Free-text Rationale Generation under Readability Level Control

Anonymous ACL submission

Abstract

Free-text rationales justify model decisions in natural language and thus become likable and accessible among approaches to explanation across many tasks. However, their effectiveness can be hindered by misinterpretation and hallucination. As a perturbation test, we investigate how large language models (LLMs) perform rationale generation under the effects of readability level control, i.e., being prompted for an explanation targeting a specific expertise level, such as sixth grade or college. We find that explanations are adaptable to such instruction, though the requested readability is often misaligned with the measured text complexity according to traditional readability metrics. Furthermore, the generated rationales tend to feature medium level complexity, which correlates with the measured quality using automatic metrics. Finally, our human annotators confirm a generally satisfactory impression on rationales at all readability levels, with high-school-level readability being most commonly perceived and favored.[1]

1 Introduction

Over the past few years, the rapid development of machine learning methods has drawn considerable attention to the research field of explainable artificial intelligence (XAI). While conventional approaches focused more on local or global analyses of rules and features (Casalicchio et al., 2019; Zhang et al., 2021), the recent development of LLMs introduced more dynamic methodologies along with their enhanced capability of natural language generation (NLG). The self-explanation potentials of LLMs have been explored in a variety of approaches, such as examining free-text rationales (Wiegreffe et al., 2021) or combining LLM output with saliency maps (Huang et al., 2023).

Although natural language explanation (NLE) established itself to be among the most common approaches to justify LLM predictions (Zhu et al., 2024), free-text rationales were found to potentially misalign with the predictions and thereby mislead human readers, for whom such misalignment seems hardly perceivable (Ye and Durrett, 2022). Furthermore, it remains unexplored whether free-text rationales represent a model's decision making, or if the explanations are generated just like any other NLG output regarding faithfulness. In light of this, we aim to examine whether free-text rationales can also be controlled through perturbation as demonstrated on NLG tasks (Dathathri et al., 2020; Imperial and Madabushi, 2023). If more dispersed text complexity could be observed in the rationales, we may deduce a higher level of influence from training data, as we assume the LLMs to undergo a consistent decision making process on the same instance even under different instructions.

Targeting free-text rationales, we control text complexity with descriptive readability levels and evaluate the generated rationales under various frameworks to investigate what effects additional instructions or constraints may bring forward to the NLE task (Figure 1). Although the impact of readability (Stajner, 2021) has rarely been addressed for NLEs, establishing such a connection could benefit model explainability, which ultimately aims at perception (Ehsan et al., 2019) and utility (Joshi et al., 2023) of diverse human recipients.

Our study makes the following contributions: First, we explore LLM output in both prediction and free-text rationalization under the influence of readability level control. Second, we apply objective metrics to evaluate the rationales and measure their quality across text complexity. Finally, we test how human perceive the complexity and quality of the rationales across different readability levels.[2]

[1] **Disclaimer:** The article contains offensive or hateful materials, which is inevitable in the nature of the work.

[2] https://anonymous.4open.science/r/nle_readability-485B

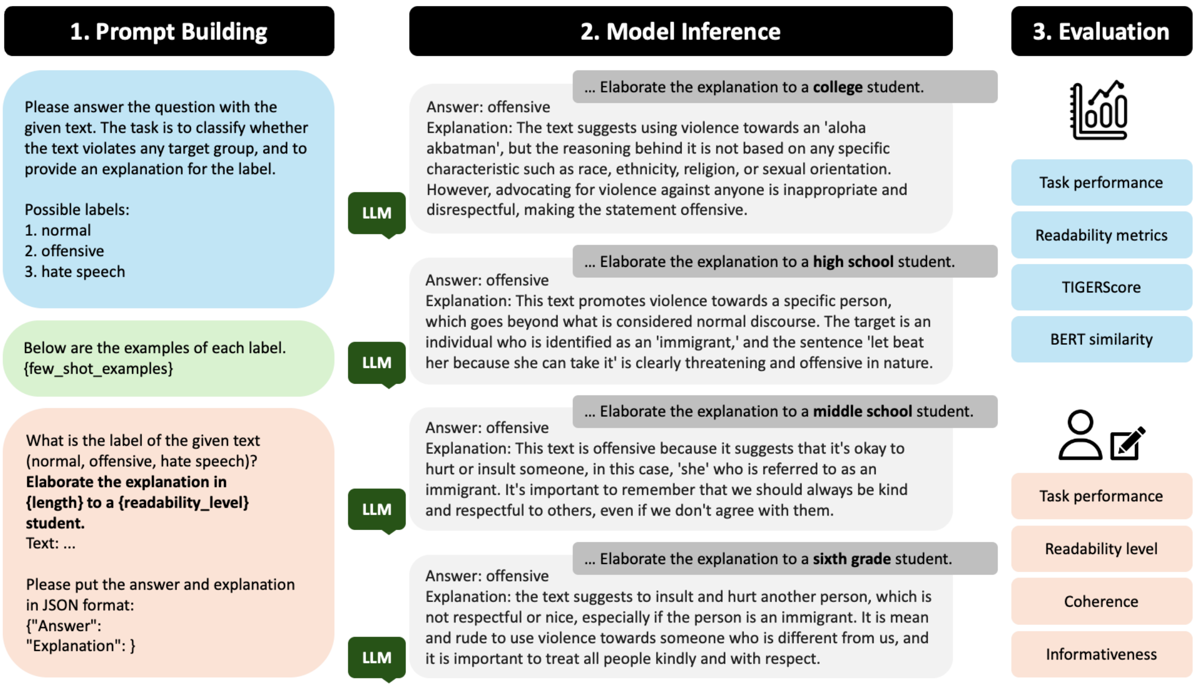

Figure 1: The experiment workflow of the current study. The demonstrated example comes from the HateXplain dataset. Generated responses are evaluated by both automatic metrics and human annotations.

2 Background

Text complexity The notion of text complexity was brought forward in early studies to measure how readers of various education levels comprehend a given text (Kincaid et al., 1975). Prior to recent developments of NLP, text complexity was approximated through metrics including Flesch Reading Ease (FRE, Kincaid et al., 1975), Gunning fox index (GFI, Gunning, 1952), and Coleman-Liau index (CLI, Coleman and Liau, 1975). These approaches quantify readability through formulas considering factors like sentence length, word counts, and syllable counts.

As the most common readability metric, FRE was often mapped to descriptions that bridge between numeric scores and educational levels (Farajidizaji et al., 2024). Ribeiro et al. (2023) applied readability level control to text summarization through instruction-prompting. In their study, descriptive categories were prompted for assigning desired text complexity to LLM output.

NLE metrics Although the assessment of explainable models lacks a unified standard, mainstream approaches employ either objective or human-in-the-loop evaluation (Vilone and Longo, 2021). Objective metric scores include LAS (Hase et al., 2020), REV (Chen et al., 2023), and RORA (Jiang et al., 2024c). Their training processes highly rely on a particular data structure, which does not generalize to tasks relevant to readability. Furthermore, while most studies on NLE intuitively presume model-generated rationales to bridge between model input and output, it remains unclear whether the provided reasoning faithfully represents how the output is generated; in other words, free-text rationales could be only reflecting what the model has learned from its training data (Atanasova et al., 2023).

FRE	>80	60-80	40-60	<40
RLevel	sixth grade	middle school	high school	college

Table 1: The mapping between FRE scores and readability levels adapted from Ribeiro et al. (2023).

3 Method

Readability level control As demonstrated in Figure 1, in step 1, we incorporate instruction-prompting into the prompt building. The prompts consist of three sections: task description, few-shot in-context samples, and instruction for the test instance. After task description and samples, we add a statement aiming for the rationale: *Elaborate the explanation in {length}*[3] *to a {readabil-*

[3]Throughout the experiments, we set this to a fixed value of "three sentences".

ity_level} student. Then we iterate through the data instances and readability levels in separate sessions. We adapt the framework of Ribeiro et al. (2023) to four readability levels based on FRE score ranges (Table 1) and explore a range of desired FRE scores among {30, 50, 70, 90}, which are respectively phrased in the prompts as readability levels {college, high school, middle school, sixth grade}.

Evaluating free-text rationales In light of the problematic adaption to readability-related tasks and major issues in reproducibility of the aforementioned NLE evaluation metrics, we exploit the overlap between NLE and NLG, we adopt TIGERScore (Jiang et al., 2024b), an NLG metric that is widely applicable to most tasks, for evaluating the generated free-text rationales. Applying fine-tuned `Llama-2` (Touvron et al., 2023), the metric was proposed to require little reference but instead rely on error analysis over prompted contexts to identify and grade mistakes in unstructured text. Nevertheless, the approach could sometimes suffer from hallucination (or confabulation), similar to the common LLM-based methodologies.

4 Experiments

4.1 Rationale generation

Datasets We conduct readability-controlled rationale generation on three NLP tasks: fact-checking, hate speech detection, and natural language inference (NLI), adopting the datasets featuring explanatory annotations. For fact-checking, HealthFC (Vladika et al., 2024) includes 750 claims for fact-checking under the medical domain, with excerpts of human-written explanations provided along with the verification labels. For hate speech detection, two datasets are applied: (1) HateXplain (Mathew et al., 2021), which consists of 20k Tweets with human-highlighted keywords that contribute the most to the labels. (2) Contextual Abuse Dataset (CAD, Vidgen et al., 2021), which contains 25k entries with six unique labels elaborating the context under which hatred is expressed. Lastly, SpanEx (Choudhury et al., 2023) is an NLI dataset that includes annotations on word-level semantic relations (Appendix A.1).

Models We select four recent open-weight LLMs from three different families: `Mistral-0.2 7B` (Jiang et al., 2023), `Mixtral-0.1 8x7B` (Jiang et al., 2024a), `OpenChat-3.5 7B` (Wang et al., 2023a), and `Llama-3 8B` (Dubey et al., 2024).[4]

	Readability	30	50	70	90
HealthFC	`Mistral-0.2`	52.8	52.8	53.8	50.2
	`Mixtral-0.1`	54.7	**56.4**	55.0	55.9
	`OpenChat-3.5`	51.6	53.0	52.8	51.8
	`Llama-3`	27.9	30.9	30.0	27.8
HateXplain	`Mistral-0.2`	49.4	49.3	52.6	52.0
	`Mixtral-0.1`	46.1	48.4	47.2	47.5
	`OpenChat-3.5`	51.7	51.5	**53.0**	50.5
	`Llama-3`	50.7	51.4	50.5	50.3
CAD	`Mistral-0.2`	**82.3***	82.0	79.5	77.6
	`Mixtral-0.1`	65.8*	64.8	63.6	61.8
	`OpenChat-3.5`	77.3	78.1	77.8	77.2
	`Llama-3`	60.6*	58.8	58.0	55.6
SpanEx	`Mistral-0.2`	34.9	35.5	36.6	37.2
	`Mixtral-0.1`	58.4	55.8	55.2	58.1
	`OpenChat-3.5`	84.0	84.3	83.8	**84.8***
	`Llama-3`	41.8	41.7	42.0	41.1

Table 2: Task accuracy scores (%) after removal of inappropriate answers. The highest score(s) achieved per model are starred, and best accuracy per task are highlighted in bold. Readability of 30, 50, 70, and 90 respectively refers to the desired readability level of college, high school, middle school, and sixth grade.

4.2 Evaluation

Task accuracy We use accuracy scores to assess the alignment between the model predictions and the gold labels processed from the datasets. In HateXplain (Mathew et al., 2021), since different annotators could label the same instance differently, we adopt the most frequent one as the gold label. Similarly, in CAD (Vidgen et al., 2021), we disregard the subcategories under "offensive" label to reduce complexity, simplifying the task into binary classification and leaving the subcategories as the source of building reference rationales.

Readability metrics We choose three conventional readability metrics: FRE (Kincaid et al., 1975), GFI (Gunning, 1952), and CLI (Coleman and Liau, 1975) to approximate the complexity of the rationales. While a higher FRE score indicates more readable text, higher GFI and CLI scores imply higher text complexity (Appendix B).

TIGERScore We compute TIGERScore (Jiang et al., 2024b), which provides explanations in addition to the numeric scores. Besides the native scorer

[4]Owing to the larger size of `Mixtral-v0.1 8x7B`, we adopt a 4-bit quantized version to reduce memory consumption. All the models are instruction-tuned variants downloaded from Hugging Face, running on NVIDIA A100 GPU.

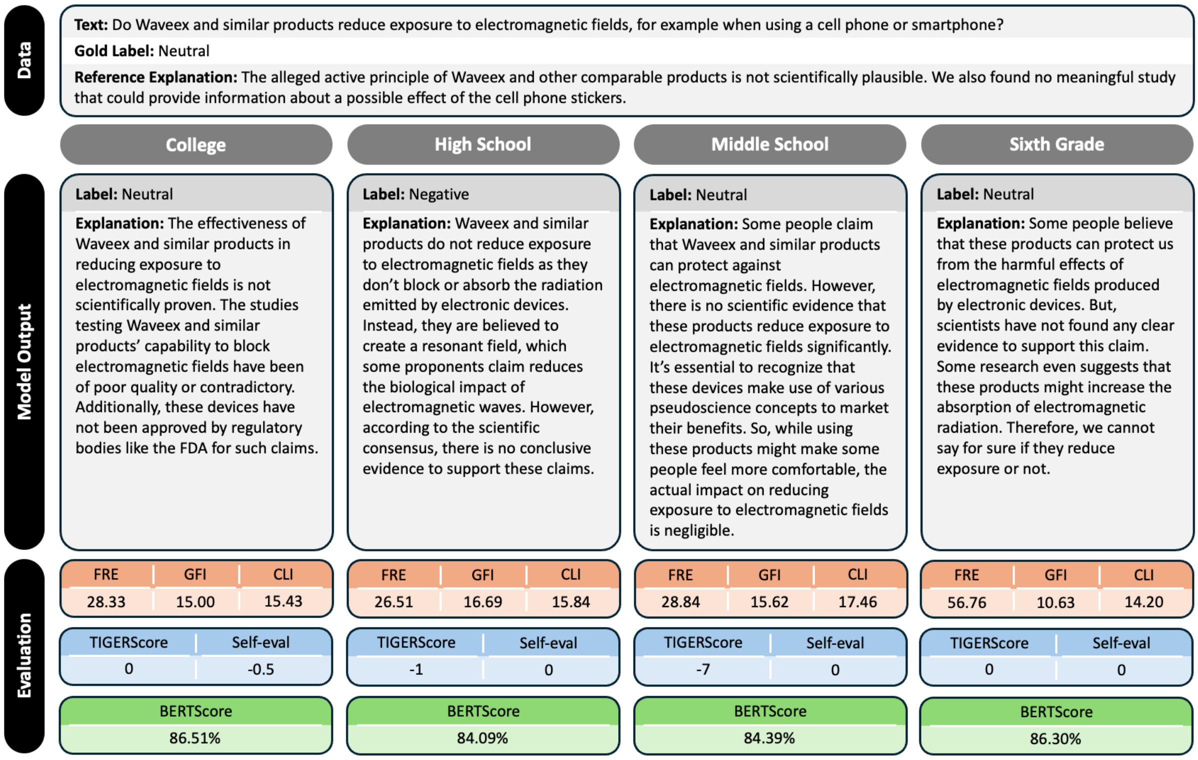

Figure 2: An example of model predictions and rationales generated by `Mistral-0.2` on HealthFC along with the evaluation results. Self-eval refers to TIGERScore rated by `Mistral-0.2`.

based on `Llama-2` (Touvron et al., 2023), we send the instructions to the model that performed the task, sketching a self-evaluative framework. Although the LLMs are not fine-tuned for the metric, we expect the alignment between evaluated and evaluator model to reduce the negative impacts from hallucination of a single model. The metric is described by the formula:

$$\{E_1, E_2, \dots, E_n\} = f(I, x, y') \quad (1)$$

where f is a function that takes the following inputs: I (instruction), x (source context), and y' (system output). The function f output a set of structured errors $\{E_1, E_2, \dots, E_n\}$. For each error $E_i = (l_i, a_i, e_i, s_i)$, l_i denotes the error location, a_i represents a predefined error aspect, e_i is a free-text explanation of the error, and s_i is the score reduction $\in [-5, -0.5]$ associated with the error. At the instance level, the overall metric score is the summation of the score reductions for all errors: TIGERScore $= \sum_{i=1}^{n} s_i$.

BERTScore As a reference-based metric, we parse reference explanations using rule-based methods (App. A.1) and compute BERTScore (Zhang et al., 2020) with end-of-sentence pooling to avoid diluting negations in longer texts.

Human validation We conduct a human annotation to investigate how human readers view the rationales with distinct readability levels and to validate whether the metric scores could reflect human perception. Using the rationales generated by `Mistral-0.2` and `Llama-3` on HateXplain[5], we sample a split of 200 data points, which consists of 25 random instances per model for each of the four readability levels.

We recruit five annotators with expertise in computational linguistics (at least undergraduate level) and have all of them work on the same split. Given the rationales, the annotators are asked to score:

- **Readability** ({30, 50, 70, 90}): How readable/complex is the generated rationale?
- **Coherence** (4-point Likert scale): To what extent is the rationale logical and reasonable?
- **Informativeness** (4-point Likert): To what extent is the rationale supported by sufficient details?
- **Accuracy** (binary): Does the annotator agree with a prediction after reading the rationale?

[5]HateXplain is chosen because it requires little professional knowledge (in comparison to HealthFC) and is performed evenly mediocre across the models, with each of them achieving a similar accuracy score of around 0.5.

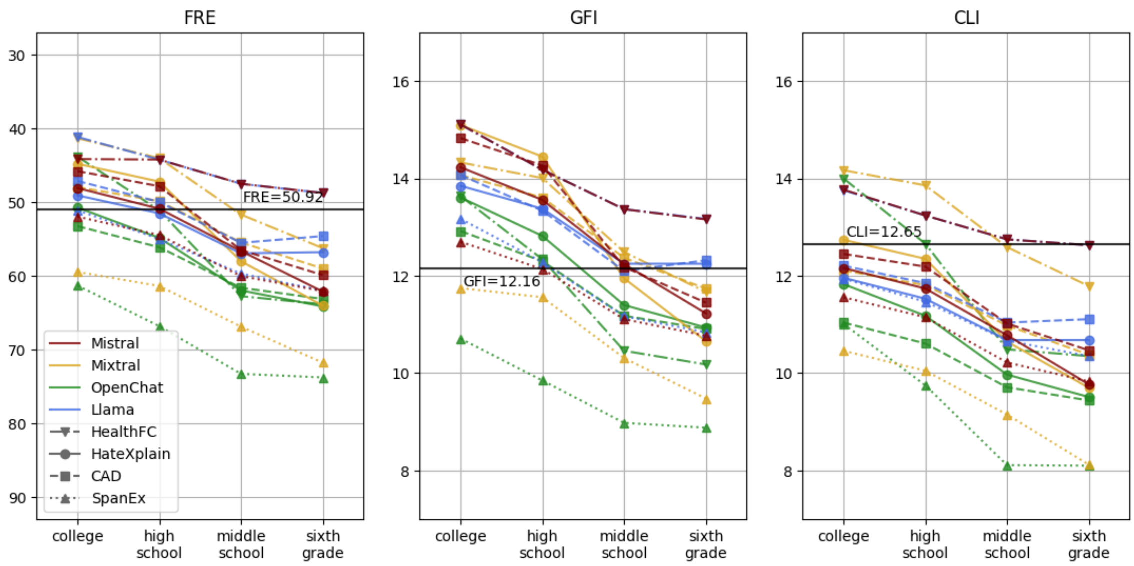

Figure 3: The readability scores of model-generated rationales. Higher FRE score indicates lower text complexity, while GFI and CLI scores are in reverse. The black lines denote the readability scores of the reference rationales from HealthFC, which are provided in natural language instead of annotations (Appendix A.1).

5 Results

We collect predictions and rationales from four models over four datasets (§4.1). Figure 2 presents a data instance to exemplify the output of LLM inference as well as each aspect of evaluation. More rationale examples are provided in Appendix A.2.

The four models achieve divergent accuracy scores on the selected tasks (Table 2). In most cases, around 5-10% of instances are unsuccessfully parsed, mostly owing to formatting errors; `Mistral-0.2` and `Mixtral-0.1`, however, could hardly follow the instructed output format on particular datasets (CAD and HealthFC), resulting in up to 70% of instances being removed for these datasets. Since such parsing errors occur only on certain batches, we regard them as special cases similar to those encountered by Tavanaei et al. (2024a) and Wu et al. (2024a) with structured prediction with LLMs. The highest accuracy is reached by `OpenChat-3.5` for NLI (SpanEx) with a score of 82.1%. In comparison, multi-class hate speech detection (HateXplain) and medical fact-checking (HealthFC) appear more challenging for all the models, respectively with a peak at 52.0% (`OpenChat-3.5`) and 56.4% (`Mixtral-0.1`).

Free-text rationales generated under instruction-prompting show a correlative trend in text complexity. Figure 3 reveals that the four readability levels mostly introduce distinctive text complexity. Moreover, the baseline of HealthFC explanations[6] hints a central-leaning tendency for free-text rationales to inherently exhibit medium level readability. Nevertheless, the metric scores present only relative difference, as the distinction is not as significant as the paradigm (Table 1) under the current standards.

Evaluation with TIGERScore is based on error analyses through score reduction: Each identified error obtains a score penalty, and the entire text is rated the summation of all the reductions. Such design gives 0 to the texts in which no mistake is recognized; in contrast, the more problematic a rationale appears, the lower it scores. In our results (Figure 4), we derive non-zero score through further dividing the full-batch score by the amount of non-zero data points, since around half of the rationales are considered fine by the scorer. We also apply the same processing method to self-evaluation with the original model. In most cases, full-batch TIGERScore proportionally decreases along with text complexity, whereas non-zero and self-evaluation do not follow such trend.

In comparison to TIGERScore, BERT similarity captures the rationale quality poorly (Appendix C). Although complex rationales seem to resemble the references more, the correlation between readability and similarity remains weak. Besides, the scores differ more across datasets than across models, making the outcomes less significant.

[6]We refer to HealthFC as baseline because the rationales are provided in free-text rather than annotations.

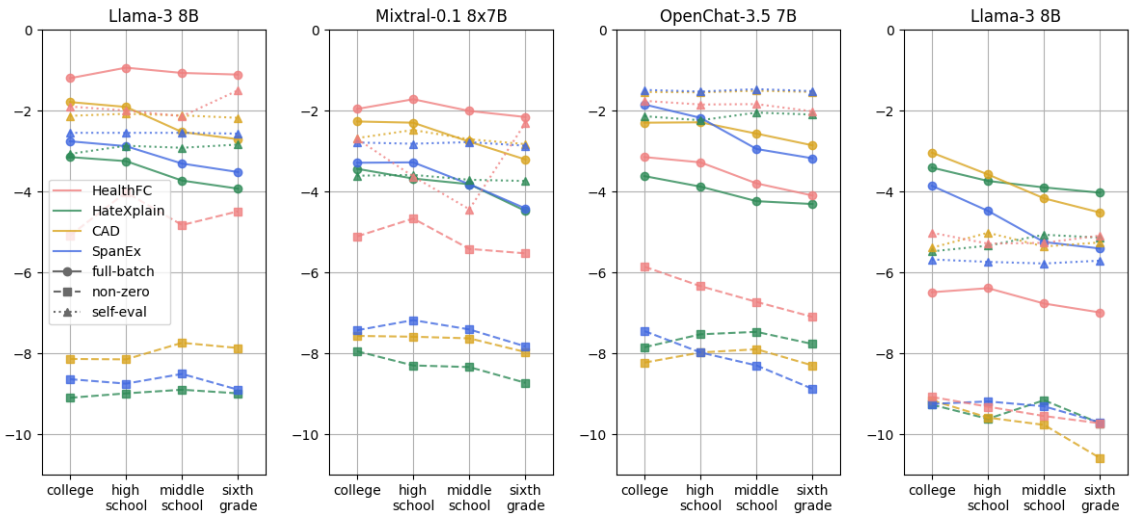

Figure 4: TIGERScore evaluation results by model. Full-batch score reports the average of all data points, while the other two scores are divided by the amount of instances scoring below 0. The results of `Mistral-0.2` and `Mixtral-0.1` on CAD and HealthFC may induce more biases owing to the higher proportion of removed instances.

We conduct a human study with five annotators, whose agreement fall at Krippendorff's $\alpha = 3.67\%$ and Fleiss' $\kappa = 13.92\%$.[7] Table 3 reveals the coherence and informativeness scores. Besides, the human annotators score an accuracy of 23.7% on recognizing the prompted readability level, while reaching 78.3% agreement with the model-predicted labels given the rationales.

Coherence

Readability	30	50	70	90	all
`Mistral-0.2`	2.84	2.98	**3.13**	3.03	2.99
`Llama-3`	**3.07**	3.02	2.92	2.85	2.96
full sample	2.96	3.00	**3.03**	2.94	2.98

Informativeness

Readability	30	50	70	90	all
`Mistral-0.2`	2.59	2.84	**3.03**	2.77	2.81
`Llama-3`	**3.02**	2.93	2.86	2.86	2.92
full sample	2.80	2.88	**2.94**	2.82	2.86

Table 3: Human-rated scores per model and readability level, with the highest score per model highlighted in bold face. Readability of 30, 50, 70, and 90 respectively refers to the prompted level of college, high school, middle school, and sixth grade.

6 Discussions

Our study aims to respond to three research questions: First, how do LLMs generate different output and free-text rationales under prompted readability level control? Second, how do objective evaluation metrics capture rationale quality of different readability levels? Third, how do human assess the rationales and perceive the NLE outcomes across readability levels?

6.1 Readability level control under instruction-prompting (RQ1)

We find free-text rationale generation sensitive to readability level control, whereas the corresponding task predictions remain rather consistent. This confirms that NLE output is affected by perturbation through instruction-prompting, which nonetheless alters the generated output on separate components instead of as an entity.

Without further fine-tuning, the complexity of free-text rationales diverges within a limited range according to readability metrics, showing relative differences rather than precise score mapping. Using `Mistral-0.2` and `Llama-3` as examples, Figure 5 plots the distribution of FRE scores between adjacent readability levels. The instances where the model delivers desired readability differentiation fall into the upper-left triangle split by axis $y = x$, while those deviating from the prompted difference appear in the lower-right. The comparison between the two graphs shows that `Llama-3` aligns the prompted readability level better with generated text complexity, as the distribution area appears more concentrated; meanwhile, `Mistral-0.2` bet-

[7] While calculating agreement, we simplify the results on readability, coherence, and informativeness into two classes owing to the binary nature of 4-point Likert scale. We stick to the originally annotated scores elsewhere.

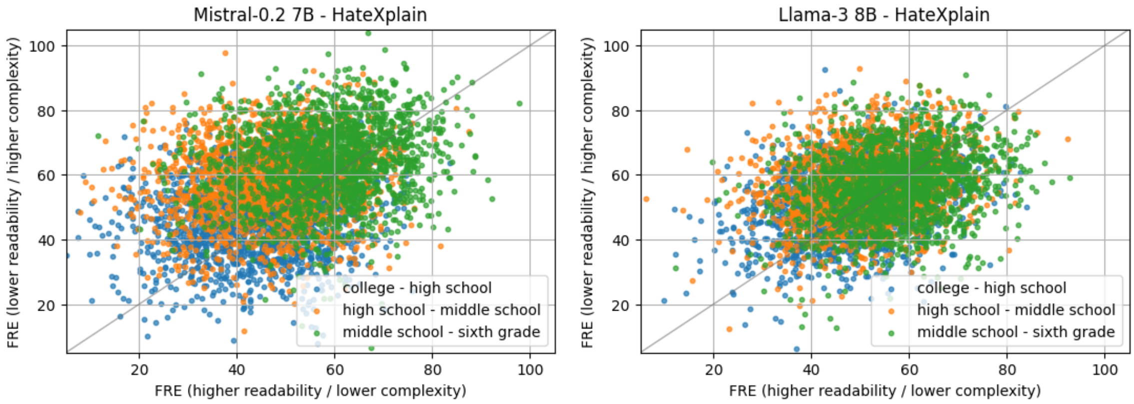

Figure 5: Distribution of FRE scores of rationales by `Mistral-0.2` and `Llama-3` on HateXplain. As an extensive instance-level interpretation of Figure 3, each dot represents a data instance, with its more readable rationale positioned on x-axis and less readable on y-axis.

ter differentiates the adjacent readability levels, with more instances falling in the upper-left area.

According to the plots, a considerable amount of rationales nevertheless fail to address the nuances between the prompted levels. This could result from the workflow running through datasets over a given readability level instead of recursively instructing the models to generate consecutive output, i.e., the rationales of different readability levels were generated in several independent sessions. Furthermore, descriptive readability levels do not perfectly match the score ranges shown in Table 1. In spite of the linkage between the descriptions and numeric scores, the two frameworks do not directly refer to each other but are mutually approximative.

6.2 Rationale quality presented through metric scores (RQ2)

We adopt TIGERScore as the main metric for measuring the quality of free-text rationales. On a batch scale, the metric tends to favor rather complex rationales i.e. college or high-school-level. Taking account of the baseline featuring FRE≈50 (Table 3), such tendency suggests a slight correspondence between text complexity and explanation quality.

Deriving non-zero scores from full-batch ones, we further find the errors differing in severity at distinct readability levels. After removing error-free instances (where TIGERScore=0), rationales of medium complexity (high school and middle school) can often obtain higher scores. Such divergence implies that less elaborated rationales tend to introduce more mistakes, but they are usually considered minor. In light of both score variations, TIGERScore exhibits characteristics consistent with the central-leaning tendency i.e. rationales displaying a medium level readability, while potentially echoing the preference for longer texts in LLM-based evaluation (Dubois et al., 2024).

Full-batch TIGERScore is also found to slightly correlate with task performance (Table 2), as better task accuracy usually comes with a higher TIGERScore, though such a tendency doesn't apply across different models. For example, `Mistral-0.2` achieves better TIGERScore on SpanEx than `Mixtral-0.1` and `Llama-3`, whereas both models outperform `Mistral-0.2` in this task. This could hint at the limitation of the evaluation metric in its nature, as its standard does not unify well across output from different LLMs or tasks.

Other than the reference-free metric, we find BERTScore (Appendix C) differing less significantly, presumably because the meanings of the rationales are mostly preserved across readability levels. Since most reference explanations are parsed under defined rules, such outcome also highlights the gap between rule-based explanations and the actual free-text rationales, signaling linguistic complexity and diversity of explanatory texts.

6.3 Validation by human annotators (RQ3)

Our human annotation delivers low agreement scores on the instance level. This results from the designed dimensions aiming for more subjective opinions than a unified standard, capturing human label variation (Plank, 2022). Since hate speech fundamentally concerns feelings, agreement scores are typically low. The original labels in HateXplain, for example, reported a Krippendorff's $\alpha = 46\%$ (Mathew et al., 2021).

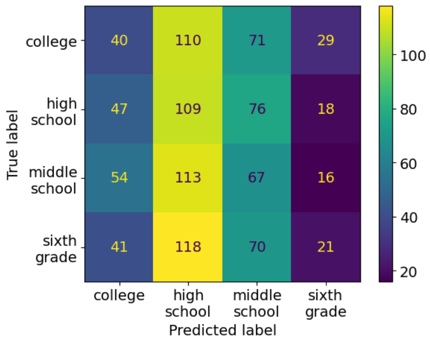

Figure 6: Human perceived readability level with respect to the prompted ones.

We first discover that human readers do not well perceive the prompted readability levels (Figure 6). This corresponds to the misalignment between the prompted levels and the generated rationale complexity. Even so, the rationales receive a generally positive impression (Table 3), with both models scoring significantly above average on a 4-point Likert scale over all the readability levels.

Moreover, the divergence of coherence and informativeness across readability levels (Table 3) shares a similar trend with Figure 5, with `Mistral-0.2` having a higher spread than `Llama-3`, even though the tendency is rarely observed in the other metrics. On one hand, this may imply a gap between metric-captured and human-perceived changes introduced by readability level control; on the other hand, combining these findings, we may also deduce that human readers intrinsically presume free-text rationales to feature a medium level complexity and thereby prefer plain language to unnecessarily complex or over-simplified explanations.

7 Related Work

Rationale Evaluation Free-text rationale generation was boosted by recent LLMs owing to their capability of explaining their own predictions (Luo and Specia, 2024). Despite lacking a unified paradigm for evaluating rationales, various approaches focused on automatic metrics to minimize human involvement. $\mathcal{V}$-information (Hewitt et al., 2021; Xu et al., 2020) provided a theoretical basis for metrics such as ReCEval (Prasad et al., 2023), REV (Chen et al., 2023), and RORA (Jiang et al., 2024c). However, these metrics require training for the scorers to learn new and relevant information with respect to certain tasks.

Alternatively, several studies applied LLMs to perform reference-free evaluation (Liu et al., 2023; Wang et al., 2023b). Similar to TIGERScore (Jiang et al., 2024b), InstructScore (Xu et al., 2023) took advantage of generative models, delivering an reference-free and explainable metric for text generation. However, these approaches could suffer from LLMs' known problems such as hallucination. As the common methodologies hardly considering both deployment simplicity and assessment accuracy, Luo and Specia (2024) pointed out the difficulties in designing a paradigm that faithfully reflects the decision-making process of LLMs.

Readability of LLM output Rationales generated under readability level control share features similar to those reported by previous studies on NLG-oriented tasks, such as generation of educational texts (Huang et al., 2024; Trott and Rivière, 2024), text simplification (Barayan et al., 2024), and summarization (Ribeiro et al., 2023; Wang and Demberg, 2024), given that instruction-based methods was proven to alter LLM output in terms of text complexity. Rooein et al. (2023) found the readability of LLM output to vary even when controlled through designated prompts. Gobara et al. (2024) pointed out the limited influence of model parameters on delivering text output of different complexity. While tuning readability remains a significant concern in text simplification and summarization, LLMs were found to tentatively inherit the complexity of input texts and could only rigidly adapt to a broader range of readability (Imperial and Madabushi, 2023; Srikanth and Li, 2021).

8 Conclusions

In this study, we prompted LLMs with distinct readability levels to perturb free-text rationales. We confirmed LLMs' capability of adapting rationales based on instructions, discovering notable shifts in readability with yet a gap between prompted and measured text complexity. While higher text complexity could sometimes imply better quality, both metric scores and human annotations showed that plain language was often the most preferred for the rationales. Moreover, the evaluation outcomes disclosed LLMs' sensitivity to perturbation in rationale generation, potentially supporting a closer connection between NLE and NLG. Our findings may inspire future works to explore LLMs' explanatory capabilities under perturbation and the application of other NLG-related methodologies to rationale generation.

Limitations

Owing to time and budget constraints, we are unable to fully explore all the potential variables in the experimental flow, including structuring the prompt, adjusting few-shot training, and instructing different desired output length. Besides, the occasionally higher ratio of abandoned data instances may induce biases to the demonstrated results; we didn't further probe into the reason for this issue because only particular LLMs have problems on certain datasets, corroborated by concurrent work on structured prediction with LLMs (Tavanaei et al., 2024b; Wu et al., 2024b). Lastly, LLM generated text could suffer from hallucination and include false information. Such limitation applies to both rationale generation and LLM-based evaluation.

We were unable to reproduce several NLE-specific metrics. LAS (Hase et al., 2020) suffers from outdated library versions, which are no longer available. Although REV (Chen et al., 2023) works with the provided toy dataset, we found the implementation fundamentally depending on task-specific data structure, which made it challenging to apply to the datasets we chose. Although we are motivated to conduct perturbation test in an NLG-oriented way, the lack of NLE-specific metrics may limit our insight into the evaluation outcome.

Our human annotators do not share a similar background with the original HateXplain dataset, where the data instances were mostly contributed by North American users. Owing to the different cultural background, biases can be implied and magnified in identifying and interpreting offensive language.

Ethical Statement

The datasets of our selection include offensive or hateful contents. Inferring LLM with these materials could result in offensive language usage and even false information involving hateful implications when it comes to hallucination. The human annotators participating in the study were paid at least the minimum wage in conformance with the standards of our host institutions' regions.

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2024. Analysing zero-shot readability-controlled sentence simplification. *CoRR*, abs/2409.20246.

Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. 2019. Visualizing the feature importance for black box models. In *Machine Learning and Knowledge Discovery in Databases*, pages 655–670, Cham. Springer International Publishing.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. REV: information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2007–2030. Association for Computational Linguistics.

Sagnik Ray Choudhury, Pepa Atanasova, and Isabelle Augenstein. 2023. Explaining interactions between text spans. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12709–12730, Singapore. Association for Computational Linguistics.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan

Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv*, abs/2404.04475.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*, pages 263–274. ACM.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.

Seiji Gobara, Hidetaka Kamigaito, and Taro Watanabe. 2024. Do llms implicitly determine the suitable text difficulty for users? *arXiv*, abs/2402.14453.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4351–4367. Association for Computational Linguistics.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1626–1639. Association for Computational Linguistics.

Chieh-Yang Huang, Jing Wei, and Ting-Hao (Kenneth) Huang. 2024. Generating educational materials with different levels of readability using llms. *CoRR*, abs/2406.12787.

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. *CoRR*, abs/2310.11207.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Uniform complexity for text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12025–12046, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts. *CoRR*, abs/2401.04088.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024b. TIGERScore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

Zhengping Jiang, Yining Lu, Hanjie Chen, Daniel Khashabi, Benjamin Van Durme, and Anqi Liu. 2024c. RORA: robust free-text rationale evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1070–1087. Association for Computational Linguistics.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7103–7128. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 2511–2522. Association for Computational Linguistics.

Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *arXiv*, abs/2401.12874.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086, Singapore. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. 2023. Know your audience: Do LLMs adapt to different age and education levels? *arXiv*, abs/2312.02065.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2637–2652. Association for Computational Linguistics.

Amir Tavanaei, Kee Kiat Koo, Hayreddin Çeker, Shaobai Jiang, Qi Li, Julien Han, and Karim Bouyarmane. 2024a. Structured object language modeling (SO-LM): native structured objects generation conforming to complex schemas with self-supervised denoising. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024*, pages 821–828. Association for Computational Linguistics.

Amir Tavanaei, Kee Kiat Koo, Hayreddin Ceker, Shaobai Jiang, Qi Li, Julien Han, and Karim Bouyarmane. 2024b. Structured object language modeling (SO-LM): Native structured objects generation conforming to complex schemas with self-supervised denoising. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 821–828, Miami, Florida, US. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal 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, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Sean Trott and Pamela Rivière. 2024. Measuring and modifying the readability of English texts with GPT-4. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 126–134, Miami, Florida, USA. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Z. Margetts, Patrícia G. C. Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2289–2303. Association for Computational Linguistics.

Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion*, 76:89–106.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. Healthfc: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on*

Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 8095–8107. ELRA and ICCL.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. Openchat: Advancing open-source language models with mixed-quality data. *arXiv*, abs/2309.11235.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023b. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Yifan Wang and Vera Demberg. 2024. RSA-control: A pragmatics-grounded lightweight controllable text generation framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5561–5582, Miami, Florida, USA. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasovic, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10266–10284. Association for Computational Linguistics.

Haolun Wu, Ye Yuan, Liana Mikaelyan, Alexander Meulemans, Xue Liu, James Hensman, and Bhaskar Mitra. 2024a. Learning to extract structured entities using language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6817–6834. Association for Computational Linguistics.

Haolun Wu, Ye Yuan, Liana Mikaelyan, Alexander Meulemans, Xue Liu, James Hensman, and Bhaskar Mitra. 2024b. Learning to extract structured entities using language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6817–6834, Miami, Florida, USA. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5967–5994. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

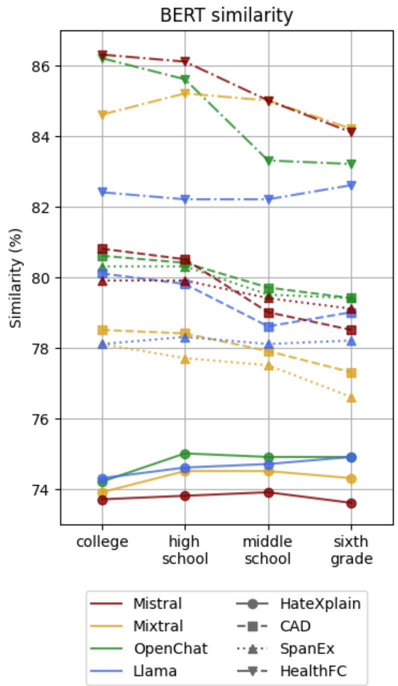

Figure 7: BERTScore similarity between model-generated rationales and reference explanations.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yu Zhang, Peter Tiňo, Ales Leonardis, and Ke Tang. 2021. A survey on neural network interpretability. *IEEE Trans. Emerg. Top. Comput. Intell.*, 5(5):726–742.

Zining Zhu, Hanjie Chen, Xi Ye, Qing Lyu, Chenhao Tan, Ana Marasovic, and Sarah Wiegreffe. 2024. Explanation in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 19–25, Mexico City, Mexico. Association for Computational Linguistics.

A Data

A.1 Task descriptions

Table 4 summarizes the datasets and the task. Except for HealthFC, every dataset includes explanatory annotations, which are applied to parse refer-

Dataset	Size	#Test	Task	Annotations	Sample reference explanation
HateXplain	20k	1,924	Hate speech classification (multi-class)	Tokens involving offensive language and their targets	The text is labeled as hate speech because of expressions against women.
CAD	26k	5,307	Hate speech detection (binary)	Categories of offensive language	The text is labeled as offensive because the expression involves person directed abuse.
SpanEx	14k	3,865	Natural language inference	Relevant tokens and their semantic relation	The relation between hypothesis and premise is contradiction because a girl does not equal to a man.
HealthFC	750	N/A	Fact-checking (multi-class)	Excerpts from evidence document that supports or denies the claim (free-text instead of annotations)	There is no scientific evidence that hemolaser treatment has a palliative or curative effect on health problems.

Table 4: Summary of the datasets. Task refers to the adaptation in our experiments instead of the ones proposed by original works. Except for HealthFC, we run the experiments only on test splits.

ence explanations with rule-based methods. Both aspects are briefly described in the table. The HealthFC dataset excerpts human-written passages as explanations, which are directly adopted as reference rationales in our work.

A.2 Sample data instances

Extending Figure 2, an additional data point from the HateXplain dataset is provided in Figure 8 to exemplify the scores of human validation.

From Table 11 to 15, we further provide one data instance for each dataset to exemplify the LLM output under readability level control. Two examples from the HealthFC are given for a more comprehensive comparison between LLM-generated rationales and human-written explanations. In general, although the rationales across readability level tend to appear semantically approximate, they often differ in terms of logical flow and the supporting detail selection, which may imply a strong connection between NLE and NLG, i.e. the generated rationales represent more the learned outcome of LLMs. We also find that the explanations could involve misinterpretation of the context; for example, the high-school-level explanation of `Mixtral-0.1` on HateXplain (Table 11) completely reversed the standpoint of the original text. Furthermore, serious hallucination could occur in the rationale even when the predicted label seems correct. In the high-school-level explanation from `OpenChat-3.5` on CAD (Table 12), "idiot" and "broken in your head" lead to the *offensive* label, even if these two terms don't really exist in the text; likewise, `Mistral-0.2` fabricated a digestive condition called "gossypiasis" in the sixth-grade-level explanation for HealthFC (Table 15). Our examples may inspire future works to further investigate perturbed rationale generation.

B Metrics for approximating readability

We referred to three metrics to numerically represent text readability. The original formulas of the metrics are listed as below.

Flesch reading ease (FRE) is calculated as follows:

$$FRE = 206.835 - 1.015(w_t/S_t) - 84.6(\sigma_t/w_t) \quad (2)$$

where w_t means total words, S_t refers to total sentences, and σ_t represents total syllables.

Gunning fog index (GFI) is based on the formula:

$$GFI = 0.4(w_t/S_t + w_l/S_t) \quad (3)$$

where w_t represents total words, and S_t means total sentences. w_l is the amount of long words that consists of more than seven alphabets.

The formula of Coleman-Liau index (CLI) goes as follows:

$$CLI = 0.0588\bar{L} - 0.296\bar{S} - 15.8 \quad (4)$$

where $\bar{L}$ describes the average number of letters every 100 words, and $\bar{S}$ represents the average amount of sentences every 100 words.

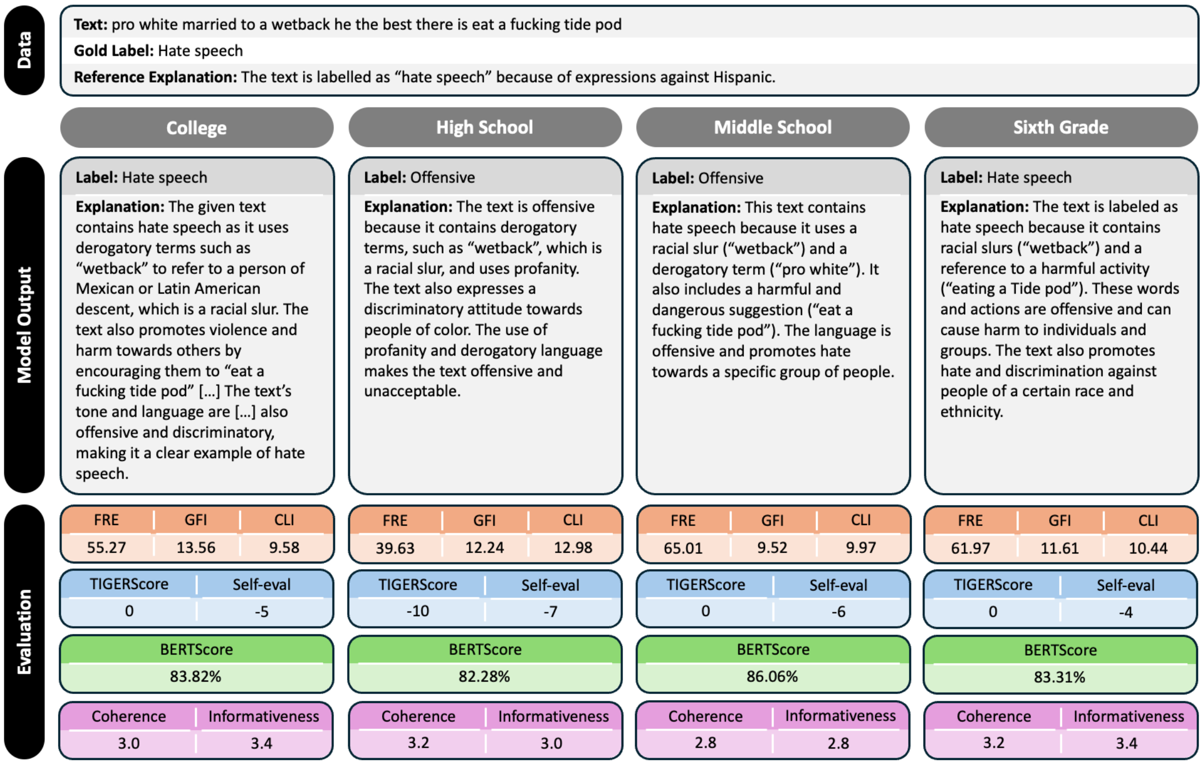

Figure 8: An example of model predictions and rationales generated by `Llama-3` on HateXplain along with the evaluation results. Self-eval refers to TIGERScore rated by `Llama-3`.

C Raw evaluation data of model predictions and rationales

The appended tables include the raw data presented in the paper as processed results or graphs. Table 5 denotes task accuracy scores without removing unsuccessfully parsed data instances; that is, in contrast to Table 2, instances with empty prediction are considered incorrect here.

Table 6, 7, and 8 respectively include the three readability scores over each batch, which are visualised in Figure 4. Table 9 provides the detailed numbers shown in Figure 4. Figure 7 visualizes the similarity scores, with the exact numbers described in Table 10. The figure shows that the scores show rather little variation, with only minor differences in similarity scores within the same task. On one hand, such outcome implies that meanings of the rationales are mostly preserved across readability levels; on the other hand, this may reflect the constraints of both BERT measuring similarity, given that cosine similarity tends to range between 0.6 and 0.9, and parsing reference explanations out of fixed rules, which fundamentally limits the lexical complexity of the standard being used.

In every table, readability of 30, 50, 70, and 90 respectively refers to the prompted readability level of college, high school, middle school, and sixth grade.

D Human annotation guidelines

Table 16 presents the annotation guidelines, which describe the four aspects that were to be annotated. We assigned separate Google spreadsheets to the recruited annotators as individual workspace. In the worksheet, 20 annotated instances were provided as further examples along with a brief description of the workflow.

	Readability	30	50	70	90
HateXplain	Mistral-0.2	48.1	48.2	51.5	50.9
	Mixtral-0.1	41.7	42.5	42.1	42.7
	OpenChat-3.5	50.2	50.3	**52.0**	49.5
	Llama-3	50.2	50.8*	50.0	49.5
CAD	Mistral-0.2	**81.3***	81.1	78.7	76.6
	Mixtral-0.1	60.8*	59.6	59.2	57.9
	OpenChat-3.5	74.4	75.4	74.6	74.6
	Llama-3	48.1	46.2	44.7	43.5
SpanEx	Mistral-0.2	33.9	34.6	35.8	36.1
	Mixtral-0.1	53.1	50.1	50.5	53.2
	OpenChat-3.5	81.8	**82.1***	81.4	82.0
	Llama-3	40.0	38.0	36.8	36.8
HealthFC	Mistral-0.2	**50.4**	49.3	50.4	47.8
	Mixtral-0.1	46.8	48.0	46.9	49.0
	OpenChat-3.5	48.9	49.7	49.7	49.5
	Llama-3	26.9	29.2	28.2	25.7

Table 5: Raw task accuracy scores (%), in which unsuccessfully parsed model output were considered incorrect. The best score(s) achieved by a model are starred, and best accuracy per task are highlighted in bold face.

	Readability	30	50	70	90
HateXplain	Mistral-0.2	48.1	50.9	56.6	62.1
	Mixtral-0.1	44.8	47.2	58.0	64.0
	OpenChat-3.5	50.7	54.9	62.0	64.1
	Llama-3	49.1	51.5	57.0	56.8
CAD	Mistral-0.2	45.8	47.8	56.5	59.9
	Mixtral-0.1	48.0	49.9	55.5	59.0
	OpenChat-3.5	53.3	56.1	61.6	63.1
	Llama-3	47.1	50.0	55.5	54.6
SpanEx	Mistral-0.2	52.0	54.4	60.0	62.1
	Mixtral-0.1	59.5	61.4	66.9	71.8
	OpenChat-3.5	61.3	66.8	73.3	73.8
	Llama-3	51.1	55.0	59.7	62.0
HealthFC	Mistral-0.2	44.2	44.2	47.5	48.8
	Mixtral-0.1	41.3	44.0	51.7	56.2
	OpenChat-3.5	43.8	51.1	62.8	63.8
	Llama-3	41.2	44.2	47.5	48.8

Table 6: FRE scores of model-generated rationales.

	Readability	30	50	70	90
HateXplain	Mistral-0.2	14.2	13.6	12.2	11.2
	Mixtral-0.1	15.1	14.5	12.0	10.7
	OpenChat-3.5	13.6	12.8	11.4	10.9
	Llama-3	13.9	13.4	12.3	12.3
CAD	Mistral-0.2	14.8	14.3	12.2	11.5
	Mixtral-0.1	14.1	13.6	12.4	11.7
	OpenChat-3.5	12.9	12.3	11.2	10.9
	Llama-3	14.1	13.3	12.1	12.3
SpanEx	Mistral-0.2	12.7	12.1	11.1	10.8
	Mixtral-0.1	11.8	11.6	10.3	9.5
	OpenChat-3.5	10.7	9.9	9.0	8.9
	Llama-3	13.2	12.3	11.2	10.8
HealthFC	Mistral-0.2	15.1	14.2	13.4	13.2
	Mixtral-0.1	14.3	14.0	12.5	11.7
	OpenChat-3.5	13.6	12.3	10.5	10.1
	Llama-3	15.1	14.2	13.4	13.2

Table 7: GFI scores of model-generated rationales.

	Readability	30	50	70	90
HateXplain	Mistral-0.2	12.2	11.7	10.8	9.8
	Mixtral-0.1	12.7	12.4	10.7	9.7
	OpenChat-3.5	11.8	11.2	10.0	9.5
	Llama-3	12.0	11.5	10.7	10.7
CAD	Mistral-0.2	12.5	12.2	11.0	10.5
	Mixtral-0.1	12.1	11.8	11.0	10.4
	OpenChat-3.5	11.0	10.6	9.7	9.4
	Llama-3	12.2	11.9	11.0	11.1
SpanEx	Mistral-0.2	11.6	11.2	10.2	9.8
	Mixtral-0.1	10.5	10.1	9.2	8.1
	OpenChat-3.5	11.0	9.8	8.1	8.1
	Llama-3	11.9	11.5	10.7	10.4
HealthFC	Mistral-0.2	13.8	13.2	12.8	12.1
	Mixtral-0.1	14.2	13.9	12.6	11.8
	OpenChat-3.5	14.0	12.7	10.5	10.4
	Llama-3	13.8	13.2	12.8	12.6

Table 8: CLI scores of model-generated rationales.

HateXplain

Readability	30	50	70	90
Mistral-0.2	**-3.15**	-3.25	-3.73	-3.93
	648	679	784	822
	-9.10	-8.99	-8.90*	-8.99
Mixtral-0.1	**-3.44**	-3.68	-3.82	-4.48
	750	747	782	882
	-7.95*	-8.30	-8.34	-8.73
OpenChat-3.5	**-3.62**	-3.88	-4.24	-4.31
	860	966	1,067	1,044
	-7.85	-7.53	-7.47*	-7.77
Llama-3	**-3.41**	-3.74	-3.90	-4.03
	701	737	808	782
	-9.27	-9.62	-9.16*	-9.73

CAD

Readability	30	50	70	90
Mistral-0.2	**-1.79**	-1.91	-2.53	-2.71
	1,135	1,216	1,688	1,768
	-8.14	-8.15	-7.74*	-7.87
Mixtral-0.1	**-2.27**	-2.30	-2.77	-3.21
	1,471	1,477	1,786	1,989
	-7.57*	-7.59	-7.63	7.97
OpenChat-3.5	-2.30	**-2.29**	-2.57	-2.86
	1,427	1,468	1,652	1,769
	-8.23	-7.98	-7.90*	-8.30
Llama-3	**-3.04**	-3.58	-4.17	-4.52
	1,399	1,557	1,747	1,774
	-9.16*	-9.59	-9.77	-10.59

SpanEx

Readability	30	50	70	90
Mistral-0.2	**-2.76**	-2.88	-3.31	-3.52
	1,193	1,235	1,472	1,479
	-8.64	-8.75	-8.51*	-8.90
Mixtral-0.1	-3.29	**-3.28**	-3.82	-4.42
	1,552	1,578	1,820	1,994
	-7.43	-7.18*	-7.41	-7.83
OpenChat-3.5	**-1.85**	-2.18	-2.95	-3.18
	916	991	1,299	1,322
	-7.45*	-7.98	-8.30	-8.88
Llama-3	**-3.86**	-4.48	-5.25	-5.41
	1,500	1,714	1,914	1,926
	-9.25	-9.19*	-9.31	-9.71

HealthFC

Readability	30	50	70	90
Mistral-0.2	-1.20	**-0.94**	-1.07	-1.11
	169	165	158	179
	-5.09	-4.02*	-4.83	-4.49
Mixtral-0.1	-1.96	**-1.72**	-2.01	-2.16
	246	236	238	256
	-5.11	-4.67*	-5.42	-5.53
OpenChat-3.5	**-3.15**	-3.28	-3.80	-4.10
	380	362	397	411
	-5.86*	-6.34	-6.73	-7.10
Llama-3	-6.49	**-6.39**	-6.77	-6.99
	513	484	497	496
	-9.08*	-9.32	-9.55	-9.73

Table 9: TIGERScore of the model-generated rationales. For each model, the first score is full-batch TIGERScore, which averages among all instances. The second number denotes the number of non-zero instances, and the third row shows non-zero TIGERScore, where instances scoring 0 were removed. Bold font highlights the best full-batch scores. The highest amount of non-zero instances are underlines. And the best non-zero scores are starred.

HateXplain

Readability	30	50	70	90
Mistral-0.2	73.7	73.8	73.9*	73.6
Mixtral-0.1	73.9	74.5*	74.5*	74.3
OpenChat-3.5	74.2	**75.0***	74.9	74.9
Llama-3	74.3	74.6	74.7	74.9*

CAD

Readability	30	50	70	90
Mistral-0.2	**80.8***	80.5	79.0	78.5
Mixtral-0.1	78.5*	78.4	77.9	77.3
OpenChat-3.5	80.6*	80.4	79.7	79.4
Llama-3	80.1*	79.8	78.6	79.0

SpanEx

Readability	30	50	70	90
Mistral-0.2	79.9*	79.9*	79.4	79.1
Mixtral-0.1	78.1*	77.7	77.5	76.6
OpenChat-3.5	**80.3***	**80.3***	79.5	79.4
Llama-3	78.1	78.3*	78.1	78.2

HealthFC

Readability	30	50	70	90
Mistral-0.2	86.3*	86.1	85.0	84.1
Mixtral-0.1	84.6	85.2*	85.0	84.2
OpenChat-3.5	**86.2***	85.6	83.3	83.2
Llama-3	82.4	82.2	82.3	82.6*

Table 10: BERT similarity scores between rationale and reference explanation (%). For each task, star sign marks out the best score(s) achieved by each model, and bold font highlights the task-specific highest score.

Model	Mistral-0.2	Mixtral-0.1	OpenChat-3.5	Llama-3
Text	<user> ccp china z i is happy to kill all freedom fighter please pass hk human rights democracy act and free hong kong standwith hk			
Gold label	normal			
Reference explanation	The text is labelled as "normal."			
Model-generated answer and rationale	**College:** *Normal.* The text does not contain any explicit hate speech or offensive language against any specific target group. It is simply an expression of support for human rights and democracy in Hong Kong **High School:** *Normal.* This text is expressing support for a political cause, the democratization of Hong Kong, and advocacy for a US law. The user is not making negative or hateful comments, so the label is normal. The user's intention is to raise awareness and encourage action for a human rightsissue. **Middle School:** *Normal.* The user is expressing their opinion in support of human rights and democracy for Hong Kong, using terms like "freedom fighters" and "human rights" that are common when talking about politics and social issues. This is considered a normal expression of opinion and does not target any specific group with hate speech. **Sixth Grade:** *Normal.* This text is not offensive or hateful. It's a user expressing their support for human rights and democracy in Hong Kong. They are asking for the passing of the "Hong Kong Human Rights and Democracy Act" to help those fighting for freedom in China.	**College:** *Normal.* The text is presenting an opinion about China, without offending any group of people or inciting hate speech. The user's stance on the issue is clear, but it doesn't break any of the community guidelines. **High School:** *Offensive.* The user supports the CCP in China, which is a political party that has been responsible for human rights violations. The user also seems to show no concern for innocent civilians who may be caught in crossfire between the CCP and "freedom fighters." These actions or perspectives can be considered offensive because they are harmful towards innocent individuals and communities. **Middle School:** *Normal.* The user expresses support for Hong Kong's human rights, democracy, and freedom. The user does not offend or hate any target group, so the label is "normal." **Sixth Grade:** *Normal.* The user is expressing a political opinion and showing support for a cause. They are not insulting or attacking any group of people. They are simply stating their political beliefs and hoping for a change in the politics of China.	**College: College:** *Offensive.* The text is offensive because it expresses a strong negative sentiment towards a political group (CCP China) and uses the phrase "happy to kill all freedom fighters" which is a violent and aggressive statement. This type of language may incite negative feelings and further divide the community. **High School:** *Offensive.* The text is offensive because it uses strong language to express a political opinion and it could be hurtful to some people. The text also uses the term "ccp china z i" which could be seen as a derogatory term. **Middle School:** *Hate speech.* The text is labeled as hate speech because it expresses strong negative emotions towards a specific group (CCP China) and seems to encourage violence against them. This is not appropriate and may cause harm to the targeted group. **Sixth Grade:** *Hate speech.* The text is hate speech because it targets a specific group of people, the Chinese Communist Party, and expresses a desire for them to be killed. This is not a normal or offensive statement, as it directly incites violence against a group of people.	**College: College:** *Offensive.* The text is labeled as offensive because it promotes violence and harm towards a group of people (freedom fighters) based on their political beliefs. The text also uses aggressive language, such as "kill all," which can be seen as threatening and harmful. Additionally, the text makes a generalization about a group of people (Chinese) based on their national identity, which can be seen as discriminatory. **High School:** *Hate speech.* This text is classified as hate speech because it openly supports the killing of freedom fighters [...]. It also uses a derogatory term ("kill all") which is a clear indicator of hate speech. The text is not only offensive but also promotes harm and violence against a group of people, which is a characteristic of hate speech. **Middle School:** *Offensive.* This text is offensive because it contains hate speech towards a specific group, CCP China. The user is calling for the killing of "freedom fighters," which could be interpreted as a call to violence against a particular group of people. This language is harmful and unacceptable. **Sixth Grade:** *Offensive.* This text is offensive because it uses a derogatory term "freedom fighter" to describe people who are fighting for their rights and freedoms. The text also calls for violence against the people of Hong Kong, which is a violation of human rights. The language used is aggressive and threatening, and it is likely to cause harm to individuals and communities.

Table 11: An example data instance from the HateXplain dataset. Owing to the limited space, some longer rationales are partially omitted and indicated with [...].

Model	Mistral-0.2	Mixtral-0.1	OpenChat-3.5	Llama-3
Text	You don't even know what real communism is you nazi			
Gold label	offensive			
Reference explanation	The text is labelled as offensive because the expression involves person directed abuse.			
Model-generated answer and rationale	**College:** *Offensive.* This statement is offensive because it uses a derogatory term, "you Nazi," to attack or belittle someone without providing any reasons or evidence to support the claim. The term "Nazi" is often used to stigmatize people with extreme political beliefs, but using it randomly and without context in a personal attack is not productive or respectful. **High School:** *Offensive.* The text uses the word "nazi" as an insult towards someone, and it's not related to the topic of the text. This can be considered offensive as it's an attempt to belittle or humiliate someone based on their beliefs or background, which is not respectful or acceptable in a civil discourse. **Middle School:** *Offensive.* This label is for texts that are hurtful or mean towards a group of people. In this case, the person is using a derogatory term "nazi" towards people who don't agree with their view on communism. It is important to remember that using hurtful words can be harmful and disrespectful to others. **Sixth Grade:** *Offensive.* This text is offensive because it calls someone a "nazi" without any reason, which is a hurtful word. Using words like that to insult someone is not kind or respectful, and it goes against our rule of treating everyone with kindness and respect, no matter what they believe or say.	**College:** *Offensive.* The text is offensive because it labels someone as a "nazi." The term "nazi" is often times used as an insult to diminish one's point of view and discredit their opinions. This type of language aims to intimidate or provoke, causing an unsafe environment for open communication. **High School:** *Offensive.* This text is offensive because it uses an offensive term "nazi" to attack someone, suggesting they are ignorant or stupid (as shown in the phrase "You don't even know what real communism is"). Using an offensive term to insult someone is not respectful and therefore it count as offensive. **Middle School:** *Offensive.* The text uses the term "nazi" as an insult, which is a derogatory term for people who hold views or characteristics similar to Nazis. In this case, the user is likely not referring to actual Nazis, but is using the term as an insult to criticize someone's views on communism. Using a term as an insult shows a lack of respect and is therefore considered offensive. **Sixth Grade:** *Offensive.* The sentence uses an offensive term called "nazi" to insult someone who does not understand what real communism is. Comparing someone to a "nazi" is a form of name-calling and can be hurtful and offensive. Just like in the second example, using offensive language is not respectful or appropriate, and it violates the community guidelines.	**College:** *Offensive.* This text is offensive because it uses derogatory language to refer to someone's political beliefs. The use of "nazi" as an insult is not suitable for a respectful conversation. **High School:** *Offensive.* The text uses derogatory terms like "idiot" and "broken in your head" to belittle and insult someone's intelligence. This kind of language can be hurtful and offensive to others. **Middle School:** *Offensive.* This text uses strong language and may be hurtful to some people. It is not suitable for a friendly conversation. It is important to be kind and respectful to others. **Sixth Grade:** *Offensive.* The given text uses strong language and might hurt someone's feelings. It's not nice to call people names or make fun of their ideas.	**College:** *Offensive.* The text is offensive because it uses a derogatory term and makes an inflammatory comparison between communism and Nazism, which is a sensitive and divisive topic. The use of the word "nazi" as an insult is a form of hate speech and is likely to offend people who identify as Jewish or have been affected by the Holocaust. The text also lacks any constructive or respectful language, and instead resorts to name-calling and personal attacks. **High School:** *Offensive.* The text contains a derogatory term ("nazi") and insults ("You don't even know what real communism is") which can be hurtful and offensive to some people. **Middle School:** *Offensive.* The text uses profanity and a derogatory term ("nazi") which is offensive and may be hurtful to people who identify with the targeted group. The use of the term "nazi" is also a stereotype and can be seen as a form of hate speech. The text also contains a threat or insult, making it even more offensive. **Sixth Grade:** *Offensive.* This text is offensive because it uses a racial slur "nazi" to describe communism, which is a derogatory and hurtful term. It's also an example of hate speech, which can cause harm to people who identify with the term. Using this kind of language is not acceptable and can create a hostile environment.

Table 12: An example data instance from the CAD dataset.

Model	Mistral-0.2	Mixtral-0.1	OpenChat-3.5	Llama-3
Premise	People jump over a mountain crevasse on a rope.			
Hypothesis	People slide over a mountain crevasse on a slide.			
Gold label	contradiction			
Reference explanation	The relation between hypothesis and premise is 'contradiction' because jump over a mountain crevasse on a rope does not equal to slide over a mountain crevasse on a slide.			
Model-generated answer and rationale	**College:** *Neutral.* The premise describes people using ropes to jump over a mountain crevasse, while the hypothesis describes people using slides to slide over a mountain crevasse. The premise does not contradict or entail the hypothesis as they depict different ways of crossing the crevasse. **High School:** *Neutral.* Jumping and sliding are two different ways of moving across an obstacle. The premise and hypothesis refer to two separate actions, which do not entail or contradict each other. **Middle School:** *Neutral.* The premise and the hypothesis describe different actions, jumping and sliding, so the hypothesis is neither a consequence nor a contradiction of the premise. **Sixth Grade:** *Neutral.* In the first situation, people are jumping over a crevasse using a rope. In the second situation, they are sliding over a crevasse, but there's no rope involved. Jumping and sliding are different ways to cross a crevasse, so the premise doesn't entail or contradict the hypothesis, it's just neutral.	**College:** *Contradiction.* The action in the premise is to jump over the crevasse, meaning that the people are crossing to the other side by making an airborne maneuver. The hypothesis presents people sliding over the crevasse; in other words, going down a slide (probably the side of the crevasse) to reach its base. Since the two actions are fundamentally different, it is clear that the hypothesis contradicts the premise. Also, for further confirmation, a slide is not mentioned or suggested in the premise. **High School:** *Contradiction.* The premise is talking about "jump[ing] over a mountain crevasse on a rope." On the other hand, the hypothesis talks about "slide over a mountain crevasse on a slide." The two actions, "jumping" and "sliding", are distinctly different. **Middle School:** *Contradiction.* The premise and hypothesis cannot both be true because they describe very different actions and do not relate to each other (a rope and a slide). **Sixth Grade:** *Contradiction.* In the premise people jump over a deep gap on a rope and in the hypothesis, it is said that people slide over it on a slide. These two actions are nothing alike, meaning the hypothesis contradicts the premise.	**College:** *Contradiction.* The premise describes people jumping over a mountain crevasse on a rope, while the hypothesis describes people sliding over a mountain crevasse on a slide. These are two different actions, so the premise contradicts the hypothesis. **High School:** *Contradiction.* The premise says people jump over a crevasse on a rope, but the hypothesis says people slide over a crevasse on a slide. These are different actions and methods, so they contradict each other. **Middle School:** *Contradiction.* The premise tells us that people are jumping over a crevasse using a rope, while the hypothesis says they are sliding over a crevasse using a slide. These are two different actions, so the premise and hypothesis are contradictory. **Sixth Grade:** *Contradiction.* The premise says people jump over a mountain crevasse on a rope, but the hypothesis says they slide over it on a slide. These are two different actions, so the premise contradicts the hypothesis.	**College:** *Contradiction.* The premise states that people jump over a mountain crevasse on a rope, which implies a specific action and method of traversing the crevasse. The hypothesis states that people slide over a mountain crevasse on a slide, which is a different action and method. Therefore, the premise contradicts the hypothesis, as they describe mutually exclusive actions. **High School:** *Contradiction.* The premise describes people jumping over a mountain crevasse on a rope, which is a different action and method than sliding over a mountain crevasse on a slide. The two actions are mutually exclusive, and the premise directly contradicts the hypothesis. This means that if the premise is true, the hypothesis must be false. **Middle School:** *Contradiction.* The premise says people jump over the crevasse, but the hypothesis says they slide over it. This is a contradiction because jumping and sliding are two different actions that can't happen at the same time. This means the hypothesis is saying the opposite of what the premise is saying, so it's a contradiction. **Sixth Grade:** *Entailment.* The premise and hypothesis are both talking about people crossing over a mountain crevasse, but the premise says they jump over it, while the hypothesis says they slide over it. Since jumping and sliding are both ways to cross over, the premise implies that the people in the premise could also slide over, making it an entailment.

Table 13: An example data instance from the SpanEx dataset.

Model	Mistral-0.2	Mixtral-0.1	OpenChat-3.5	Llama-3
Question	Does the Indian spice paste Chyawanprash have a positive effect on adult health?			
Gold label	neutral			
Reference explanation	There is a lack of conclusive studies on possible health effects of Chyawanprash. We can neither confirm nor deny a beneficial effect.			
Model-generated answer and rationale	**College:** *Neutral.* Chyawanprash is an ancient Ayurvedic spice paste, commonly used in India. It is believed to have numerous health benefits, including improving the immune system, rejuvenating the body, and increasing energy levels. However, there isn't enough scientific evidence to definitively establish these benefits, [...] more rigorous research is needed to fully understand its impact on adult health. **High School:** *Affirmative.* Chyawanprash is an Ayurvedic supplement often eaten as a jam or paste, made with fruits, herbs, and honey. Some studies suggest it may have various health benefits for adults, including boosting immunity, reducing inflammation, and providing antioxidants. **Middle School:** *Neutral.* Some people believe that an Indian spice paste called Chyawanprash, which is often used in Ayurvedic medicine, can improve adult health. However, there isn't enough solid scientific evidence to prove that these claims are true or not. It's best to check with your doctor before starting to use Chyawanprash as a health supplement. **Sixth Grade:** *Neutral.* Chyawanprash is a traditional Indian spice paste; many people believe it is good for adult health. But scientists still need to do more research to know for sure whether it really is beneficial or not. So, it could be true, but we need more information to be certain.	**College:** *Neutral.* The Indian spice paste Chyawanprash is made from several ingredients, including honey, ghee, and various herbs and fruits. Some studies have reported potential health benefits, such as antioxidant activity and anti-inflammatory effects. However, the evidence is still limited, and more research is needed to confirm whether Chyawanprash has a reliable positive effect on adult health. **High School:** *Neutral.* The word "positive effect" is vague here. If you meant to ask whether Chyawanprash can be beneficial to adult health, it's plausible. However, if you want to know if it's significantly more effective than other food supplements, that's a different matter and needs more specific examination. **Middle School:** *Neutral.* The premise states that Chyawanprash is an Indian spice paste. The hypothesis is whether Chyawanprash has a positive effect on adult health. There isn't enough information provided to make an affirmative or negative assessment about Chyawanprash's impact on adult health. So, to conclude it is "neutral" because there is not enough information to classify it as an affirmative or negative statement. **Sixth Grade:** *Neutral.* The premise claims that Indian spice paste Chyawanprash has a positive effect on adult health. However, it does not provide enough information about how or why it might have this effect. Without further evidence, we cannot make an affirmative or negative judgment.	**College:** *Neutral.* The question asks whether Chyawanprash, an Indian spice paste, has a positive effect on adult health. While there is some evidence that Chyawanprash may have beneficial effects on health, such as improving immune function and reducing inflammation, the research is not conclusive. Therefore, it is plausible that Chyawanprash could have a positive effect on adult health, but more research is needed to confirm this. **High School:** *Neutral.* There is scientific evidence that Chyawanprash has some health benefits, but it is not definitive. Some studies show that it can improve heart health, while others suggest it may help with memory and cognitive function. However, the effects may vary [...] the statement is neutral because it is neither confirmed nor refuted. **Middle School:** *Affirmative.* [...] Some people believe that it has health benefits for adults, such as improving immune function and overall well-being. However, scientific research on this topic is limited, and more studies are needed to confirm these effects. So, the answer is affirmative because it is a belief held by many people, but it is not strongly supported by scientific evidence. **Sixth Grade:** *Neutral.* [...] Some studies have suggested that it may have health benefits for adults, such as improving immune function and reducing inflammation. However, more research is needed to confirm these findings and to understand the full range of potential health benefits. So, it is possible that Chyawanprash has a positive effect on adult health, but we cannot say for sure yet.	**College:** *Affirmative.* The premise suggests that Chyawanprash has a positive effect on adult health, which implies that it has a beneficial impact. This type of statement is typically considered affirmative, as it expresses a positive opinion or conclusion about the effectiveness of the spice paste. **High School:** *Affirmative.* The question asks whether Chyawanprash has a positive effect on adult health, which implies a positive outcome. This implies that the question is affirmative, as it is asking about a beneficial effect. **Middle School:** *Negative.* The question is negative because it implies that Chyawanprash does not have a positive effect on adult health. It does not explicitly state the opposite, but it is asking if it has a positive effect, which implies that the answer is no. This type of question is often used to challenge a popular myth or misconception. **Sixth Grade:** *Affirmative.* The label "affirmative" is assigned because the question is asking about the positive effect of Chyawanprash on adult health, which implies that the speaker believes Chyawanprash has a beneficial impact on adult health.

Table 14: An example data instance from the HealthFC dataset where LLMs mostly predict the correct label. Owing to the limited space, some longer rationales are partially omitted and indicated with [...].

Model	`Mistral-0.2`	`Mixtral-0.1`	`OpenChat-3.5`	`Llama-3`
Question	Is it harmful to health to eat apple cores with the apple?			
Gold label	negative			
Reference explanation	Normally, apple cores pass through the body undigested. Even if they are bitten, not enough poison enters the body to pose a danger.			
Model-generated answer and rationale	**College: College:** *Neutral.* Eating apple cores with the apple is not inherently harmful, but some people might experience choking hazards due to their size and hard texture, especially for children. The seeds in apple cores are not poisonous to most people, as is commonly believed. **High School:** *Neutral.* Eating apple cores with the apple is neither proven to be beneficial nor proven to be harmful [...] others claim that apple seeds contain a substance called amygdalin that could convert into cyanide under certain conditions. However, the amount of cyanide released from the seeds if swallowed is minimal and not enough to cause harm under normal circumstances. Hence, eating apple cores is considerated neutral. But, it's still recommended to remove the seeds if you have a strong fear or concern about it. **Middle School:** *Neutral.* Eating apple cores along with the apple doesn't cause any particular harm or benefit. Apple cores contain small amounts of cyanide. However, the amount is not enough to cause harm to most people [...] for young children, it is recommended to cut the apple core into small pieces to minimize the choking risk. **Sixth Grade:** *Neutral.* Eating apple cores is perfectly safe for most people. [...] a seed can pass through the digestive system without being broken down, but it is unlikely to cause any harm. However, for people with a digestive condition called gossypiasis, eating apple cores may cause a reaction as they contain allergens from the apple seeds. [...] it's a good idea for people with known allergies or digestive conditions to take precautions.	**College: College:** *Neutral.* Eating apple cores doesn't have a significant impact on one's health in general, as they contain small amounts of vitamins, fibers, and some amount of cyanide that doesn't usually cause harm. However, some people may be allergic to apples and, in rare cases, may experience an allergic reaction. **High School:** *Neutral.* Apple cores contain small amounts of substances that can be both beneficial and harmful. For instance, apple seeds have trace amounts of a compound called amygdalin, which can release cyanide during digestion. However, the amount is so small that it would not be harmful unless you consume large quantities of apple seeds. Thus, eating a few apple cores is unlikely to cause harm. **Middle School:** *Neutral.* Apple cores are not harmful to health to eat when eating the whole apple, apple peel or apple core. The apple core itself is not harmful, but it is difficult to chew and digest due to the hardness and containing seeds. Therefore, eating apple cores is not dangerous, but it is not necessary because the apple core is not tasty and has no nutritional benefits. **Sixth Grade:** *Neutral.* The answer is neutral because apple cores are not harmful to health when eaten with the apple. But there is also no special benefit to it. Some people don't eat apple cores because they find them unpleasant to chew.	**College:** *Affirmative.* Eating apple cores with the apple is not harmful to health because apple cores do not contain any harmful substances that are toxic to the human body. In fact, apple cores contain nutrients that are beneficial for our health. Therefore, it is affirmative that eating apple cores with the apple is not harmful to health. **High School:** *Neutral.* Eating apple cores is not necessarily harmful to health, but they can be difficult to chew and swallow. Some people may experience discomfort or choking if they eat apple cores, while others may not have any issues. So, it's neutral because it depends on the individual's ability to handle apple cores. **Middle School:** *Neutral.* Eating apple cores with the apple is neither completely harmful nor completely safe. Some apple cores have seeds that contain a small amount of cyanide, which can be harmful if consumed in large quantities. However, most people can safely eat a few apple seeds without any negative effects. So, it's not harmful, but also not completely safe. **Sixth Grade:** *Neutral.* Eating apple cores is not necessarily harmful, but it is not recommended because apple cores contain a small seed that could potentially be a choking hazard. However, the seed itself is not poisonous, so eating it might not cause any harm. Therefore, the statement is neither affirmative nor negative.	**College: College:** *Negative.* This question is classified as negative because it implies that eating apple cores with the apple is harmful to health, which is a harmful or undesirable outcome. The question is not neutral as it is not asking about a neutral or neutral topic, and it is not affirmative as it is not asking about a beneficial or positive outcome. **High School:** *Negative.* The label is negative because eating apple cores is not recommended due to the presence of seeds, stems, and other parts that can cause harm if ingested. Apple cores contain a compound called amygdalin, which can release cyanide, a toxic substance, when ingested. Therefore, it is not safe to eat apple cores with the apple. **Middle School:** *Affirmative.* Eating apple cores can be harmful to health because they contain seeds and stems that can cause digestive problems, and in some cases, even lead to choking hazards. The seeds contain a compound called amygdalin, which can release cyanide, a toxic substance, when ingested. Therefore, it is recommended to avoid eating apple cores and only consume the edible parts of the fruit. **Sixth Grade:** *Negative.* Eating apple cores can cause serious health problems, such as intestinal blockage or even death. Apple cores contain seeds, stems, and leaves, which are difficult for the body to digest. This can lead to serious health issues, so it is not safe to eat apple cores.

Table 15: An example data instance from the HealthFC dataset where LLMs tend to make wrong predictions. Owing to the limited space, some longer rationales are partially omitted and indicated with [...].

Readability: Which readability level best describes the rationale?

Level	Description	Example
College	Advanced English, often featuring elaborative discourse or terms.	The senators said allowing Russia to host the World Cup "inappropriately bolsters the prestige of the (Russian President Vladimir) Putin regime."
High school	Plain English, typical for ordinary documents.	The senators say Russia's role in the Ukraine crisis and occupation of Crimea should be condemned.
Middle school	Fairly easy to read, with some details but in informal language.	US senators call on FIFA to move the 2018 World Cup from Russia. 13 Democratic and Republican US lawmakers signed the letter.
Sixth grade	Easy to read, colloquial as consumer conversation	13 US senators call for Russia to be removed as hosts of the 2018 World Cup. They say the decision should be made in 2017.

Coherence: How reasonable is the logical flow of the rationale put together and connected?

Level	Example
4. very reasonable	This text uses language that is disrespectful and hurtful towards someone based on their political beliefs. It is considered offensive language, even though there is no hate speech towards a specific group mentioned in the text.
3. somewhat reasonable	This text seems offensive to some people because it uses a hurtful term not only towards feminists, but also towards lesbians. It might make some people feel insulted or disrespected.
2. somewhat unreasonable	This text is offensive because it makes light of a serious issue, rape. It's not a joke and it's not OK to make fun of something like that. It's important to treat serious issues with respect and sensitivity.
1. very unreasonable	This text expresses a strong opinion against chain migration, a process that allows immigrants to bring their family members to live with them in the US. This type of expression is common in everyday conversations and debates, it is not considered offensive or hate speech as it does not include any derogatory language towards any specific group of people.

Informativeness: How much sufficient information is provided to support the model-predicted answer?

Level	Example
4. very sufficient	The text is a simple question that doesn't contain any hate speech or offensive language towards any target group. It concerns the Marvel Cinematic Universe and a report by Matthew del Tobo, and there's no indication of any community standards violations in the text.
3. somewhat sufficient	The word "towel head" is not a kind or respectful way to talk about people with Middle Eastern or Muslim heritage. Using this name can hurt and stereotype people based on their background. We should always use kind and respectful words when talking about people.
2. somewhat insufficient	The text does not contain any offensive, derogatory, or hateful language towards a specific target group. It is a simple request for information.
1. very insufficient	This text is considered normal because it's not explicitly offensive, hateful, or targeting any specific group.

Accuracy: Do you agree with the model predicted answer? (Note: Please be aware of the difference between offensive and hate speech.)

Label	Description	Example
normal	Texts that hardly violate individuals or groups.	you know what i like to do when i come to california get raped by the gas prices
offensive	Expressions that discomfort or annoy certain individuals or groups, often through labels or remarks.	Twitter is full of tween dikes who think they're superior because of "muh oppression." News flash: No one gives a shit.
hate speech	Languages that incite discrimination, violence, or hostility against particular individuals or groups.	A nigress too dumb to fuck has a scant chance of understanding anything beyond the size of a dick

Table 16: Annotation guidelines provided to the annotators.