

# Comparing zero-shot self-explanations with human rationales in multilingual text classification

Stephanie Brandl\*

Department of Computer Science  
University of Copenhagen  
brandl@di.ku.dk

Oliver Eberle\*

Machine Learning Group  
Technische Universität Berlin  
oliver.eberle@tu-berlin.de

## Abstract

Instruction-tuned LLMs are able to provide *an explanation* about their output to users by generating self-explanations that do not require gradient computations or the application of possibly complex XAI methods. In this paper, we analyse whether this ability results in a *good explanation* by evaluating self-explanations in the form of input rationales with respect to their plausibility to humans as well as their faithfulness to models. For this, we apply two text classification tasks: sentiment classification and forced labour detection. Next to English, we further include Danish and Italian translations of the sentiment classification task and compare self-explanations to human annotations for all samples. To allow for direct comparisons, we also compute post-hoc feature attribution, i.e., layer-wise relevance propagation (LRP) and apply this pipeline to 4 LLMs (Llama2, Llama3, Mistral and Mixtral). Our results show that self-explanations align more closely with human annotations compared to LRP, while maintaining a comparable level of faithfulness.

## 1 Introduction

Providing model explanations in order to increase trust and transparency into their decision making has been a key motivation for the field of Explainable AI (XAI) even before LLMs have found their way into everyday life. Nowadays, LLMs are being used for tasks like creative writing, help with homework but also for advice giving and translation while providing self-generated explanations at the same time.<sup>1</sup> This makes it all the more crucial to understand the quality of those self-explanations, how reliable they are and whether their faithfulness to the model and their plausibility to humans compare with other, widely investigated, post-hoc XAI methods. In this paper, we evaluate self-explanations

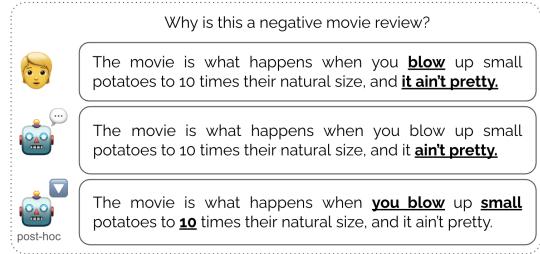


Figure 1: An example from the SST sentiment classification dataset. With rationale annotations by humans, generated by *Llama3* and computed post-hoc with LRP.

from two text classification tasks for which human rationale annotations are available: sentiment classification and forced labour detection. We instruct 4 different LLMs (Mistral, Mixtral, Llama2 and Llama3) to solve the respective task and generate explanations based on the input text in a zero-shot experiment. We compare these with rationales provided on the same samples from various annotation studies and with state-of-the-art post-hoc explanations for Transformers, calculated based on the layer-wise relevance propagation (LRP) framework (Ali et al., 2022) for each model respectively. For sentiment classification, we consider two different subsets from two different annotation studies, one also including Italian and Danish translations next to the English original. We evaluate on plausibility to the human annotations and faithfulness to the model decision, two established evaluation methods in the XAI literature (DeYoung et al., 2020; Jacovi and Goldberg, 2020). We further carry out qualitative analyses, for instance on the distribution of POS tags among the most frequently selected tokens, the differences between languages and the ability to follow precise instructions. We see this controlled study as a first step into a deeper understanding of the reliability and quality of self-explanations and we believe that by covering four language models, three languages and two very different domains of text classification tasks, with different levels of difficulty and text length we can

\* Equal contribution.

<sup>1</sup>[www.washingtonpost.com/technology/2024/08/04/chatgpt-use-real-ai-chatbot-conversations](http://www.washingtonpost.com/technology/2024/08/04/chatgpt-use-real-ai-chatbot-conversations)

provide valuable insights. We release our code on GitHub to enable reproducibility and potential future research.

## 2 Related Work

Generated self-explanations come with both new opportunities and challenges. Prior work in self-explanations for text has focused on new evaluation strategies and model improvements. [Ye and Durrett \(2022\)](#) evaluate whether including self-explanations can improve model performance on in-context learning while [Madsen et al. \(2024\)](#) proposed several instruction-based self-consistency checks to measure faithfulness in generated explanations.

Another line of work by [Wiegreffe et al. \(2022\)](#) seeks to improve free text self-explanations with the help of human-written explanations that are included in the instruction. Similarly to [Kunz and Kuhlmann \(2024\)](#), self-explanations are evaluated on a variety of properties by the means of human annotation. They are found to be generally true, grammatical and factual ([Wiegreffe et al., 2022](#)) and further selective, to contain illustrative examples and rarely subjective according to [Kunz and Kuhlmann](#). For those two papers GPT-3 on CommonsenseQA/NLI and GPT-4 on the Alpaca dataset were analysed, respectively. In our study, we consider human rationales as the ground truth for explaining a decision, against which we compare model self-explanations and post-hoc attributions.

Recent work by [Huang et al. \(2023\)](#) investigates self-explanations by ChatGPT on sentiment classification for SST, comparing faithfulness of self-explanations against different features attribution methods. They experiment with different settings by swapping the order of classification and explanations within a single instruction prompt, asking the model for top-k rationale tokens or continuous token scores, but find no method that stands out in faithfulness while observing significant disagreement across explainability approaches.

Our work focuses on a direct comparison of plausibility and faithfulness, using binary rationales and comparing them to post-hoc LRP attributions, which have been shown to faithfully reflect LLM predictions ([Ali et al., 2022; Achitbat et al., 2024](#)). Additionally, we extend our analysis to a more complex text classification task—forced labor detection—and to multilingual settings, further broadening the scope and applicability of our study.

## 3 Experimental Setup

### 3.1 Datasets

We selected two text classification datasets for sentiment analysis and forced labor detection, for which human rationale annotations have been collected. With those two dataset we cover different aspects and levels of difficulty in both classification and rationale annotation. SST has been widely used for binary sentiment classification, with rationales available in English, Italian and Danish subsets. Texts are rather short and language models have been shown to solve this task successfully, while the second dataset of longer news articles on forced labour detection is more challenging for both classification and rationale extraction, and is also less likely to have been part of the models’ pre-training.

**SST** We use two different subsets from the Stanford Sentiment Treebank (SST2, [Socher et al. 2013](#)) for binary sentiment classification on movie reviews. The first subset contains 263 samples from the validation and test split from SST2 with an average sentence length of 18 tokens. Human rationale annotations have been published for that subset by [Thorn Jakobsen et al. \(2023\)](#) where each sample has been annotated by multiple annotators, 8 on average, who were recruited via Prolific. Annotators were first asked to classify the sample into one of three classes: *positive*, *neutral* or *negative* where none of the sentences was assigned *neutral* as a gold label. In a second step, annotators should choose the parts of the input that support their label choice. We select the rationale annotations with the correct labels from the first step for further analysis. We averaged the binary rationales across all annotators (with correct label classification) and set a threshold of 0.5 (after averaging) for the token selection. We additionally analyse the rationale annotations collected by [Jørgensen et al. \(2022\)](#) on a subset of 250 samples from the validation set of SST2. All samples were translated into Danish and Italian with an average sentence length of 15-17. Rationale annotation was carried out by 2 annotators per language (including English), who were native speakers with linguistic training. In contrast to the annotations collected by [Thorn Jakobsen et al.](#), the correct sentiment (*positive* or *negative*) was provided and the annotators were asked to select parts of the input that supported the gold label.

**RaFoLa** The authors of [Mendez Guzman et al. \(2022\)](#) published a **Rationale**-annotated corpus for

**Forced-Labour detection.** This multi-class and multi-label dataset contains 989 English news articles that have been labelled and annotated according to 11 risk indicators defined by the International Labour Organization. Rationale annotations were carried out by two annotators who selected parts of the input to justify their label decision if they found evidence for any of the 11 labels. A subset of 100 articles was annotated by both annotators with a label agreement of 0.81 (micro F1) and a rationale agreement of 0.73 (intersection-over-union). The remaining articles were only annotated by one of the annotators. Each news article was assigned 1.2 labels on average while 43% were assigned with at least one label. For our analysis, we selected the 4 most frequent classes with occurrences between 117-256 out of the 989 articles. As we carry out zero-shot experiments on models that have not been fine-tuned on this task, we further convert this task into a binary classification task where we ask for a specific label once at a time. We are providing the definition of the respective forced labour indicator as part of the instruction, see Figure 9 & 10.

### 3.2 Rationale Extraction

For our experiments, we apply the same pipeline to the following 4 instruction fine-tuned LLMs: Llama2-13B,<sup>2</sup> Llama3.1-8B<sup>3</sup>, Mistral-7B<sup>4</sup> and quantized Mixtral-8x7B<sup>5</sup>.

In a first step, we ask the model to classify the given text into positive/negative for SST and into yes/no depending on evidence for a specific risk indicator for the RaFoLa dataset. If the model manages to generate the correct answer, we ask it to generate rationales based on the relevant provided context of the input. In case of RaFoLa, we follow the original data collection and only request rationales if the respective risk indicator is present.

For the subsets in Italian and Danish, we have manually translated the prompts to the respective language with the help of native speakers.

### 3.3 LRP Relevances

To extract input attribution scores, we use layer-wise relevance propagation (LRP); a widely used and state-of-the-art XAI method to compute feature attributions (Ali et al., 2022; Achitbat et al., 2024). Following the proposed propagation rules for Trans-

former models, we compute relevance scores for Llama and Mistral by backpropagating the logit for the correctly generated class token. While self-explanations and human annotations provide binary rationales, LRP assigns continuous scores to each token. To allow direct comparison, we used the number of tokens annotated by humans as the cut-off for LRP relevance scores in both tasks. Tokens were ranked by their scores, and the top- $k$  tokens—where  $k$  is the number of rationales from the human annotation—were selected for further analysis.

### 3.4 Constraining Self-Explanations

Initial experiments showed that without precise instructions, the model would return 80% of the input tokens as rationales for SST where humans had annotated approximately 30%. This made comparisons difficult, so we chose to request a maximum number of tokens based on the number of annotated tokens by humans for each sample. Language models did not always follow this request but we could reduce the ratio of tokens to a comparable level with human annotations. For RaFoLa, this problem did not occur on the same level since the input texts were much longer and humans annotated entire phrases. We thus decided not to include an upper bound for the RaFoLa rationales. We will discuss ratios and instruction following in more details.

## 4 Results

We first show and discuss the main results of pair-wise agreement between the different kinds of rationales, i.e., plausibility scores, before further analysing their faithfulness, the differences between languages and a qualitative analysis.

### 4.1 Plausibility

We show pair-wise comparisons between human-annotated, model-generated and post-hoc computed rationales by calculating sample-wise Cohen’s Kappa scores between the binary scores and averaging across samples for different models. When considering human annotations as the ground truth, this is usually referred to as *plausibility* (DeYoung et al., 2020). We here also show the comparison between post-hoc and self-generated rationales, i.e., self-explanations.

Cohen’s Kappa (Cohen, 1960) is a well-established method to measure inter-annotator agreement (IAA) between two annotators, in our case those are either the averaged human annotations or the two different types of model rationales

<sup>2</sup>[meta-llama/Llama-2-13b-chat-hf](https://meta-llama/Llama-2-13b-chat-hf)

<sup>3</sup>[meta-llama/Meta-Llama-3.1-8B-Instruct](https://meta-llama/Meta-Llama-3.1-8B-Instruct)

<sup>4</sup>[mistralai/Mistral-7B-Instruct-v0.3](https://mistralai/Mistral-7B-Instruct-v0.3)

<sup>5</sup>[mistralai/Mixtral-8x7B-Instruct-v0.1](https://mistralai/Mixtral-8x7B-Instruct-v0.1)

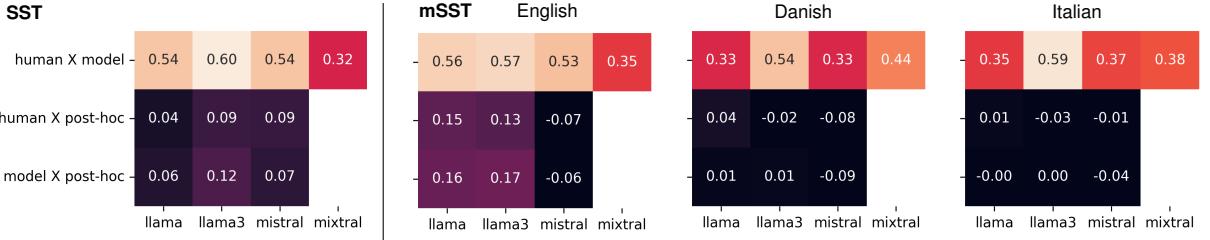


Figure 2: Pair-wise comparison scores (Cohen’s Kappa) between rationales on SST and multilingual SST (English, Danish and Italian). We compare human-annotated, model-generated, and post-hoc LRP rationales.

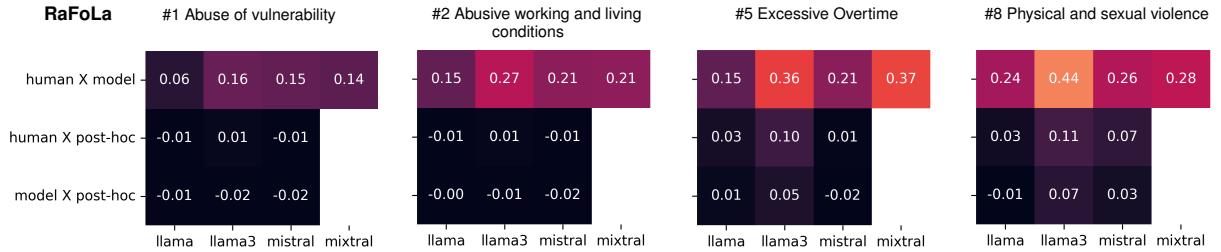


Figure 3: Pair-wise scores (Cohen’s Kappa) between rationales on RaFoLa for 4 different forced labour indicators.

(generated and post-hoc). We chose Kappa over F1 scores, which is also often used to evaluate IAA but comes with two obstacles. It is (i) driven by the imbalance of classes (here selected and not selected tokens) leading to a higher offset for annotations with a ratio of selected tokens closer to 0.5 and it (ii) does not consider randomness as a confounding factor. Both of these obstacles are taken into account when computing Cohen’s Kappa scores, leading to overall lower but more robust scores in comparison to F1. We report F1 scores in the Appendix in Section B.

Results for SST averaged across 3 seeds are shown in Figure 2. For both English subsets, we see moderate level of agreement ( $> 0.4$ ) for the comparison between human annotation and self-explanations (*human × model*) in the range  $0.53 – 0.6$  except for *Mixtral* where agreement only reaches  $0.32 – 0.35$ .<sup>6</sup> For both comparisons between post-hoc rationales and human annotations/self-explanations, we see no agreement or slight agreement in some cases, i.e., both *Ilama* models for mSST English show scores  $> 0.13$ . For Danish and Italian, we see a fair level of agreement for the *human × model* comparison, only *Mixtral* for Danish (0.44) and *Ilama3* on both languages (0.54 – 0.59) show a moderate level of agreement. For both languages, the comparisons with post-hoc show scores around 0, i.e., no clear effects of agree-

ment can been measured. Overall we see highest scores for *Ilama3* in comparison to other models, in particular for the *human × model* comparisons, where scores for all SST subsets are  $> 0.5$ .

**Thorn Jakobsen et al.** who first published the SST human annotations (most left subplot) report plausibility F1 scores below 0.4 between post-hoc explanations for different models and humans and F1 inter-annotator scores in the range of  $0.5 – 0.6$  between different demographics groups. Our results, (see F1-scores in Figure 11) thus confirm the agreement between post-hoc rationales and humans and exceed the human-human agreement when comparing human annotations with self-explanations (*human × model*) for all models except *Mixtral*.

Results for RaFoLa, averaged across 3 seeds, are shown in Figure 3. We here see overall lower levels of agreement for the *human × model* comparisons, reaching from only slight levels of agreement for indicator #1 (0.06 – 0.16), to a fair level of agreement for indicator #2 (0.15 – 0.27) and #5 (0.15 – 0.37) and moderate agreements in indicator #8 (0.24 – 0.44). Similar to SST, we see highest agreements for *Ilama3* followed by *Mixtral*. Comparisons with post-hoc rationales show scores around 0, indicating low agreement with human rationales.

## 4.2 Task accuracy and rationale ratios

Table 1 presents model accuracies and ratios of identified rationale tokens for SST, multilingual

<sup>6</sup>We follow Landis (1977) when classifying the levels of agreement.

SST (mSST), and RaFoLa datasets in humans and models. Accuracies for SST and mSST are generally high across models, while RaFoLa shows more variation, particularly lower accuracies and rationale ratios. The ratio of identified rationale tokens by humans tends to be higher for the shorter movie reviews in SST and mSST, and lower for the longer paragraphs in RaFoLa, suggesting that the human annotators have more sparsely identified relevant phrases when presented with longer texts. This can be attributed to the differences in samples across the datasets. While movie reviews in SST tend to be short and were written with the purpose of expressing a sentiment, the articles in RaFoLa were more descriptive of a specific situation or incident and might or might not include the violation of one (or more) of the 11 risk indicators. We further can assume that most models nowadays have seen the original English version of SST during training and are thus more familiar with this type of data.

Table 1: Model accuracies for SST, multilingual SST and RaFoLa, alongside the ratio of identified rationale tokens by humans and those of self-generated rationales across models.

	acc				ratio human	ratio model			
	llama	llama3	mistral	mixtral		llama	llama3	mistral	mixtral
SST	0.88	0.98	0.98	0.98	0.29	0.22	0.25	0.32	0.14
mSST (EN)	0.85	0.98	0.98	1.00	0.33	0.23	0.27	0.32	0.15
mSST (DK)	0.84	0.87	0.97	0.97	0.38	0.18	0.29	0.40	0.21
mSST (IT)	0.88	0.95	0.97	0.99	0.37	0.26	0.31	0.39	0.25
RaFoLA #1	0.50	0.47	0.65	0.39	0.05	0.05	0.23	0.12	0.16
RaFoLA #2	0.58	0.58	0.58	0.48	0.06	0.04	0.21	0.11	0.12
RaFoLA #5	0.88	0.92	0.89	0.81	0.04	0.02	0.08	0.06	0.05
RaFoLA #8	0.90	0.90	0.90	0.90	0.07	0.04	0.12	0.08	0.10

### 4.3 Faithfulness

Besides plausibility, faithfulness is the most commonly used evaluation approach to judge the quality of model explanations. Especially for feature attribution approaches, removal of most relevant features has been used to assess how faithful a feature subset is with respect to the model prediction, i.e. if removing a highly relevant subset will lead to a strong decrease of the prediction. We evaluated faithfulness by measuring the change in probability after masking the tokens as identified by the different rationales (human, self-generated and post-hoc).

Compared to human and self-generated rationales, post-hoc attributions provide a relevance score for each token in the input prompt, requiring to binarize post-hoc attributions to allow for direct comparison as described in Section 3.3. Addition-

ally, we included a baseline that randomly removes as many token as identified by humans. Our results for SST and RaFoLa are summarized in Figure 4. Average initial probability for the correct answer token is given as a dashed line for each model.

We find that levels of faithfulness for humans, generated, and post-hoc explanations are comparable, with human explanations being overall as faithful as those generated by models or post-hoc methods. For some cases, e.g. class #5 for RaFoLA, human rationales provide more faithful feature attributions than mistral explanations.

Self-generated model explanations can be more faithful than post-hoc ones, and the reverse case can also be true, depending on the task and model. Overall, they provide similarly faithful model rationales. We further investigated the limited impact of removing features on the class token probability in the post-hoc setting and found that the most relevant tokens are typically part of the provided task instruction, such as the class definition or question. Contrastive explanations have been proposed to provide more task-specific attributions (Yin and Neubig, 2022) and we provide an additional analysis of the plausibility of post-hoc explanations in Section 4.8.

### 4.4 Languages

From the set of models we consider for this study, all have been pre-trained on English data while *Llama3* and *Mixtral* also have been pre-trained on Italian, but none of the model officially supports Danish. Considering this difference in language exposure, our results show that all models are able to solve the sentiment classification task (see Table 1) and are also able to extract meaningful rationales with plausibility on the same (or even superior) level than English. Plausibility scores between self-explanations and humans are ranging 0.46 – 0.74 for Danish and Italian versus 0.4 – 0.69 for English. *Llama2* shows the biggest gap between English (0.67) and Danish/Italian (0.46/0.52) while *Mixtral* shows an increase in plausibility from English (0.4) to (0.57) for Danish. *Llama3* shows high and stable plausibility scores across all languages ( $\sim 0.7$ ). It seems all models can handle both Danish and Italian, whether this is based on data contamination during pre-training, e.g., training data for *Llama2* has been reported to contain 0.11% Italian (Touvron et al., 2023), is unclear. Previous work has focused on zero-shot abilities in unseen languages but has mostly followed ap-

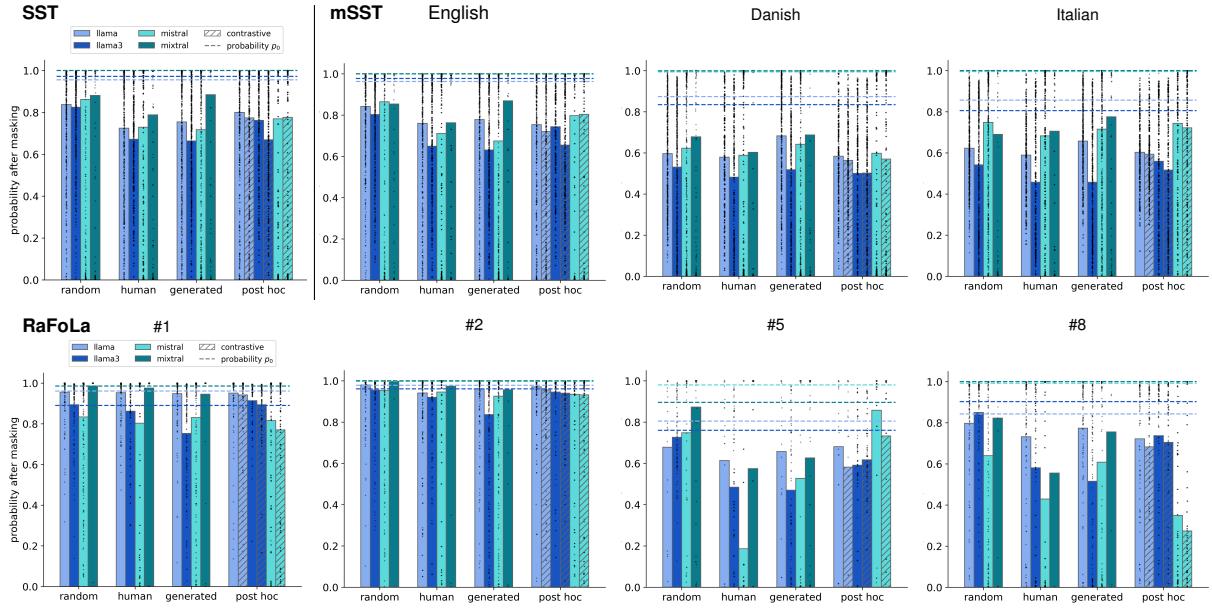


Figure 4: Faithfulness evaluation for SST and mSST (top row) and RaFoLa (bottom row). The probability after masking the tokens extracted from human rationale annotations, self-generated model rationales and post-hoc model attributions is compared across models. Lower probability indicates more faithful identification of rationales.

proaches with either multilingual models or translating test sets into English (Ebrahimi and von der Wense, 2024; Artetxe et al., 2023). To the best of our knowledge, there has not been any published work in a truly zero-shot fashion, we will leave that for future work.

Table 2: Most frequent tokens in the RaFoLa corpus(first row) and in rationales identified by human annotators, as well as self-generated and post-hoc explanations of Llama3 and Mistral for #1 and #8.

	#1 Abuse of vulnerability	#8 Physical and sexual violence
corpus	said, workers, labour, human, work, forced, rights, slavery	
human	workers, work, forced, women, children, labour, said, exploitation	sexual, abuse, harassment, women, violence, said, verbal, physical
llama3	workers, work, said, labour, forced, women, children, working	said, women, sexual, abuse, harassment, workers, violence, physical
llama3-post-hoc	labour, said, slavery, vulnerable, workers, according, trafficking, forced	violence, said, harassment, abuse, report, based, sexual, physical
mistral	workers, work, forced, labour, children, women, working, conditions	sexual, women, harassment, said, abuse, physical, children, workers
mistral-post-hoc	said, trafficking, forced, labour, slavery, abuse, workers, 2020	sexual, violence, said, abuse, harassment, based, report, women

## 4.5 Selected tokens

Plausibility scores between humans and self-explanations (human  $\times$  model) for RaFoLa vary by a magnitude of 2-3 between the different indicators, for instance *Llama3* shows an agreement of 0.16 for #1 *Abuse of vulnerability* and 0.44 for #8

*Physical and sexual violence* according to Figure 3. When considering the ratios of tokens selected, we also see clear differences between the indicators. *Llama3* selected 0.23 of all tokens for #1 and only 0.08 and 0.12 for #5 and #8, respectively, while the human annotators selected in the range of 0.02 – 0.05 across the different classes (indicators), see Table 1. We extracted the most frequent tokens for indicators #1 and #8, which show the lowest and highest agreement between human annotators and self-explanations, and present them in Table 2. When comparing the most frequent tokens as selected by human annotators and different models, we see that they differ more from the overall most frequent tokens in the corpus (first row) for #8 than for #1. It seems that keywords for #8 *Physical and sexual violence*, like *sexual* and *women* are easier to detect than keywords for #1 *Abuse of vulnerability*, which might contain a less clear definition for an untrained model. We present all instructions with class definitions in Appendix A.2. We further observe more overlap in tokens selected by models directly (self-explanations) and human annotations than with post-hoc rationales.

## 4.6 POS analysis

In order to gain a deeper insight into the differences between rationale selection, we are extracting part-of-speech (POS) tags of each selected tokens. We show results for the monolingual SST dataset in Figure 5. We see that for humans and self-

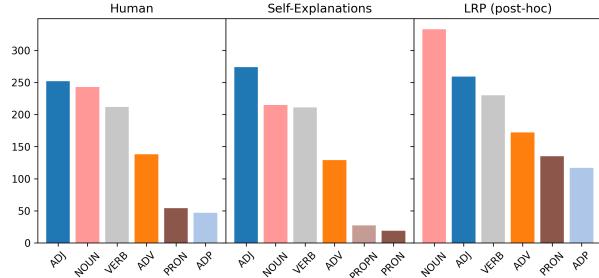


Figure 5: Top-6 POS tags of extracted rationales from SST monolingual for the human annotated rationales (left), self-explanations by *llama3* (center) and post-hoc explanations based on *llama3*.

	llama2	llama3	mistral	mixtral
#tokens	0.11	0.04	0.34	0.16
json-syntax	0.86	0.02	0.05	0.37

Table 3: Ratio of mismatches when prompting models for a max. number of tokens and correct json syntax. Averaged scores for SST across all subsets & languages.

explanations *adjectives* are most common, closely followed by *nouns*, *verbs* and with a bigger gap *adverbs*. For the post-hoc rationales, *nouns* were most often selected followed by *verbs* and *adjectives*. This means, the top-3 POS tags are the same across the different selection methods but differ in ranking between post-hoc explanations and the others. We see similar effects across models and languages, i.e., there is higher agreement in POS rankings between humans and self-explanations in comparison to post-hoc rationales. There are a few variations in ranking across languages, which has been seen in eye-tracking fixation times and attention scores before (Brandl and Hollenstein, 2022).

#### 4.7 Instruction following

For generating and processing the self-explanations, we instructed the models to return rationales in a *json* format. Since many of those outputs resulted in *SyntaxErrors*, we included a syntax check based on *llama3* where we instructed the model in a separate step to correct the json syntax in case such an error occurred. The ability to return correct syntax varied across models. We also saw differences when following the instruction of returning the correct number of rationales for which we set an upper bound for SST based on the human annotations. We show results for SST for all 4 models, averaged across subsets and languages in Table 3. The results show that *llama2* has a lot of difficulties with respect to json syntax with syntax errors occurring in 86% of the

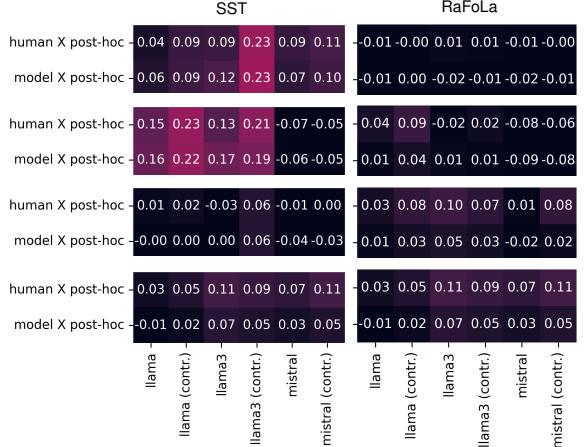


Figure 6: Comparing plausibility scores for non-contrastive and contrastive post-hoc approaches using Kappa agreement scores. Left: SST and multilingual SST for English, Danish and Italian (top to bottom rows). Right: RaFoLa for classes #1, #2, #5, #8 (cf. Figure 3).

case. Both *llama3* and *Mistral* have a low error rate with 2% and 5% respectively. At the same time, *Mistral* returns more than the maximum number of requested rationale tokens in 1 out of 3 instructions where *llama3* follows the instruction in most cases. Analysing and evaluating the ability to follow instructions has previously been discussed in Qin et al. (2024); Zeng et al. (2024).

#### 4.8 Contrastive LRP

In a complementary analysis, we test the alignment of model rationales with post-hoc attributions, examining whether there is a difference in the plausibility of contrastive and non-contrastive post-hoc explanations. Prior work has suggested that contrastive explanations are more aligned with human reasoning and are thus considered more valuable for humans to understand the model’s decisions (Lipton, 1990; Miller, 2019; Jacovi et al., 2021).

Comparing Cohen’s Kappa scores shown in Figure 6 suggests that contrastive post-hoc approaches do not generally result in higher plausibility than non-contrastive ones. Similarly, contrastive explanations overall exhibit a similar level of faithfulness as non-contrastive explanations (cf. Figure 4). Although contrastive explanations can be more faithful and plausible in certain cases, such as SST and mSST (English) for *llama3*, there is no consistent difference between the two approaches. This aligns with previously reported high correlation between them (Eberle et al., 2023) and the minor observed performance differences for generating the correct label based on contrastive or non-contrastive post-hoc approaches (Krishna et al., 2023).

## 5 Discussion

In this paper, we investigated explanations generated by 4 different instruction-tuned LLMs for 2 text classification tasks in English but also in Italian and Danish. Those generated explanations, i.e., self-explanations, were limited to the input tokens of the respective text samples. We compared those self-explanations on the tasks of sentiment classification (SST) and forced labour classification (RaFoLa) with human annotations for the same samples and post-hoc feature attribution with layer-wise relevance propagation (LRP). Pairwise comparison between the three different types of explanations (humans, generated, post-hoc) shows that human annotations and generated explanations agree on a much higher level than post-hoc with any of them. We further saw that *Llama3* has shown the highest level of agreement across both tasks and all languages and followed the instructions more closely in comparison to the other models. Our POS analysis confirmed this finding and showed higher agreement with the “type” of tokens selected by humans compared to post-hoc rationales.

Besides established post-hoc attribution approaches, the ability of language models to provide self-generated explanations, i.e., self-explanations, has offered a direct and human-understandable communication between user and model. This not only enhances usability in particular for lay people but, as presented here, also results in explanations that are similarly faithful and align more closely with human rationales compared to binarized post-hoc attributions. Herein, the generation process behind self-explanations remains obfuscated and may suffer from counterfactualty, enabling the model to give untruthful explanations for correct predictions (Ji et al., 2023). We further find that current language models require careful instructions to provide useful self-generated explanations.

While human plausibility is a desired property of predictions and explanations made by machine learning systems, LLMs are capable of identifying alternative solutions to task-solving that may not be intuitive to humans. Good explanations should thus also highlight the learned prediction strategy that is faithful to the one used by the model—even if it is not directly plausible (Agarwal et al., 2024). In contrast to post-hoc attributions, which are directly derived from the prediction score (e.g., using explanatory gradients as for the LRP attributions considered here), the self-generation process is not

explicitly tied to the prediction in a clear manner.

Attribution approaches herein consider the entire input to rank relevant features, including both the provided context and all instructions. Instruction templates are crucial for understanding and solving tasks, and we find that tokens related to task instructions often play a more significant role than the context itself. While this effectively identifies relevant task features from the perspective of the model, users are typically more interested in understanding which parts of the provided evidence are most relevant. Disentangling task instructions from contextual evidence is essential for aligning insights with user expectations and should inform the development of future interpretability methods.

Our study represents a first step toward understanding and building more intuitive model explanations by directly comparing human annotations with those generated by models. Evaluating free-text explanations for factuality, usability, and faithfulness is crucial for ensuring their practical and intuitive application, especially given the growing use of increasingly complex LLMs by lay people who may not fully understand their mechanisms

We currently do not know why self-explanations align more closely with human annotations than post-hoc attributions. While training procedures such as reinforcement learning from human feedback (Ziegler et al., 2019) may incentivize more human-like explanations (Agarwal et al., 2024), limited access to models, procedures, and datasets, restricts detailed analysis.

The ability of the models to solve the sentiment classification task in unseen languages like Danish is remarkable. Not only is the model able to understand the Danish prompt, it also is able to return specific input tokens that are aligned with the human selection. Future work could investigate whether this zero-shot performance is specific to certain language families and tasks or if it extends to more complex instructions and datasets.

**Conclusion** As language models have advanced and become widely adopted, understanding the mechanisms behind their predictions has emerged as a critical challenge. Using human rationale annotations across three datasets, we have investigated the human plausibility and faithfulness of self-generated explanations to post-hoc feature attributions. We find that self-generated explanations agree more closely with human rationales than post-hoc feature attribution, while remaining faithful to the model prediction.

## Limitations

We acknowledge that annotations may be affected by annotator bias, varying guidelines, and differing expertise, impacting the consistency of rationales. Also the number of annotators and the level of details in the instructions varied across the annotation studies we have considered for this paper. Furthermore, for the forced labour detection, annotations by legal scholars might differ from the ones provided and would also be interesting to compare with model rationales.

We focus our study on rationales based on the input while free text explanations might provide more useful information and pose the more realistic scenario.

While agreement between human and model rationales may be desired, it has been shown in previous work, that humans do not necessarily prefer human-written explanations in comparison to the ones generated by LLMs in the case of free text explanations (Wiegreffe et al., 2022).

The high zero-shot performance, especially with SST, may be an effect of data contamination, which is likely part of the training data. We can further not exclude the possibility that rationales or task explanations have been included in the training corpus.

## Acknowledgments

We thank our colleagues at the CoAStaL NLP group for constructive feedback in the beginning of the project. In particular, we would like to thank Alice Schiavone and Anders Søgaard for helping us with the prompt translations.

SB received funding by the European Union under the Grant Agreement no. 10106555, FairER. OE is funded by the German Ministry for Education and Research (under refs 01IS18056A and 01IS18025A) and BIFOLD. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor REA can be held responsible for them.

## References

- Reduan Achitbat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 135–168. PMLR.

- study of llm-generated self-explanations. *Preprint*, arXiv:2310.11207.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.
- Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).
- Rasmus Jørgensen, Fiammetta Caccavale, Christian Igel, and Anders Søgaard. 2022. Are multilingual sentiment models equally right for the right reasons? In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 131–141, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Satyapriya Krishna, Jiaqi Ma, Dylan Z Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. Post hoc explanations of language models can improve language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Jenny Kunz and Marco Kuhlmann. 2024. Properties and challenges of LLM-generated explanations. In *Proceedings of the Third Workshop on Bridging Human-Computer Interaction and Natural Language Processing*, pages 13–27, Mexico City, Mexico. Association for Computational Linguistics.
- JR Landis. 1977. The measurement of observer agreement for categorical data. *Biometrics*.
- Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266.
- Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? *Preprint*, arXiv:2401.07927.
- Erick Mendez Guzman, Viktor Schlegel, and Riza Batista-Navarro. 2022. RaFoLa: A rationale-annotated corpus for detecting indicators of forced labour. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3610–3625, Marseille, France. European Language Resources Association.
- Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.
- Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13025–13048, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Terne Sasha Thorn Jakobsen, Laura Cabello, and Anders Søgaard. 2023. Being right for whose right reasons? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1033–1054, Toronto, Canada. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Bin Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.
- Sarah Wiegrefe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.
- Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.
- Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Pro-*

*ceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

## A Instructions

### A.1 SST



Consider the following movie review: [...]



Choose one of the following options and start your answer with the respective letter:



(a) This review is overall rather negative  
(b) This review is overall rather positive

Answer:



Considera la seguente recensione di un film: [...]



Scegli una delle seguenti opzioni e inizia la tua risposta con la rispettiva lettera:



(a) Questa recensione è nel complesso piuttosto positiva  
(b) Questa recensione è nel complesso piuttosto negativa

Risposta:



Overvej følgende filmanmeldelse: [...]



Vælg en af følgende muligheder, og start dit svar med det pågældende bogstav:



(a) Denne anmeldelse er generelt ret positiv  
(b) Denne anmeldelse er generelt ret negativ

Svar:

Figure 7: Prompts in all 3 languages to solve sentiment classification.



Return a maximum of x individual words from the text that show evidence for [...]. Format your output as a list of json entries with each entry having one field 'rationales' with the relevant word from the text. Output only the list and no other text than a maximum of x individual words of the input text, separate different words by a ".



Restituisci un massimo di x singole parole del testo che motivano la scelta di [...]. Formatta il tuo output come una lista di elementi json per cui ogni elemento contiene il campo 'rationales' con la parola pertinente del testo. Restituisci solo la lista e nessun altro testo oltre a un massimo di x singole parole del testo in ingresso, separando le diverse parole con un ".



Returner maksimalt x individuelle ord fra teksten, der viser evidens for [...]. Formater dit output som en liste af json-poster, hvor hver post har et felt 'rationales' med det relevante ord fra teksten. Output kun listen og ingen anden tekst end maksimalt x individuelle ord i inputteksten, adskil forskellige ord med et ".

Figure 8: Follow-up prompts in all 3 languages to extract rationales.

## A.2 RaFoLa

Consider the following definition of `{article_label}`: `{article_definition}`

Does the following text show any evidence for `{article_label}`?

Choose one of the following options and start your answer with the respective letter:

(a) Yes, this text contains evidence for `{article_label}`

(b) No, this text does not contain any evidence for `{article_label}`

Answer:

Return the parts of the text that show evidence for [...]. Format your output as a list of json entries with each entry having one field 'rationales' with the relevant part of the text. Output only the list and no other text than parts of the input text, separate different parts by a ":".

Figure 9: Prompts for classification and rationale extraction for the RaFoLa dataset.

**#1 Abuse of vulnerability:**  
Referring to people who lack knowledge of the local language or laws, have few livelihood options, belong to a minority religious or ethnic group, have a disability or have other characteristics that set them apart from the majority population.

**#2 Abusive working and living conditions:**  
Forced labour victims may endure living and working in conditions that workers would never freely accept. Work may be performed under conditions that are degrading or hazardous and in severe breach of labour law.

**#5 Excessive overtime**  
Referring to the obligation of working excessive hours or days beyond the limits prescribed by national law or collective agreement.

**#8 Physical and sexual violence**  
Violence can include forcing workers to take drugs or alcohol to have greater control over them. Violence can also be used to force a worker to undertake tasks that were not part of the initial agreement.

Figure 10: Indicators defined by the International Labour Organization and published by [Mendez Guzman et al.](#).

## B F1-Plausibility scores

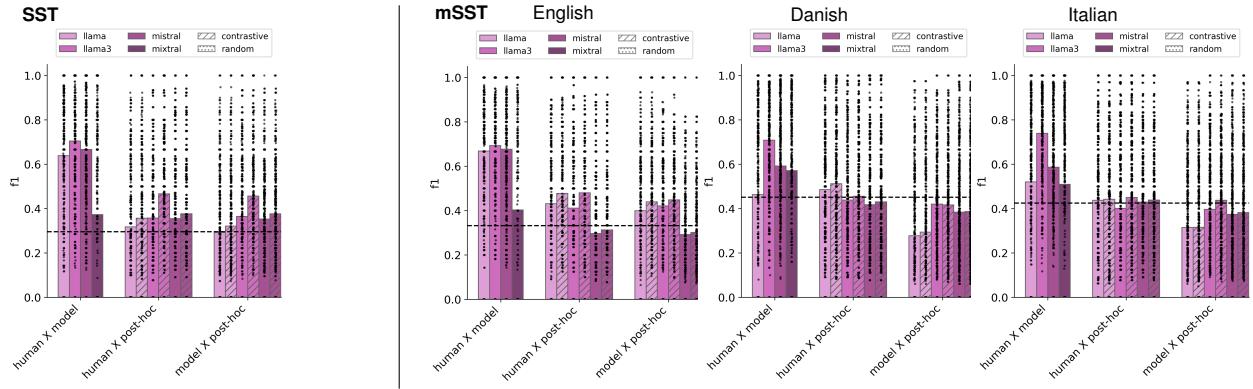


Figure 11: Pair-wise F1 comparison scores between rationales on SST and multilingual SST (English, Danish and Italian). We compare rationales annotated by humans, generated by models, and computed post-hoc with LRP.

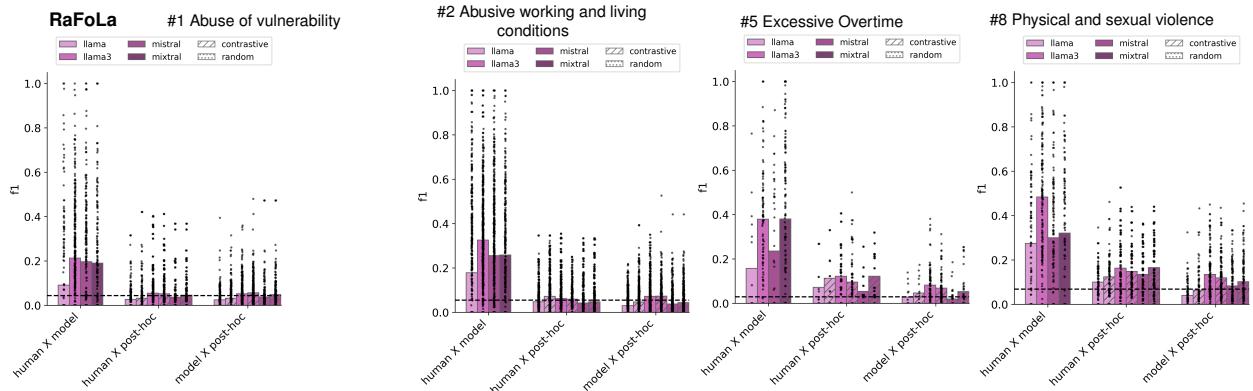


Figure 12: Pair-wise F1 comparison scores between rationales on RaFoLa.

## C Most frequent rationale tokens

Table 4: List of top-8 most frequent tokens in the RaFoLa corpus (first row) together with the most frequent rationales as identified by human annotators, as well as self-generated and post-hoc explanations.

	#1 Abuse of vulnerability	#2 Abusive working and living cond.	#5 Excessive overtime	#8 Physical and sexual violence
corpus	said, workers, labour, human, work, forced, rights, slavery			
human	workers, work, forced, women, children, labour, said, exploitation	workers, conditions, work, forced, water, little, said, working	hours, work, week, days, long, claimed, worked, like	sexual, abuse, harassment, women, violence, said, verbal, physical
llama2	workers, work, migrant, women, exploitation, forced, labour, children	workers, work, conditions, forced, day, working, children, water	pay, work, ahmad, received, little, breaks, \$, 600	abuse, sexual, women, harassment, retaliation, verbal, physical, advances
llama2 post-hoc	trafficking, world, china, children, said, forced, child, exploitation	trafficking, forced, said, world, children, conditions, labour, covid-19	covid-19, workers, thailand, thai, kingdom, california, hours, basi	said, violence, women, based, walmart, global, guardian, jennifer
llama3	workers, work, said, labour, forced, women, children, working	workers, work, said, conditions, working, labour, forced, day	hours, working, day, work, pay, said, workers, days	said, women, sexual, abuse, harassment, workers, violence, physical
llama3 post-hoc	labour, said, slavery, vulnerable, workers, according, trafficking, forced	said, workers, labour, according, conditions, slavery, work, trafficking	hours, working, work, overtime, day, workers, forced, said	violence, said, harassment, abuse, report, based, sexual, physical
mistral	workers, work, forced, labour, children, women, working, conditions	workers, work, said, forced, conditions, working, labour, children	said, day, mr, work, hours, days, delivery, home	sexual, women, harassment, said, abuse, physical, children, workers
mistral post-hoc	said, trafficking, forced, labour, slavery, abuse, workers, 2020	said, covid-19, 2019, workers, trafficking, labour, forced, world	employer, covid-19, said, years, police, cotton, mr, paying	sexual, violence, said, abuse, harassment, based, report, women
mixtral	workers, work, said, forced, children, labour, women, abuse	workers, work, said, conditions, forced, labour, working, paid	hours, work, day, forced, workers, working, days, week	sexual, women, said, workers, harassment, abuse, violence, reported

Table 5: List of top-8 most frequent tokens in the SST and mSST corpus together with the most frequent rationales as identified by human annotators, as well as self-generated and post-hoc explanations.

	SST	mSST English	mSST Danish	mSST Italian
full corpus	movie, film, like, comedy, -, characters, work, romantic	film, movie, characters, bad, like, performances, funny, story	film, filmen, karakterer, bare, sjov, 'se, præstationer	film, i, divertente, personaggi, interpretazioni, trama, avvincente, commedia
human	funny, movie, beautifully, bad, best, hilarious, stupid, wonderful	bad, performances, funny, good, dull, film, compelling, dumb	film, sjov, overbevisende, plot, vittig, bare, dårligt, præstationer	divertente, avvincente, noioso, interpretazioni, ben, brutto, film, intelligente
llama2	bad, beautifully, best, fun, stupid, wonderful, funny, worst	bad, funny, dull, compelling, witty, good, long, little	sjov, spændende, underholdende, præstationer, bedste, overbevisende, spænding, vittig	divertente, film, avvincente, noioso, interpretazioni, ben, intelligente, senso
llama2 post-hoc	movie, comedy, film, little, like, far, funny, stupid	movie, film, bad, funny, dull, little, dumb, compelling	film, filmen, 'sjov, præstationer, karakterer, dårlig, dårligt	film, divertente, noioso, avvincente, senso, umorismo, brutto, i
llama3	bad, best, beautifully, compelling, funny, love, little, hilarious	funny, bad, performances, dull, compelling, good, little, best	sjov, bedste, overbevisende, humor, dårlig, dårligt, kedelig, spænding	divertente, avvincente, noioso, umorismo, senso, brutto, intelligente, spiritoso
llama3 post-hoc	comedy, like, far, bad, beautifully, best, little, love	comedy, like, good, little, big, high, bad, dull	i, påv, vā, film, sāv, re, spā, sjov	film, divertente, cosa-, avvincente, piā, noioso, interpretazioni, brutto
mistral	bad, funny, comedy, best, beautifully, compelling, love, far	bad, performances, funny, characters, dull, compelling, good, little	sjov, filmen, præstationer, film, overbevisende, sjovt, spænding, humor	film, divertente, personaggi, interpretazioni, avvincente, noioso, trama, brutto
mistral post-hoc	movie, film, bad, -, like, love, feels, year	film, bad, funny, dull, good, dumb, characters, comedy	film, filmen, filmens, sjov, 'præstationer, blanding, spændende	film, divertente, interpretazioni, personaggi, commedia, noioso, cinema, avvincente
mixtral	best, film, laugh, like, year, love, pretty, good	bad, funny, compelling, great, good, witty, performances, intelligent	sjov, sjovt, film, præstationer, overbevisende, humor, spændende, tilfredsstillende	divertente, film, interpretazioni, noioso, umorismo, i, commedia, trama