics helped users to communicate what the model had learned and to identify next steps for debugging it. Rönnqvist et al. (2022) aggregated attribution scores from multiple documents to find top-ranked keywords for classes.

In early explainability literature, van Lent et al. (2004) already used **template filling**. Templates in NLE frameworks were engineered by Camburu et al. (2020) to find inconsistencies in generated explanations. While their templates were designed to mimic commonsense logic patterns present in the e-SNLI dataset (Camburu et al., 2018), our templates are a means to verbalize arbitrary saliency maps. Paranjape et al. (2021) crafted templates and used a mask-infilling approach to produce contrastive explanations from pre-trained language models. Donadello and Dragoni (2021) utilized a template system to render explanation graph structures as text. Recently, Tursun et al. (2023) used templates together with ChatGPT prompts to generate captions containing verbalized saliency map explanations in the computer vision domain. However, they did not conduct an automated or human evaluation.

8 Conclusion

We conducted a comparative study on explanation representations. We formalized the task of translating feature attributions into natural language and proposed two kinds of saliency map verbalization methods. Instruction-based verbalizations outperformed all other saliency map representations on human ratings, indicating their summarization and contextualization capabilities are a necessary component in making saliency maps more accessible to humans, but they are still unreliable in terms of ensuring faithfulness and are dependant on a closed-source black-box model. We find that template-based saliency map verbalizations reduce the cognitive load for humans and are a viable option to improve on the ease of understanding of heatmaps without the need for additional resources.

Limitations

Our experimental setup excludes free-text rationales explaining the decisions of a model (Wiegreffe et al., 2022; Camburu et al., 2018), because their output is not based on attribution scores or highlighted spans of the input text, so we argue that they are not trivially comparable. However, there are end-to-end rationalization frameworks that can accommodate arbitrary saliency methods (Jain et al., 2020; Chrysostomou and Aletras, 2021; Ismail et al., 2021; Atanasova et al., 2022; Majumder et al., 2022), but require large language models that are expensive to train and perform inference with, so this is out of scope for this study. However, we also see that high-quality free-text rationales can be more easily generated with LLMs (Wang et al., 2023; Ho et al., 2023), and a comparison between them and our attribution-based explanations is an interesting avenue for future work.

Inferring high-quality explanations from large language models necessitates excessive amounts of compute and storage. Although GPT verbalizations are most promising, we urge the research community to look into more efficient ways to achieve similar results. In the future, we will explore if training a smaller model on top of the collected rationale-augmented verbalizations is feasible.

Emphasizing the concerns of Rogers (2023), we do not recommend the black-box model GPT-3.5 as a baseline for interpretability, because the model's training data or internal parameters can not be accessed and the dangers of deprecation as well as the lack of reproducibility are serious concerns. However, we do think it has revealed great potential as a surface realization and contextualization tool for the task of saliency map verbalization.

The causality problem explained in Jacovi et al. (2023a) is not solved by our verbalizations, as it is an inherent problem with feature attribution and rationalization. Future work includes verbalizations alongside counterfactuals, e.g. in interactive setups (Feldhus et al., 2022; Shen et al., 2023).

Although multiple models and explanation-generating methods are available, we specifically focus on one pair for both datasets (BERT and Integrated Gradients), because the focus of our investigation is on the quality of the representation rather than the model.

Finally, explicitly modelling expected highlights to mitigate misalignments as reported on in Schuff et al. (2022), Jacovi et al. (2023b) and Prasad et al. (2021) is still unexplored.

Acknowledgments

We would like to thank Aleksandra Gabryszak for her help setting up and conducting the human evaluation. We are indebted to Elif Kara, Ajay Madhavan Ravichandran, Akhyar Ahmed, Tatjana Zeen, Shushen Manakhimova, Raia Abu Ahmad, Ekaterina Borisova, João Lucas Mendes de Lemos Lins, Qianli Wang, and Sahil Chopra for their valuable work as annotators and evaluators. We thank Gabriele Sarti, Arne Binder, Yuxuan Chen, and David Harbecke as well as the reviewers of the NLRSE Workshop at ACL 2023 for their feedback on the paper. This work has been supported by the German Federal Ministry of Education and Research as part of the project XAINES (01IW20005).

Contributions

- NF: Writing, implementation of baselines and summarized template-based verbalization, prompt design and data generation, evaluation of user study, illustrations.
- LH: Writing and supervision.
- MDN: Conception, implementation and illustration of search and scoring methods for template-based verbalization.
- CE: Data curation, implementation of instruction-based verbalization pipeline, validation and empirical results on search and scoring methods.
- RS: Initial idea and outline.
- SM: Supervision and funding.

References

David Alvarez-Melis, Hal Daumé III, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2019. Weight of evidence as a basis for human-oriented explanations. In *Human-Centric Machine Learning (HCML) Workshop @ NeurIPS 2019*.

Isaac Ampomah, James Burton, Amir Enshaei, and Noura Al Moubayed. 2022. Generating textual explanations for machine learning models performance: A table-to-text task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3542–3551, Marseille, France. European Language Resources Association.

Siddhant Arora, Danish Pruthi, Norman Sadeh, William W Cohen, Zachary C Lipton, and Graham Neubig. 2022. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Thirty-Six AAAI Conference on Artificial Intelligence*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Diagnostics-guided explanation generation. *In Proceedings of the 36th AAAI Conference on Artificial Intelligence*.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for non-monotonic reasoning with distant supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12592–12601.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! Adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1042, Brussels, Belgium. Association for Computational Linguistics.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. REV: information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, Online and Dubrovnik, Croatia. Association for Computational Linguistics.

Michael Chromik. 2021. Making SHAP Rap: Bridging local and global insights through interaction and narratives. In *Human-Computer Interaction – INTERACT 2021*, pages 641–651, Cham. Springer International Publishing.

George Chrysostomou and Nikolaos Aletras. 2021. Enjoy the salience: Towards better transformer-based faithful explanations with word salience. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8189–8200, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. 2022. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc.

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.

Ivan Donadello and Mauro Dragoni. 2021. Bridging signals to natural language explanations with explanation graphs. In *Proceedings of the 2nd Italian Workshop on Explainable Artificial Intelligence*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*.

Upol Ehsan and Mark O. Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, pages 449–466, Cham. Springer International Publishing.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: A technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 263–274, New York, NY, USA. Association for Computing Machinery.

Nils Feldhus, Ajay Madhavan Ravichandran, and Sebastian Möller. 2022. Mediators: Conversational agents explaining NLP model behavior. In *IJCAI 2022 - Workshop on Explainable Artificial Intelligence (XAI), Vienna, Austria*. International Joint Conferences on Artificial Intelligence Organization.

Nils Feldhus, Robert Schwarzenberg, and Sebastian Möller. 2021. Thermostat: A large collection of NLP model explanations and analysis tools. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 87–95, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James Forrest, Somayajulu Sripada, Wei Pang, and George Coghill. 2018. Towards making NLG a voice for interpretable machine learning. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 177–182, Tilburg University, The Netherlands. Association for Computational Linguistics.

Ana Valeria González, Anna Rogers, and Anders Søgaard. 2021. On the interaction of belief bias and explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2930–2942, Online. Association for Computational Linguistics.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.

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

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online. Association for Computational Linguistics.

Denis Hilton. 2017. Social attribution and explanation. *The Oxford Handbook of Causal Reasoning*.

Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. 2023. WikiWhy: Answering and explaining cause-and-effect questions. In *The Eleventh International Conference on Learning Representations*.

Fred Hohman, Arjun Srinivasan, and Steven M. Drucker. 2019. Telegam: Combining visualization and verbalization for interpretable machine learning. In *2019 IEEE Visualization Conference (VIS)*, pages 151–155.

Chao-Chun Hsu and Chenhao Tan. 2021. Decision-focused summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 117–132, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. 2021. Improving deep learning interpretability by saliency guided training. *Advances in Neural Information Processing Systems*, 34.

Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. 2023a. Diagnosing AI explanation methods with folk concepts of behavior. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, New York, NY, USA. Association for Computing Machinery.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 624–635, New York, NY, USA. Association for Computing Machinery.

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

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

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, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural NLP: A survey. *ACM Comput. Surv.*

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2022. Knowledge-grounded self-rationalization via extractive and natural language explanations. In *International Conference on Machine Learning*. PMLR.

Ettore Mariotti, Jose M. Alonso, and Albert Gatt. 2020. Towards harnessing natural language generation to explain black-box models. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pages 22–27, Dublin, Ireland. Association for Computational Linguistics.

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.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Alex Okeson, Rich Caruana, Nick Craswell, Kori Inkpen, Scott Lundberg, Harsha Nori, Hanna Wallach, and Jennifer Wortman Vaughan. 2021. Summarize with caution: Comparing global feature attributions. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. Association for Computational Linguistics.

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. 2021. To what extent do human explanations of model behavior align with actual model behavior? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 1–14, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ehud Reiter. 2019. Natural language generation challenges for explainable AI. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7. Association for Computational Linguistics.

Anna Rogers. 2023. Closed AI Models Make Bad Baselines. *Hacking Semantics*.

Samuel Rönnqvist, Aki-Juhani Kyröläinen, Amanda Myntti, Filip Ginter, and Veronika Laippala. 2022. Explaining classes through stable word attributions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1063–1074, Dublin, Ireland. Association for Computational Linguistics.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Toronto, Canada. Association for Computational Linguistics.

Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human interpretation of saliency-based explanation over text. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul, South Korea. Association for Computing Machinery.

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4596–4608, Online. Association for Computational Linguistics.

Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel A. Keim, and Mennatallah El-Assady. 2018. Going beyond visualization : Verbalization as complementary medium to explain machine learning models. In *Workshop on Visualization for AI Explainability at IEEE (VIS)*.

Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao 'Kenneth' Huang. 2023. ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. *arXiv*, 2305.09770.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2022. TalkToModel: Explaining machine learning models with interactive natural language conversations. *Trustworthy and Socially Responsible Machine Learning (TSRML) Workshop @ NeurIPS 2022*.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: The effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*, page 109–119, New York, NY, USA. Association for Computing Machinery.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, United States. Association for Computational Linguistics.

Marcos Treviso and André F. T. Martins. 2020. The explanation game: Towards prediction explainability through sparse communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 107–118, Online. Association for Computational Linguistics.

Osman Tursun, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2023. Towards self-explainability of deep neural networks with heatmap captioning and large-language models. *CoRR*, abs/2304.02202.

Michael van Lent, William Fisher, and Michael Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the 16th Conference on Innovative Applications of Artifical Intelligence*, IAAI'04, page 900–907. AAAI Press.

Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, Online. Association for Computational Linguistics.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. 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): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2023. PINTO: Faithful language reasoning using prompted-generated rationales. In *International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark O. Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. In *Proceedings of NeurIPS*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Jiannan Xiang, Zhengzhong Liu, Yucheng Zhou, Eric Xing, and Zhiting Hu. 2022. ASDOT: Any-shot data-to-text generation with pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1886–1899, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. Refining language models with compositional explanations. In *Advances in Neural Information Processing Systems*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Wencan Zhang and Brian Y Lim. 2022. Towards relatable explainable ai with the perceptual process. In *Proceedings of the 2022 CHI Conference on Human

Factors in Computing Systems, CHI '22, New York, NY, USA. Association for Computing Machinery.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

A Token ranks

Figures 5 and 6 show the coverage of the verbalizations, which makes up one aspect of explainer-faithfulness.

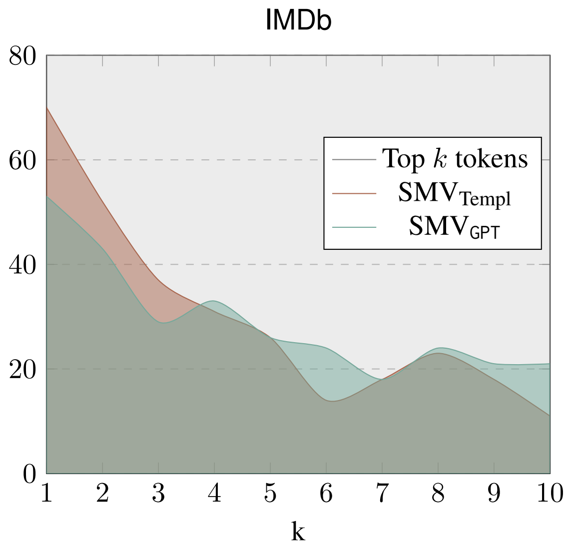

Figure 5: Number of SMVs mentioning top k attributed tokens in IMDb.

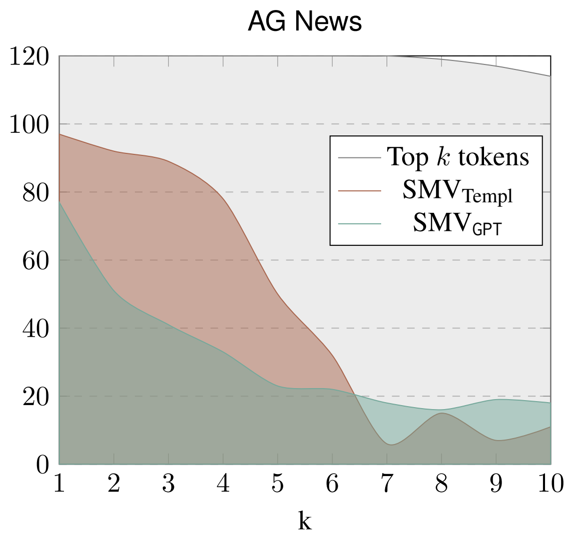

Figure 6: Number of SMVs mentioning top k attributed tokens in AG News.

	AG News		IMDb	
	S_V	$W + S_V$	S_V	$W + S_V$
Conv. Search	91.73	94.10	86.08	**96.00**
Span Search	87.16	**94.39**	89.08	95.90
Top $k = 5$ tokens	**92.54**	93.93	92.38	95.60
$\text{SMV}_{\text{Templ}}$	91.94	94.10	**94.26**	94.90
SMV_{GPT}	69.16	70.00	81.25	81.25

Table 4: Automated simulatability evaluation (Accuracy in %) using a T5-large model (Accuracy on original input: AG News – 92.58%; IMDb – 97.62%) to reproduce the underlying BERT model's prediction based on only seeing one of the verbalizations S_V (prepended by the original input W).

B Automated simulatability evaluation

We follow Wiegreffe et al. (2021) and Hase et al. (2020) and train a second language model to simulate the behavior of the explained BERT model. Table 4 shows the simulation accuracy of a `T5-large` receiving various types of verbalizations (plus the original input). We can see for both datasets that SMV_{GPT} induces the most noise and thus results in the lowest accuracy, while the raw output of the search methods (Conv/Span) are most faithful in combination with the original input.

C Efficiency

First, we measure a runtime of less than two minutes on a CPU (i5-12600k) to generate template-based verbalizations for all 25k instances of IMDb. Given pre-computed saliency maps from any explainer, this is considerably faster than using an end-to-end model for extractive rationales, e.g. Treviso and Martins (2020), which takes several hours for training and then more than 10 minutes for inference on an RTX 3080 GPU. `GPT-3.5` with at least 175B parameters, which obliterates the other two setups. This means that there is a considerable carbon footprint associated with using it. Future work has to look into training considerably smaller models on the generated verbalizations.

D Subset selection heuristics

- We restrict our experiments to explaining a single outcome – the predicted label $\hat{y}$ – and thus modify our metric (Eq. 3): Cov_+ only considers the positive attributions $s_i > 0$.
- We select instances achieving at least a Cov_+ score of 15% (indicating the attribution mass is not too evenly distributed, making interpretations of saliency maps challenging).

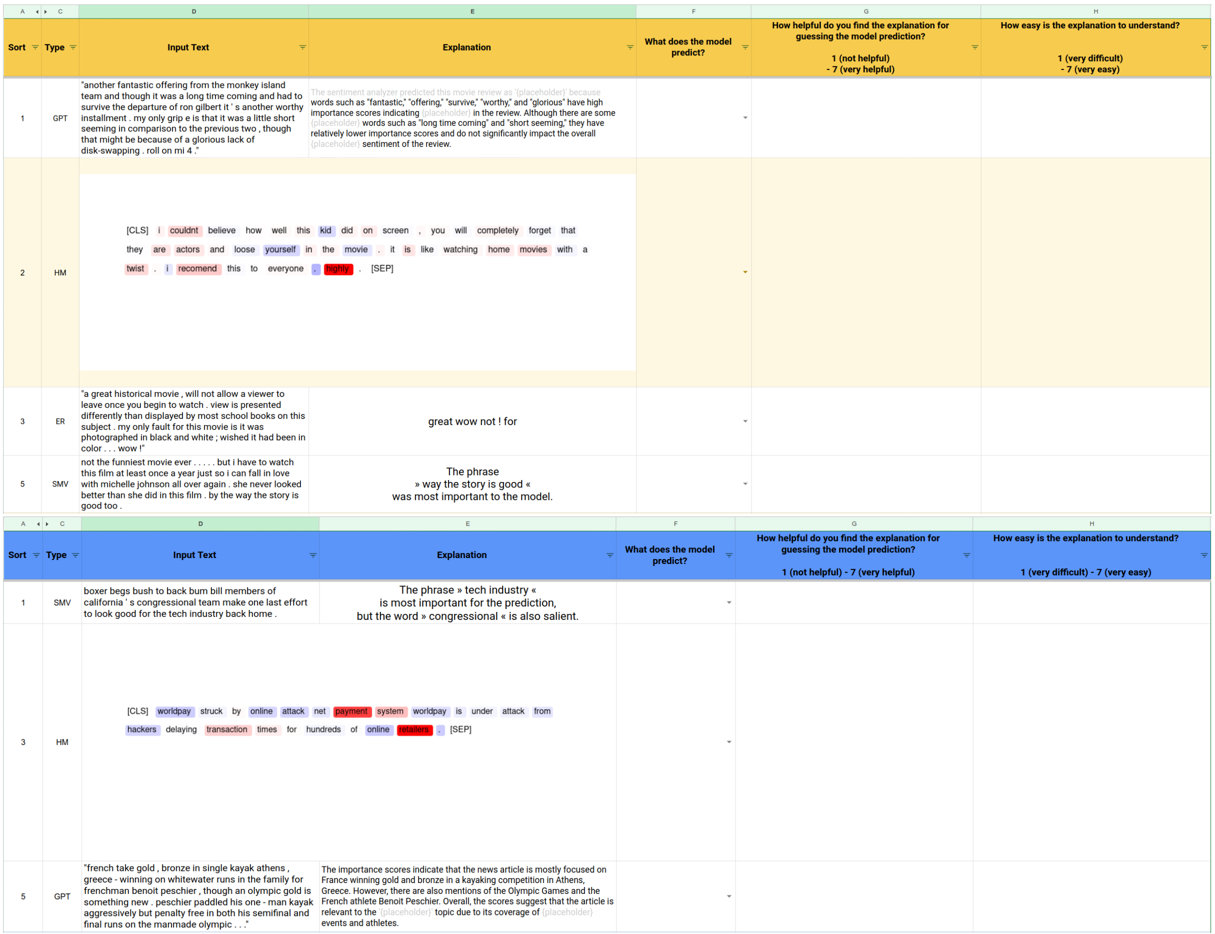

Figure 7: Annotation spreadsheet including one instance from every type of explanation representation in IMDb and AG News, as used in the human evaluation described in §3.1.

- We find values for q (§4.1.3) of $0.5 \leq q \leq 0.75$ to produce the right amount of candidates in the end, s.t. there almost always is at least one candidate in the q-th quantile and the resulting verbalization is not longer than most text inputs.
- We only consider instances with a maximum token length of 80, s.t. the human evaluation is more manageable for annotators.
- We select equal amounts of instances for every true label y (IMDb: 40 positive + 40 negative sentiment; AG News: 30 World + 30 Sports + 30 Business + 30 Sci/Tech) in each dataset.
- We select 25% of IMDb and 46.67 % of AG News to be false predictions by the BERT model ($y \neq \hat{y}$).

We apply the weighted average for IMDb-BERT-IG ($\beta = 0.4$) and the quantile scoring metric for AG News-BERT-IG ($n = 3$). We chose the number of candidates to be $k = 5$ in all cases and the threshold q to be .75 for IMDb and AG News as the average length of the input is lower for the latter which results in too few candidates with higher qs.

E Templates for Verbalizing Explanations

We design our templates as atomic expressions with constraints and blanks that can be filled with words from W. In the most basic cases, we refer to spans, phrases, words and characters as salient or important for some prediction. We design the templates to express saliency information concisely and enable users to reproduce the model's decision process (simulatability). The set of templates is depicted in Table 8.

Our template-based methodology is task- and model-invariant by design, because no task-specific model or NLG component is involved. Achieving sufficiency (measured by coverage) is harder, because a full translation of any saliency map is too verbose and thus not helpful.

F List of LLM prompts

At first, we treated this as table-to-text task – which has recently been tackled with prompt-based large language models (Chen, 2023; Xiang et al., 2022) –

Examples for leading sentence	
The words $\{w_1\}$, $\{\ldots\}$, and $\{w_n\}$ are most important.	Most important is $\{\ldots\}$
The most salient features are $\{\ldots\}$	The model predicted this label, because $\{\ldots\}$
... is the span that was most important.	
Features or linguistic units	**More than one unit**
feature(s)	The two phrases $\{\ldots\}$ and $\{\ldots\}$
word(s)	Both phrases $\{\ldots\}$ and $\{\ldots\}$
token(s)	... are both salient.
phrase(s)	The (top) three most important tokens ...
punctuation	... words such as $\{\ldots\}$ and $\{\ldots\}$
Synonyms for *important*	**Conjunctions & Adverbs**
salient	$\{\ldots\}$, while $\{\ldots\}$
influential	$\{\ldots\}$, whereas $\{\ldots\}$
key	... also salient
impactful	with the word $\{\ldots\}$ also being salient.
Additions for *important* $\{\ldots\}$	**Variations of *important***
... for (the/this) prediction.	... focused on the most for this prediction.
... (to the model) in (making/predicting choosing/producing/shaping) this outcome.	... used by the model to make its prediction.
	... caused the model to predict this outcome.
... with respect to the outcome.	indicate the model's predicted label.
... in this text.	... shaped the model's outcome (the most).
Synonyms for *prediction*	**Polarity**
outcome	$\{\ldots\}$ is least important.
model('s) prediction	$\{\ldots\}$ is more salient than $\{\ldots\}$.
model's judgment	$\{\ldots\}$ is less influential than $\{\ldots\}$.
model('s) behavior	
prediction of the classifier	
(model's) predicted label	
decision	
Dataset-specific	
IMDb	AG News
$\{\ldots\}$ for the sentiment label.	$\{\ldots\}$ indicative of the model's topic classification.
$\{\ldots\}$ most indicative of the sentiment.	$\{\ldots\}$ in this article.
$\{\ldots\}$ most indicative for the sentiment analysis.	The most salient words in this article are $\{\ldots\}$.
$\{\ldots\}$ used by the model to predict this sentiment label.	$\{\ldots\}$, because $\{\ldots\}$ appeared in the article.

Figure 8: Templates for model-free saliency map verbalization.

where we provided a list of attribution scores and, separate from that, a list of tokens. However, we registered less hallucinations (the model incorrectly mapping between words and their scores) when we provided the input as a joint representation as shown in Fig. 3.

For the two datasets, we then used the token+score representation as `{sample}` and a `{label_str}` being the predicted label (IMDb: *positive* or *negative*; AG News: *Worlds*, *Sports*, *Business*, or *Sci/Tech*) and wrote the instructions in Fig. 9.

G Post-processing of GPT outputs

AG News In order to prevent label leakage, we employed the string replacements listed in Tab. 5. In our human evaluation, they were replaced with "{placeholder}", so annotators could perform the simulatability task without cheating.

IMDb Movie review with importance scores: `{sample}`.
A sentiment analyzer has predicted this text as '`{label_str}` sentiment'. The scores behind each word indicate how important it was for the analyzer to predict '`{label_str}` sentiment'. The scores have been determined after the sentiment analyzer has already made its prediction. The sentiment analyzer cannot base its prediction on the scores, only on the movie review itself. Based on the importance scores, briefly explain why the sentiment analyzer has predicted this movie review as '`{label_str}` sentiment':

AG News (Figure 1, r.)
News article with importance scores: `{sample}`.
A topic classifier has predicted this text as '`{label_str}`'. The scores behind each word indicate how important it was for the classifier to predict '`{label_str}`'. The scores have been determined after the topic classifier has already made its prediction. The topic classifier cannot base its prediction on the scores, only on the news article itself.
Based on the importance scores, briefly explain why the topic classifier has predicted this news article as '`{label_str}`':

Figure 9: Task instructions applied to IMDb and AG News used by GPT-3.5 (see App. F for details).

IMDb	AG News			
Classes	**Sports**	**Business**	**World**	**Sci/Tech**
positivity (+)	sport	businesses	global	science
negativity (-)	the world of sports	business and economics	global politics	science and technology
		business and finance	international	scientific
		economics	all over the world	tech
		finance	global issues	technical
		financial	global affairs	technology
		the business world	international relations	technological
		the economy	a global issue or event	the tech industry
		corporate finance		the technology industry

Table 5: Post-processing of GPT-3.5 verbalizations for human evaluation.