Human Interpretation of Saliency-based Explanation Over Text

HENDRIK SCHUFF*, Bosch Center for Artificial Intelligence and University of Stuttgart, Germany
ALON JACOVI*, Bar Ilan University, Israel
HEIKE ADEL, Bosch Center for Artificial Intelligence, Germany
YOAV GOLDBERG, Bar Ilan University and the Allen Institute for Artificial Intelligence, Israel
NGOC THANG VU, University of Stuttgart, Germany

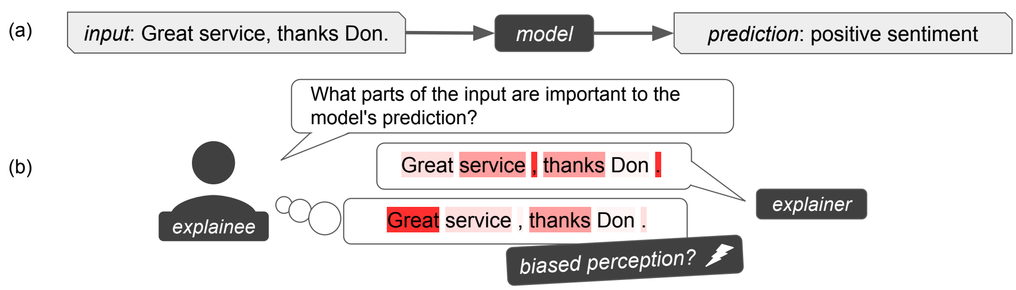


Fig. 1. A saliency explantion is generated to answer the human's need to understand the model. We investigate whether the saliency explanation can be systematically mis-perceived by humans and which factors influence its perception.

While a lot of research in explainable AI focuses on producing effective explanations, less work is devoted to the question of how people understand and interpret the explanation. In this work, we focus on this question through a study of saliency-based explanations over textual data. Feature-attribution explanations of text models aim to communicate which parts of the input text were more influential than others towards the model decision. Many current explanation methods, such as gradient-based or Shapley value-based methods, provide measures of importance which are well-understood mathematically. But how does a person receiving the explanation (the explainee) comprehend it? And does their understanding match what the explanation attempted to communicate? We empirically investigate the effect of various factors of the input, the feature-attribution explanation, and visualization procedure, on laypeople's interpretation of the explanation. We query crowdworkers for their interpretation on tasks in English and German, and fit a GAMM model to their responses considering the factors of interest. We find that people often mis-interpret the explanations: superficial and unrelated factors, such as word length, influence the explainees' importance assignment despite the explanation communicating importance directly. We then show that some of this distortion can be attenuated: we propose a method to adjust saliencies based on model estimates of over- and under-perception, and explore bar charts as an alternative to heatmap saliency visualization. We find that both approaches can attenuate the distorting effect of specific factors, leading to better-calibrated understanding of the explanation.

CCS Concepts: • **Human-centered computing → Empirical studies in visualization**; • **Computing methodologies → Natural language processing**; **Machine learning**.

Additional Key Words and Phrases: feature attribution, text, saliency, explainability, interpretability, human, perception, cognitive bias, generalized additive mixed model

1 INTRODUCTION

Machine learning models' application in various domains (e.g., criminal justice and healthcare) has motivated the development of explanation methods to understand their behavior. One popular class of explanation methods explains

*Both authors contributed equally to this research.

model decisions by specifying the parts of the input which are most salient in the model's decision process [6, 18, 48]. In natural language processing (NLP), this refers to which words, phrases or sentences in the input contributed most to the model prediction [10, 36]. While much research exists on developing and verifying such explanations [1, 4, 32, 35, 50, 51], less is known about the information that human explainees actually understand from them [2, 12, 19, 39].

In the explainable NLP literature, it is generally (implicitly) assumed that the explainee interprets the information "correctly", as it is communicated [4, 17, 20]: e.g., when one word is explained to be influential in the model's decision process, or more influential than another word, it is assumed that the explainee understands this relationship [28]. We question this assumption: research in the social sciences describes modes in which the human explainee may be biased—via some cognitive habit—in their interpretation of processes [15, 37, 39, 52]. Additional research shows this effect manifests in practice in AI settings [11, 14, 22, 25, 40]. This means, for example, that the explainee may underestimate the influence of a punctuation token, even if the explanation reports that this token is highly significant (Figure 1), because the explainee is attempting to understand how the model reasons *by analogy to the explainee's own mind* which is an instance of *anthropomorphic bias* [8, 29, 61] and *belief bias* [16, 22].

We identify three different such biases which may influence the explainee's interpretation: (i) *anthropomorphic bias* and *belief bias*: influence by the explainee's self projection onto the model; (ii) *visual perception bias*: influence by the explainee's visual affordances for comprehending information; (iii) *learning effects*: observable temporal changes in the explainee's interpretation as a result of interacting with the explanation over multiple instances.

We thus address the following question in this paper: *When a human explainee observes feature-attribution explanations, does their comprehended information differ from what the explanation "objectively" attempts to communicate? If so, how?*

We propose a methodology to investigate whether explainees exhibit biases when interpreting feature-attribution explanations in NLP, which effectively distort the objective attribution into a subjective interpretation of it (Section 4). We conduct user studies in which we show an input sentence and a feature-attribution explanation (i.e., saliency map) to explainees, ask them to report their subjective interpretation, and analyze their responses for statistical significance across multiple factors, such as word length, total input length, or dependency relation, using GAMMs (Section 5).

We find that word length, sentence length, the position of the sentence in the temporal course of the experiment, the saliency rank, capitalization, dependency relation, word position, word frequency as well as sentiment can significantly affect user perception. In addition to *whether* a factor has a significant influence, we also investigate *how* this factor affects perception. We find that, for example, short words overall decrease importance ratings while short sentences or intense sentiment polarities increase them.

Finally, we propose two visualization interventions to mitigate learning effect and visual perception biases: model-based color correction and bar charts. We conclude that (a) model-based color correction can predict and mitigate distorting temporal effects and (b) bar charts can successfully remove the influence of word length.

Overall, our results show that *supposedly irrelevant factors such as word length do affect how explainees perceive the influence of words in feature-attribution explanations, despite the explanations explicitly communicating this influence.* This is a surprising result, which raises important questions for explainability in NLP, and in general, about the ability of feature-attribution tools available today to convey the information that they intend to communicate: even in the case of a relatively straightforward explanation, such as directly informing importance regions in the input, cognitive biases of explainees run deep, and may erroneously affect the understanding of the given information.

We show that bar charts and color correction result in better-aligned human assessments in our setting on multiple bias factors. We urge researchers to not blindly trust that users perceive explanations as communicated, and to use methodologies similar to the one we present here, in order to validate our findings on how a given audience comprehends the explanation in-context. The collected data and analysis code will be released upon publication.

2 FEATURE-ATTRIBUTION EXPLANATIONS

Feature-attribution explanations aim to convey which parts of the input to a model decision are "important", "responsible" or "influential" to the decision [3, 7, 36, 42, 60]. This class of explanation methods is a prevalent mode of describing NLP processes [10, 31, 36, 47], due to two main strengths: (1) it is flexible and convenient, with many different measures developed which communicate some aspect of feature importance; (2) and it is intuitive, with—seemingly, as we discover—straightforward interfaces of relaying this information. Here we cover background on feature-attribution explanations on two fronts in alignment with these strengths: the underlying technologies (Section 2.1) and the information which they communicate to humans (Section 2.2).

2.1 Attribution Methods

We consider feature-attribution explanations generally as scoring (or ranking) functions that map portions of the input to scores that communicate some aspect of importance about the aligned portion: $E_f(f(\mathbf{x})) : \Sigma^n \to \mathbb{R}^n$, where E_f is the explanation method with respect to f, f is the model and $\mathbf{x} \in \Sigma^n$ the input text to the model, i.e., the input consists of n tokens which are are elements of an alphabet Σ.[1] For simplicity, we assume that a high score implies high importance.

The loose definition proposed above for feature-attribution explanations as communicating "important" portions of the input (words, sub-words, or characters) is often interpreted with causal lens: that by intervening on the tokens assigned a high score, the model behavior will change more than by intervening on the tokens assigned a low score [3, 23, 28]. This perspective is relaxed in various ways to produce various softer measures of importance: for example, *gradient-based methods* measure the change required in the embedding space to cause change in model output, while *Shapley-value methods* measure the change with respect to the "average case" in the data.

The granularity provided in the scoring function may vary greatly, from a binary measure—important or not important—to a complete saliency map, depending on the tokenization granularity, the method and visualization. Most commonly, the explanation is given as a colorized saliency map over word tokens [e.g., 2, 4, 5, 47, 51]. Note that this work is *not* concerned with a particular feature-attribution method, but rather how feature-attribution explanations generally communicate information to human explainees, and what the explainees comprehend from them.

2.2 Social Attribution: The Case of Text Marking

Is it really possible for the explainee to comprehend feature-attribution explanations differently from what they objectively communicate? What is the nature of any discrepancy in this perception?[2] As Miller [39] writes, literature in the social sciences about how humans comprehend explanations and behavior can help illuminate this problem.

In particular, we assume that the human explainee comprehends the explanation with respect to their own reasoning. By assigning human-like reasoning to the model behavior being explained [39], the explainee may fill any incompleteness in the explanation with assumptions from their own priors about what is plausible to them [9, 22].

To demonstrate, consider the case of binary feature-attribution—marking parts of the input as "important" and "not important", also known as *highlighting* or *extractive rationalization* [33]. Even this simple format of communicating

[1] E can potentially be agnostic to f, known as a black-box explanation method [24].
[2] This question is distinct from the question of whether the explanation faithfully communicates information about the model [27, 53]: even if the feature-attribution information is entirely faithful, discrepancies may still arise in how humans comprehend this information.

information can be assigned human-like reasoning by the explainee, on account of "*who marked this text*" and "*for what purpose*": Marzouk [38] identifies various objectives that humans follow when marking or observing marked text, e.g., marking forgettable secions (for memorization); marking as a summary (for subsequent reading); marking exemplifying text; marking contradicting or surprising text, etc. In the context of NLP models, Jacovi and Goldberg [28] note two possible central objectives: reducing the input to a summary which comprehensively informs the decision, or identifying influential evidence in the input which non-comprehensively supports the decision.

These many different objectives can influence the choice of marking, and the information that it communicates. This means that both the marked text, and the choice of what text to mark, are information which the explainee comprehends when observing the explanation. Therefore, how the explanation is perceived is influenced by both factors.

Text marking is a special case of feature-attribution. The above demonstrates how the explainee's interpretation is potentially shaped by aspects of the explanation which are implicit or unintended—leading to an "erroneous" interpretation of the explanation. We identify three biases that may cause this effect, as motivation for our investigation: (i) anthropomorphic bias and belief bias, via the explainee's a-priori opinion on human-like or plausible reasoning; (ii) visual perception bias, via characteristics of the explainee's visual affordances for comprehending information; (iii) learning effects, as observable influence in the explainee's interpretation by previous explanation attempts in-context.

3 STUDY OVERVIEW

Research Question. The core research question in this work is to probe into which, if any, factors in the explanation process—aside from the saliency itself—may influence the explainee's interpretation of the saliency information. Formally, we view the saliency explanation as a process whose result is the explainee's interpretation of the saliency scores. The "input" to this process is the original text as well as the saliency information and the visualization method. Then, we ask which factors in the original text have statistically significant effects on the explainee's interpretation and how properties of the saliency score and the visualization method affect it.

Notably, a key challenge in analyzing the explainees' saliency understanding is that we want to identify influencing factors on the explainee's ratings without the existence of an inherently correct ground truth perception.

Proposed Methodology. We propose a combination of study design and statistical analysis to quantify the influence of arbitrary factors such as word length, sentiment polarity or dependency relations. We collect explainees' subjective interpretations of the saliency scores in a crowdsourcing setup. We relate this interpretation to the original explanation considering various potentially influencing factors using an ordinal generalized additive mixed model (GAMM). The result from this comparison is an answer on *which* of the a-priori candidate explanatory factors indeed have significant effect on the explainee's interpretation and *how* these factors functionally affect interpretation.

4 STUDY METHODOLOGY SPECIFICATION

The study consists of two phases: collecting subjective importance interpretations (Section 4.1), and analyzing responses with an adequate statistical model (Section 4.3). The collected data and analysis code will be released upon publication.

4.1 Collecting Self-Reported Importance Ratings

In our main study, we investigate the interpretation of color-coding saliency visualization of the feature-attribution by crowdsource laypeople (variations on this study will be described later). We measure the perceived importance of a word within a saliency score explanation by directly probing human self-reported word importance. In this instance, we

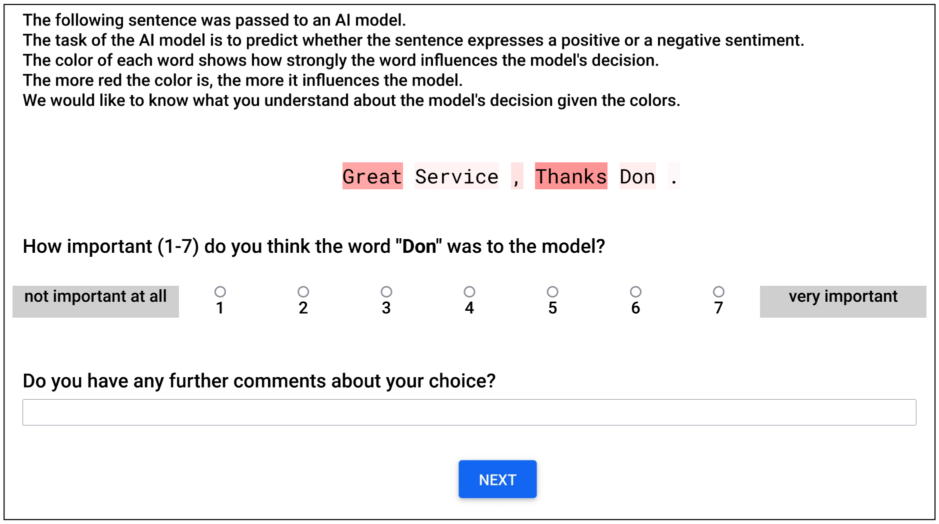


Fig. 2. Screenshot of the importance rating interface for English sentiment sentences using saliency visualization.

ask “How important (1-7) do you think the word "X" was to the model?” (Figure 2). We collect answers on a single-item unipolar 7-point Likert scale ranging from *not important at all* to *very important*.

Texts. We use sentences from the Universal Dependencies English Web Treebank [44].[3] This treebank contains comprehensive annotation, including dependency relations of sentences, stemming from various domains such as newsgroups or online reviews. We use sentences from the reviews group for a plausible framing of a sentiment analysis task.[4] We randomly select 150 sentences to be used.

Saliency Scores. We assign random saliency scores to each token to uniformly sample the space of saliency intensities. We are, at this stage, not interested in using a “real” model or saliency score (e.g., attention or integrated gradients), as we investigate general perception of arbitrary scores. It is therefore useful to create saliency scores that “do not make sense” because a saliency score should reflect the model’s reasoning which might very well not make sense at all. We study an instance of “real” saliency scores (integrated gradients) later in Section 5.3.

Study Interface. See Figure 2 for the rating collection interface. We display all sentences using monospaced font and fixed whitespaces to obtain a direct mapping between the number of characters and the color area for each word.[5]

Procedure. We ask participants to rate the importance of a randomly-selected word in the sentence.[6] We show all 150 sentences from the described review dataset to each participant, displayed in a randomized order per participant. Saliency scores for all tokens are randomized for each participant (such that we collect responses to many different saliency maps, rather than numerous responses for the same set). We do so because our aim is not to obtain accurate

[3] `https://github.com/UniversalDependencies/UD_English-EWT`

[4] We choose sentences without sub-token dependency relations (e.g., excluding “it’s” because displaying it as two tokens breaks the orthography) and with unique word occurrences (i.e., excluding sentences that contain a word several times). From this subset we remove length outliers: sentences with number of tokens longer than one standard deviation above the mean (concretely, 11 tokens).

[5] Ligatures and other typographic attributes of non-monospaced fonts would break this mapping.

[6] Alternatively, one can imagine a setting in which participants rate all words within the sentence. We choose to ask for single-word ratings to (i) avoid carry-over effects from ratings of the first to the last words and (ii) collect ratings of more sentences within the same experiment time compared to splitting the set of sentences over participants which would introduce further difficulty in the statistical analysis.

(mean) estimates of single ratings as one would do in a corpus annotation, but to collect rich data to build an accurate model describing the underlying general phenomenon. For each sentence, we collect the participant's importance rating, the completion time and a voluntary free-text comment. We choose to not include a dedicated training phase, e.g., showing the participants ten explanation instances before starting the data collection: we are explicitly interested in potential learning effects. These can be crucial in real-world applications: for example, should we find a decaying learning effect, an effective model audit should make sure to include a sufficient number of model predictions.[7]

Participants. We recruit 50 crowdworkers on Mechanical Turk. One crowdworker failed all of the trap sentences, so we exlude this worker's responses and recruit one additional worker. All other participants successfully passed all trap sentences. In total, this yields 7500 importance ratings.

4.2 Factors of Saliency Perception

For our set of possible candidate factors, we model factors which are motivated by the three types of biases: anthropomorhic and belief biases, visual biases, and learning effects. Each factor will be tested for statistical significance on the explainees' interpretations. Table 1 lists the factors we investigate in this work.

Selected factors in Table 1 include: (i) *word length* as longer words correspond to a larger colored draw area, which we hypothesize influences visual perception bias; (ii) *word polarity* as we present participants a sentiment classification task and expect that the participants' own assessment of word importance influences their perception of how important it is to the model, which we hypothesize is an instance of belief bias; (iii) *display index* as we hypothesize that participant ratings are affected by temporal effects such as learning; (iv) *word position* as we hypothesize that, e.g., words at the center of a sentence might be perceived more strongly due to the center bias (visual perception bias) which was observed in various eye-tracking studies, i.a. for natural scenes [49].[8]

4.3 Statistical Analysis Using GAMMs

Given a set of text instances for which there are available (1) the feature-attribution scores, (2) the interpreted importance scores, we describe the analysis methodology aiming to derive the possible factors in the input that will cause discrepancy between (1) and (2).

4.3.1 Ordinal Generalized Additive Mixed Model. We analyze the collected ratings of perceived importance using a ordinal generalized additive mixed model (GAMM). Its key properties are that it (i) models the ordinal response variable (i.e., the importance ratings in our setting) on a continuous latent scale (*ordinal generalized*), which is (ii) modeled as a sum of smooth functions of covariates (*additive*) and (iii) accounts for random effects (*mixed*). The continuous latent scale is linked to ordinal categories by estimating threshold values that separate neighboring categories. The smooth functions can comprise single covariates (*univariate* smooths) such as $f_1(x_1)$ or combinations of multiple covariates such as $f_2(x_2, x_3)$. Random effects allow to account for, e.g., systematic differences in individual participants' rating behaviour. For example, a specific participant might have a tendency to give overall higher ratings than other participants. Including a *random effect* allows to disentangle this influence on the response variable from the influence of the covariates in question (such as word length) and thereby offers a clearer view on these *fixed effects*. The GAMM analysis enables us (i) to make statements about which factors significantly influence saliency perception, without

[7]In order to filter-out participants that just "click through" the interface to obtain the study reward, we insert three trap sentences at random positions in the last two thirds of the real sentences. See example and more integration details in Figure 9 in the appendix.
[8]We derive word frequencies from the WikiMatrix corpus [43] and sentiment polarities from SentiWords [21].

Table 1. List of factors that presupposedly affect saliency explanation perception along with the findings of our three user studies. EN refers to the English sentiment classification study, DE to the German fact checking study and EN-IG to the English sentiment classification study using integrated gradients as feature attribution method (without correction visualizations).

Factor	**Description**	**Significant Effects**		
		EN	*DE*	*EN-IG*
Saliency	The color intensity specified as the saturation value ($S \in [0, 1]$) in a (H, S, V) color triple [45], e.g., (0°,0.5,1.0) (■) and (0°,0.25,1.0) (■).	✓	✓	✓
Word length	The number of characters in a word, e.g. 7 for "example".	✓	✓	✓
Word frequency	The word's normalized frequency, estimated on a large corpus.			✓
Sentence length	Number of words in the sentence.	✓	✓	
Display index	The sentence's position within a sequence of sentences (e.g. the third sentence in the sequence of 150 sentences). This relates to temporal effects such as learning.	✓	✓	
Sentiment polarity	The sentiment polarity of a word (defined via its lemma) $\in [-1, 1]$.	✓	–	
Saliency rank	Normalized rank of a word's saliency score (i.e. color intensity) in comparison to the other words in its sentence $\in [0, 1]$.	✓		✓
Word position	The index of the token's position within its sentence.		✓	
Capitalization	The word's capitalization, e.g. "example", "Example" or "EXAMPLE".		✓	
Dependency relation	Dependency relation to its parent within the dependency graph (36 types for *EN*).		✓	

prescribing any notion of "correct perception" and (ii) to study the relation between these factors and participants' importance ratings in detail, via an interpretation of the model's parametric terms (categorical factors) as well as smooth terms (numeric factors). We provide a description of each of the ordinal GAMMs components starting from a linear model over linear mixed models, generalized linear models and generalized additive models in Appendix A.[9]

4.3.2 Model Details. We include all factors listed in Table 1 into our model formula. We use smooth terms for numeric factors and parametric terms for categorical factors. Additionally, we include tensor product interactions for all pairs of smooth terms.[10] In order to statistically account for potentially confounding effects of individual participants or sentences, we include random intercepts as well as random slopes for each participant and each sentence. Before fitting the model, we remove a small amount of outlier ratings.[11] We use fast REML for smoothness selection and apply variable selection via double-penalty shrinkage (i.e., additionally penalizing the splines' null space). We fit the model using discretized covariates as described in Wood et al. [59] and Li and Wood [34].[12]

5 STUDY RESULTS, INTERPRETATION AND GENERALIZATIONS

In the following, we conduct three user studies. The first study (Section 5.1) investigates saliency perception for English and a sentiment classification task. The second study (Section 5.2) extends the investigation to German language and a fact checking task to evaluate generalization of the findings.[13] Since these two studies use random saliency scores so as

[9]For further information on ordinal GAMMs, we refer to Divjak and Baayen [13], who provide a comprehensive introduction. For detailed information on GAM(M)s as well as explanations of implementations and analyses, we recommend the textbook by Wood [58].
[10]Such a functional ANOVA decomposition is supported by mgcv and allows to study, e.g., the interaction between word length and sentiment polarity in addition to the isolated main effects of word length and sentiment polarity.
[11]We remove outliers from the intially 7500 importance ratings by excluding words with 20 or more characters (8 ratings) and ratings with a completion time of 60 seconds or more (50 ratings), leaving 7442 ratings left for analysis. We apply the identical filters to the study described in Section 6. For the German study described in Section 5.2, we only apply the completion time filter.
[12]We use R and mgcv [54–58] (version 1.8-38) to fit all our models.
[13]*Task* refers to the AI's task which operation is communicated to the explainee via the saliency explanation.

Table 2. Effective degrees of freedom (edf), reference degrees of freedom and Wald test statistics for the uniariate smooth terms of the first user study.

	edf	ref. df	F	p
s(saliency)	12.0967	19	728.8738	< 0.0001
s(display index)	1.0921	9	2.0872	0.0001
s(word length)	2.5416	9	4.1826	< 0.0001
s(sentence length)	0.9200	9	1.7531	0.0001
s(word frequency)	0.0011	9	0.0001	0.1082
s(sentiment polarity)	2.1281	9	1.6156	0.0065
s(saliency rank)	0.9580	9	4.4417	< 0.0001
s(word position)	0.0005	9	0.0000	0.7882

to not prescribe a specific feature-attribution method, we report a third study (Section 5.3) which uses the wide-spread integrated gradient scores as a generalization to practically-used attribution methods.

5.1 Sentiment Analysis in English

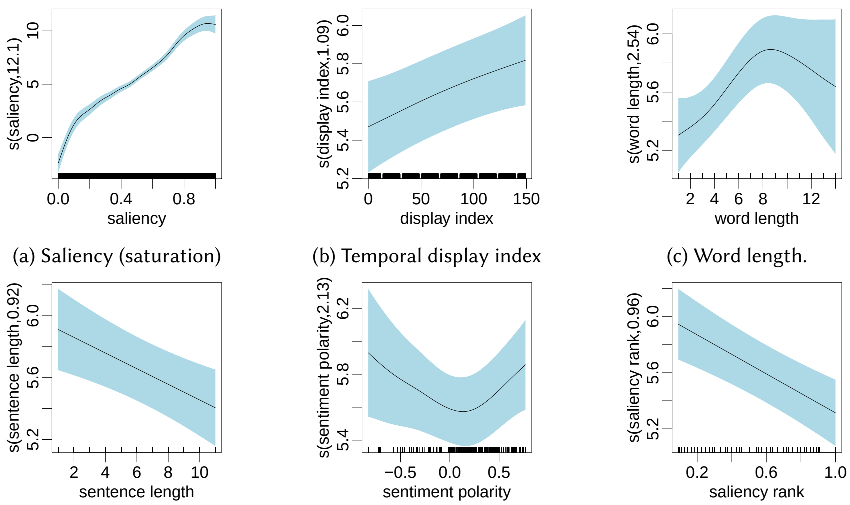
We discuss quantitative results based on the fitted GAMM (Section 5.1.1) as well as qualitative findings based on the participants written feedback (Section 5.1.2).

5.1.1 Quantitative. Table 2 shows statistics for the univariate smooth terms in the fitted GAMM. Figure 3 shows partial effect plots of the respective significant smooth terms. Regarding the parametric terms, neither a words's capitalization (df=2, F=1.84, p=0.16) nor its dependency relation (df=35, F=1.17, p=0.24) show a significant effect on perceived importance. Regarding the smooth terms, we observe that saliency score, display index, word length, sentence length, word sentiment polarity and saliency rank show significant effects on perceived importance. In the following, we discuss each effect in detail.

Saliency (Figure 3a): The saliency (i.e., the color saturation) has the strongest impact on perceived importance as the graph spans the by-far widest y-axis range of all plots in Figure 3. Except for the saliency scores around 1, the entire graph shows a monotonous relation between saliency score and perceived importance.

Display Index (Figure 3b): Participants' ratings increased over the course of the experiment. We hypothesize that the participants report more conservative ratings in the beginning of the experiment to "leave enough room" for more extreme sentences and adapt their ratings to a more "calibrated" level over the course of the experiment. Interestingly, this trend does not seem to stop after our maximum number of 150 sentences. We leave the study of sufficient amount of training required for the effect to reach a peak to future work.

Word Length (Figure 3c): With increasing word length, importance ratings rise up until a length of approximately eight characters and decrease again afterwards. We hypothesize that the initial increase corresponds to an increase of the colored area that a longer word directly causes, as the saliency score is visualized within a box which is proportional to the number of characters. To interpret the subsequent decrease of perceived importance, we consider the interactions between word length and other factors. We find significant pairwise interactions of word length with (i) saliency, (ii) display index and (iii) word frequency (Appendix C). For the interaction with display index, we observe that the decreasing effect of high word lengths grows with increasing display index up until around the 55th sentence. After



(a) Saliency (saturation) (b) Temporal display index (c) Word length.

(d) Sentence length (e) Word sentiment (f) Saliency rank

Fig. 3. Partial effect plots for all significant smooth terms (note that y-axes are scaled per effect). Numbers in y-axis labels are estimated degrees of freedom (edf) of the respective smooth. The shaded area displays confidence intervals (plus and minus one standard error) including uncertainty about the overall mean.

this point, the effect decreases. While the latter decrease can be explained with the partial effect of increasing ratings with higher display indices (as shown in Figure 3b), the former decrease demands detailed investigation in future work.

Sentence Length (Figure 3d): Importance ratings decrease for words in longer sentences. A longer sentence leads to a higher number of color samples and therefore also to a larger expected color range. We argue that such an increased color range inhibts users to make very high importance ratings due to a missing "maximum color" anchor.

Sentiment Polarity (Figure 3e): The effect of a word's lemma's sentiment polarity on importance ratings. We observe a parabola-shaped curve with a minimum at slightly-positive sentiment. To the left, importance ratings increase with increasingly negative polarity and to the right importance ratings increase with increasingly positive polarity. This indicates that users ratings of "what was important to the model when classifying the sentence" are biased by their answer to "what is important to me when classifying the sentence myself". Such a substitution of a presumably complex-to-compute target attribute with a simpler heuristic attribute is a known cognitive bias and often referred to as *attribute substitution* or *substitution bias* [30]

Saliency Rank (Figure 3f): The partial effect of a word's normalized saliency rank on participants' importance ratings. We normalize the rank by dividing by sentence length, as low ranks (i.e., larger numbers) would otherwise be strongly correlated to sentence length, and potentially cause stability issues within the model estimation. We observe that an increased rank (a value of one corresponds to the last rank, i.e., the lowest saliency score) corresponds to a decrease in rated importance. In contrast to the effect of saliency score shown in Figure 3a, the saliency rank is not only a property of a word but of a word in context of its sentence. A word's saliency score can remain unchanged while at the same time its rank can be arbitrarily modified by changing the saliency scores of the other words in its sentence. We argue that the significant effect of saliency rank indicates that users interpret saliencies *in relation* to each other, i.e., their judgements are relative and lack a fixed anchoring point. This is supported by qualitative analysis in Section 5.1.2.

Table 3. Comments of the participants of the English sentiment study. Participants were asked to rate the underlined word or symbol.

Type	Sentence & Saliency	Rating	Comment
relative judgement	Best Electrician in Florence	2	"Best" highlighted in the light pink was not scored as high as the other words in deeper shades of red, so I assume the model didn't find it very important. (P11)
	Absolutely amazing job !	3	I see 4 different levels of highlights. Absolutely seems to be the third darkest so that's why I chose 3 (P20)
	Love Hop City	4	There are only 3 words but they are all highlighted differently. And there is a big difference between the darkest color and the lightest color so it doesn't seem right to put Love as number 6. It's more so in the middle because of how much lighter it is than the darkest color. (P20)
own opinion	Room was amazing .	3	I am uncertain why the period at the end of the sentence would be important, so I choose a 3, even though the AI coded it as red color. (P26)
	best	2	I would think that if it's one word then the word should be important. But I don't think it is important because it's such a light color (P20)
light color	Would do business with them again .	1	the symbol has no color code around it at all so I chose 1 (P26)
	David Bundren is the Tire GooRoo .	1	It probably didn't even notice the last name (P44)
other	Listened to my problem and took care of it .	7	Now I understand the range of red colors better. "it" outside of the phrase "care of it" is meaningless, but since blanks between words are NOT colored, I have to think that AI is judging "it" by itself. (P39)
	Great Place !	7	well you state that the redder the word is, the more influence it has...that's pretty red. (P44)

In addition to the significant partial effects, we also find numerous significant interactions. We provide the statistics of Wald tests for all pairwise tensor product interactions (following a functional ANOVA decomposition) as well as summed effect plots of all significant pairwise interactions in Table 5 and Figure 10 in Appendix C.

5.1.2 Qualitative. In addition to the statistical evaluation, we also evaluate the participants' voluntary free-text comments. Table 3 shows a selection of comments grouped into four categories:

Relative Judgement: Participants explicitly state that they make relative importance judgements. This supports our argumentation of relative judgments discussed for the effects of sentence length and saliency rank.

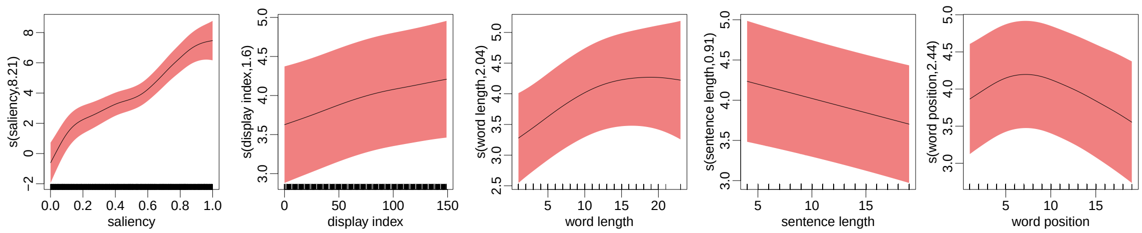
Own Opinion: Similarly, participant comments support our hypothesis that users' ratings are subject to the cognitive bias of attribute substitution as discussed for the effect of word sentiment polarity.

Light Color: Participants seem to make a categorial distinction between *very light color* and *seemingly no color* although this distinction does not exist in terms of the attribution score. This can be important when communicating very low influences and should be addressed in more detail in future work.

Other: Miscellaneous comments on, e.g., issues of word-level attribution and the resulting ambiguity in interpretation.

5.2 Generalization Across Tasks and Languages: Fact Checking in German

So far, we found indication that numerous factors (word length, saliency rank, etc.) significantly influence users' subjective importance ratings. Two important limitations are that (i) the findings are limited to English, and (ii) they are



(a) Saliency (saturation) (b) Temporal display index (c) Word length (d) Sentence length (e) Word position

Fig. 4. Partial effect plots for all significant smooth terms (note that y-axes are scaled per effect) for the German experiment. Numbers in y-axis labels are estimated degrees of freedom (edf) of the respective smooth. The shaded area displays confidence intervals (plus and minus one standard error) including uncertainty about the overall mean. (a) refers to color saturation.

limited to one AI task (sentiment classification). To assess whether the findings do generalize to another language and another task, we repeat the study identically with German sentences from the PUD Corpus[14] with a fact checking AI task. We collect responses from 25 German-speaking participants from a participant pool including Germany, Austria and Switzerland. In total, this corresponds to 3750 ratings.

5.2.1 Confirmed Effects. Our analysis confirms the significants effects of saliency, display index, word length and sentence length. Figure 4 displays the respective partial effect plots. While the smooths for saliency (Figure 4a) and sentence length (Figure 4d) show high similarity to the respective smooths of the English study (see Figures 3a and 3d), we observe slight differences for display index (Figures 3b and 4b) and word length (Figures 3c and 4c). While the English display index smooth grows more or less linearly (edf=1.09), the respective German smooth reaches a plateau after around half the sentences (edf=1.60). We hypothesize that such a saturation effect will also be visible for English, but requires a larger number of sentences. We argue that this is caused by the fact that the sentences in the German study are longer than in the English study, which makes participants of the German study see more colored words and thereby "calibrates" their ratings faster in terms of number of sentences. Similarly, the German word length smooth saturates after around 15 characters, while the English smooth decreases after around 8 characters. We hypothesize that this difference can be attributed to the overall longer words in German as well as the difference in compounding.

The effect of saliency rank cannot be confirmed in the German experiment. Like in the English study, we find no indication that word frequency has a significant effect on importance ratings. We provide test statistics of paramatric and smooth terms (univariate smooths and pairwise interactions) as well as coefficient estimates in Appendix D.

As in the English study, we additionally qualitatively analyse the participants' free text comments and observe (as in the English study) numerous instances for which participants mix their own estimate of importance with the communicated importance. We provide exemplary instances in Table 9 in Appendix D.

5.2.2 Additional Effects. In addition to the effects that we already observed in the English study, we also find that the word's position within the sentence (Figure 4e) as well as capitalization and dependency relation have significant effects on importance ratings. A full list of coefficient estimates along with further details is provided in Table 8 in Appendix D.

The estimate for completely capitalized words is 1.91 (SE=0.9638), the respective estimate for words with the first letter capitalized is 0.41 (SE=0.12).[15] This confirms the intuition that words that are completely capitalized receive the

[14] `https://universaldependencies.org/treebanks/de_pud/index.html`.
[15] The estimate for lower-cased words is fixed to zero as the reference level. For dependency relations, we choose the (most frequent) punctuation relation.

highest importance ratings followed by words for which the first letter is capitalized. We argue that this effect, and in particular the effect for first-letter-capitalized words are more visible in the German experiment as German uses more frequent capitalization (e.g., for all nouns).

Regarding dependency relations, the highest estimate can be observed for temporal modifiers (obl:tmod, $\beta = 1.70$, SE=0.55) like “today” and numerical modifiers (nummod, $\beta = 1.39$, SE=0.36) like “one”. The lowest estimate can be observed for clausal modifier of nouns (acl, $\beta = -1.22$, SE=0.64) like “sees” in “the issues as he sees them” and indirect objects (iobj, $\beta = -0.48$, SE=0.52) like “me” in “she gave me the book”. We hypothesize that the grammatical function effect is larger here than in the previous experiment because properties such as the use of temporality, numerals and embedded clauses are more important for determining factuality than for determining sentiment.

5.3 Generalization to Model-based Saliencies (derived via Integrated Gradients)

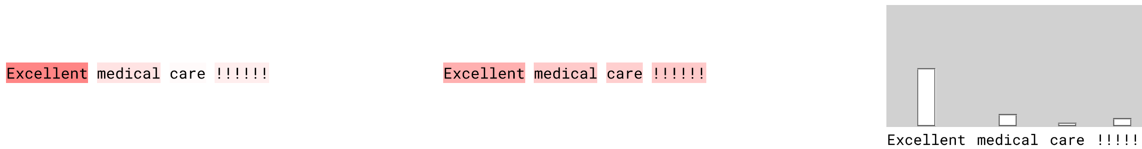
We want to assess whether our findings on the random saliency scores used in the previous two studies also hold for practically-used feature attribution scores. Therefore, we conduct an additional user study using integrated gradients [46] instead of random saliencies.[16]

5.3.1 Study Modification: Within-Subject Design. We combine the evaluation of integrated gradient scores with a within-subject evaluation of three visualization methods which we detail in Section 6. In this section, we focus on the unmodified visualization as it is used in the two previously described studies. In the remainder of this paper, this visualization method is referred to as *saliency*. We sample another 150 sentiment sentences from the sentence pool described in Section 4.1 and present them in the same sentiment classification context. Instead of using one saliency visualization method for all 150 sentences, we now use the three visualizations and show each participant 50 sentences per visualization.[17] We collect 9000 importance ratings from 60 participants and exclude participants of the previous study to avoid carry-over effects from previous exposures.

5.3.2 Model Modification: Factor-Smooth Interactions. We again use an ordinal GAMM model using the same covariates as described in Section 4.3. In addition, we add a parametric term for the visualization condition to account for overall differences in rating intensities between the visualization conditions.[18] We use factor-smooth interactions for each variable which leads to separate estimates for each variable per visualization (e.g., three smooths for word length, one per visualization). First, this yields smooths for the “orginal” saliency visualization, i.e., the heatmap visualization without saliency corrections. In contrast to our first study, these smooths now correspond to effects on integrated gradient attribution scores instead of random scores. First, comparing the smooths allows us to compare how factors influence importance ratings across the three visualizations, e.g., to assess whether the bar visualization did mitigate the biasing effect of word length. We discuss the respective results in Section 6.3. Second, analyzing the smooths relating to the original saliency visualization allows us to evaluate which of the effects we observed in the first study do generalize to the integrated gradients attribution scores. We discuss the respective results in the following paragraph.

5.3.3 Results. We find significant effects of saliency score, word length, relative word frequency and saliency rank. We provide details and test statistics on all parametric coefficients as well as smooth terms in Table 12 in Appendix F. All of these variables except relative word frequency were also found to be significant in our first study and all of them except

[16] We make use of the Language Interpretability Toolkit [47] to obtain normalized integrated gradient scores with respect to the SST2-base sentiment model and 30 interpolation steps.
[17] The order of visualization methods is balanced across participants. Sentence order is fixed to ensure identical ordering effects for the three visualizations.
[18] We additionally include a random intercept to account for visualization order.



(a) Original saliency (b) Corrected saliency (c) Bars

Fig. 5. The three different visualization methods we compare in our third study (Section 6).

relative word frequency and saliency rank were confirmed in our German study. The significant influence of relative word frequency was observed for the first time.

Overall, three studies confirmed the presumably biasing influence of word length, (pairs of) two studies respectively confirmed the effect of sentence length, display index and saliency rank, and one study (each) found significant effects of word position, sentiment polarity, word frequency, capitalization and dependency relation. Together, these reflect the three sources of bias: anthropomorphism and belief bias, visual perception, and learning effects.

6 MITIGATING VISUAL PERCEPTION AND LEARNING EFFECT BIASES

So far, we observed that various seemingly irrelevant factors influence human perception in unintended ways from the explicit and objective saliency information across different languages, tasks and feature-attribution scores. Next, we explore two methods to decrease the bias in human perception (Figure 5): (i) controlling for the bias by modifying the color-coding to account for over-estimation and under-estimation of importance (over-estimated tokens will receive decreased color saturation, and vice versa); (ii) replacing the color-coding visualization with bar chart visualization.

6.1 Model-Based Color Correction Technique

We compute an alternative color-coding visualization which a-priori accounts for over-estimation and under-estimation of tokens based on the data collected in the previous experiments. Here we investigate whether it is possible to “correct” the explainees’ saliency perception by superimposing the initial saliency values with a correction signal.

We require a procedure which increases the saliency scores for words which are predicted to be under-perceived (e.g., short words and words that appear in long sentences) and decrease the saliency scores for words that are predicted to be over-perceived (e.g., words with a high sentiment polarity or words that appear in short sentences). Briefly, the trained GAMM model from the English study (Section 5.1) allows us to map a combination of a saliency score together with word/sentence properties to a perceived importance score (on a continuous latent scale). By grounding this prediction of perceived importance to a prediction conditioned on a particularly chosen reference level, we can iteratively, globally correct the explained scores over the sentence such that the (predicted) perception bias is decreased in each iteration. Table 4 displays examples of the application of this correction. In Appendix E, we discuss the full algorithm including its components and motivating details and provide an extended list of example applications in Table 11 as well as an example of the gradual correction of one sentence over the course of 100 correction steps in Table 10.

6.2 Bar Chart Visualization

For an alternative to color-coding visualization, we consider bar charts (Figure 5c). Here we investigate whether a sufficiently distinct visualization method will result in different perception in our experiments. We hypothesize that this is related to visual perception bias.

Table 4. Examples of the bias reduction procedure. The *saliency* column shows the saliency explanations (how users would see them) before and after the bias correction procedure. The *bias* column shows the color-coded bias estimates. Predicted over-estimations are in red whereas predicted under-estimations are in blue.

	Saliency	Bias	Removed Bias
original	Great people !	Great people !	94.9%
corrected	Great people !	Great people !	
original	Horrible service .	Horrible service .	100.0%
corrected	Horrible service .	Horrible service .	
original	I remain unhappy .	I remain unhappy .	84.3%
corrected	I remain unhappy .	I remain unhappy .	
original	many thanks 2scompany ...	many thanks 2scompany ...	100.0%
corrected	many thanks 2scompany ...	many thanks 2scompany ...	

We note two visual qualities of bars which differentiate it from color-coding, and therefore make it a relevant alternative visualization candidate: (i) The bars are communicated with objective reference points of 0 and 1 (the top and bottom of the draw area), while the results in Section 5.1 indicate that participants perceive colored saliency in relation to each other, instead of in reference to 0 and 1 (pure white and pure red, respectively); (ii) The draw area for the bars is separate from the draw area for the input text, in contrast to color-coding, where they occupy the same space. This means that in color-coding, for example, a word with more characters will receive a larger area of color, in comparison to a shorter word with the same color. As our studies in Section 5 show, word length influenced explainee perception. In the bar chart visualization scheme, all words are treated identically within the draw area which communicates importance, and this draw area is disconnected from the text display area.

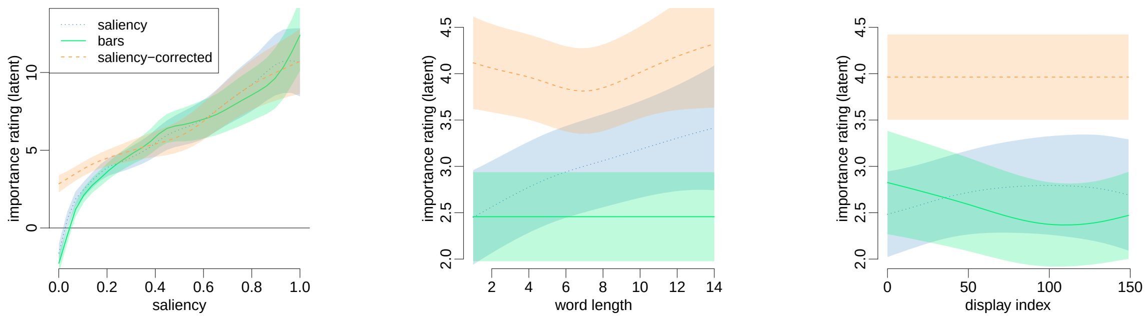
6.3 Results

We investigate how well the two proposed visualization alternatives counteract bias in user perception within the study described in Section 5.3. We find that visualization has a significant effect on importance ratings (df=2, F=35.45, p<0.0001) where the bar visualization leads to lower importance ratings ($\beta = -0.5991$, SE=0.1579) and the correction visualization leads to higher importance ratings ($\beta = 1.1102$, SE=0.2515). Regarding the visualizations' effect on smooth terms, we focus on the smooths for color saturation, word length and display index in Figure 6.

Figure 6a shows that the saliency scores' effect on importance ratings is very similar for the unmodified saliency visualizations and the bar visualizations, while the corrected saliency visualization leads to higher importance ratings in the lower end of the color saturation spectrum. These differences are neither "good" nor "bad", but we argue that the similarity between the original saliency visualization and the bar visualization is remarkable as the two visualizations are fundamentally different.

Figure 6b shows that the biasing effect of word length in the original visualization is successfully eliminated using the bar visualization as shown by the nearly constant smooth of the bar visualization (edf=0.0009). This confirms our hypothesis that a bar visualization evades word length bias. The correction visualization leads to a different effect than the original visualization, however, this effect indicates a different but equally distorting bias of word length.

Figure 6c indicates a successful application of our color correction technique. While the original visualization as well as the bar visualization show a biasing effect regarding the model smooths, the saliency correction visualization leads



(a) Saliency (b) Word length (c) Temporal display index

Fig. 6. Selected summed-effects comparison plots of the visualization alternatives.

to a nearly constant smooth (edf=0.0009). Regarding the original and the bar visualizations, the smooths indicate that, in contrast to the original visualization, the bar visualization leads to an initial overestimation of importances which decreases over time, while the original visualization lead to a respective underestimation. However, a difference plot between the two conditions (see Figure 12c in Appendix F) shows no significant differences.

While these examples demonstrate indications for successful bias mitigation, we want to note that this mitigation cannot be observed for most of the other variables, in particular not for the effect of saliency rank, which we expected to be mitigated by the bar visualization. We provide summed-effect comparison plots for all effects under investigation in Figure 11, difference plots between all conditions in Figures 12 and 13 as well as details and test statistics on all parametric coefficients as well as smooth terms in Table 12 in Appendix F.

Tying back to our initial categorization of biases, we observe that our proposed visualization alternatives can successfully remove instances of visual bias (word length) and learning effect bias (display index).[19]

7 CONCLUSION

We analyze feature-attribution for text from a novel, arguably under-explored perspective: we investigate how humans interpret saliency explanations over text and which factors affect their perception. We show that there can be discrepancy between the communicated information and how it is interpreted, even for a straight-forward and explicit explanation medium of feature-attribution for text. This is achieved through a general methodology for investigating which factors in the input may cause this discrepancy. We demonstrate the methodology for a lay-people audience of crowd-workers over multiple tasks, languages and visualizations, showing different setups yield similar but distinct distortions. We find that word length, sentence length, learning effects, and within-sentence saliency relations affect human importance ratings across multiple user studies. The methodology can and should be used on other audiences and tasks, before trusting a saliency visualization for this audience/task pair. We present two bias correction methods and demonstrate their ability to compensate the distorting influence of word length and repeated exposure. Our findings inform future design of saliency visualizations towards closing the gap between communicated and interpreted saliency explanations, and call for further research in the human factors in interpretation methods of AI, that study not only how the AI operates, but how humans perceive the communicated information.

[19] We hypothesize that belief biases (such as sentiment polarity) exhibit more distinct expression across indiviuals, which requires subject-adaptive correction methods and should be addressed by online estimation of individual participant slopes and intercepts within our GAMM model in future work.

ACKNOWLEDGMENTS

We are grateful to Paul Roit, Diego Frassinelli, Yanai Elazar, and the members of the BCAI NLP&KRR research group, especially Sophie Henning and Stefan Grünewald, for valuable advice and feedback. Alon Jacovi and Yoav Goldberg received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 802774 (iEXTRACT). Ngoc Thang Vu is funded by Carl Zeiss Foundation.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 9525–9536. https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

[2] Siddhant Arora, Danish Pruthi, Norman M. Sadeh, William W. Cohen, Zachary C. Lipton, and Graham Neubig. 2021. Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations. *CoRR* abs/2112.09669 (2021). arXiv:2112.09669 https://arxiv.org/abs/2112.09669

[3] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. *CoRR* abs/1612.07843 (2016). arXiv:1612.07843 http://arxiv.org/abs/1612.07843

[4] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. " What is relevant in a text document?": An interpretable machine learning approach. *PloS one* 12, 8 (2017), e0181142. Publisher: Public Library of Science San Francisco, CA USA.

[5] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, Alexandra Balahur, Saif M. Mohammad, and Erik van der Goot (Eds.). Association for Computational Linguistics, 159–168. https://doi.org/10.18653/v1/w17-5221

[6] Nadia Burkart and Marco F. Huber. 2021. A Survey on the Explainability of Supervised Machine Learning. *J. Artif. Intell. Res.* 70 (2021), 245–317. https://doi.org/10.1613/jair.1.12228

[7] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (Jul 2019), 832. https://doi.org/10.3390/electronics8080832

[8] Mike Dacey. 2017. Anthropomorphism as Cognitive Bias. *Philosophy of Science* 84, 5 (2017), 1152–1164. https://doi.org/10.1086/694039

[9] Mike Dacey. 2017. Anthropomorphism as cognitive bias. *Philosophy of Science* 84, 5 (2017), 1152–1164. Publisher: University of Chicago Press Chicago, IL.

[10] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, Kam-Fai Wong, Kevin Knight, and Hua Wu (Eds.). Association for Computational Linguistics, 447–459. https://aclanthology.org/2020.aacl-main.46/

[11] Kate Darling. 2015. 'Who's Johnny?' Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. *SSRN Electronic Journal* (01 2015). https://doi.org/10.2139/ssrn.2588669

[12] Jonathan Dinu, Jeffrey P. Bigham, and J. Zico Kolter. 2020. Challenging common interpretability assumptions in feature attribution explanations. *CoRR* abs/2012.02748 (2020). https://arxiv.org/abs/2012.02748 arXiv: 2012.02748.

[13] Dagmar Divjak and Harald Baayen. 2017. Ordinal GAMMs: a new window on human ratings. In *Each venture, a new beginning: Studies in Honor of Laura A. Janda*. Slavica Publishers, 39–56.

[14] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael J. Muller, and Mark O. Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *CoRR* abs/2107.13509 (2021). arXiv:2107.13509 https://arxiv.org/abs/2107.13509

[15] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review* 114, 4 (Oct. 2007), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

[16] J St BT Evans, Julie L Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition* 11, 3 (1983), 295–306. Publisher: Springer.

[17] Thomas Fel, Rémi Cadène, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. 2021. Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis. *CoRR* abs/2111.04138 (2021). arXiv:2111.04138 https://arxiv.org/abs/2111.04138

[18] Thomas Fel, Julien Colin, Rémi Cadène, and Thomas Serre. 2021. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. *CoRR* abs/2112.04417 (2021). arXiv:2112.04417 https://arxiv.org/abs/2112.04417

[19] Thomas Fel, Julien Colin, Remi Cadene, and Thomas Serre. 2021. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. arXiv:2112.04417 [cs.CV]

[20] Shi Feng and Jordan L. Boyd-Graber. 2019. What can AI do for me?: evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*, Wai-Tat Fu, Shimei Pan,

Oliver Brdiczka, Polo Chau, and Gaelle Calvary (Eds.). ACM, 229–239. https://doi.org/10.1145/3301275.3302265

[21] Lorenzo Gatti, Marco Guerini, and Marco Turchi. 2016. SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. *IEEE Trans. Affect. Comput.* 7, 4 (2016), 409–421. https://doi.org/10.1109/TAFFC.2015.2476456

[22] Ana Valeria Gonzalez, 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, Online Event, August 1-6, 2021 (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 2930–2942. https://doi.org/10.18653/v1/2021.findings-acl.259

[23] Christopher Grimsley, Elijah Mayfield, and Julia R.S. Bursten. 2020. Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1780–1790. https://aclanthology.org/2020.lrec-1.220

[24] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *CoRR* abs/1805.10820 (2018). arXiv:1805.10820 http://arxiv.org/abs/1805.10820

[25] Woodrow Hartzog. 2015. UNFAIR AND DECEPTIVE ROBOTS. *Maryland Law Review* 74 (2015), 785.

[26] Trevor J Hastie and Robert J Tibshirani. 1990. *Generalized additive models*. Vol. 43. CRC press.

[27] 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, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

[28] Alon Jacovi and Yoav Goldberg. 2021. Aligning Faithful Interpretations with their Social Attribution. *Trans. Assoc. Comput. Linguistics* 9 (2021), 294–310. https://transacl.org/ojs/index.php/tacl/article/view/2635

[29] David Kyle Johnson. 2018. *Anthropomorphic Bias*. John Wiley & Sons, Ltd, Chapter 69, 305–307. https://doi.org/10.1002/9781119165811.ch69 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119165811.ch69

[30] Daniel Kahneman and Shane Frederick. 2002. Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In *Heuristics and Biases: The Psychology of Intuitive Judgment*, Thomas Gilovich, Dale Griffin, and DanielEditors Kahneman (Eds.). Cambridge University Press, 49–81. https://doi.org/10.1017/CBO9780511808098.004

[31] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[32] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (Un)reliability of Saliency Methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (Eds.). Lecture Notes in Computer Science, Vol. 11700. Springer, 267–280. https://doi.org/10.1007/978-3-030-28954-6_14

[33] Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 107–117. https://doi.org/10.18653/v1/d16-1011

[34] Zheyuan Li and Simon N. Wood. 2020. Faster model matrix crossproducts for large generalized linear models with discretized covariates. *Stat. Comput.* 30, 1 (2020), 19–25. https://doi.org/10.1007/s11222-019-09864-2

[35] Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2021. Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. *CoRR* abs/2110.08412 (2021). https://arxiv.org/abs/2110.08412 arXiv: 2110.08412.

[36] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc Interpretability for Neural NLP: A Survey. *CoRR* abs/2108.04840 (2021). arXiv:2108.04840 https://arxiv.org/abs/2108.04840

[37] Bertram F. Malle. 2003. Folk Theory of Mind: Conceptual Foundations of Social Cognition. http://cogprints.org/3315/

[38] Zahia Marzouk. 2018. Text Marking: A Metacognitive Perspective.

[39] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[40] Manisha Natarajan and Matthew Gombolay. 2020. Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/3319502.3374839

[41] John Ashworth Nelder and Robert WM Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* 135, 3 (1972), 370–384. Publisher: Wiley Online Library.

[42] Marco Túlio 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, San Francisco, CA, USA, August 13-17, 2016*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 1135–1144. https://doi.org/10.1145/2939672.2939778

[43] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main*

Volume, EACL 2021, Online, April 19 - 23, 2021, Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, 1351–1361. https://aclanthology.org/2021.eacl-main.115/

[44] Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

[45] Alvy Ray Smith. 1978. Color gamut transform pairs. *ACM Siggraph Computer Graphics* 12, 3 (1978), 12–19. Publisher: ACM New York, NY, USA.

[46] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3319–3328. http://proceedings.mlr.press/v70/sundararajan17a.html

[47] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. , 107–118 pages. https://www.aclweb.org/anthology/2020.emnlp-demos.15

[48] Erico Tjoa and Cuntai Guan. 2021. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE transactions on neural networks and learning systems* 32, 11 (November 2021), 4793—4813. https://doi.org/10.1109/tnnls.2020.3027314

[49] Po-He Tseng, Ran Carmi, Ian GM Cameron, Douglas P Munoz, and Laurent Itti. 2009. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision* 9, 7 (2009), 4–4. Publisher: The Association for Research in Vision and Ophthalmology.

[50] David Tuckey, Krysia Broda, and Alessandra Russo. 2019. Saliency Maps Generation for Automatic Text Summarization. *CoRR* abs/1907.05664 (2019). http://arxiv.org/abs/1907.05664 arXiv: 1907.05664.

[51] Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based Analysis of NLP Models is Manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 247–258. https://doi.org/10.18653/v1/2020.findings-emnlp.24

[52] David Watson. 2020. *The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence.* 45–65. https://doi.org/10.1007/978-3-030-29145-7_4

[53] 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)*. Association for Computational Linguistics, Hong Kong, China, 11–20. https://doi.org/10.18653/v1/D19-1002

[54] S.N. Wood, N., Pya, and B. S"afken. 2016. Smoothing parameter and model selection for general smooth models (with discussion). *J. Amer. Statist. Assoc.* 111 (2016), 1548–1575.

[55] S. N. Wood. 2003. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)* 65, 1 (2003), 95–114.

[56] S. N. Wood. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Assoc.* 99, 467 (2004), 673–686.

[57] S. N. Wood. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73, 1 (2011), 3–36.

[58] Simon N Wood. 2017. *Generalized additive models: an introduction with R.* CRC press.

[59] Simon N Wood, Zheyuan Li, Gavin Shaddick, and Nicole H Augustin. 2017. Generalized additive models for gigadata: modeling the UK black smoke network daily data. *J. Amer. Statist. Assoc.* 112, 519 (2017), 1199–1210. Publisher: Taylor & Francis.

[60] Yu Zhang, Peter Tiño, Ales Leonardis, and Ke Tang. 2021. A Survey on Neural Network Interpretability. *IEEE Trans. Emerg. Top. Comput. Intell.* 5, 5 (2021), 726–742. https://doi.org/10.1109/TETCI.2021.3100641

[61] Jakub Zlotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *International Journal of Social Robotics* 7 (2015), 347–360.

A BRIEF INTRODUCTION TO ORDINAL GAMMS

For an intuitive understanding, we sketch how one arrives at ordinal generalized additive mixed models (GAMMs) starting from linear models. We follow notation by Wood [58].

Linear Model. In a linear model, the response variable $\mathbf{y}$ (e.g., a numeric rating of importance) is modeled as a function of explanatory variables $\mathbf{X}$ which are related to $\mathbf{y}$ *linearly* via parameters β assuming additional noise ϵ:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon. \tag{1}$$

Linear Mixed Model. In many scenarios, there are *random effects* which one wants to account for in the model. For example, we collect 150 word importance ratings per participant, i.e., we collect *repeated measures* and are at danger of violating the independence assumption and introducing a confounding effect of the variable *participant ID* because specific participants might have a tendency to give overall higher ratings than other participants. Like the linear model, linear mixed models estimate *fixed effects* but in addition they also model *random effects* (e.g., of the participant ID) to disentangle their influence on the response variable and thereby offer a clearer view on the fixed effects. The general formulation of a linear mixed model reads

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon, \tag{2}$$

where $\mathbf{Z}$ corresponds to the random effects and $\mathbf{b}$ to the respective weights.

Generalized Linear Model (GLM). While linear models require the response distribution to be normal, *generalized* linear models [41] generalize to non-normal (exponential family) response distributions such as categorical responses (e.g., dog or cat) or ordinal responses (e.g., Likert item ratings). To achieve this generalization, GLMs link values on the response scale (e.g., categorial ratings) to a latent scale (e.g., logits) via a *link function* $g(\cdot)$ (e.g., logit function): For a row i, the general formulation reads:

$$g(\mu_i) = \mathbf{X}_i\beta. \tag{3}$$

Generalized Additive Model (GAM). While a generalized linear model only allows to model linear relationships between the the explanatory variables and $g(\mu_i)$, a GAM [26] generalizes the linear relationship to a *sum of smooth functions* of explanatory variables using:

$$g(\mu_i) = \mathbf{X}^*{}_i\theta + f_1(x_{1i}) + f_2(x_{2i}, x_{3i}) + \ldots, \tag{4}$$

where f_1 and f_2 are smooth functions that typically are chosen to be a sum of basis functions, such as splines and $\mathbf{X}^*$ corresponds to strictly parametric model components. A regularized estimation of these functions allows GAMs to model complex functions, but also to fall back to simpler, e.g., constant or linear functions when an increase in model complexity is not sufficiently warranted by improved model fit.

Ordinal Generalized Additive Mixed Model (ordinal GAMM). Having introduced the previous models, an ordinal GAMM can be described as a generalized additive model that additionally accounts for random effects and models ordinal ratings via a continuous latent variable that is separated into the ordinal categories via estimated threshold values. For further details, Divjak and Baayen [13] provide a practical introduction to ordinal GAMs in a linguistic context and Wood [58] offers a detailed textbook on GAM(M)s including implementation and analysis details.

B STUDY INTERFACES

In addition to the screenshot shown in Figure 2, Figure 7 shows the interface of the German study and Figure 8 shows an interface that uses the alternative bar chart visualization. Figure 9 displays one of the three trap questions we use to detect participants that do not pay attention to the task.

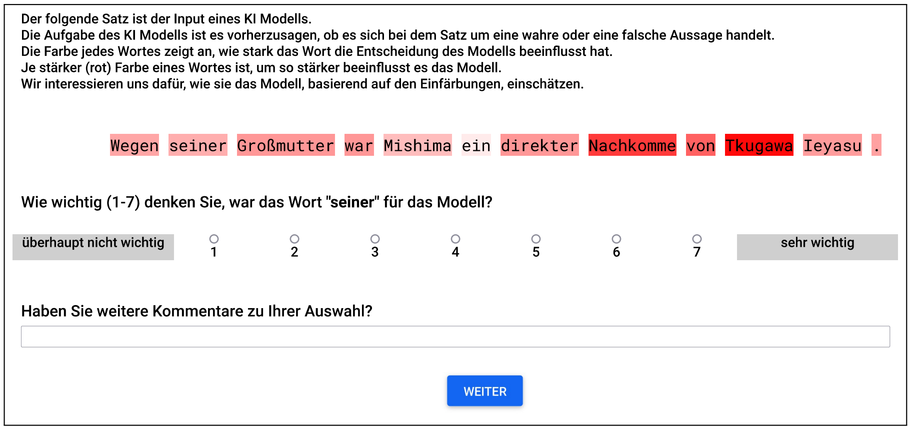


Fig. 7. Screenshot of the importance rating interface for German fact checking sentences using saliency visualization.

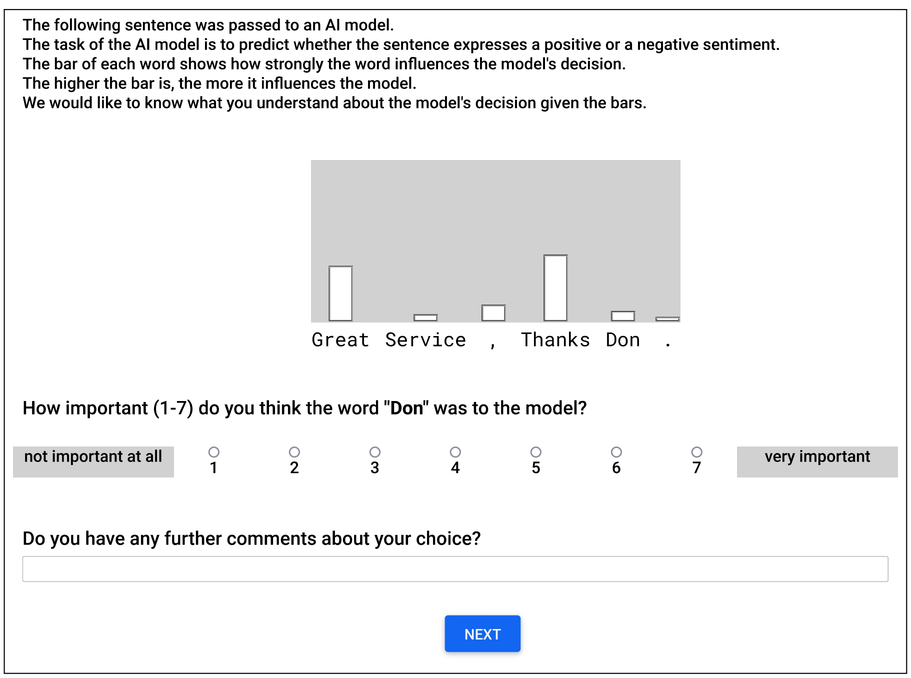


Fig. 8. Screenshot of the importance rating interface for English sentiment sentences using bar visualization.

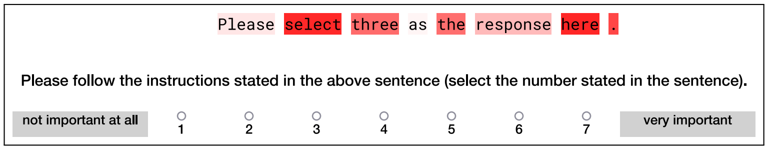


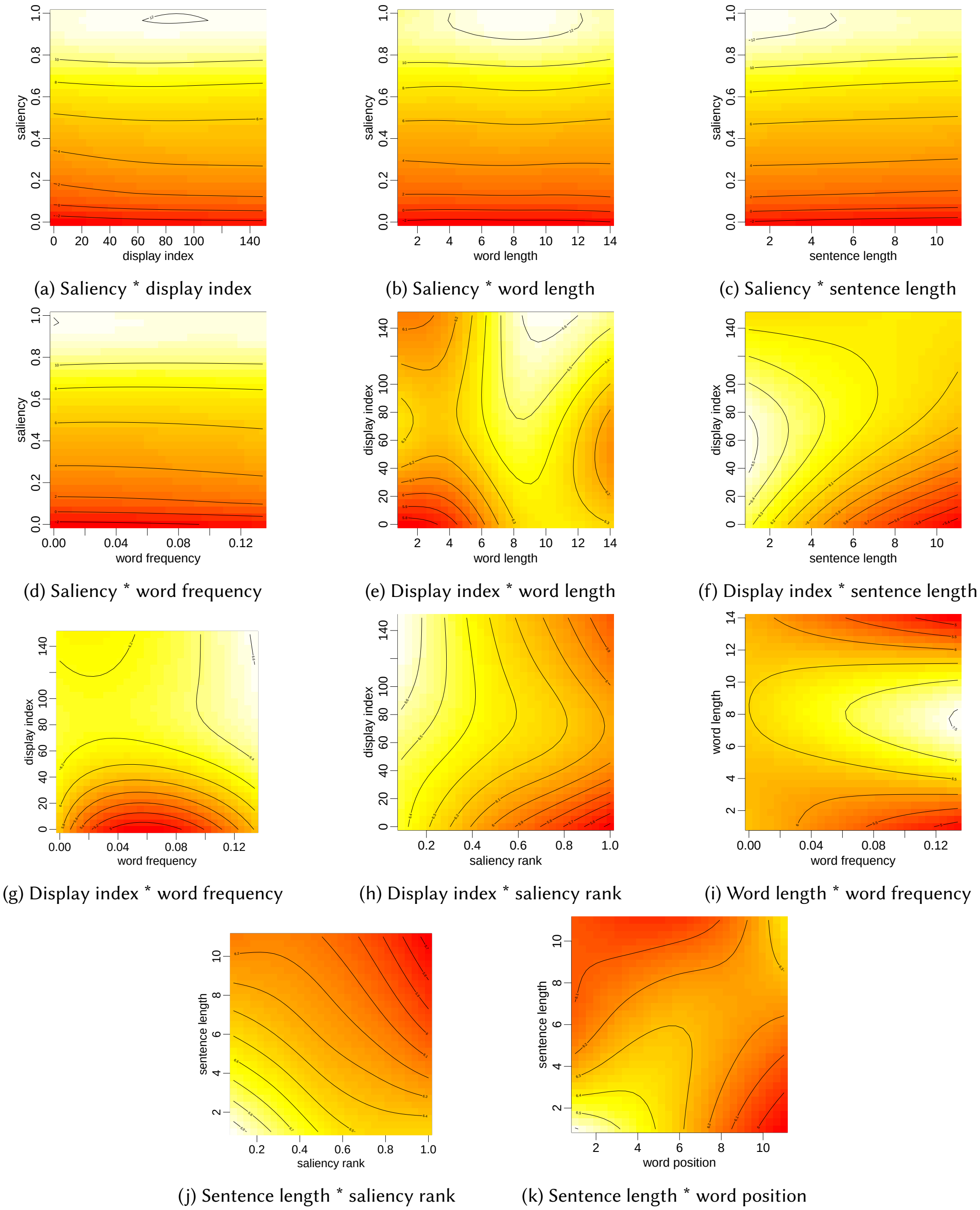
Fig. 9. Screenshot of one of three trap sentences used to validate that the participant pays attention to the task.

C ENGLISH STUDY DETAILS

Table 5 displays test statistics for all smooth pairwise interactions. We make use of tensor interaction smooths following a functional ANOVA decomposition. Figure 10 shows summed effect plots for the respective significant interactions. Ordered categorial cut points are located at -1, 1.31, 3.29, 5.15, 7.1 and 9.22.

Table 5. Wald tests for the pairwise interactions (tensor interactions) (upper) and random effects (lower) of the English user study.

	edf	ref. df	F	p
ti(saliency,display index)	2.5102	16	2.1075	0.0001
ti(saliency,word length)	6.0566	16	2.2698	0.0001
ti(saliency,sentence length)	3.1609	16	1.1203	0.0020
ti(saliency,word frequency)	0.9176	12	1.8325	0.0004
ti(saliency,sentiment polarity)	2.9357	16	0.5553	0.0814
ti(saliency,saliency rank)	0.0004	16	0.0000	0.5864
ti(saliency,word position)	0.6254	16	0.1144	0.1276
ti(display index,word length)	1.5112	16	0.6637	0.0026
ti(display index,sentence length)	1.2776	16	1.0159	0.0010
ti(display index,word frequency)	2.6938	16	1.7810	0.0001
ti(display index,sentiment polarity)	0.5386	16	0.0853	0.1678
ti(display index,saliency rank)	1.3966	16	0.5272	0.0174
ti(display index,word position)	3.3649	16	0.6625	0.0520
ti(word length,sentence length)	0.0004	16	0.0000	0.9236
ti(word length,word frequency)	2.1540	16	6.5510	< 0.0001
ti(word length,sentiment polarity)	0.0014	16	0.0000	0.6790
ti(word length,saliency rank)	2.2175	16	0.3503	0.0573
ti(word length,word position)	1.0296	16	0.1270	0.1222
ti(sentence length,word frequency)	0.0005	16	0.0000	0.8608
ti(sentence length,sentiment polarity)	0.0013	16	0.0001	0.5113
ti(sentence length,saliency rank)	1.3045	16	0.2651	0.0453
ti(sentence length,word position)	3.1995	16	0.8487	0.0067
ti(word frequency,sentiment polarity)	0.0015	16	0.0001	0.1969
ti(word frequency,saliency rank)	0.0022	15	0.0001	0.3230
ti(word frequency,word position)	2.0375	16	0.3168	0.0924
ti(sentiment polarity,saliency rank)	0.0006	16	0.0000	0.8407
ti(sentiment polarity,word position)	0.0005	16	0.0000	0.9558
ti(saliency rank,word position)	0.0006	16	0.0000	0.6542
s(sentence_id)	0.0006	150	0.0000	0.9276
s(saliency,sentence_id)	9.1441	150	0.0676	0.2305
s(worker_id)	48.1065	49	10640.8475	< 0.0001
s(saliency,worker_id)	48.0654	50	6593.7769	< 0.0001



(a) Saliency * display index
(b) Saliency * word length
(c) Saliency * sentence length
(d) Saliency * word frequency
(e) Display index * word length
(f) Display index * sentence length
(g) Display index * word frequency
(h) Display index * saliency rank
(i) Word length * word frequency
(j) Sentence length * saliency rank
(k) Sentence length * word position

Fig. 10. Summed effect plots of all significant pairwise interactions.

D GERMAN STUDY DETAILS

In this section, we provide details on the analysis of the German experiment. Table 7 and Table 6 display test statistics for the smooth and parametric terms of the fitted GAMM model. Table 8 shows statistics regarding parametric coefficient estimates. Cut points are located at -1, 0.86, 2.42, 3.75, 5.53 and 7.67. Table 9 lists exemplary participant comments.

Table 6. Wald tests for the parametric terms of the German user study.

	df	F	p
capitalization	2	7.62	0.0005
dependency relation	33	2.57	< 0.0001

Table 7. Wald tests for the smooth terms of the German user study.

	edf	ref. df	F	p
s(saliency)	8.2052	19	148.1115	< 0.0001
s(display index)	1.5999	9	2.4742	< 0.0001
s(word length)	2.0440	9	3.9174	< 0.0001
s(sentence length)	0.9073	9	1.7657	0.0003
s(word frequency)	0.0017	9	0.0002	0.2816
s(saliency rank)	0.0004	9	0.0000	0.7016
s(word position)	2.4429	9	2.8142	0.0002
ti(saliency,display index)	0.0007	16	0.0000	0.5846
ti(saliency,word length)	2.4114	16	1.1662	0.0013
ti(saliency,sentence length)	1.8496	16	0.7410	0.0125
ti(saliency,word frequency)	0.6953	11	0.3084	0.0549
ti(saliency,saliency rank)	1.4340	16	0.4958	0.0142
ti(saliency,word position)	0.0765	16	0.0053	0.2970
ti(display index,word length)	0.3529	16	0.0477	0.1968
ti(display index,sentence length)	0.1902	16	0.0171	0.2622
ti(display index,word frequency)	0.0005	15	0.0000	0.7096
ti(display index,saliency rank)	0.2967	16	0.0332	0.2325
ti(display index,word position)	1.1440	16	0.4244	0.0168
ti(word length,sentence length)	0.9858	16	0.3138	0.0290
ti(word length,word frequency)	0.9622	11	1.0293	0.0050
ti(word length,saliency rank)	0.0005	16	0.0000	0.8581
ti(word length,word position)	0.8285	16	0.5132	0.0091
ti(sentence length,word frequency)	0.0009	15	0.0001	0.3536
ti(sentence length,saliency rank)	0.0005	16	0.0000	0.9945
ti(sentence length,word position)	0.0005	16	0.0000	0.6862
ti(word frequency,saliency rank)	0.0003	16	0.0000	0.9438
ti(word frequency,word position)	0.0005	15	0.0000	0.6085
ti(saliency rank,word position)	0.0004	16	0.0000	0.8379
s(sentence ID)	0.0004	149	0.0000	0.9007
s(saliency,sentence ID)	36.6567	150	0.3534	0.0087
s(worker ID)	23.5324	24	8128.6327	< 0.0001
s(saliency,worker ID)	23.6122	25	5645.0812	< 0.0001

Table 8. Capitalization and dependency relation coefficients for the German user study.

Coefficients	β	**SE**	**t**	**p**
capitalization: all capital	1.9051	0.9638	1.9767	0.0481
capitalization: first capital	0.4074	0.1151	3.5390	0.0004
dependency relation: acl	-1.2155	0.6428	-1.8910	0.0587
dependency relation: acl:relcl	1.3605	0.5947	2.2878	0.0222
dependency relation: advcl	0.8647	0.7154	1.2087	0.2269
dependency relation: advmod	0.3741	0.2369	1.5790	0.1144
dependency relation: amod	0.4794	0.2653	1.8072	0.0708
dependency relation: appos	0.2823	0.4119	0.6852	0.4932
dependency relation: aux	0.6395	0.3138	2.0379	0.0416
dependency relation: aux:pass	-0.0679	0.3798	-0.1789	0.8581
dependency relation: case	0.1169	0.2082	0.5613	0.5746
dependency relation: cc	0.1126	0.2571	0.4379	0.6615
dependency relation: cc:preconj	0.8039	1.1491	0.6996	0.4842
dependency relation: ccomp	1.1850	0.5206	2.2763	0.0229
dependency relation: compound	0.8738	0.4488	1.9470	0.0516
dependency relation: compound:prt	0.4114	0.4577	0.8989	0.3688
dependency relation: conj	0.1673	0.2900	0.5769	0.5640
dependency relation: cop	0.4169	0.2598	1.6043	0.1087
dependency relation: csubj	1.0154	0.7533	1.3480	0.1777
dependency relation: det	-0.1604	0.2088	-0.7682	0.4424
dependency relation: expl	-1.0130	0.4605	-2.1998	0.0279
dependency relation: flat:name	0.3786	0.5401	0.7010	0.4833
dependency relation: iobj	-0.4807	0.5162	-0.9312	0.3518
dependency relation: mark	0.1537	0.3646	0.4216	0.6734
dependency relation: nmod	0.4656	0.2787	1.6707	0.0949
dependency relation: nmod:poss	-0.0658	0.3251	-0.2025	0.8395
dependency relation: nsubj	0.4443	0.2369	1.8755	0.0608
dependency relation: nsubj:pass	0.4296	0.4305	0.9979	0.3184
dependency relation: nummod	1.3866	0.3609	3.8419	0.0001
dependency relation: obj	0.2406	0.2649	0.9082	0.3638
dependency relation: obl	0.3126	0.2679	1.1668	0.2434
dependency relation: obl:tmod	1.7042	0.5544	3.0739	0.0021
dependency relation: parataxis	-0.2780	0.8595	-0.3234	0.7464
dependency relation: root	0.5463	0.2432	2.2460	0.0248
dependency relation: xcomp	0.6718	0.4494	1.4948	0.1351

Table 9. Comments of the participants of the German study. Participants were asked to rate the underlined word or symbol.

Sentence with Saliency Explanation	Rating	Comment
Durch den Deal zwischen Aoun und Hariri kommen sich die beiden verfeindeten Bündnisse (vorerst) näher .	4	Wenn dann müssen beide Klammern weg (P4)
Jedes Gedicht erzählt nur von einem Teil des Krieges .	2	Das Symbol wird von der KI zu hoch bewertet. (P10)
Gewitterstürme sind selten , die Stadt berichtet nur an sieben Tagen pro Jahr von Gewittern .	7	Auch hier: "Gewitterstürme" viel zu gering gewichtet, "Jahr" zu hoch bewertet (P10)
Die Geschichte von Doss hat auch etwas Unglaubhaftes an sich , das sie nur umso attraktiver macht .	3	Der Artikel ist sicher wichtig, jedoch nicht zwingend für den Sinn verantwortlich. (P16)
Frau Hopley fügte hinzu : „ Der starke Anstieg des politischen Risikos sollte nicht unbeachtet bleiben. “	1	Es ist nur eine grammatische Kennzeichnung. Diese ist für KI meines Erachtens wenig bis garnicht relevant. (P16)
Wasser aus den Flüssen wird in über 500 Wasserkraftwerken genutzt , wobei 2900 Kilowatt Elekrizität generiert werden .	3	Die KI sollte schon den Wert einer Aussage kennen, die erst in der Zukunft eintritt und diese gegenüber aktuell bereits eingetretenen Ereignissen bewerten können. (P16)
Der Kunde kann die Forderung nach Veränderung verstärken .	5	Das Verb gibt dem Satz seinen Sinn. (P16)
Ich glaube , darum haben sie sich mit so vielen Mustern und Farben umgeben .	5	Das Adjektiv beschreibt eine wichtige Eigenschaft und ist für die Satzbewertung relevant. (P16)

E MODEL-BASED BIAS CORRECTION

Our second approach to bias mitigation is to leverage the previously described GAMM model of human saliency perception and to *correct* saliency perception by superimposing the initial saliency values with a correction signal.

Concretely, we want to increase the saliency scores for words which are predicted to be under-perceived (e.g., short words and words that appear in long sentences) and decrease the saliency scores for words which are predicted to be over-perceived (e.g., word's with a high polarity score or words that appear in very short sentences).

When we want to *correct* a user perception via the saliency scores, we cannot say whether a subjective user rating of importance is right or wrong. However, the previously described GAMM model allows us to map a combination of a saliency score together with word/sentence properties to a perceived importance score (on a continuous latent scale). In the following, we denote this mapping as:

$$u(s, \mathbf{x}) : [0, 1] \times \mathbb{R}^d \to \mathbb{R}, \tag{5}$$

where s is a saliency score and $\mathbf{x}$ is a d-dimensional feature vector representing the word/sentence properties. This function allows us to take a fixed saliency score s (e.g., 0.7) and predict its perceived importance given word and sentence features $\hat{\mathbf{x}}$ (corresponding to, e.g., a word length of 5 characters and a sentence length of 4). We define this predicted importance score as

$$p := u\left(s, \hat{\mathbf{x}}\right) \tag{6}$$

Additionally, it allows us to predict the perceived importance of that same saliency (0.7) in a hypothetical reference context $\mathbf{x}_{\text{ref}}$ (corresponding to, e.g., a word length of 3 and a sentence length of 6). We define this second predicted importance score as

$$p_{\text{ref}} := u\left(s, \mathbf{x}_{\text{ref}}\right) \tag{7}$$

We can now define a *bias score* $b \in \mathbb{R}$ as the difference between the the importance score for the saliency in the observed context and the importance score for the same saliency in the reference context

$$b := p - p_{\text{ref}}. \tag{8}$$

The predicted bias score b is positive if the saliency in the observed context is *over-perceived* with respect to the reference level and negative if it is *under-perceived* with respect to the reference level. A bias score of zero corresponds to an *unbiased* predicted perception. Intuitively, this formalization allows us to answer the question "*In which direction do I have to change the saliency such that the predicted bias with respect to the reference context is decreased?*".

To gain an executable process for bias mitigation we still lack (a) way to handle the random effects in the model, i.e., participant IDs and sentence IDs, (b) a definition of the reference context and (c) a procedure to minimize the absolute value of the bias score. We detail these three aspects in the following.

E.1 Including Random Effects

So far, our definition of the model function u ignores the random effects of the GAMM model, i.e., we did not specify which worker ID and which sentence ID should be used in predicting the importance score. However, the choice of the respective levels directly influences the model predictions not only via a the random intercepts but also via the random slopes for each worker and sentence ID. We see two options to address this. While a first, intuitive remedy is to use an arbitrary worker ID and an arbitrary sentence ID for all predictions, this approach has the disadvantage of introducing an arbitrary bias. Therefore, we choose to make each model prediction not only for one participant ID and one sentence

ID, but instead for all combinations of participant IDs and sentence IDs ($50 \times 150 = 7500$ combinations). Thereby, we consider each combination of participant sentence as equally relevant for the prediction on unseen participants and sentences and smooth-out extreme influences of single participants or sentence IDs. Formally, we thus update our definition of Equation (5) to:

$$u(s, \mathbf{x}, w, v) : [0, 1] \times \mathbb{R}^d \times W \times V \to \mathbb{R}, \tag{9}$$

where W is the set of participant (or crowdworker) IDs ($|W| = 50$) and V is the set of sentence IDs ($|V| = 150$). Consequently, a single evaluation of $u\,(s, \mathbf{x})$ is now replaced with

$$\frac{1}{|W||V|} \sum_{w \in W} \sum_{v \in V} u\,(s, \mathbf{x}, w, v)\,. \tag{10}$$

E.2 Choosing the Reference Context

So far, our definitions in Equations (6) to (8) do not impose any constraints on the choice of reference context. Why can we not just use an arbitrary reference context with, e.g., a word length of eight and a sentence length of one (and respective choices for all remaining covariates such as sentiment polarity etc.)? The problem that arises for that concrete context is that the model assigns a very high importance prediction to words with eight characters within a sentence with length one. Consequently, p_{ref} will be larger than p for most words and the bias score b would get negative, indicating an under-perception for most words. If we then increase all these words' saliency scores in order to minimize the absolute bias score, we, overall, have to make large changes to the saliencies. In other words, this specific reference contexts corresponds to an, overall, raised level of saliency intensities. While this is not bad per-se, we favor a reference context that is as neutral as possible regarding its impact on predicted importance ratings.

In order to find such a reference context, we sample 10001 random points from the space of possible contexts defined as the cross product of intervals of observed values (e.g., 1-37 characters word length) per variable if the variable is numeric (e.g., word length) and the set of possible values if the variable is categorial (e.g., dependency relation). Each point is a candidate context. We evaluate the term in Equation (10) for a saliency score of 0.5 and each candidate context. Among all predicted importance scores, we select the median score and choose the corresponding candidate context as our reference context $\mathbf{x}_{\text{ref}}$.[20]

E.3 Iterative Bias Minimization

In order to minimize the absolute predicted bias score, we have to modify each word's original saliency score $s^{(i)}_{\text{orig}} \in [0, 1]$ into a corrected saliency score $s^{(i)}_{\text{corr}} \in [0, 1]$. While this seems to be a straight-forward minimization at first glance there is one covariate in the model that complicates optimization. The value of the saliency rank variable depends on the saliencies of all words in the sentence. Thus, changing one word's saliency can impact all other word's saliency rank. We therefore propose an iterative minimization that (i) sequentially picks a token in the sentence (one after the other, round-robin) and (ii) updates this token's saliency score into the direction of a decreased absolute bias score while keeping all other tokens' saliencies fixed. Algorithm 1 shows the complete correction procedure, Table 10 shows

[20] The concrete $\mathbf{x}_{\text{ref}}$ corresponds to a "flat" dependency relation, a "first letter capitalized" capitalization, a display index of 129.7, a word length of 24.6, a sentence length of 4.1, a relative word frequency of 0.04, a sentiment polarity of -0.78, a normalized saliency rank of 0.11 and a word position index of 1.08. While non-integer values for, e.g. word length cannot occur in any prediction, this does not limit the utility of $\mathbf{x}_{\text{ref}}$ as the reference context as it only serves as an arbitraty, but neutral reference point.

the procedure's impact on an example sentence over the course of 100 optimization steps. Besides the examples shown in Table 4, we provide additional examples in Table 11.

Algorithm 1: Saliency color correction procedure.

```
Input: s_orig^(i): Original saliency scores for each word of the sentence with length l.
Input: x_ref: Feature representation of the reference input.
Output: s_corr^(i): Corrected saliency scores for each word of the sentence.
s_corr^(i) ← s_orig^(i) for all i. // Initialization
// Iterate for a fixed number of steps
for k ← 1 to n_steps do
    // Each iteration goes over all tokens in the sentence
    for i ← 1 to l do
        x̂ ← feature representation of the i-th word (also depends on all other s_corr^(i) via the saliency rank feature)
        p ← 1/(|W||V|) Σ_{w∈W} Σ_{v∈V} u(s_corr^(i), x̂, w, v) // Model-predicted perceived importance (on the
            latent continuous scale) averaged over participant IDs W and sentence IDs V.
        p_ref ← 1/(|W||V|) Σ_{w∈W} Σ_{v∈V} u(s_orig^(i), x_ref, w, v) // Model-predicted perceived importance if the
            word would be the reference level word (in the reference level sentence).
        b ← p − p_ref // Define bias.
        s_corr^(i) ← s_corr^(i) − α · (1 − (k−1)/n_steps)^2 · sgn(b) // Update saliency with quadratically-decaying step
            size (starting from α) into the direction of reduced predicted bias.
        s_corr^(i) ← max(0, min(s_corr^(i), 1)) // Make sure we stay within [0, 1].
    end
end
return s_corr^(i) for all i.
```

F INTEGRATED GRADIENTS AND CORRECTION STUDY

We report detailed estimates and test statistics regarding our third user study in Table 12. Figure 11 shows comparison plots for each smooth term and Figure 12 as well as Figure 13 visualize the respective difference functions between visualizations along with highlighted regions of significant differences. Cut points are located at -1, 0.95, 2.37, 3.67, 5.06 and 6.83.

Table 10. Evolution of saliency scores and corresponding bias estimates across 100 optimization steps of our bias correction procedure. The first row corresponds to the initial saliency scores. The first row of the right column shows that our method predicts that the word “thanks” is perceived as overly important, while the other parts of the sentence (especially “...”) are under-perceived. After 100 optimization steps, the saliencies of “many”, “2scompany” and “...” have been increased while the saliency of “thanks” is decreased resulting in a removal of nearly all predicted bias.

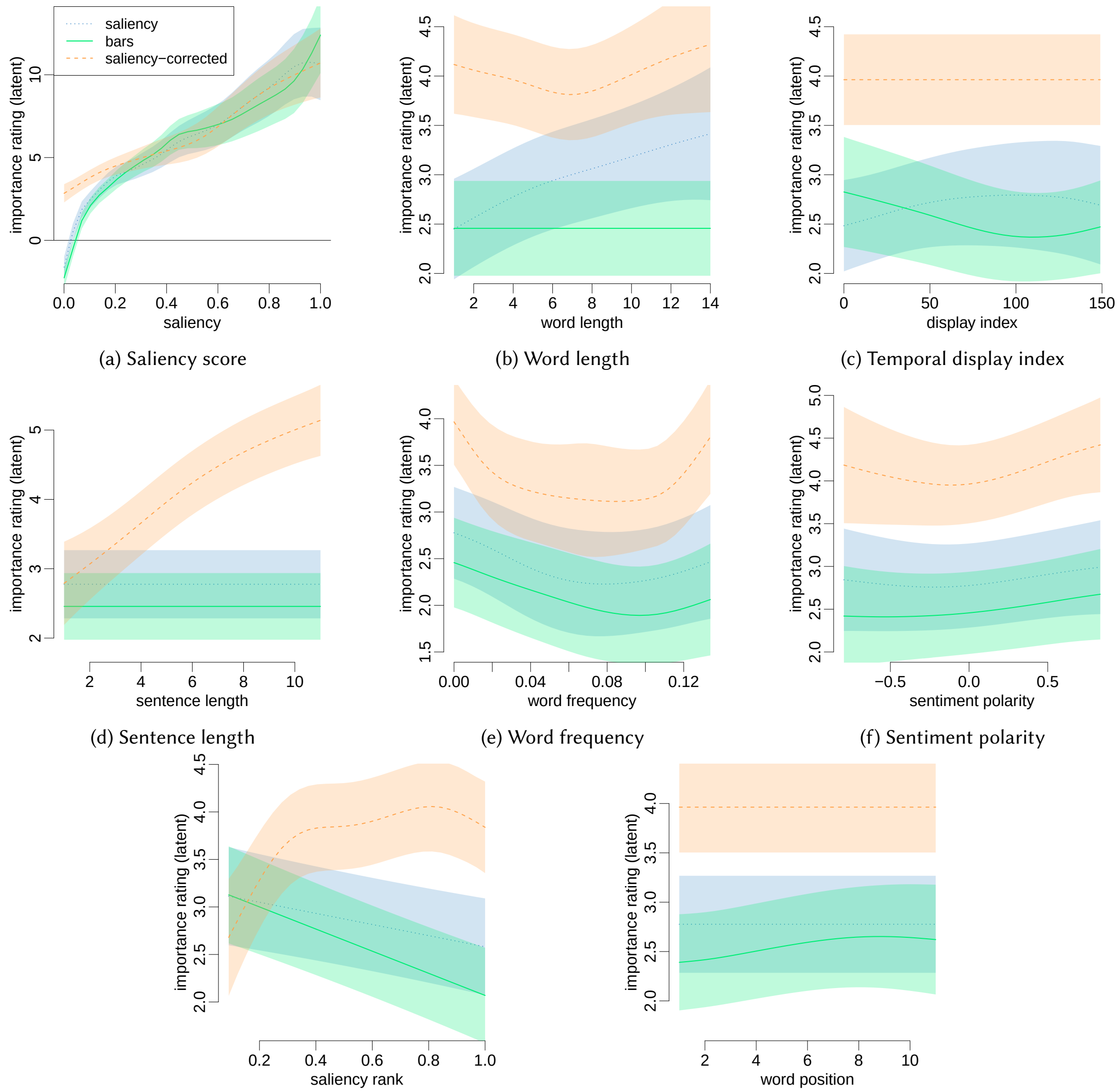
Step	Saliency				Bias			
1	many	thanks	2scompany	...	many	thanks	2scompany	...
10	many	thanks	2scompany	...	many	thanks	2scompany	...
21	many	thanks	2scompany	...	many	thanks	2scompany	...
41	many	thanks	2scompany	...	many	thanks	2scompany	...
61	many	thanks	2scompany	...	many	thanks	2scompany	...
81	many	thanks	2scompany	...	many	thanks	2scompany	...
100	many	thanks	2scompany	...	many	thanks	2scompany	...

Table 11. Examples of our proposed bias reduction method. The table shows sentences along with their initial saliency scores and the respective corrected saliency scores in the *saliency* column. The *bias* column shows the color-coded bias estimates as defined in our method. Predicted overestimations are colored in red whereas predicted underestimations are colored in blue. For each example, we scale the range of biases to use the full color spectrum in one direction. The column *removed bias* lists how many percent of the initial bias were removed in the corrected saliencies.

	Saliency	Bias	Removed Bias
original	Wonderful Atmosphere	Wonderful Atmosphere	100.0%
corrected	Wonderful Atmosphere	Wonderful Atmosphere	
original	Craig and Nate are wonderful .	Craig and Nate are wonderful .	95.3%
corrected	Craig and Nate are wonderful .	Craig and Nate are wonderful .	
original	Love this place !!	Love this place !!	91.6%
corrected	Love this place !!	Love this place !!	
original	But not so .	But not so .	98.5%
corrected	But not so .	But not so .	
original	Usually very quick and timely .	Usually very quick and timely .	92.7%
corrected	Usually very quick and timely .	Usually very quick and timely .	
original	Just ask American Express	Just ask American Express	100.0%
corrected	Just ask American Express	Just ask American Express	
original	Rubbish	Rubbish	76.3%
corrected	Rubbish	Rubbish	
original	Great Manicure	Great Manicure	100.0%
corrected	Great Manicure	Great Manicure	
original	Fantastic couple of days .	Fantastic couple of days .	86.6%
corrected	Fantastic couple of days .	Fantastic couple of days .	
original	They are especially rude to women .	They are especially rude to women .	80.7%
corrected	They are especially rude to women .	They are especially rude to women .	
original	Not enough seating .	Not enough seating .	89.4%
corrected	Not enough seating .	Not enough seating .	
original	Not impressed .	Not impressed .	100.0%
corrected	Not impressed .	Not impressed .	
original	The food was incredibly bland .	The food was incredibly bland .	86.8%
corrected	The food was incredibly bland .	The food was incredibly bland .	
original	Dessert was good .	Dessert was good .	92.9%
corrected	Dessert was good .	Dessert was good .	
original	Horrible !	Horrible !	100.0%
corrected	Horrible !	Horrible !	

Table 12. Parametric and smooth coefficients of the GAMM corresponding to the third user study comparing the three visualizations.

Parametric Coefficients	β	**SE**	**t**	**p**
(Intercept)	2.1119	0.1994	10.5909	< 0.0001
bars	-0.5991	0.1578	-3.7974	0.0001
saliency-corrected	1.1102	0.2515	4.4135	< 0.0001
Smooth Terms	**edf**	**ref. df**	**F**	**p**
s(saliency):saliency	11.4304	19	283.3393	< 0.0001
s(saliency):bars	11.0767	19	321.0314	< 0.0001
s(saliency):saliency-corrected	5.5202	19	113.9321	< 0.0001
s(display index):saliency	1.4830	9	7.2492	0.2575
s(display index):bars	1.7044	9	15.3135	0.0254
s(display index):saliency-corrected	0.0009	9	0.0001	0.6438
s(word length):saliency	1.7724	9	4.1550	< 0.0001
s(word length):bars	0.0009	9	0.0001	0.3775
s(word length):saliency-corrected	2.3645	9	1.3936	0.0213
s(sentence length):saliency	0.0005	9	0.0001	0.2313
s(sentence length):bars	0.0004	9	0.0000	0.8967
s(sentence length):saliency-corrected	2.4024	9	22.4406	< 0.0001
s(word frequency):saliency	1.8086	9	2.3192	< 0.0001
s(word frequency):bars	1.7381	9	2.7043	< 0.0001
s(word frequency):saliency-corrected	2.8913	9	7.2153	< 0.0001
s(sentiment polarity):saliency	1.0751	9	0.4727	0.0633
s(sentiment polarity):bars	1.0022	9	0.5076	0.0507
s(sentiment polarity):saliency-corrected	1.6991	9	2.2243	0.0020
s(saliency rank):saliency	0.9279	9	2.0901	0.0002
s(saliency rank):bars	0.9764	9	6.5779	< 0.0001
s(saliency rank):saliency-corrected	4.1893	9	6.8094	< 0.0001
s(word position):saliency	0.0004	9	0.0000	0.9754
s(word position):bars	1.2970	9	0.7165	0.0167
s(word position):saliency-corrected	0.0005	9	0.0000	0.9615
s(capitalization):saliency	0.0009	2	0.0003	0.4268
s(capitalization):bars	0.0003	2	0.0001	0.4525
s(capitalization):saliency-corrected	1.0644	2	3.2665	0.0245
s(dependency relation):saliency	0.0057	29	0.0002	0.3443
s(dependency relation):bars	0.0010	28	0.0000	0.5819
s(dependency relation):saliency-corrected	1.4715	28	0.0731	0.1955
s(condition order):saliency	3.7653	6	30.7306	0.0044
s(condition order):bars	0.0007	6	0.0001	0.5619
s(condition order):saliency-corrected	4.4665	6	150.1092	< 0.0001
s(sentence ID)	12.7259	150	0.1028	0.2236
s(saliency,sentence ID)	68.0861	150	1.7605	< 0.0001
s(worker ID)	55.7637	59	313.9570	< 0.0001
s(saliency,worker ID)	53.3619	60	230.3436	< 0.0001

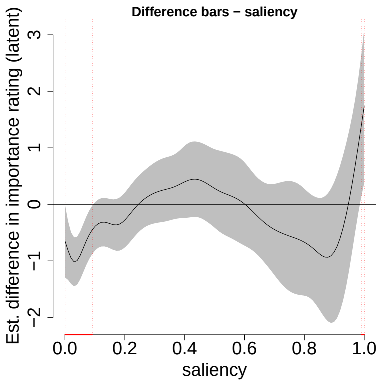


(a) Saliency score (b) Word length (c) Temporal display index

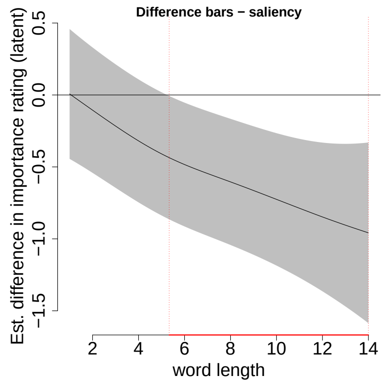
(d) Sentence length (e) Word frequency (f) Sentiment polarity

(g) Saliency rank (h) Word position

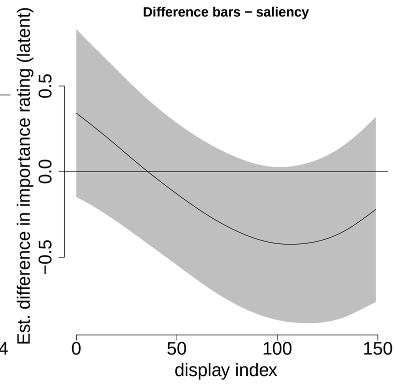
Fig. 11. Summed-effects comparison plots of the correction methods.



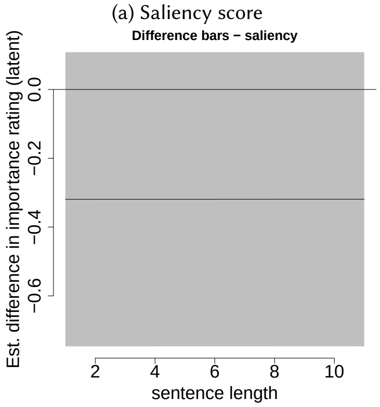
(a) Saliency score



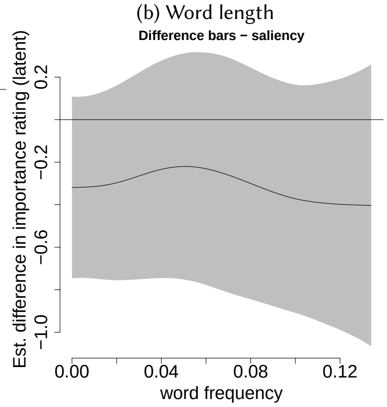
(b) Word length



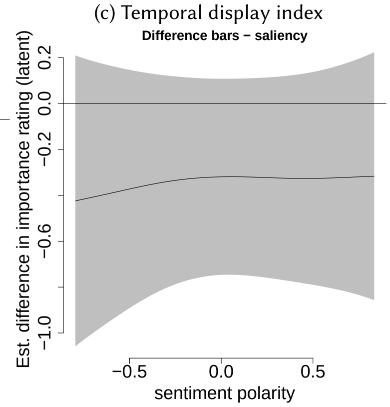
(c) Temporal display index



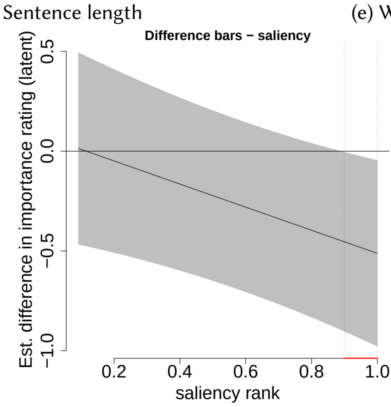
(d) Sentence length



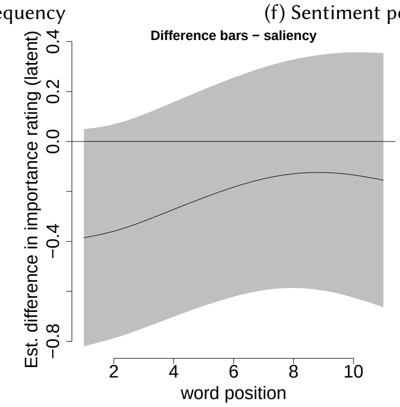
(e) Word frequency



(f) Sentiment polarity

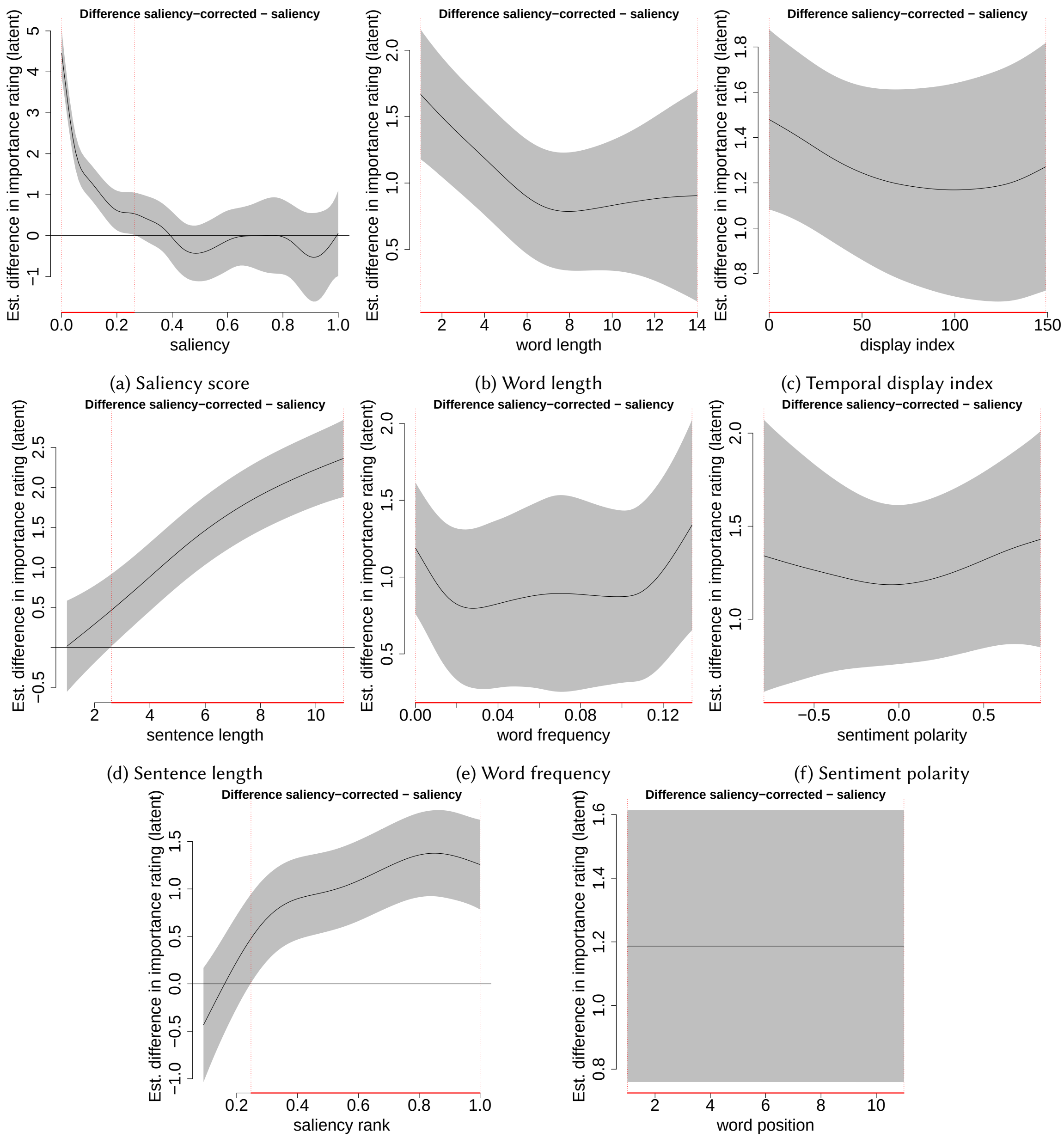


(g) Saliency rank



(h) Word position

Fig. 12. Difference plots between the bar visualization and the original visualization. Areas of significant differences are marked red.



(a) Saliency score
(b) Word length
(c) Temporal display index
(d) Sentence length
(e) Word frequency
(f) Sentiment polarity
(g) Saliency rank
(h) Word position

Fig. 13. Difference plots between the model-corrected saliencies and original saliencies. Areas of significant differences are marked red.