A Systematic Comparison between Extractive Self-Explanations and Human Rationales in Text Classification

Stephanie Brandl*
Center for Social Data Science
University of Copenhagen
stephanie.brandl@sodas.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, without requiring the application of complex interpretability techniques. In this paper, we analyse whether this ability results in a *good* explanation. We evaluate self-explanations in the form of input rationales with respect to their plausibility to humans. We study three text classification tasks: sentiment classification, forced labour detection and claim verification. We include Danish and Italian translations of the sentiment classification task and compare self-explanations to human annotations. For this, we collected human rationale annotations for Climate-Fever, a claim verification dataset. We furthermore evaluate the faithfulness of human and self-explanation rationales with respect to correct model predictions, and extend the study by incorporating post-hoc attribution-based explanations. We analyse four open-weight LLMs and find that alignment between self-explanations and human rationales highly depends on text length and task complexity. Nevertheless, self-explanations yield faithful subsets of token-level rationales, whereas post-hoc attribution methods tend to emphasize structural and formatting tokens, reflecting fundamentally different explanation strategies.

1 Introduction

Providing model explanations to increase trustworthiness and transparency is a key goal of interpretability research, with LLMs offering new ways to trace model decision-making. As AI-based analyses increasingly have the power to influence decisions and perceptions, systematically tracing and comparing explanations of these decisions is essential for fostering trust, transparency and regulatory compliance. Today, LLMs are being used for a wide variety of tasks, ranging from creative writing and homework assistance to offering advice, summarizing reports and translation, while providing self-generated explanations of their outputs, i.e., self-explanations, in the process.[1] This makes it even more important to understand the quality of those self-explanations, how reliable they are and to evaluate their faithfulness to the model and their plausibility to humans. In this paper, we evaluate self-explanations from two text classification tasks for which human rationale annotations are available: sentiment classification and forced labour detection, i.e., identifying pre-defined risk indicators of forced labour. We also collect and release new rationale annotations by three personally instructed annotators for a subset of Climate-Fever, a claim verification dataset with claims related to climate change. We instruct 4 different LLMs (Gemma3, Llama3, Mistral and Qwen3) to solve the respective tasks and generate rationales based on the input text in a zero-shot experiment. For sentiment classification, we consider two different subsets from two different annotation studies, one also including Italian and Danish translations alongside the English original. Following established evaluation methods in the interpretability and Explainable AI literature (DeYoung et al., 2020; Jacovi and Goldberg, 2020), we assess the plausibility of model rationales by measuring their agreement with human annotations and evaluate faithfulness via interventions on input tokens to determine their importance for the

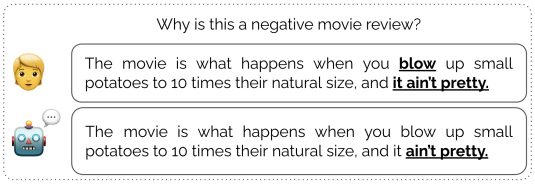

Figure 1: An example from the SST sentiment classification dataset. With rationale annotations by humans and generated by *Llama3*.

* Equal contribution.

[1] www.washingtonpost.com/technology/2024/08/04/chatgpt-use-real-ai-chatbot-conversations

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 563–583

model's decision. Our study takes an initial step towards a better understanding of the reliability and quality of self-explanations, for which we analyzed four language models, three languages, and three distinct text classification domains. Those domains widely vary in text length and task complexity and are more realistic in how they are composed than the datasets in many other explainability studies. While most studies focus on short texts and keyword-based tasks, we also include tasks such as forced labour detection on news articles and claim verification on automatically retrieved text snippets from Wikipedia where neither the task itself nor the rationale annotation is trivial, requiring domain-specific terminology and non-trivial evidence assessment. In order to ensure a fair comparison, we consider state-of-the-art gradient-based attribution methods to compute post-hoc explanations which have not been systematically compared with self-explanations so far. Our findings provide relevant insights for model interpretability and user trust in self-explanations, and we further support reproducibility and future research by openly releasing our code and the collected rationale annotations.[2][3]

Contributions In this work, we

(i) collect human rationales for a subset of ClimateFever, a climate claim verification dataset.

(ii) conduct a *controlled study comparing human annotations with LLM-generated explanations*.

(iii) evaluate *plausibility*, i.e.,the level of agreement between model and human rationales and *faithfulness*, i.e., the relevance of selected rationale tokens for the task (model decision).

(iv) study three different text classification tasks: *sentiment classification*, *forced labour detection* and *claim verification*.

(v) systematically compare human, model, and post-hoc attribution-based rationales, showing self-explanations do yield faithful token-level explanations, whereas post-hoc attributions emphasize structural and formatting tokens, reflecting different explanation strategies.

Extractive vs. abstractive explanations. Extractive explanations such as token-level rationales, have been the standard setting in the interpretability and Explainable AI literature, e.g., in the context of sentence labeling, (closed-book) question answering, or factual recall. In this study, we have chosen to focus on extractive explanations, as the explanation is grounded in the provided input, and thus can be clearly evaluated with respect to human rationale annotations. Relevant evaluation methods for such extractive settings have been established over many years, whereas the evaluation of free-text explanations for faithfulness and plausibility are still being developed (Ye and Durrett, 2022; Wiegreffe et al., 2022; Kunz and Kuhlmann, 2024; Madsen et al., 2024). We see this controlled analysis as a first step into building reliable self-explanations. Abstractive evaluation of generative AI presents a much broader challenge in evaluating generative AI outputs and explanations, see also Ross et al. (2021); Sarti et al. (2023).

2 Related Work

Generated self-explanations present 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 Resck et al. (2024) incorporate human-annotated rationales during model training to improve plausibility of post-hoc explanations in text classification. Other work focuses on instruction-based self-consistency checks to measure faithfulness in different types of generated explanations (Madsen et al., 2024; Parcalabescu and Frank, 2024; Zhao and Iii, 2025).

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. Wang and Atanasova (2025) explore different self-feedback strategies to improve free text explanations and find that extracting rationales from the input to be the best among their strategies.

In our study, we consider human rationales as the ground truth for explaining a decision, against which we compare model self-explanations in order to evaluate how plausible they are.

Recent work by Huang et al. (2023) investigates self-explanations by ChatGPT on sentiment clas-

[2] https://github.com/oeberle/self_explanations_human_rationales

[3] https://huggingface.co/datasets/stephaniebrandl/climate_fever_rationales

sification for SST, comparing faithfulness of self-explanations against different feature 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. Randl et al. (2025) compare extractive self-explanations from three text classification tasks with rather short texts with saliency-based explanations and human rationale annotations. They exclude the more sophisticated gradient-based methods in their work.

Instead, we ground our evaluation in human-annotated rationales to assess plausibility, and use post-hoc LRP attributions as well as GradientxInput specifically designed for Transformer models to measure faithfulness (Ali et al., 2022; Achtibat et al., 2024). This systematic approach defines a clear evaluation setting to effectively assess explanation quality by comparing attribution-based explanations, self-explanations, and human rationales alongside a random baseline. The three text classification tasks we include cover a wider range of difficulty in terms of ambiguity, multilinguality and terminology than previous work.

3 Experimental Setup

3.1 Datasets

We select two text classification datasets for sentiment analysis and forced labour detection, for which human rationale annotations had been collected. We also collect human rationales for a third dataset which contains a claim verification task. With those three datasets 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. We collect new human rationales for a subset of *Climate-Fever*, an English claim verification dataset of 1.535 real-world claims about climate change with 5 evidences each (7.675 in total), automatically retrieved from Wikipedia. Here, claims are often ambiguous and it is sometimes unclear if evidences refer to the claim or a semantically similar topic which makes it challenging to classify both evidences and claims for models and humans.

SST/mSST We use two different subsets from the **S**tanford **S**entiment **T**reebank (SST2, Socher et al. 2013) for binary sentiment classification on movie reviews. The first subset (SST) contains 263 English 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 (mSST). 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 **Ra**tionale-annotated corpus for **Fo**rced-**La**bour detection. This multi-class and multi-label dataset contains 989 English news articles that were labeled and annotated according to 11 risk indicators defined by International Labour Organization (2012). 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 indicators. 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 provide the definition of the respective forced labour indicator as part of the instruction, see Figures 8 & 9. Grounded in internationally defined labor standards, the forced labour risk indicators in this dataset serve as a socially relevant case study and valuable benchmark for evaluating how well meaningful and faithful explanations can be extracted from language models in a challenging real-world classification task.

Climate-Fever First published by Diggelmann et al. (2020), this dataset contains claim-evidence pairs with real-world claims retrieved from the web and 5 evidences for each automatically retrieved as text snippets from Wikipedia. The original evidences were labeled by annotators (*support*, *refute* or *not enough info*) and based on a majority vote across all 5 evidences, the overall claim was labeled as either *support*, *refute*, *not enough info* or *dispute*. We manually select 104 claims with 520 evidences, aiming for a balanced label set and claims that are easy to understand. Three personally recruited annotators then annotated rationales for the 520 evidences that either support or refute the claim or are left without label if not relevant. Pairwise inter-annotator agreement across all tokens for the 3-class settings reaches an average of 0.36 ± 0.05 (Cohen's Kappa). For further analysis, rationales were averaged with the same strategy as for SST with thresholds of 0.5 and -0.5. Figure 5 in the Appendix illustrates the difficulty of annotating this dataset. The claim itself *The polar bear population has been growing* is under-specified, it neither refers to a specific time frame nor a geographic location. On the other hand, among the automatically retrieved evidences, we find specific time frames (#2 and #3) but can sometimes only assume that a particular statement refers to the original claim (#3 mentions *bears* but not *polar bears*). More details on the annotation study can be found in AppendixA.

3.2 Rationale Extraction

For our experiments, we evaluate the following 4 instruction fine-tuned LLMs: Gemma3-12B (language-only), Llama3.1-8B, Qwen3-8B (non-thinking mode) and Mistral-7B with more details in Appendix B.

We first 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 returns the correct answer, we ask it to generate rationales based on the relevant 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 manually translated the prompts to the respective language with the help of native speakers.

For Climate-Fever, we follow the original annotation approach for both model experiments and human annotations where we first ask for rationales on the 5 provided statements, i.e., evidences with respect to the claim and based on those, ask for an overall claim label.

Experimental Details The experiments are based on the `transformers` library. We set the repetition penalty to 1.0 and adjust the maximum length of generated text with respect to the task and expected output. We ensure reproducibility of our results by consistently using the same set of 3 seeds across our experiments and will release our code upon publication, including all parameters and the libraries used. Instructions with class definitions are presented in the Appendix in Figures 6 - 9.

Data \ Acc.	Gemma3	Llama3	Qwen3	Mistral
SST	0.98	0.98	0.98	0.99
mSST (EN)	**1.00**	0.98	0.99	0.98
mSST (DA)	0.94	0.84	**0.96**	**0.96**
mSST (IT)	**1.00**	0.95	**1.00**	0.97
RaFoLA #1	0.25	0.47	0.38	**0.57**
RaFoLA #2	0.37	**0.60**	0.47	0.58
RaFoLA #5	**0.79**	0.73	0.74	0.60
RaFoLA #8	0.65	**0.76**	0.67	0.73
Claim	**0.45** $\pm$ .04	0.33 $\pm$ .04	0.38 $\pm$ .02	0.24 $\pm$ .01
Evidence	**0.54** $\pm$.02	0.40 $\pm$ .03	0.46 $\pm$ .00	0.45 $\pm$ .02

Table 1: Model accuracies (macro F1) for SST, multilingual SST, RaFoLa and Climate-Fever (claim verification task and evidence classification), with highest scores shown in bold. Scores are averaged across 3 seeds, standard deviation for the first three datasets is ≤ 0.01.

4 Main Results

In the following, we will present results on respective task performances and plausibility, i.e., pairwise agreement between human annotations and generated rationales.

4.1 Task performance

Table 1 shows task accuracies (macro-F1) for SST, mSST, RaFoLa and Climate-Fever. Macro-F1 scores for SST and mSST are generally high across models ($0.84 - 1.0$). We 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. From the set of models we consider for this study, only *Mistral* is considered an English-only model. *Qwen3* was pre-trained on Italian and Danish, *Llama3* was pre-trained on Italian and for *Gemma3* we only know that it supports 140+ languages but the technical report does not reveal which languages are included (Team, 2025). Despite this difference in language exposure, our results show that all models are able to solve the sentiment classification task in Danish and Italian with accuracies comparable to English.

RaFoLa shows more variation across articles and overall lower performance ($0.25 - 0.79$) than SST/mSST. Performance is lower for Articles #1/#2 than for Articles #5/#8 (best performances $0.57/0.60$ vs. $0.79/0.76$). We also observe that, although released more recently, *Gemma3* and *Qwen3*, on average, perform worse than *Mistral* and *Llama3* in this task, $0.52/0.57$ vs. $0.62/0.64$.

We present claim verification performance and evidence classification for Climate-Fever in the lower part of Table 1, both as macro-F1 scores. *Gemma* performs best in both cases, in the 4-class claim verification task it reaches 0.45 and 0.54 in the 3-class evidence classification task.

4.2 Plausibility

Following the interpretability literature, we assess *plausibility* of rationales to humans by considering human annotations as the ground truth and compute their agreement to generated rationales (DeYoung et al., 2020). For this, we calculate sample-wise Cohen's Kappa scores between the binary scores and average across samples for different models. Results (averaged across 3 seeds) are shown in Figure 2.

Metric Cohen's Kappa (Cohen, 1960) is a well-established method to measure inter-annotator agreement (IAA) between two annotators, in our case the averaged human annotations and the model rationales. We choose 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. Cohen's Kappa scores account for both issues, leading to overall lower but more robust scores than F1.

SST Results for SST are shown in the left part of Figure 2. For both English subsets, we mostly see a moderate level of agreement ($0.4 - 0.6$) for the comparison between human annotation and self-explanations (generated by LLMs) except for *Mistral* (DA: 0.32, IT: 0.31) and *Gemma* (DA: 0.33).[4]

RaFoLa Results for RaFoLa, are shown in the middle part of Figure 2. We here see overall lower levels of agreement but also a high variance by a magnitude of up to 4 between different articles. Plausibility scores reach from only lower levels of agreement for article #1 ($0.12 - 0.17$), to a fair level of agreement for article #2 ($0.19 - 0.27$) and moderate agreements for articles #5 ($0.21 - 0.48$) and #8 ($0.27 - 0.41$). Similar to SST, we see highest agreements for *Llama3* and *Gemma3* which is surprising given the low performance for *Gemma3*.

Climate-Fever The right panel of Figure 2 shows Kappa scores for Climate-Fever. Here, we concatenate correctly classified evidence statements per claim and compute agreement scores with human annotations for the respective evidence statements. In contrast to the other datasets, rationales are annotated and generated as either supportive (1), contradictive (-1) or not relevant (0). Only *Gemma3* reaches a fair level of agreement (0.24) while the other models reach only slight levels of agreement ($0.12 - 0.18$).

4.3 Token Statistics

We extract the top-8 tokens from all datasets as well as from the rationale annotations. For SST/mSST (Table 6) and Climate-Fever (Table 5), we do not find any meaningful differences between humans and models, neither across models nor languages. For Climate-Fever we find mostly *nouns* among the selected tokens whereas for SST/mSST we find an even distribution among nouns, adjectives and adverbs. For RaFoLa (Table 7), we see that articles #1 (Abuse of vulnerability) & #2 (Abusive working and living conditions) mostly contain descriptive nouns covering the general topic of the dataset such as *workers*, *work*, *labour* etc. Whereas selected tokens for articles #5 (Excessive overtime) & #8 (Physical and sexual violence) deviate more from

[4]We follow Landis (1977) to classify levels of agreement.

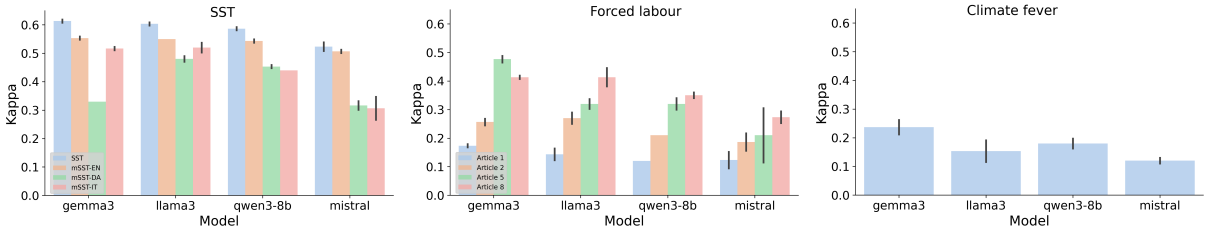

Figure 2: Human-model agreement as Cohen's Kappa scores for all datasets. Scores were computed for correctly classified samples and then averaged across datasets, standard deviation across seeds is shown as error bars.

the corpus' most frequent tokens (first row) with keywords such as *hours*, *day* and *sexual*, *women*, *violence*, respectively for the two articles. Those keywords are easier to identify which supports the higher performance for #5/#8 observed in Table 1.

Summary We find that (i) task performance and human-model agreement vary a lot across models and articles for RaFoLa with *Llama3* and *Mistral* performing better on average than *Gemma3* and *Qwen3* where the former can be explained by more indicative keywords for some articles than for others, (ii) *Gemma3* outperforms the other models on claim verification and evidence classification for Climate-Fever and (iii) *Gemma3* shows highest agreement with human annotations on all datasets followed by *Llama3*.

5 Analyses

In this section, we compare human and model rationales with post hoc interpretability methods to gain a deeper understanding of (i) the extent to which human and model rationales provide faithful token identification, i.e., whether models are sensitive to changes in their predictions when these rationales are masked, (ii) the degree to which human and model rationales align with post hoc attributions, and (iii) how these approaches differ in their strategies for information extraction and token selection. We exclude Climate-Fever from this analysis as there is no clear protocol for applying gradient-based methods on a collection of independent statements.

5.1 Faithfulness

Besides plausibility, faithfulness is the most commonly used criterion for evaluating model explanations. It is assessed by removing the most relevant features and measuring the resulting change in the model prediction, which can be viewed as an interventional probe of the model's decision process. Faithful rationales are characterized by a strong decrease in the prediction score when the associated

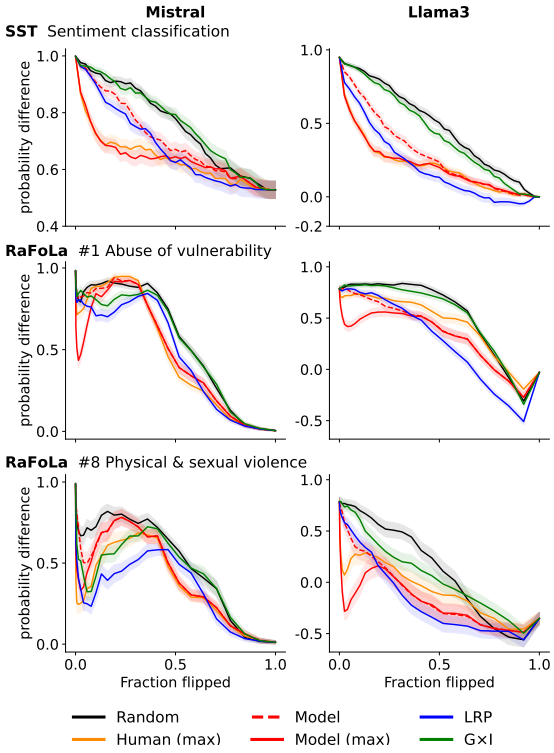

Figure 3: Faithfulness evaluation for SST and RaFoLa (articles #1 and #8). Model probability difference after masking tokens extracted from human rationales, model self-explanation rationales and post-hoc attributions (LRP, GxI) for Mistral and Llama3 with full results in Figure 10 (Appendix). Shaded bands indicate standard errors across samples. Faster drop in probability for early fractions indicates more faithful identification of task-relevant rationales. Human/Model (max) refers to rationales selected via greedy maximization of next-token probability difference.

tokens are iteratively masked. Accordingly, we evaluate faithfulness by measuring the *probability difference* between the correctly predicted answer token and the alternative answer token after masking tokens identified by human rationales, model self-explanations, and post-hoc attributions.

Experimental Setup To extract input attribution scores, we use two widely adopted gradient-based attribution methods that enable efficient computation of feature attributions in LLMs: Gradient×Input (Baehrens et al., 2010; Shrikumar et al.,

2017) and layer-wise relevance propagation (LRP) (Ali et al., 2022; Achtibat et al., 2024). Both methods have been successfully applied in recent interpretability work on information retrieval (Rezaei Jafari et al., 2024) and causal circuit discovery (Syed et al., 2024; Jafari et al., 2025), and scale effectively to larger models and longer input sequences, allowing us to study more complex datasets like RaFoLa. Details on implementation of attribution techniques are given in Section D in the Appendix.

We follow a perturbation-based masking approach and rank binary human and model rationales using a k-greedy importance ordering algorithm that prioritizes tokens according to their maximal impact on the probability difference in descending order. Iteratively selecting the top-k tokens that produce the largest drop in probability difference and flipping them first yields a ranking that captures the cumulative contribution of tokens to the model prediction, which we denote as *Human/Model (max)*. We set $k = 1$ for the shorter SST sentences, and $k = 3$ for all other datasets. Our faithfulness results on SST and RaFoLa subsets for *Mistral* and *Llama3* are summarized in Figure 3, with results for all models provided in Appendix Figure 10.

Ranked self-explanations offer faithful model interventions. Steepest drop in probability difference occurs for k-greedy ranked self-explanations (Model (max)), especially for the first 5-10% of tokens perturbed. Flipping of model self-explanations in random order (dashed red line) results in less pronounced impact on probability difference without a pronounced dip in probability difference.

Self-explanations can identify more faithful rationales than human annotators. We find that model rationales extracted from self-explanations are overall at least as faithful as human rationales, in particular, considering the first 0-10% of token fractions flipped. Specifically, we observe a transient drop followed by a rebound in probability difference for long-text classification settings in RaFoLa compared to SST sentiment classification. This *faithfulness dip* can also occur for human rationales, but is less pronounced. As RaFoLa requires weighting potentially contradictory evidence in news articles, removing highly influential phrases can alter the meaning of other evidence. This suggests more complex text interactions than in short sentiment classification, potentially reflecting higher-order semantic and epistemic processes that drive the observed rebound.

Post-hoc explanations differ from both humans and model rationals. Depending both on task and model, we see good to moderate agreement with self-explanations, in particular, for SST and RaFoLa #8, and especially for the first 10% of flipped tokens. However, we find clear differences in faithfulness curves, as models react differently to interventions on model versus post-hoc (LRP) rationales for RaFoLa. LRP-based rationales provide consistently more faithful token identification than GradientxInput, while overall, post-hoc attributions being less faithful in detecting the first few most relevant tokens than model self-explanation rationales. While post-hoc LRP rationales can, depending on the model, result in fast drops in probability difference, its ordering remains less faithful than both model and human rationales, hinting at differences in how text importance is assigned.

These differences may reflect the **distinct interpretability approaches**, with model rationales relying on coherent evidence communicated through natural language explanations while post-hoc methods assign relevance to the entire input. The most attributed tokens are often not task-specific content (e.g., RaFoLA news articles) but structural tokens from the system prompt, such as "<|begin_of_text|>" in Llama3, "<s>" in Mistral, or "<bos>" in Gemma3. As these tokens are required for the model's internal processing and sequence delimitation, attribution methods mark them as highly important despite their limited semantic role. Consequently, post-hoc rationales tend to emphasize structural or formatting tokens rather than evidence, highlighting a key difference between evidence-focused explanations and natural-language model rationales.

5.2 Analyzing Rationale Strategies in RaFoLa

We examine the top 5% of faithful tokens obtained via intervention-based analysis on the RaFoLA dataset, where faithfulness curves differ most across human, model, and post-hoc rationales. Through an initial LLM-assisted exploratory analysis of top flipped tokens using GPT-5 (see D.2 in the Appendix for details), we identified recurring themes and preliminary distinctions across these groups. We validate those findings with a statistical analyses and standard NLP pipelines. This reveals a variety of strategies that reflect systematic differences in focus and function across rationales.

Overall, we find that **human rationales contain more tokens that convey narrative content**, em-

phasizing lived experiences and the broader significance of exploitation through story-like language. This is reflected in higher lexical diversity, with type–token ratios ranging from 29% to 44%, and high proportions of stop words, typically 34–39% across models (see Table 9 in the Appendix). Typical rationales include descriptive and narrative words like *vulnerable, victims, poverty, children, working, factory, men, women* and phrases like *One year I was pregnant [...]*, *girl describes how her boss [...]*, or *She told me*, reflecting a focus on human contexts and experiences. Tables 10 and 11 in the Appendix show additional samples.

Model-generated rationales provide more factual and analytical detail, using denser, technical phrasing to document mechanisms and procedures, as well as quantitative evidence. Their lexical diversity is similar or slightly higher than humans (type–token ratios 28–47%), while stop word fractions remain comparable (35–38%), indicating natural-language style usage. Example (sub)tokens include *trafficking, labour, bail, conditions, hundreds, $, Johannesburg, Xinjiang* and phrases like *The Global Slavery Index* or *million labourers [...] is the biggest recruitment programme anywhere in the world according to the International Labor Organization ILO*, highlighting their tendency to record events, locations, and procedural aspects in a factual manner.

Post-hoc rationales emphasize more isolated evidence tokens and structural elements, including publisher artifacts such as source labels, dates, locations, and editorial markers. These rationales contain lower stop word fractions (12–25%), lower lexical diversity (type–token ratios 24–37%), and higher proportions of named entities—for example, GPE up to 1.69%, ORG up to 2.74%, and PERSON up to 2.23% compared to human and model rationales. Typical tokens include *Reuters, CNN, "By", Swiss, Lisa Krist, Ghana, "https:/"*, reflecting their focus on evidence extraction and sources.

Differences relative to human rationales are quantified by the Δ values reported in Table 9, which show both model vs. human and post-hoc vs. human difference patterns. Positive or negative Δ values indicate increases or decreases, respectively, in token-type ratios, stop word fractions, formatting tokens, and named entity mentions. For instance, post-hoc rationales typically exhibit 12–24% lower percentage points of stop word fractions and higher fractions of named entities (up to 1.51% for ORG and 1.66% for PERSON) compared to human rationales, reflecting their evidence-focused role in extracting and isolating token patterns. In contrast, model rationales differ less from human ones, with only modest Δ values across most linguistic and entity features, emphasizing their adherence to natural-language.

5.3 Corpus Statistics

Table 4 summarizes key corpus statistics to allow analyzing potential implications regarding the rationale agreement scores.

Document and sentence length. RaFoLa contains substantially longer documents (945 tokens on average) than Climate-Fever (200) and SST (21), suggesting that increased length and contextual complexity may contribute to annotation difficulty. In contrast, the short length of SST likely constrains contextual variation, which may partially explain its higher agreement scores (0.4–0.6).

Lexical diversity and ambiguity. By computing token-type rations (TTR), we find that SST exhibits the highest lexical diversity (37% TTR), whereas RaFoLa and Climate-Fever show much lower diversity (3–5%), while lexical ambiguity remains similar across datasets (54%). This suggests that neither lexical ambiguity nor diversity alone explains the notably low agreement observed in Climate-Fever (0.12–0.24).

Syntactic complexity. Climate-Fever has lower mean dependency depth (MDD = 1.87) than SST (2.75) and RaFoLa (2.92), indicating simpler sentence structures, which combined with moderate document length, suggests that syntactic complexity is not the primary source of disagreement and instead points toward task-specific effects.

Entities and POS. Dataset-specific distributions of named entities and POS tags further differentiate the corpora: Climate-Fever contains more quantitative expressions (e.g., cardinals and dates), SST includes relatively few named entities (primarily persons), and RaFoLa shows higher densities of geopolitical, organizational, and demographic entities. These patterns suggest that entity distributions may influence when post-hoc explanations align with or diverge from self-explanations, consistent with our observations in Section 5.2.

In summary, our analysis indicates that low agreement scores on Climate-Fever is not a result of surface-level textual properties and rather relates to its intrinsic task complexity, with SST and RaFoLa exhibiting patterns consistent with their structural and lexical profiles.

6 Discussion

In this paper, we evaluate self-explanations, i.e., explanations generated by instruction-tuned LLMs, based on their plausibility to humans and their faithfulness to models. We instruct 4 open-weight LLMs: *Gemma3*, *Llama3*, *Qwen3* and *Mistral* for three text classification tasks in English but also in Italian and Danish. We analyse the sentiment classification (SST/mSST) and forced labour classification (RaFoLa) tasks for which human annotations are available. We collect new human rationale annotations for the third dataset, Climate-Fever (Diggelmann et al., 2020), a claim verification dataset with claims related to climate change. We further include post-hoc attribution-based explanation methods in order to evaluate faithfulness of rationales to the correct model predictions.

For the **plausibility** analysis, we compute agreement scores between human and model rationales. Overall, we find highest plausibility scores for the English subsets of SST. For RaFoLa, we find higher variance across the different articles which aligns with the task performance. Agreement scores for Climate-Fever, the only dataset with 3-class rationale annotations, reach overall lower agreement scores than the other datasets.

There are several potential factors that influence those plausibility scores and their differences. The three datasets considered here were originally composed with varying incentives, see also (Eberle et al., 2023) for a related discussion. The movie reviews in SST/mSST were written with the clear purpose of expressing an emotion and justify a given movie rating. Sentences are quite short and usually contain clear keywords that are easy to identify by both humans and models. The RaFoLa dataset consists of news articles whose main purpose is to report on a given event. Those articles are much longer and information about the potential risk indicators might be implicit and subtle as this was not necessarily the main purpose while writing them. As we see in the analysis of the most frequent words, articles #5 and #8 contain much clearer and more distinct keywords than articles #1 and #2 which makes both classification and rationale annotation more predictable and consistent. Climate-Fever, a claim verification benchmark pairing real-world web claims with five semantically relevant Wikipedia statements, involves longer evidence contexts and greater ambiguity, making classification and rationale selection more demanding overall. Incorporating such diverse text sources and types of real-world datasets therefore remains essential for obtaining broader and more realistic evaluation benchmarks.

Our analyses further revealed **distinct faithfulness patterns** for human, self-explanations and post-hoc rationales: human and, in particular, self-explanations are consistently able to intervene strongly on the model's ability to predict correct answer tokens, especially when considering early fractions of text tokens. In comparison, LRP-based post-hoc rationales also reduce the probability difference, but the drop is generally less pronounced. These differences reflect distinct approaches to interpretability: self-explanations communicate coherent evidence through natural language, directly summarizing the model's decision process, whereas post-hoc methods assign relevance across the entire input, including system and task prompts, often highlighting structural or formatting cues rather than semantically critical phrases. This makes them less effective at disrupting coherent units of meaning but can provide a complementary, more detailed view of the model's low-level processing compared to self-explanations. To bridge this gap, *interpretability agents* (Han et al., 2025; Lermen et al., 2025) that are designed to utilize model activation or attribution patterns to provide natural language explanations, which may provide more grounded and faithful insights into the mechanisms underlying model predictions.

Good explanations should faithfully reflect the model's learned strategy, even if not fully plausible or unintuitive to humans (Agarwal et al., 2024). Further, self-explanations can suffer from counterfactuality, producing untruthful rationales for correct predictions (Ji et al., 2023), highlighting the need for rigorous evaluation. We therefore concentrated on extractive rationales grounded in the input text, as this setting enables controlled measurements of plausibility and faithfulness against human annotations. Our results constitute an initial step toward characterizing the reliability of self-explanations in realistic text classification. Future research should examine how these findings transfer to abstractive free-text explanations and to process-oriented analyses of how models integrate and resolve complex, potentially contradictory evidence to arrive at a final decision.

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.

References

Reduan Achtibat, 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.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *Preprint*, arXiv:2402.04614.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. 2022. Xai for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning*, pages 435–451. PMLR.

Leila Arras, José Arjona-Medina, Michael Widrich, Grégoire Montavon, Michael Gillhofer, Klaus-Robert Müller, Sepp Hochreiter, and Wojciech Samek. 2019. Explaining and interpreting LSTMs. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 211–238.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614.

Oliver Eberle, Ilias Chalkidis, Laura Cabello, and Stephanie Brandl. 2023. Rather a nurse than a physician - contrastive explanations under investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6907–6920, Singapore. Association for Computational Linguistics.

Jiaojiao Han, Wujiang Xu, Mingyu Jin, and Mengnan Du. 2025. Sage: An agentic explainer framework for interpreting sae features in language models. *Preprint*, arXiv:2511.20820.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *Preprint*, arXiv:2310.11207.

International Labour Organization. 2012. ILO Indicators of Forced Labour. Special Action Programme to Combat Forced Labour (SAP-FL).

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.

Farnoush Rezaei Jafari, Oliver Eberle, Ashkan Khakzar, and Neel Nanda. 2025. Relp: Faithful and efficient circuit discovery via relevance patching. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

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.

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.

J. Richard Landis. 1977. The measurement of observer agreement for categorical data. *Biometrics*.

Simon Lermen, Mateusz Dziemian, and Natalia Pérez-Campanero Antolín. 2025. Deceptive automated interpretability: Language models coordinating to fool oversight systems. *Preprint*, arXiv:2504.07831.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

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.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, Cham.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2025. Mind the gap: from plausible to valid self-explanations in large language models. *Machine Learning*, 114(10):220.

Lucas Resck, Marcos M. Raimundo, and Jorge Poco. 2024. Exploring the trade-off between model performance and explanation plausibility of text classifiers using human rationales. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4190–4216, Mexico City, Mexico. Association for Computational Linguistics.

Farnoush Rezaei Jafari, Grégoire Montavon, Klaus-Robert Müller, and Oliver Eberle. 2024. Mambalrp: Explaining selective state space sequence models. In *Advances in Neural Information Processing Systems*, volume 37, pages 118540–118570. Curran Associates, Inc.

Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L Glassman, and Finale Doshi-Velez. 2021. Evaluating the interpretability of generative models by interactive reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar Van Der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International Conference on Machine Learning (ICML)*.

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.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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.

Yingming Wang and Pepa Atanasova. 2025. Self-critique and refinement for faithful natural language explanations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8492–8518, Suzhou, China. Association for Computational Linguistics.

Sarah Wiegreffe, 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.

Lingjun Zhao and Hal Daumé Iii. 2025. A necessary step toward faithfulness: Measuring and improving consistency in free-text explanations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15810–15824, Suzhou, China. Association for Computational Linguistics.

A Climate-Fever Annotation Study

The Climate-Fever dataset was first collected and published by (Diggelmann et al., 2020). The dataset consists of 1535 real-world English claims about climate change for which 7675 related evidence statements were retrieved from Wikipedia, 5 for each claim. Each evidence statement was labeled by up to 5 annotators into either SUPPORT, NOT ENOUGH INFO or REFUTE. Micro (per claim-evidence pairs) and macro labels (per claim) were aggregated afterwards. Labels for claim-evidence pairs were decided on a majority vote (or NOT ENOUGH INFO on a tie). Macro labels were labeled NOT ENOUGH INFO by default unless there is SUPPORTING or REFUTING evidence and in case of both, the claim was labeled as DISPUTE.
For our study, we are interested in token-level rationales which are not available from the initial publication of Climate-Fever. Therefore, we manually selected a subset of 102 claims (510 claim-evidence pairs) based on clarity of the claim formulation and balanced claim labels.

A.1 Annotator selection

We used Prodigy as the annotation interface and initially planned to collect data via Prolific, a crowdsourcing platform. It turned out more difficult than expected for annotators to open the correct link for Prodigy, i.e., they would need to add their Prolific-id to the link and to understand all parts of the annotation study. After several attempts, we decided to instead personally instruct 2 student annotators and one of the authors for the entire study.

A.2 Annotation tasks

Annotators were asked to solve 4 tasks, an example is shown in Figure 5.

0. reading the claim and all 5 evidences
1. token-level rationale annotation for each evidence statement
2. decide on a claim label
3. briefly explain the decision

We publish the entire study but will focus on rationale annotation and claim labels in this paper.

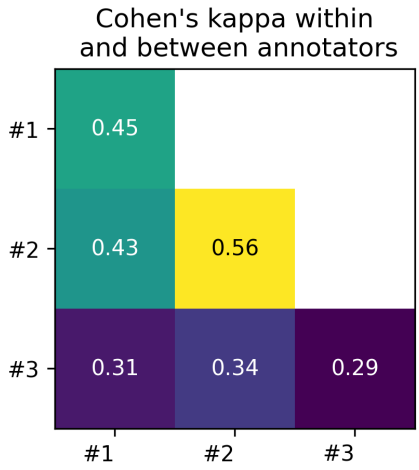

Figure 4: Kappa scores for inter and intra annotator agreement of all 3 annotators.

A.3 First and last batch

104 samples were divided into 10 batches with 12 samples each, except for the last batch which only contained 6 samples and a pre-trial which contained 3 samples. Since batches were annotated across several weeks, we repeated the 3 samples from the pre-trial to the last batch in order to compute intra-annotator agreements on those 3 samples.

A.4 Inter- and Intra- Annotator Agreements

We show Cohen's kappa scores for inter- and intra-annotator agreement in Figure 4. For the inter-annotator agreement, we concatenated all annotations and computed pairwise kappa scores. For the intra-annotator agreement, we computed scores on the overlap between pre-trial and the last batch for each annotator. Average inter annotator agreement is 0.36 ± 0.05 (kappa) and 0.57 ± 0.04 (macro F1).

A.5 Comparison to previous annotations

We compare labels on the newly annotated subset of Climate-Fever with previously annotated labels. There are several differences in the annotation process that require attention when interpreting the results. In the original study, evidences were labeled by multiple annotators and claim labels then inferred from two stages of majority voting (see previous section). In our annotation study, we annotated

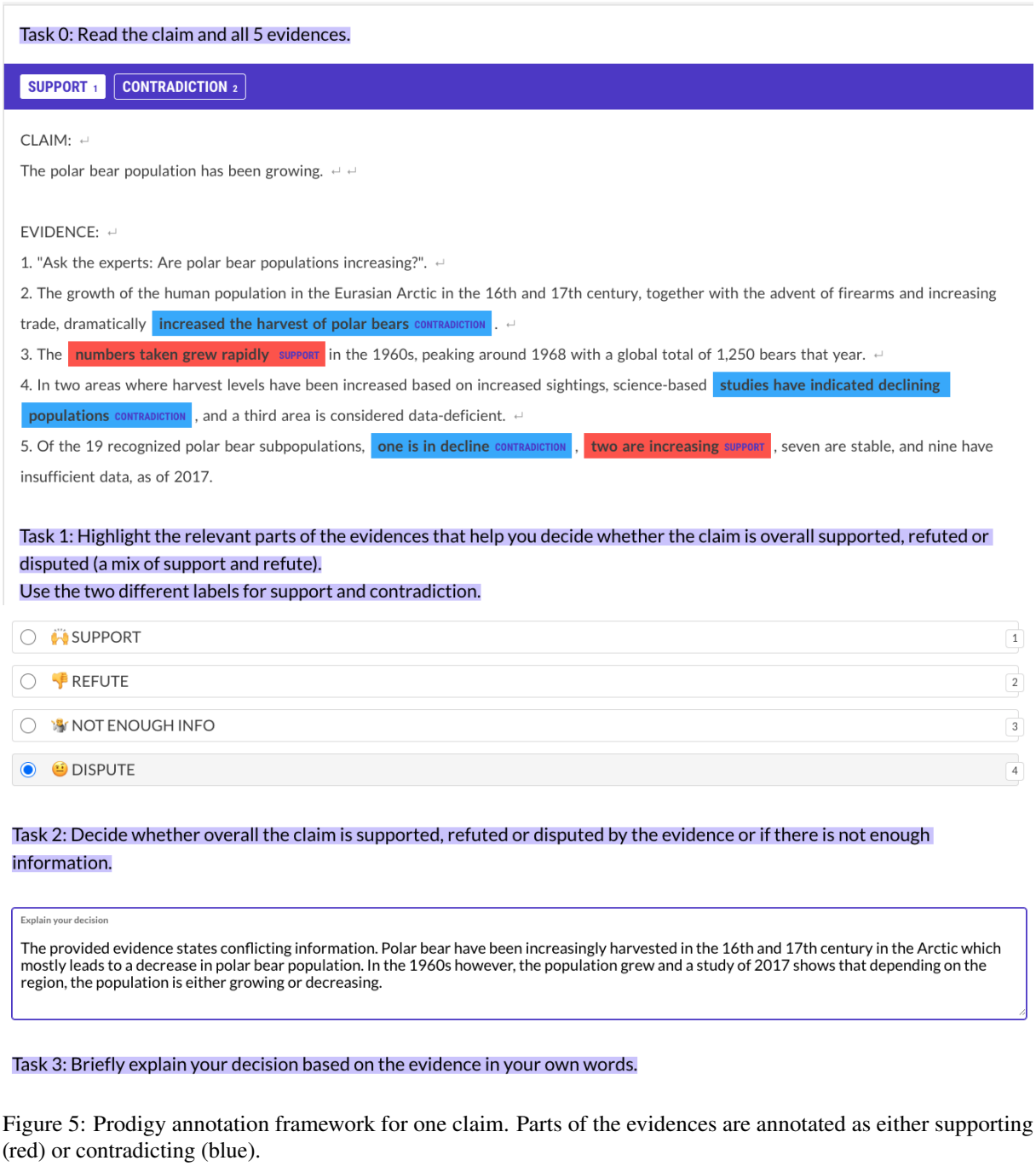

Figure 5: Prodigy annotation framework for one claim. Parts of the evidences are annotated as either supporting (red) or contradicting (blue).

	Prec.	Recall	F1-score
SUPPORT	0.75	0.83	0.79
REFUTE	0.71	0.87	0.78
NOT ENOUGH INFO	0.00	0.00	0.00
DISPUTE	0.46	0.43	0.44
Macro-F1			0.50

Table 2: Comparison between original claim labels and newly annotated claim labels

	Prec.	Recall	F1-score
SUPPORT	0.56	0.79	0.65
REFUTE	0.54	0.67	0.60
NOT ENOUGH INFO	0.74	0.60	0.66
DISPUTE	0.00	0.00	0.00
Macro-F1			0.48

Table 3: Comparison between original evidence labels and newly annotated evidence labels

evidences with rationales and inferred evidence labels based on supporting evidence (SUPPORT), contradicting evidence (REFUTE), no evidence (NOT ENOUGH INFORMATION) or disputing evidence (DISPUTE). Claim labels were annotated after the rationales annotation. This also means that in our annotation study, evidences could get a DISPUTE label which was not the case in the original data collection process. Tables 2 and 3 show classification results where *new* labels are considered as the ground truth. Results show that label agreement on the claim level is much higher for SUPPORT and REFUTE than for the other two labels. On the evidence level we see balanced F1 scores around $0.6 - 0.66$ for all overlapping labels but a score of 0 for the DISPUTE class that only exist in the new annotations. Those scores are comparable to the previously reported average kappa and macro-F1 scores for the inter- and intra-annotator agreements.

B Model experiments

B.1 Models

For our experiments, we use the following instruction-tuned LLMs: Gemma3-12b[5], Llama3.1-8B[6], Qwen3-8B[7], Mistral-7B[8] Although some of them might not be considered large in size, we follow the convention of calling them Large Language Models based on their abilities and training procedures.

B.2 Instructions

We show multilingual instructions for SST in Figures 6 - 7. Instructions for forced labour detection in RaFoLa are shown in Figure 8 and relevant definitions from the International Labour Organization in Figure 9.

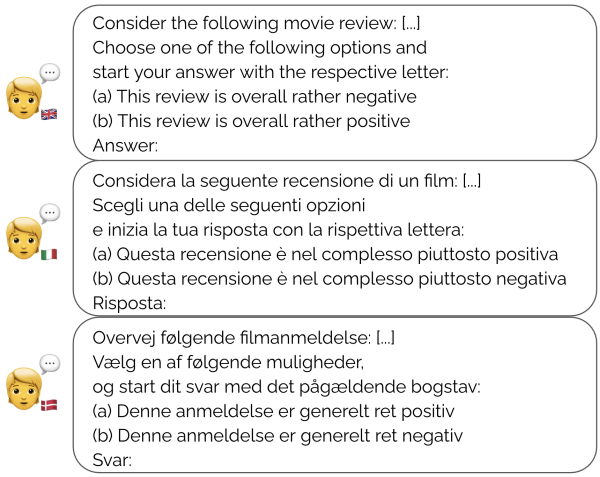

Figure 6: Prompts in all 3 languages to solve sentiment classification.

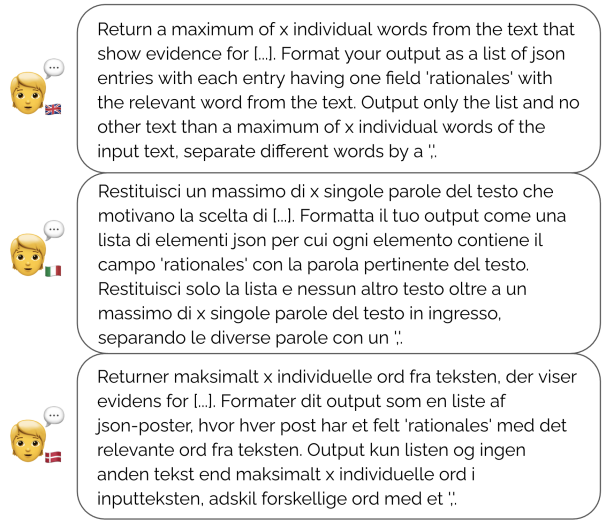

Figure 7: Follow-up prompts in all 3 languages to extract rationales.

C Corpus Statistics & Top-8 tokens

[5] huggingface.co/google/gemma-3-12b-it

[6] huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

[7] https://huggingface.co/Qwen/Qwen3-8B

[8] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

Table 4: Corpus Statistics across all three datasets. Abbreviations as follow: MDD: Mean dependency depth (syntactic complexity), TTR: Token type ratio, POS: Fraction of pos entities, GPE: Geopolitical entity, ORG: Organization entity, NORP: Nationalities or religious or political groups.

Dataset	Toks/Doc	Toks/Sent	MDD	TTR [%]	Stopwords [%]	Formatting [%]	POS [%]	Lex. Ambig. [%]	GPE [%]	Cardinal [%]	ORG [%]	Date [%]	NORP [%]	Person [%]
SST	20.86	20.71	2.75	36.78	43.03	12.65	3.52	54.07	0.36	0.35	0.67	0.40	0.24	1.49
RaFoLa	944.89	29.64	2.92	3.30	38.37	11.82	6.47	54.03	1.43	0.80	1.90	0.98	0.47	0.89
Climate-Fever	199.86	16.70	1.87	4.88	33.33	14.42	7.22	54.10	0.47	3.60	0.96	1.74	0.13	0.32

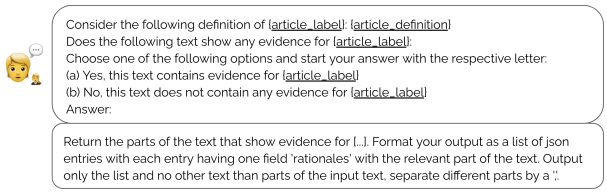

Figure 8: Prompts for classification and rationale extraction for the RaFoLa dataset.

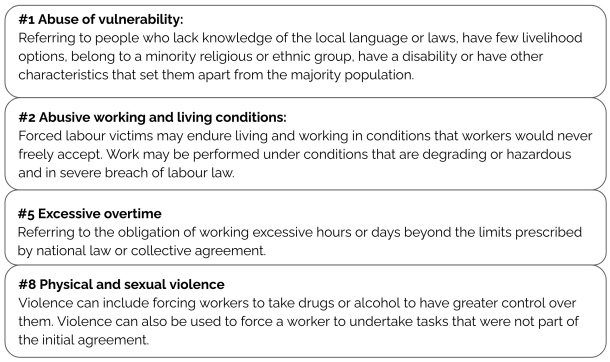

Figure 9: Indicators defined by the International Labour Organization and published by Mendez Guzman et al..

corpus	climate, global, ice, sea, warming, change, greenhouse, human
human	global, sea, climate, ice, warming, earth, greenhouse, human
gemma3	climate, global, greenhouse, change, warming, sea, ice, temperatures
llama3	global, climate, warming, ice, change, sea, temperatures, greenhouse
qwen3	ice, sea, climate, warming, global, level, temperature, mass
mistral	ice, global, warming, heat, temperatures, greenhouse, human, increased

Table 5: Top-8 tokens from Climate-Fever (first row) and from respective rationale annotations by humans (2nd row) and models.

D Faithful rationale comparison

D.1 Methodological Details

We compare human and model-based rationales provided on the same samples from various annotation studies and evaluate plausibility, i.e., agreement with human rationales and faithfulness to the model in comparison to state-of-the-art gradient-based feature attribution methods such as layer-wise relevance propagation (LRP)(Ali et al., 2022) and GradientxInput across models. We provide details on the LRP propagation procedure in the following.

LRP We applied the LN-rule (Ali et al., 2022) for LayerNorm/RMSNorm, the Identity-rule (Rezaei Jafari et al., 2024) for nonlinear activation functions, and the LRP-0 rule (Montavon et al., 2019) for linear transformations. The AH-rule was used for Mistral as it yielded better faithfulness compared to application of the Half-rule (Rezaei Jafari et al., 2024; Arras et al., 2019) (see (Rezaei Jafari et al., 2024; Jafari et al., 2025) for further discussion). A summary of implemented rules used in our experiments is given in Table 8.

D.2 LLM-assisted Analysis of Rationale Tokens

We use GPT-5 for an initial analysis of the large number of faithful token subsets derived from human rationales, model-based self-explanations, and LRP post-hoc explanations. Specifically, we prompted GPT-5 with: "Given token lists for multiple samples, summarize the main similarities and differences across Groups H, M, and P.". This task prompt was followed by up to 50 randomly selected samples per group (human, model, post-hoc), each containing the top 5% of tokens. A randomly selected subset of token lists is also presented in Tables 10 and 11. All text inputs used for the LLM-assisted analysis will be made available upon publication. This procedure was repeated for each model and used for hypothesis generation

	SST	mSST-EN	mSST-DA	mSST-IT
corpus	movie, film, like, comedy, work, -, love, funny	film, movie, performances, characters, bad, funny, like, story	film, ‘, filmen, ’, sjov, karakterer, bare, filmens	film, i, divertente, personaggi, interpretazioni, storia, trama, avvincente
human	funny, best, bad, movie, beautifully, compelling, film, performance	performances, bad, funny, good, characters, dull, film, compelling	sjov, film, overbevisende, præstationer, bedste, plot, vittig, sjovt	divertente, avvincente, noioso, interpretazioni, film, ben, brutto, assolutamente
gemma3	best, funny, bad, love, beautifully, compelling, hilarious, fun	funny, bad, performances, dull, compelling, good, long, best	sjov, overbevisende, spændende, humor, tilfredsstillende, klodset, dårlig, dårligt	divertente, avvincente, film, noioso, ben, brutto, intelligente, umorismo
llama3	best, funny, bad, beautifully, year, little, compelling, hilarious	bad, performances, funny, good, dull, best, compelling, little	sjov, overbevisende, bare, dårlig, dårligt, kedelig, spænding, spændende	divertente, avvincente, umorismo, noioso, senso, ben, interpretazioni, intelligente
qwen3	bad, best, love, movie, funny, beautifully, compelling, film	funny, bad, performances, dull, compelling, good, comedy, little	sjov, spændende, præstationer, overbevisende, vittig, sjovt, komedie, dårlig	divertente, avvincente, film, noioso, interpretazioni, personaggi, intelligente, intelligenti
mistral	comedy, best, bad, beautifully, funny, little, fun, stupid	bad, performances, funny, dull, characters, best, film, intelligent	sjov, filmen, præstationer, plot, film, spændende, sentimentalitet, giver	divertente, film, interpretazioni, i, trama, personaggi, avvincente, ben

Table 6: Top-8 tokens from SST/mSST splits (first row) and from respective rationale annotations by humans (2nd row) and models.

	#1 Abuse of vulnerability	#2 Abusive working and living conditions	#5 Excessive overtime	#8 Physical and sexual violence
corpus	workers, said, labour, work, rights, labor, human, children			
human	work, workers, children, forced, women, labour, said, vulnerable	workers, conditions, work, little, water, working, forced, said	hours, day, working, 12, work, worked, days, week	abuse, sexual, harassment, said, physical, women, violence, verbal
gemma3	workers, work, labour, forced, said, children, women, rights	workers, work, forced, conditions, labour, said, working, children	hours, work, day, days, working, workers, forced, week	workers, sexual, abuse, said, women, forced, violence, harassment
llama3	workers, work, said, labour, forced, children, women, working	workers, work, said, working, conditions, forced, labour, day	hours, day, working, said, work, workers, 12, days	said, sexual, workers, abuse, women, harassment, violence, work
qwen3	workers, work, forced, labour, said, children, women, working	workers, work, said, forced, labour, conditions, working, children	work, hours, workers, said, day, working, labour, forced	workers, said, abuse, sexual, forced, women, work, violence
mistral	workers, work, said, forced, conditions, working, children, women	workers, work, said, conditions, forced, labour, working, day	day, contracts, working, hours, work, said, days, supervisors	said, sexual, women, factory, abuse, woman, report, physical

Table 7: Top-8 tokens from RaFoLa (first row) and from respective rationale annotations by humans (2nd row) and models.

Propagation Rule	Mistral	Llama3	Qwen3	Gemma3
LN-rule (Ali et al., 2022)	✓	✓	✓	✓
Identity-rule (Rezaei Jafari et al., 2024)	✓	✓	✓	✓
0-rule (Montavon et al., 2019)	✓	✓	✓	✓
Half-rule (Rezaei Jafari et al., 2024)	×	✓	✓	✓
AH-rule (Ali et al., 2022)	✓	×	×	×

Table 8: Propagation rules used in our LRP implementation across model families. A checkmark (✓) indicates the rule was applied, and a cross (×) indicates it was not.

and initial exploratory corpus analysis. We then validated this qualitative analysis with statistical methods and standard automated NLP pipelines, including entity identification, stopword filtering, and lexical diversity assessment as presented in Table 9.

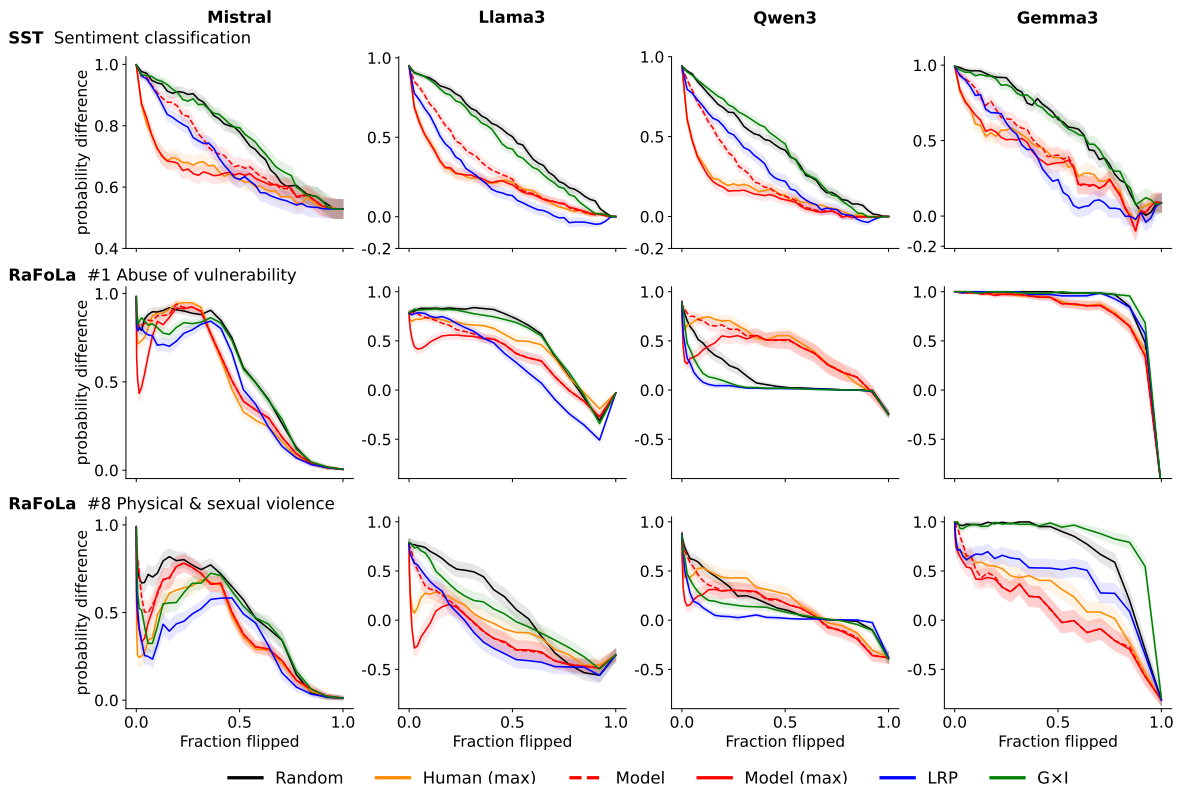

Figure 10: Faithfulness evaluation for SST and RaFoLa (articles #1 and #8). Model probability difference after masking tokens extracted from human rationales, model self-explanation rationales and post-hoc attributions (LRP, GxI) across models. Shaded bands indicate standard errors across samples. Faster drop in probability for early fractions indicates more faithful identification of task-relevant rationales. Human/Model (max) refers to rationales selected via greedy maximization of next-token probability difference.

Table 9: Token-level statistics for different models and sources. Numeric values are percentages (%). TTR refers to Type-Token ratios, measuring lexical diversity, and entity columns correspond to the following semantic categories: GPE (countries, cities, locations), Cardinal (numbers), ORG (organizations), Date (temporal expressions), NORP (nationalities, religious or political groups), and PERSON (individuals). Human, model, and post-hoc rows correspond to different rationale sources, with Δ rows showing differences relative to the human baseline.

Model	Source	TTR	Stopwords	Formatting	GPE	Cardinal	ORG	Date	NORP	Person
RaFoLa: # 1										
Llama3	human	31.91	38.35	2.87	0.89	1.15	0.71	0.52	0.64	0.92
Llama3	model	34.43	35.28	3.25	1.04	1.34	1.39	0.54	0.87	0.95
Llama3	post-hoc	31.87	14.02	17.44	1.69	2.15	2.22	0.33	0.73	2.23
Mistral	human	32.16	36.41	2.35	0.50	1.08	0.87	0.48	0.46	0.71
Mistral	model	32.05	37.23	1.66	0.48	1.45	0.87	0.46	0.35	0.64
Mistral	post-hoc	25.11	12.57	31.97	0.56	2.06	0.98	0.54	0.25	0.91
Qwen3	human	44.43	37.92	2.78	0.62	1.36	0.90	0.51	0.73	0.34
Qwen3	model	47.33	38.42	2.34	1.01	1.57	1.35	0.45	0.51	1.12
Qwen3	post-hoc	37.47	19.60	16.98	1.07	0.95	2.74	0.36	0.83	1.90
Gemma3	human	29.19	38.03	1.79	0.87	1.27	0.53	0.41	0.73	0.75
Gemma3	model	28.68	37.36	1.35	1.01	1.37	0.80	0.60	0.70	0.63
Gemma3	post-hoc	24.18	25.56	6.19	0.87	2.06	1.82	0.63	0.57	1.37
Llama3	Δ human - model	-2.52	3.07	-0.38	-0.15	-0.19	-0.68	-0.02	-0.23	-0.03
Mistral	Δ human - model	0.11	-0.82	0.69	0.02	-0.37	0.00	0.02	0.11	0.07
Qwen3	Δ human - model	-2.90	-0.50	0.44	-0.39	-0.21	-0.45	0.06	0.22	-0.78
Gemma3	Δ human - model	0.51	0.67	0.44	-0.14	-0.10	-0.27	-0.19	0.03	0.12
Llama3	Δ human - post-hoc	0.04	24.33	-14.57	-0.80	-1.00	-1.51	0.19	-0.09	-1.31
Mistral	Δ human - post-hoc	7.05	23.84	-29.62	-0.06	-0.98	-0.11	-0.06	0.21	-0.20
Qwen3	Δ human - post-hoc	6.96	18.32	-14.20	-0.45	0.41	-1.84	0.15	-0.10	-1.56
Gemma3	Δ human - post-hoc	5.01	12.47	-4.40	0.00	-0.79	-1.29	-0.22	0.16	-0.62
RaFoLa: # 8										
Llama3	human	32.93	38.57	2.93	0.36	0.97	0.69	0.40	0.47	0.97
Llama3	model	35.07	38.85	3.18	0.47	0.87	0.98	0.47	0.54	0.83
Llama3	post-hoc	34.36	15.70	16.80	1.25	1.33	2.00	0.94	0.43	2.63
Mistral	human	31.40	34.99	2.29	0.24	0.87	1.02	0.35	0.24	1.06
Mistral	model	33.18	37.63	2.68	0.28	0.91	1.38	0.39	0.12	0.99
Mistral	post-hoc	25.65	13.63	32.94	0.83	1.22	0.91	0.63	0.28	0.87
Qwen3	human	40.80	38.76	2.33	0.59	1.72	1.08	0.34	0.49	0.69
Qwen3	model	40.70	38.85	3.40	0.59	2.21	0.79	0.39	0.49	0.69
Qwen3	post-hoc	32.98	17.97	17.24	1.20	0.89	1.99	0.58	0.84	2.09
Gemma3	human	29.79	36.07	2.05	0.66	1.21	0.80	0.25	0.73	0.93
Gemma3	model	30.38	37.14	1.96	0.82	1.12	0.80	0.59	0.77	0.64
Gemma3	post-hoc	25.42	23.37	6.55	0.73	2.44	1.41	0.84	0.25	1.25
Llama3	Δ human - model	-2.14	-0.28	-0.25	-0.11	0.10	-0.29	-0.07	-0.07	0.14
Mistral	Δ human - model	-1.78	-2.64	-0.39	-0.04	-0.04	-0.36	-0.04	0.12	0.07
Qwen3	Δ human - model	0.10	-0.09	-1.07	0.00	-0.49	0.29	-0.05	0.00	0.00
Gemma3	Δ human - model	-0.59	-1.07	0.09	-0.16	0.09	0.00	-0.34	-0.04	0.29
Llama3	Δ human - post-hoc	-1.43	22.87	-13.87	-0.89	-0.36	-1.31	-0.54	0.04	-1.66
Mistral	Δ human - post-hoc	5.75	21.36	-30.65	-0.59	-0.35	0.11	-0.28	-0.04	0.19
Qwen3	Δ human - post-hoc	7.82	20.79	-14.91	-0.61	0.83	-0.91	-0.24	-0.35	-1.40
Gemma3	Δ human - post-hoc	4.37	12.70	-4.50	-0.07	-1.23	-0.61	-0.59	0.48	-0.32

Table 10: Qualitative examples of extracted tokens across models (rows) and human-annotated rationales, model-generated rationales, and post-hoc attribution-based (LRP) rationale for the RaFoLa dataset (here: article #1). Tokens are selected from the top 5% most faithful tokens in randomly selected samples. Irregular spacing can occur due to differences in tokenizers' subtoken processing strategies.

model	human	model	post-hoc
Llama3	- fearful Workers looking fearful could be a sign - with He said that in the past two years victims of trafficking had been found in hotels in North Wales where organised crime gangs had set up - serv itude . When choosing their victims traff ickers target the most vulnerable - monitoring to ensure children and other vulnerable groups - People working as cooks bus staff and wait staff might be exploited with traff ickers often taking advantage of language barriers between exploited workers and patrons	- million labour ers the Uzbek harvest is the biggest recruitment programme anywhere in the world according to the International Labor Organization I LO). U zbek istan - and girls are disproportionately affected accounting - Fatal accidents and child labour are common - from China and South Korea . - every where . X in jiang the north western region of China is home to minority	- Three - .aspx - forced - Swiss -backed - BBC - UK - Officers - Gang - appeal - Wales - modern slavery - trafficking
Mistral	- Work ers looking fear ful could be a sign of - He said that in the past two years victims of traff icking had been found in hotels in North Wales where organ ised crime gang s had set up - Bul gar ian nation als from disadv - When choosing their victims traff ick ers target the most vulnerable - The U igh urs had to come because they	- Work ers looking fear ful could be a sign of - I remember thinking What is it that they have ? Because I needed that in my own - Work ers in the Indian capital build buildings for about - ly co er ced to work against their will were members of the Muslim U igh ur minority in north western - t ann ery workers do not have	- Modern - S la - published - 0x0A - An appeal has - rural businesses - gang - modern slavery - Bulgar - ’“’ - http - 0x0A Acc ording
Qwen3	- serv itude . When choosing their victims traff ickers target the most vulnerable - foreign workers most of whom are women have very little means to defend themselves should the employer abuse them in any way - others . "One year I was pregnant but it - If they are undocumented migrants they may also fear that coming forward will result in their deportation . They - Some never go to school or learn to read and write Others are smugg led across borders and left vulnerable to trafficking or sexual abuse . The	- Forum for Human Rights reported that - serv itude . When choosing their victims traff ickers target - ok hid akh on K has anova was signed out - false job prospects and cash loans with ex orbit ant interest rates - ranks of workers without social protections including low wages and employer controls	- Swiss - programme - female - In the - speak - photographer Lisa Krist - Lisa Krist - Disclaimer - film director - Twitter - By - exploitation
Gemma3	- In the past schoolchildren university students doctors teachers and other public sector professionals were required to participate in the harvest for little or no pay and under threat of punishment - fearful Workers looking fearful could be a sign of the various methods used to control workers who are being exploited - with .” He said that in the past two years victims of trafficking had been found in hotels in North Wales where organised crime gangs had set up - network Bul garian nationals from disadvantaged regions were recruited by - servitude When choosing their victims traffickers target the most vulnerable	- of workers were forced labourers according to a - dies Officers found two people working at the site who they believed were the victims of modern slavery A - trafficking These victims of organised crime have been found on remote cannabis farms in - network Bul garian nationals from disadvantaged regions were recruited by the traffickers to - 7 2 hours deep in the mines shafts ,	- citizens - forced to - but not - cotton - Modern Day Slavery - UK - https :/ - forced - An appeal has gone - rural - Wales - gang masters

Table 11: Qualitative examples of extracted tokens across models (rows) and human-annotated rationales, model-generated rationales, and post-hoc attribution-based (LRP) rationale for the RaFoLa dataset (here: article #8). Tokens are selected from the top 5% most faithful tokens in randomly selected samples.

model	human	model	post-hoc
Llama3	- the blood ran out of the g ashes Fast - Many of the children also undergo physical and sexual abuse from the traff ickers as well as sometimes being forced into drug addiction - He threatened to kill my whole - investigators observed children working in the fields interviewed women who were paid - supervisors . The research show that in Jordan woman migrants routinely face sexual harassment and physical assaults by male supervisors . All	- Many of the children also undergo physical and sexual abuse from the traff ickers as well as sometimes being forced into drug addiction - fields and sexually assaulted by plantation fore - reported being raped while working . The - Naz ma Ak ter told New Age that in - will quite often be exposed to physical	- humiliation - later - them - from - sexual violence - Speaking - workers - Meanwhile - 16 - girl describes how - raped her - .com
Mistral	- Many of the children also under go physical and sexual abuse from the traff ick ers as well as sometimes being forced into drug addiction - girl describes how her boss rap ed her amid the tall - AP investig ators observed children working in the fields interviewed women who were paid nothing and women and - The research show that in Jordan woman migr ants rout inely face sexual harass ment and physical assault s by male super vis ors 0x0A All - reported harsh pun ish ments for min ers not comp lying with the rules imposed by the criminal	- Many of the children also under go physical and sexual abuse from the traff ick ers - girl who described being rap ed by her boss in - AP investig ators observed children working in the fields - The research show that in Jordan woman migr ants rout inely face sexual harass ment - In addition to severe beat ings other san ctions have included being shot in the hands or having a hand cut off as well as kill ings	- Beg - Iran - earn - 0 - abuse - year - describes how - boss rap ed her - 0x0A - worked - child labour - slavery
Qwen3	- Many of the children also undergo physical and sexual abuse from the traff ickers as well as sometimes being forced into drug addiction - work . A quarter of those surveyed had experienced verbal - Come sleep with me I will give you a baby Now - They to il for hours a day with little to no food and face abuse at the hands of their masters - girl describes how her boss raped	- Many of the children also undergo physical and sexual abuse from the traff ickers as well as sometimes being forced into drug addiction - 9 8 0 . Four in ten reported - 1 2 being taken into the fields and sexually assaulted by plantation fore men . While - can hear the sound of tools hitting stone of men cough ing - 1 2 being taken into the fields and sexually assaulted by plantation fore men . While	- Iranian - Speaking - Iran - . - migrant - Risk - rapport - ". - the - palm oil plantation that - some - .
Gemma3	- “ he would have them tied up to a - Many of the children also undergo physical and sexual abuse from the traffickers as well as sometimes being forced into drug addiction - According to the Typ ology study traffickers more frequently use physical violence in outdoor solicitation than in other types of sex trafficking Residential - officers of some garment factories h url abusive words at female workers and even go for sexual harassment If - work A quarter of those surveyed had experienced verbal	- Many of the children also undergo physical and sexual abuse from the traffickers as well as sometimes being forced into drug addiction - It is a broad term used in a commercial sex trade referring to commercial sex acts This acts primarily occur at a temporary indoor location - officers of some garment factories h url abusive words at female workers and even go for sexual - € 9 8 0 Four in ten reported feeling unsafe at work - girl describes how her boss raped her amid the tall trees on an Indonesian palm oil plantation	- beat - blood - Copyright - forced - beggars - . Meanwhile - In - ’ - “ - – - or - does not recognise