GMEG-EXP: A Dataset of Human- and LLM-Generated Explanations of Grammatical and Fluency Edits

S. Magalí López Cortez[1*], Mark Norris[2], Steve Duman[2]
[1]University at Buffalo, [2]Grammarly
609 Baldy Hall, Buffalo, NY 14260; 548 Market Street #35410, San Francisco, CA 94104
solmagal@buffalo.edu, {mark.norris, steve.duman}@grammarly.com

Abstract

Recent work has explored the ability of large language models (LLMs) to generate explanations of existing labeled data. In this work, we investigate the ability of LLMs to explain revisions in sentences. We introduce a new dataset demonstrating a novel task, which we call **explaining text revisions**. We collected human- and LLM-generated explanations of grammatical and fluency edits and defined criteria for the human evaluation of the explanations along three dimensions: Coverage, Informativeness, and Correctness. The results of a side-by-side evaluation show an Overall preference for human explanations, but there are many instances in which annotators show no preference. Annotators prefer human-generated explanations for Informativeness and Correctness, but they show no preference for Coverage. We also examined the extent to which the number of revisions in a sentence influences annotators' Overall preference for the explanations. We found that the preference for human explanations increases as the number of revisions in the sentence increases. Additionally, we show that the Overall preference for human explanations depends on the type of error being explained. We discuss explanation styles based on a qualitative analysis of 300 explanations. We release our dataset and annotation guidelines to encourage future research.

Keywords: human and LLM explanations, text revisions, side-by-side evaluation

1. Introduction

This work addresses two aspects of model explanations. On the one hand, we explore the ability of large language models (LLMs) to generate explanations of existing labeled data. On the other hand, we approach the question of model interpretability. LLMs are effective at rewriting text to suit a variety of goals, but it is not always clear what has changed or why (Fang et al., 2023).

To explore these questions, we introduce a new dataset, GMEG-EXP, built from the GMEG (Grammarly Multi-domain Evaluation for GEC) dataset, demonstrating a novel task, which we call **explaining text revisions**. Beginning with an existing data set containing pairs of sentences with stylistic and grammatical corrections, we collected free-text explanations of those corrections from expert human annotators. We then collected the same kinds of explanations from GPT-3.5. We propose evaluation criteria for the quality of explanations along three dimensions: Coverage (whether the explanations address all revisions in the sentence), Informativeness (whether the explanations provide enough information to understand the revisions), and Correctness (whether the explanations provide a true description of what changed and valid reasons for the changes), and we share the results of a side-by-side comparison of human-generated and LLM-generated explanations of revisions in text where they are compared along these dimensions.

Our results show an Overall preference for human explanations, but there are also many instances with no preference. Annotators prefer human-generated explanations in terms of Informativeness and Correctness, but they show no preference for Coverage. We also examine the extent to which the number of revisions in a sentence influences annotators' Overall preference for the explanations. We find that the preference for human explanations increases as the number of revisions in the sentence increases. Additionally, we show that the Overall preference for human explanations depends on the type of error being explained. In summary, our paper's main contributions are (i) the release of a dataset of human- and LLM-generated explanations, guidelines, and associated scripts;[1] (ii) the definition of the evaluation criteria and the approach to evaluating explanations; and (iii) a qualitative description of the properties of explanations.

2. Related Work

Our work is at the intersection of two strands of research: (i) model-generated explanations, and (ii) explainable Grammatical Error Correction (GEC) and text revision more broadly.

*This research was performed while S. Magalí López Cortez was an intern at Grammarly.

[1]https://github.com/grammarly/gmeg-exp

LREC-COLING 2024, pages 7785–7800
20-25 May, 2024.

2.1. Model-Generated Explanations

Wiegreffe et al. (2022) use a methodology most similar to ours. They begin with existing data sets containing crowdsourced human-generated explanations of the labels in existing benchmark data sets for CommonsenseQA (a multiple choice task over commonsense questions) and Natural language inference (NLI; inferring whether a given hypothesis sentence entails, contradicts, or is neutral toward a premise), and explore LLM-generated explanations of those labels using prompting with in-context learning. They use human evaluation on the premise that "existing automatic metrics often do not correlate well with human judgments of explanation quality" (p. 634; See also Clinciu et al., 2021; Kayser et al., 2021) with a side-by-side setting: annotators are given an item, the gold label, and two explanations, and they are asked to choose their preferred explanation. They also conduct an item-by-item evaluation, in which annotators score individual explanations for defined qualitative categories such as factuality, grammaticality, validity, supportiveness and general acceptability of each explanation. They demonstrate a preference (among crowdsourced annotators) for LLM-generated explanations, and they demonstrate that improving the prompt (e.g., by using higher-quality labels for in-context examples) can improve overall evaluation scores. We follow Wiegreffe et al. (2022) in the construction of the prompt with in-context examples and in the side-by-side human evaluation setting.

Marasovic et al. (2022) also use LLMs to generate explanations for existing data sets (for NLI, commonsense QA, nonsensical sentence selection, and offensiveness classification); they differ from Wiegreffe et al. (2022) in that their prompts do not include in-context examples. In exploring whether difficult examples are likewise difficult to explain, Saha et al. (2022) use human annotators and LLMs to generate explanations of items from the Winograd schema (Levesque et al., 2012). They evaluate the explanations along three dimensions: grammaticality (whether the explanation is linguistically well-formed), supportiveness (whether the explanation helps understand the specific data point in question), and generalizability (whether the explanation is framed in such a way that it could be applied to other similar examples). They find that human- and LLM-generated explanations are comparable in terms of grammaticality, LLMs are comparable (but sometimes worse) than humans in terms of supportiveness, and LLMs are definitely worse than humans in terms of generalizability.

2.2. Explainable GEC

A common approach to explainable GEC systems is to classify revisions by error type, e.g., ERRANT (ERRor ANnotation Toolkit; Bryant et al., 2017) or GECToR (Grammatical Error Correction: Tag, Not Rewrite; Omelianchuk et al., 2020). Fei et al. (2023) build on this to annotate evidence spans—words/spans that lead to the error in question, e.g., a particular verb requires a particular preposition. They argue that these explanations—the combination of evidence spans and error types—are more useful for language learners.

Taking this one step further, Nagata (2019); Nagata et al. (2020, 2021) have pursued a task that they call feedback comment generation. They define feedback comments as hints or explanatory notes for language learners. While feedback comments have a similar goal to what we call explanations, Nagata (2019); Nagata et al. (2020, 2021) focus on preposition errors only; we explore grammatical and fluency edits in general. Additionally, our human evaluation tasks differ: they collect evaluations for whether each feedback comment is appropriate, partially appropriate, or inappropriate, whereas we focus on a side-by-side evaluation procedure. We also provide a qualitative description of explanations.

Kaneko and Okazaki (2023) also identify the lack of work exploring free-text explanations of text revisions. Like us, they introduce the task of generating such explanations with humans and LLMs. Beyond that, the goals of our studies are rather different. We focus on direct comparison of human- and LLM-generated explanations and the structure of explanations of text revisions more broadly. Kaneko and Okazaki (2023) explore prompting to generate both text revisions and explanations of those revisions with LLMs, and they investigate how generating explanations can improve LLM performance on GEC. They also compare human- and LLM-generated explanations by varying the source of in-context examples. Based on that, they conclude that human- and LLM-generated explanations are comparable. Though our comparison finds, in contrast, that human-generated explanations are preferred, the data sets and prompts we use are different; thus, we cannot directly compare results.

3. Our Approach

3.1. Data Source

We used the Grammarly Multi-domain Evaluation for GEC (GMEG) dataset (Napoles et al., 2019). GMEG contains approximately 6,000 sentences split roughly equally across three sources: informal web posts from Yahoo! Answers (`yahoo`), Wikipedia articles (`wiki`), and student First Cer-

tificate in English essays (fce).[2] Each input sentence in GMEG is matched with four versions corrected for grammar, spelling, and fluency. Because we are interested not in the revisions themselves but in explanations of them, we randomly chose one revised version from GMEG for each sentence, ignoring any with no edits. We refer to the sentences with errors as the original sentences and to the corrected sentences as revised sentences.

3.2. Data Preparation

We aligned the sentences using the align function from Lund et al. (2023).[3] We will refer to these sentences as aligned sentences. An example is in (1).

(1) With the help of {some=>} modern technology {=>,} our daily life {would=>will} not {=>be} so difficult . (fce_test_762)

The changes from the original to the revised sentences may include substitution (e.g., {would=>will}), insertion (e.g., {=>be}), and/or deletion (e.g., {some=>}).

4. Human Annotation: Explanations

In this section, we describe our process for the collection of human-generated explanations.

4.1. Annotation Details

The human-generated explanations were written by 11 professional annotators with linguistic training. Annotators were shown one aligned sentence at a time. For each sentence, we instructed annotators to provide an explanation for all the revisions in the sentence in a bulleted list. We include a sample annotation item in Appendix A. We refer to the entire bulleted list as the explanation and to each individual list item as an explanatory statement.

We chose this design because it gave annotators the option of grouping revisions together into one explanatory statement. This also allowed us to explore the extent to which human annotators and LLMs address all revisions in a sentence when presented with multiple revisions.

Annotators were instructed not to judge the revisions, so if they did not agree with one revision because it was stylistic or they would have preferred to revise the sentence differently, they were still asked to explain the revision. However, if they felt a particular revision was truly incorrect, they were instructed to mark the item as Not annotatable. Example (2) shows a sentence containing a stylistic revision that was seen during annotator training and the explanation provided by the annotator.

(2) **Sentence:** A similar sentence with visual predicates is claimed to be, “I can see it clearly {”,=>,”} or with auditory predicates, “That sounds right to {me.=>me.”} (wiki_test_537)
Human Explanation: *Putting the comma and period inside the quotations follows American English rules.*

Table 1 shows an example of an aligned sentence, a human-authored explanation for the revisions in that sentence, and an explanation generated by an LLM. We discuss how we generated explanations with an LLM in the Section 5.

4.2. Pilot Annotation Task

For the pilot annotation task, we randomly selected 200 sentences from the wiki and fce data only.[4] We provided annotators with written guidelines and trained them during a live session where we reviewed the guidelines and examples. We also encouraged questions and feedback through a Slack channel. Even though our guidelines included a section about what explanations may look like and examples, annotators asked for an explanation style guide during training and annotation. We chose not to provide strict instructions about the shape of explanations because we are interested in studying what human-generated explanations of text revisions look like, and we did not want to bias annotators towards a particular style. We discuss explanation styles in Section 8.

We reviewed the pilot explanations manually and updated the guidelines based on the results.

4.3. Full-Scale Annotation Task

After the pilot study, we trained the entire group of 11 annotators and launched the full-scale annotation task with 3804 items sampled equally from all three data sources. We discarded all items from the full-scale study marked as Not annotatable (272).

We asked two annotators to inspect approximately 10% of the annotations. Out of 304 items reviewed by annotators, 200 (65.8%) were accepted. The main reason for rejection was that the explanation was incorrect, confusing, or slightly off, which suggests that the task of writing explanations of text revisions is difficult. We provide more details about the Quality Control process in Appendix B.

[2] GMEG contains dev and test data; since we were not building a model, we sampled from both splits.

[3] Available at: https://github.com/grammarly/gender-inclusive-gec/blob/main/scripts/utils/align.py.

[4] We did not include the yahoo data in the pilot task because we had not gained access to this split at the time we ran the pilot annotation.

Aligned	Human	LLM
In March 2004, he was named by Pelé as one of the top 125 greatest living footballers and **{honoured=>honored}** as part of the **{in [=>}** 100.	- "Honoured" is replaced with "honored" to follow American English standards. - The preposition and bracket are deleted to improve fluency and grammar.	- "Honoured" is changed to "honored" to match American English spelling. - "in [" is removed as it is likely a typographical error and does not make sense in the context of the sentence.

Table 1: An example explanation from an annotator and the LLM for an aligned sentence. In this case, each explanation consists of two explanatory statements. Sentence id: `wiki_test_790`

5. LLM Data Generation

5.1. Model Details

We used GPT 3.5 turbo through Microsoft Azure's OpenAI API. We set the model's temperature to 0.7 and top_p to 1.

5.2. Prompt Engineering

We selected the best-performing prompt after manual review of the output of 15 different prompts.

Our prompt consists of the instructions "*Given an ORIGINAL sentence, a REVISED sentence with grammatical error corrections and fluency edits, and an ALIGNED sentence that shows what changed from ORIGINAL to REVISED, please explain why each revision was made*," followed by three in-context examples randomly sampled from the pilot human annotation task each showing an original sentence, a revised sentence, the aligned version, and the explanation. For in-context examples, we followed Min et al. (2022), who suggest that randomly sampling items from the actual distribution may be more effective than ensuring that all labels are represented in the in-context examples. After the in-context examples, we repeated the instructions, following Maddela et al. (2023), and then included the target item. We include a complete prompt in Appendix C.

5.3. Data Generation

102 items (out of the 3804) did not successfully go through the API[5], and we discarded 3 items because of other output issues, leaving us with 3699 items.

6. Side-by-Side Evaluation

We followed Marasovic et al. (2022) and Wiegreffe et al. (2022) in conducting human evaluation on free-text explanations. Specifically, we performed a side-by-side (SBS) evaluation of our human- and LLM-generated explanations.

[5]We got the following error message every time we sent these items: *ERROR: An Azure OpenAI InvalidRequestError fired while processing this request.*

6.1. Annotation Preparation

We excluded items that were discarded as part of either the human annotation or LLM data generation, leaving us with 3398 items for the side-by-side evaluation. Each aligned sentence was matched with its corresponding human- and LLM-generated explanations. The source of the explanations was anonymized and randomly ordered. Annotators saw the aligned sentence and two explanations that were simply labeled `Explanation 1` and `Explanation 2`.

6.2. Evaluation Criteria

For the side-by-side evaluation of explanations, we first asked annotators to choose which explanation they preferred **Overall** using a three-point scale to answer the question *Which explanation better explains the revisions?*: `Explanation 1`, `About the same`, `Explanation 2`.

We then defined three categories to evaluate explanations: **Coverage**, **Informativeness**, and **Correctness**. For these three categories, annotators also had to choose which explanation they preferred or indicate that they found both explanations equal. We define each category below and include examples in Appendix D.

6.2.1. Coverage

Coverage refers to whether the explanations address all revisions in the sentence. In cases where the sentence contains only one revision, this category is somewhat trivial, as a lack of coverage would mean no explanation at all. However, in cases with two or more revisions, an explanation could explain one or more revisions while failing to explain others.

6.2.2. Informativeness

Informativeness refers to whether the explanations provide enough information to understand the revision. An informative explanation should help the reader understand what changed in the revised sentence and why. Different types of revisions may require different degrees of detail in the explanation. For example, a spelling error may not require a detailed explanation, but a word choice

edit may require more details. Annotators were asked to judge Informativeness based on whether the explanation provided enough details to help them understand why the revision was made as compared to a scenario where they had only seen the revision itself.

6.2.3. Correctness

Correctness refers to whether the explanations provide a true description of what changed and valid reasons for the change. A correct explanation does not include revisions that did not occur in the sentence. For example, a statement explaining a hallucinated revision would be incorrect.

6.3. Annotation Procedure

The evaluation of the explanations was conducted by a different pool of 13 annotators with a mixture of annotation and copy-editing experience.[6]

Each annotation item showed the aligned sentence, the human- and LLM-generated explanations in random order, the question for Overall preference, and the questions for each subcategory. Annotators could also flag the item as `Not annotatable`, and they could leave comments. We include a sample annotation item in Appendix E.

We used a dynamic judgment approach for Overall preference. We required at least two annotations per item. If the two annotators disagreed on their Overall preference, we collected a third annotation. We did not include this constraint for the subcategories of Coverage, Informativeness, or Correctness.

6.4. Annotation Quality: Side-by-Side

In total, we collected 8271 judgments, of which 48 were marked as non-annotatable and skipped by annotators. 1953 items received 2 judgments, 1438 items received 3, and 3 items received 1 judgment.[7] This means that the first two annotators agreed on their Overall preference in a bit more than half of the items.

We calculated Krippendorff's α as a measure of inter-rater reliability for the SBS evaluation. The results across categories can be seen in Table 2. Similar to Wiegreffe et al. (2022), we find low-to-moderate agreement, which suggests the subjectivity of the task. Agreement was higher for Coverage, probably the least subjective category, and lowest for Correctness.

[6]We selected these annotators rather than linguistically-trained annotators because we believed their judgments would be similar to a less specialized but still discerning member of the general population.

[7]Because the other 2 annotators marked them as non-annotatable/skipped them.

Category	Krippendorff's α
Overall Preference	0.41
Coverage	0.64
Informativeness	0.33
Correctness	0.27

Table 2: Inter-rater reliability for the SBS evaluation

7. Side-by-Side Data Analysis

For the data analysis, we convert annotator preferences to numeric observations (-1 for Human, 0 for No Preference, and 1 for LLM) and then take an average of all annotators for each item. We then categorize the averages such that an Overall preference less than or equal to -0.5 is a preference for Human, an Overall preference greater than or equal to 0.5 is a preference for the LLM, and anything else is No Preference.

7.1. Coverage, Informativeness, and Correctness

Annotators show, on average, an Overall preference for explanations from humans compared to those from the LLM (coeff = -.54, p < .001). This effect is validated by a multinomial logistic regression model using the `nnet` package in R (Venables and Ripley, 2002). Beyond Overall preference, annotators also judged the explanations in terms of Coverage, Informativeness, and Correctness. Multinomial regression models show that annotators prefer human explanations for Informativeness (coeff. = -.36, p < .001) and Correctness (coeff. = -.74, p <.001), but they show no preference with regard to Coverage (coeff. = -.05, p = .67). In other words, human- and LLM-generated explanations are equally good at covering all revisions in the sentence, but human-generated explanations tend to be more correct and/or informative. Table 3 provides annotator means and variation.

Metric	Avg	St. Dev.
Overall Preference	**-0.18**	0.74
Coverage	0.01	0.52
Informativeness	**-0.12**	0.82
Correctness	**-0.22**	0.70

Table 3: Annotator means for side-by-side evaluation categories. 0 = no preference; <0 = preference for human

7.2. Number of Revisions

Each sentence includes at least one revision, but sentences can include more. More revisions require more robust, detailed explanations. Therefore, we examine the extent to which the number

of revisions in a sentence influences annotators' Overall preference for the explanations. Figure 1 shows the preference for human over LLM explanations by revision count.

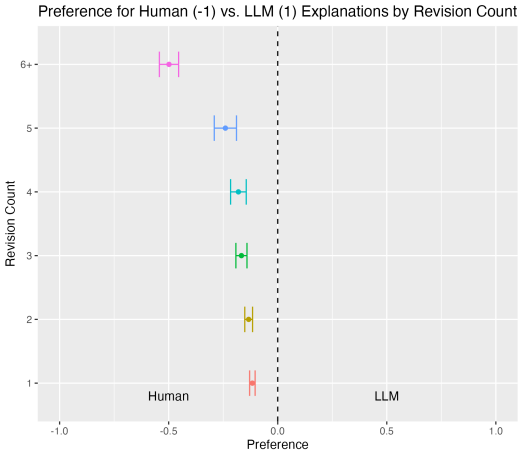

Figure 1: SBS annotator preference by revision count

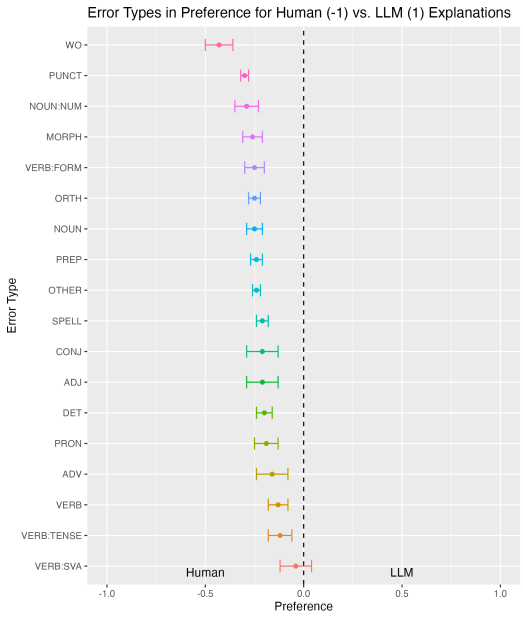

Figure 2: SBS annotator preference by included error types

A multinomial logistic regression model predicting Overall preference for human explanations with Revision Count as a predictor (and all examples with six or more revisions treated as one category) shows that average Overall preference for human explanations increases as the number of revisions in the sentences increases. For example, there is a statistically significant difference in preference for human compared to LLM explanations for sentences with one revision compared to those with 6+ revisions (coeff. = -.93, $p < .001$).

7.3. Error Types

Sentences also contained different types of corrected errors. We used ERRANT (Bryant et al., 2017) to identify the error types in the dataset. Our analysis shows that the Overall preference for human-generated explanations depends on the types of errors in the target sentence. This is demonstrated in Figure 2, which plots average preference with error types present in the sentences (including only errors that appeared 50 or more times in the dataset).[8] It shows, for example, that the preference for human explanations is strongest when the sentence contains Word Order (WO) or Punctuation (PUNCT) errors. In contrast, there is no clear preference when sentences contain Subject-Verb Agreement errors (VERB:SVA). In other words, these results show the error categories that humans explain better than the LLM.[9]

Example (3) shows a sentence with a word order error for which the LLM did worse at explaining the revision.[10] The aligned version of this revision type tends to be instantiated as two pairs of curly braces; i.e., one for a word/punctuation mark deleted from the wrong place and one for the word/punctuation mark added to the right place. In this example, the human annotator captured that the period was moved before the closing quotation marks for style reasons, whereas the LLM explained each revision separately. In other words, the LLM was worse at interpreting both pairs of curly braces being part of the same word order revision.

(3) **Sentence:** StormierZiggy alleged that the moderators had {been=>} been waiting weeks to ban him and his friends for anything , because they disliked them for " speaking their minds and telling the truth **{=>.}** " **{.=>}** (`wiki_dev_876`)
Human Explanation:
- The closing quotation marks were put after the period, which is the preferred placement in North American conventions.
LLM Explanation:
- A period was added after the quote to adhere to grammar rules.
- The extra period at the end of the sentence

[8] To simplify the analysis, we kept only the error class labels from ERRANT, ignoring the operation type (**M**issing, **R**eplacement, **U**nnecessary).

[9] This analysis is based on individual error types with ratings of entire explanations, and it does not include error frequencies or combination of errors. We leave for future work further exploration of the relationship between error types and preferred source of explanation.

[10] For space reasons, we include only the explanatory statements that illustrate the relevant phenomena. We use boldface to highlight the relevant revision(s).

was removed.

8. Explanation Styles

During human data collection, annotators often asked what explanations should look like. We provided examples using different sentence structures but did not provide a style guide for the annotators. Likewise, we did not provide the LLM with a specific style guide. In this section, we examine different qualitative aspects of human- and LLM-generated explanations in an attempt to infer the most common characteristics of explanations of text revisions.

8.1. Explanation Lengths

We compare lengths of explanatory statements from human annotators and the LLM in Table 4. Even though the maximum statement length is

	Human	LLM
Avg	18.67	17.66
Max	174	73
Min	3	3

Table 4: Token length of statements from human annotators and the LLM

longer for humans than for the LLM, the averages are quite similar.

8.2. Qualitative Analysis of Subcategories

We randomly selected 150 items and manually annotated the corresponding human- and LLM-generated explanations (300 total explanations) to gain more insights into the characteristics of the explanations generated by each source.

8.2.1. Coverage

Recall that Coverage concerns whether the explanation explains all revisions in the target item. Example (4) shows an instance where the LLM did not address a revision.

(4) **Sentence:** I have {to choose=>chosen} painting and {the=>} photography {therefore=>}. (`fce_test_729`)
LLM Explanation:
- The revised sentence simplifies the structure by using the past participle form of the verb 'choose' to show that the action has already been completed, and eliminates the unnecessary words 'to' and 'therefore.'

In (4), the LLM combines two revisions into one statement: (i) replacing *to choose* with *chosen* and (ii) the removal of *therefore*. However, it does not address the removal of *the*, and thus, it fails to cover all revisions.

The proportion of revisions covered by the explanation across the 150 items we annotated is provided in Table 5. While some of these proportions

N_revisions	Count	Human	LLM
1	51	0.99	0.97
2	40	0.93	0.94
3	30	0.92	0.93
4	12	0.93	0.95
5	11	0.95	0.88
6	3	0.94	0.81
7	1	1	0.93
8	2	1	0.88

Table 5: Results of the annotation of 150 human and LLM explanations for Coverage. The table shows the mean proportion of covered revisions for all the sentences grouped by number of revisions.

may not achieve statistical significance, they suggest a trend where sentences with more revisions have slightly worse coverage.

8.2.2. Informativeness

Recall that an informative explanation should help the reader understand what changed in the revised sentence and why. Some maximally informative explanatory statements are shown in (5).

(5) **Sentence:** Another thing was that the show I went to was {starting=>supposed to start} at {19=>7}:30{=>,} but it started at {20=>8}:15! (`fce_test_168`)
Human Explanation:
- The verb "starting" was replaced with the phrase "supposed to start" because this is in keeping with the overall meaning conveyed in the sentence. The change improves the fluency of the sentence.
- Converted the 24-hour notation of time to a 12-hour clock. So 19 is replaced with 7, and 20 is replaced with 8. This is a style edit.

In reviewing the data, we identified four categories that an explanatory statement may contain:

- **Description:** what changed in the revision, e.g., *"starting" was replaced with the phrase "supposed to start"*
- **Error type:** a general category to describe the error that was corrected, e.g., *this is a style edit*
- **Reason:** an indication of why the edit was necessary, e.g., *this is in keeping with the overall meaning of the sentence*
- **Purpose:** indicates how the sentence improves after the edit, e.g., *improves the fluency*

Example (6) further illustrates the difference between `reason` and `purpose`.

(6) **Sentence: {User=>The user}** fails to follow BRD , {=>and} restored BOLD edit while discussion was still going on {and no=>. No} consensus was achieved for his position. (`wiki_dev_173`)
Human Explanation:
- "The" is inserted as the first word of the sentence because the following noun, "user," requires a specific determiner.
LLM Explanation:
- The change from 'User' to 'The user' makes the sentence more formal and clear.

While the human explanation focuses on the noun's requirements (a reason the edit was necessary), the LLM explanation focuses on the broader results of the changes (a purpose).

Explanations from both the human and LLM vary in which pieces they use. The presence/absence of each component in the evaluated items is represented in Table 6.

	Human	LLM
Description	0.93	0.92
Error Type	0.23	0.1
Reason	0.89	0.93
Purpose	0.25	0.3

Table 6: Proportion of human and LLM explanations containing various explanation components in a sample of 150 explanations from each source

8.2.3. Correctness

The final category is Correctness— whether the explanation contains false or incorrect information. There are examples of both humans (7) and LLMs (8) providing incorrect explanatory statements in the data.

(7) **Sentence: {Understand=>}** what makes her tick {=>,} and her desires. (`yahoo_test_621`)
Human Explanation:
- "Understand" is inserted to change the sentence into a command.

(8) **Sentence:** Throughout the 1970s and 80's, it {has been=>} absorbed a huge number of Jewish immigrants from Israel. (`wiki_test_247`)
LLM Explanation:
- The phrase 'has been absorbed' is in the present perfect tense, which suggests a continuous action from the past to the present. However, the sentence is referring to a specific time period in the past, so the present perfect tense is not appropriate.
- The word 'it' is unclear in the original sentence, so the revision clarifies that it refers to a place or country.

In (7), the human annotator did not provide a true description of the revision (*understand* was deleted rather than inserted). In (8), the LLM did not provide a valid reason for one revision (the edit of *has been absorbed* is due to passive/active voice rather than tense) and hallucinated another ('it'). The proportion of explanations across the entire data set with Correctness issues is shown in Table 7.

	Human	LLM
Not a true description	0.013	0.22
Not a valid reason	0.02	0.13
Hallucination	0	0.07

Table 7: Proportion of human and LLM explanations containing certain Correctness issues in a sample of 150 explanations from each source

Because some Correctness issues are more severe than others and because some explanations may contain multiple Correctness issues, we also scored the explanations for Correctness on a scale of 1-5. The results are in Table 8. It is notable that scores for Human-generated explanations are almost exclusively 4 or 5, whereas LLM-generated explanations have a wider spread, including scores of 2 or 3. Both the scores and the proportion of output issues suggest that Human-generated explanations are more likely to be correct than LLM-generated explanations.

Score	Human	LLM
(high) 5	79	56
4	64	34
3	4	32
2	1	20
(low) 1	2	8

Table 8: Correctness scores for Human and LLM explanations in a sample of 150 items

9. Discussion

Based on our analysis, we found that the LLM was quite good at covering all the revisions in the sentence, especially when sentences contain up to 4 revisions, and that humans were better at providing informative and accurate explanations of text revisions. We note that Informativeness assumes that the explanation is also correct. In other words, a correct explanation may or may not be informative, but an incorrect explanation is also uninformative. Kaneko and Okazaki (2023) defined their evaluation criteria with two categories: coverage and validity. Validity refers to "the accuracy and usefulness of grammatical information in LLM-generated explanations for language learners" and encompasses our two categories of Informativeness and Correctness. Future research can validate the evaluation criteria, but we believe

that obtaining judgments on both Informativeness and Correctness provides more fine-grained information about the quality of explanations.

Given that we did not find a substantial difference in coverage of revisions, it might be possible to ignore the coverage question and instead focus on individual explanatory statements. We decided against the side-by-side evaluation of individual explanatory statements because they are not always directly comparable. For example, one source may group together two or more revisions in one statement, whereas the other might write one explanatory statement for each revision. However, the evaluation of explanatory statements (as well as entire explanations) can be done individually, without comparing sources. Wiegreffe et al. (2022), for example, use a combination of both. However, this would come at the expense of evaluating a smaller portion of the data. We evaluated 3398 items in a SBS manner, which would double if evaluating individual explanations and further multiply if evaluating individual explanatory statements.

Finally, while this task shares some similarities with the task of Error-Type Classification, we think it has several novel properties. First, explanations may include classification of the error, but they may also include additional content, such as a reason or a purpose. Second, the goal of this task is to produce free-text explanations that are legible to humans, whereas error-type classification may not have this particular goal. However, we note that adding a step of error-type classification to the pipeline could help humans or LLMs produce better explanations. We leave the exploration of this possibility to future work.

10. Conclusion and Future Work

We collected explanations of revised sentences from human annotators and GPT-3.5, defined criteria for human evaluation and conducted a quantitative and qualitative analysis of the explanations. In future work, we would like to investigate explanations of revisions of longer stretches of text and extend the comparison of humans and LLMs to revising a text and explaining the revisions in a combined task.

Limitations

This task focuses on revisions of sentences. However, the challenge of understanding text revisions goes beyond sentences, especially given that LLMs can rewrite paragraphs or even entire documents. It is possible that the quality of explanations could change if annotators or the LLM were asked to explain all revisions that occurred for a paragraph. Even though we would ideally collect explanations on paragraph- or short-document-level revisions, given that this is the type of revisions LLMs are doing, we decided to focus on sentences for this project, especially because of the comparison with human-generated explanations. An annotation task of explanations of entire paragraphs or document rewrites would have been too time-consuming.

Additionally, the data set we use focuses on GEC and fluency edits; of course, the general category of text revision includes stylistic edits such as tone, organization, and word choice. It is not clear that the conclusions we draw from this data set would extend to all kinds of text revisions.

It is known that the choice of prompt for LLMs can dramatically affect the result (Lu et al., 2022; Min et al., 2022). As discussed in Section 5.2, we tried many different templates, but it is nevertheless possible that substantially better prompts exist for this task.

A limitation of the evaluation we performed is that we only evaluated the explanations as a whole, not each individual explanatory statement. As we discussed, this had a number of advantages. However, we might be missing some granular details that could be captured by evaluating individual explanatory statements for each revision.

Ethics Statement

Our primary ethical concerns for this work concern the dataset that we created, which contains annotations from both humans and LLMs. Because annotators are not necessarily representative of population as a whole, it is possible that their biases are reflected in the collected data. Similarly, it is known that LLMs are capable of generating text that replicates bias from the training process. Thus, it is possible that the annotations coming from the LLM could show that bias.

Acknowledgements

We thank three anonymous reviewers for helpful comments and feedback. We would also like to thank Max Gubin, Christy Doran, Philip Dwelle, Melissa Lopez, and David Rojas, as well as the Language Research team and the 2023 Applied Research Interns at Grammarly for feedback and comments on earlier versions of this paper.

11. Bibliographical References

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.

Julian Martin Eisenschlos, Jeremy R. Cole, Fangyu Liu, and William W. Cohen. 2023. WinoDict: Probing language models for in-context word acquisition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. *arXiv preprint arXiv:2305.15676*.

Masahiro Kaneko and Naoaki Okazaki. 2023. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. *arXiv preprint arXiv:2309.11439*.

M. Kayser, O. Camburu, L. Salewski, C. Emde, V. Do, Z. Akata, and T. Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1224–1234, Los Alamitos, CA, USA. IEEE Computer Society.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Gunnar Lund, Kostiantyn Omelianchuk, and Igor Samokhin. 2023. Gender-inclusive grammatical error correction through augmentation. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 148–162, Toronto, Canada. Association for Computational Linguistics.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China. Association for Computational Linguistics.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared task on feedback comment generation for language learners. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 320–324, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating corpora for research in feedback comment generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 340–345, Marseille, France. European Language Resources Association.

Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. Enabling robust grammatical

error correction in new domains: Data sets, metrics, and analyses. *Transactions of the Association for Computational Linguistics*, 7:551–566.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Swarnadeep Saha, Peter Hase, Nazneen Rajani, and Mohit Bansal. 2022. Are Hard Examples also Harder to Explain? A Study with Human and Model-Generated Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2121–2131, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

William N Venables and Brian D Ripley. 2002. *Modern applied statistics with S*. Springer Science & Business Media.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

A. Human Annotation: Sample Item

Figure 3 shows the annotation interface for annotators who are writing explanations. They see the sentence, and a box is provided to type in the explanation. Another box is provided to write optional comments.

B. Full-scale Human Annotation Quality Control: More Details

We manually reviewed and scored 10 items per annotator. We identified issues along four categories:

1. This is a **false non-annotatable item**: the item was marked as "Not annotatable" but could have been annotated. For example, in the following case, the annotator may not have agreed with the revision, but the revision is not wrong and could have been explained:

(9) **Sentence:** I think **{it can=>that could}** make your festival more popular. (`fce_test_263`)

Human Annotation: *Not annotatable*

2. The explanations **don't cover all the revisions**. For example, in this case, the explanation misses the first revision:

(10) **Sentence:** In 1998, ICP released a **{remix=>remixed}** version of the song as a single entitled " Hokus Pokus Remix **{" .=>."}** (`wiki_test_801`)

Human Annotation:

- Periods go inside quotations according to American English conventions.

3. The explanation is **incorrect, confusing, or slightly off**. For example, in this case, the first explanatory statement mentions a parenthesis, but there is no parenthesis in the revision.

(11) **Sentence:** I am writing in order to complain about the show " Over the Rainbow **{=>,}** " **{,=>}** which was played in the **{"=>}** Circle Theatre **{" ,=>}** because the information which was given in your advertisement was not true at all. (`fce_dev_211`)

Human Annotation:

- A comma was moved after the title, inside the parentheses, to separate the main clause from the explanatory thought that follows it.

- The quotation marks around the theater name were removed as it is a location, not a title.

4. The item **does not have any explanations**.

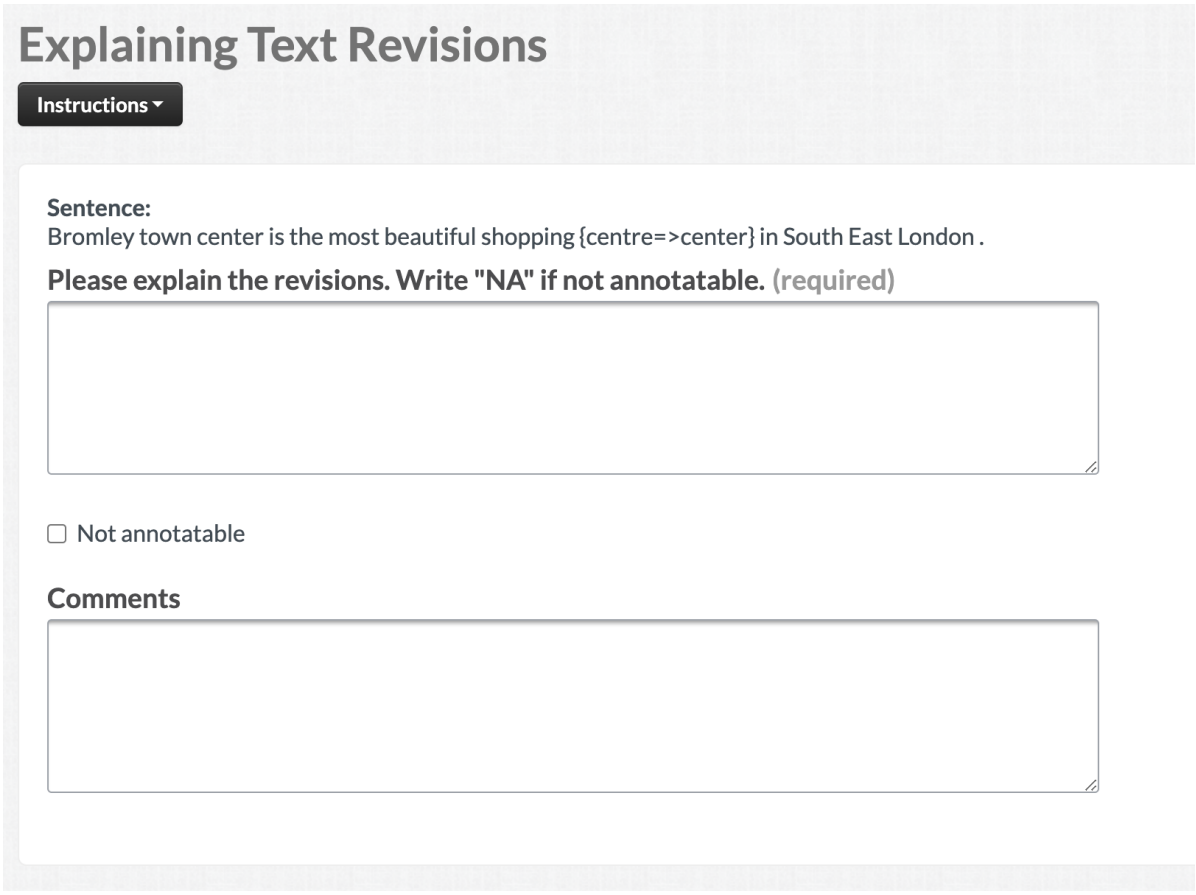

Figure 3: Sample annotation item.

Based on this review, we selected the two top-scoring annotators to perform a larger Quality Control (QC) task. We gave them 304 items randomly chosen from each annotator and asked them to accept or reject the items based on whether the explanations followed the guidelines. We asked them to select a reason for rejection based on the four issues above. We also included an Other category and a text input box for comments.

Table 9 shows the results of the quality control annotation task. Out of 304 reviewed items, 104 were rejected. The main reason for rejection was that the explanation was incorrect, confusing or slightly off, which suggests the difficulty of the task of writing explanations of text revisions. No items were rejected due to a lack of explanations or another reason.

False Non-annotatable items constitute about 21% of the rejected items. As example (9) shows, in some cases, annotators disagreed with the revisions presented and marked these items as NAs even if they had been instructed to only use this NA category if the revision was clearly wrong. Examples they were given in the guidelines as to what constitutes a wrong revision include {creat => creatte}, and whether a comma was added when not needed. However, it is important to note that this set of rejected items does not show issues with the explanations per se, but rather with how human annotators interpreted the instructions. Items marked as `Not annotatable` were not included as input in the LLM prompt.

Reason	Count
Off	54
Lacks coverage	23
False NA	22
Lacks coverage + Off	5
No explanation	-
Other	-
Total	104

Table 9: Rejected items from the Quality Control annotation task based on 304 reviewed items.

The next set of items includes those explanations rejected due to lack of coverage (27 out of a total of 104 rejected items). While it is true that human annotators sometimes miss revisions when writing explanations, Section 8.2.1 shows that LLMs also miss revisions. Our side-by-side evaluation accounts for this category. Section 7.1 shows no preference of our human evaluators with regard to

Coverage, suggesting that it is not the case that one source is significantly better at covering revisions than the other.

Out of 104 rejected items, 59 were rejected because they were incorrect, confusing, or slightly off. Manual inspection shows that, in a few cases, the items are clearly wrong. For example, in the sentence below, *the* was deleted, not added as the second explanatory statement says:

(12) **Sentence:** In the fourth year {=>,} he is visited by a spacecraft that regularly brings him supplies and news from {the=>} Earth four times a year. (`wiki_dev_760`)
Human Annotation:
- A comma was added after the introductory phrase.
- "The" was added before Earth to convey that there is only one specific Earth.

In most other cases, however, the items have one explanatory statement that includes a reason for the edit that is too vague or slightly off, for example:

(13) **Sentence:** Me {=>,} but I don't know what to do with it and probably will just leave it sitting there. (`yahoo_test_708`)
Human Annotation:
- A comma was inserted after "me" to show the beginning of a new thought.

(14) **Sentence:** In 1994, Rose fired guitarist Gilby Clarke and hired an old friend of his {,=>} named Paul Tobias. (`wiki_test_481`)
Human Annotation:
- A comma was removed before the phrase "named Paul Tobias." When information in a clause such as this is necessary for the meaning of the main clause, no comma is needed.

In example (13), "new thought" might be too vague. In (14), "when information in a clause such as this is necessary for the meaning of the main clause" might also be vague. This shows that the task of writing explanations of text revisions is difficult and subjective.

In a few additional cases, the explanation is not necessarily wrong, vague or off, but it seems the rejection could have been a matter of preference. For example, in 15 below, the annotator QC'ing this item might have found the sentence fragment structure of the statement confusing, or they might have preferred a more precise reason for the edit:

(15) **Sentence:** like=>Like import your user-made book? (`yahoo_dev_83`)
Human Annotation:
- "Like" capitalized at the beginning of a sentence.

The QC results are based on a small sample (304 items). However, this is not the only signal about the quality of evaluations, as the side-by-side human evaluation (see Section 6) also covers qualitative categories of explanations which provide deeper insights into human preferences when reading explanations of text revisions. While it is possible that modified guidelines or additional training could improve the quality of human-generated explanations, we want to emphasize that we believe it is reasonable to compare these human-generated explanations with LLM-generated explanations even though they contain errors. Humans and LLMs alike make errors when annotating data, and neither dataset needs to be perfect to be evaluated.

C. Prompt

C.1. The actual prompt

The format of the prompt we used is given below with placeholders for the actual experimental items.

Given an ORIGINAL sentence, a REVISED sentence with grammatical error corrections and fluency edits, and an ALIGNED sentence that shows what changed from ORIGINAL to REVISED, please explain why each revision was made.

Examples:
ORIGINAL SENTENCE: ```{sentence from GMEG dataset}```
REVISED SENTENCE: ```{human annotator revised version from GMEG dataset}```
ALIGNED SENTENCE: ```{aligned version}```
EXPLANATIONS:
```{human-generated explanation from the pilot structured in XML format}```

ORIGINAL SENTENCE: ```{sentence from GMEG dataset}```
REVISED SENTENCE: ```{human annotator revised version from GMEG dataset}```
ALIGNED SENTENCE: ```{aligned version}```
EXPLANATIONS:
```{human-generated explanation from the pilot structured in XML format}```

ORIGINAL SENTENCE: ```{sentence from GMEG dataset}```
REVISED SENTENCE: ```{human annotator revised version from GMEG dataset}```
ALIGNED SENTENCE: ```{aligned version}```
EXPLANATIONS:
```{human-generated explanation from the pilot structured in XML format}```

Given an ORIGINAL sentence, a REVISED sentence with grammatical error corrections and fluency edits, and an ALIGNED sentence that shows

what changed from ORIGINAL to REVISED, please explain why each revision was made.

ORIGINAL SENTENCE: ```{sentence from GMEG dataset}```
REVISED SENTENCE: ```{human annotator revised version from GMEG dataset}```
ALIGNED SENTENCE: ```{aligned version}```
EXPLANATIONS:

C.2. Discussion of Prompt Qualities

The instructions use the original, revised, and aligned sentences because we found that only including the aligned sentence made it harder for the LLM to interpret the revisions. We also asked for XML formatted output to constrain the output for simpler postprocessing.

For the three in-context examples, we included a constraint such that the last example was sampled from the sentences with more than one revision to ensure that the prompt demonstrates the possibility of more than one explanatory statement. The first two examples were sampled from the entire pilot set and could include one or more revisions.

C.3. Discussion of Other Prompts Tried

As discussed in the previous section, we tried prompts with only the aligned sentence but found that it was hard for the LLM to identify the revisions. We followed the rationale that using all three versions of the sentences could help the LLM understand the markup of the aligned sentences and find the edits.

We also tried different prompt instructions. As a baseline, we started with a plain instruction, such as "Given this revised sentence, please explain the revisions enclosed in curly braces." We then built from this baseline and tried a long description that included how to interpret the aligned sentence and types of errors that could be explained, together with an example. We found this type of prompt was not as effective and seemed to be confusing the LLM. We also tried a few-shot scenario without any specific instructions and a chain-of-thought type of prompt. We tested all these prompts on a small set, manually reviewed the outputs, and found that the prompt we used yielded better outputs than the alternatives. As acknowledged in the limitations section, there may exist some prompt that would produce better results (on this, see also Eisenschlos et al. (2023)).

D. Side-by-Side Evaluation Categories

The tables on the following pages show the examples we gave annotators for each side-by-side evaluation category: Coverage (Table 10), Informativeness (Table 11), and Correctness (Table 12).

E. Side-by-Side Evaluation Sample Item

Figure 4 on the following page shows the annotation interface for the side-by-side task comparing human-written explanations with LLM-generated explanations. Annotators saw the aligned sentence and two explanations with their source anonymized. They selected an overall preference and then a preference, if any, for the three individual subcategories of Coverage, Informativeness, and Correctness. Finally, they had an option to include additional notes.

Coverage: Whether the explanations cover all the revisions in the sentence.				
Aligned	**All revisions covered**	**Missing revisions**	**Comments**	**id**
I **{learnt=>learned}** from **{you=>your}** letter that I've won the first prize.	- "Learned" is the preferred spelling in American English. - "Your" was used to indicate possession.	- "Learned" is the preferred spelling in American English.	The explanation misses the second revision in the sentence.	`fce_test_215`

Table 10: Examples for the Coverage category in the SBS evaluation.

Informativeness: Whether the explanations have enough information to understand the revision.				
Aligned	**More informative**	**Less informative**	**Comments**	**id**
However, it is still unclear whether he will **{dawn=>don}** the mask upon his return or not.	- 'Don,' meaning to wear, is the correct word here. 'Dawn' refers to the beginning of something, especially the day.	- "Don" was substituted for "dawn."	In the less informative explanation, there is only a description of what changed but no reason why.	`wiki_test_667`

Table 11: Examples for the Informativeness category in the SBS evaluation.

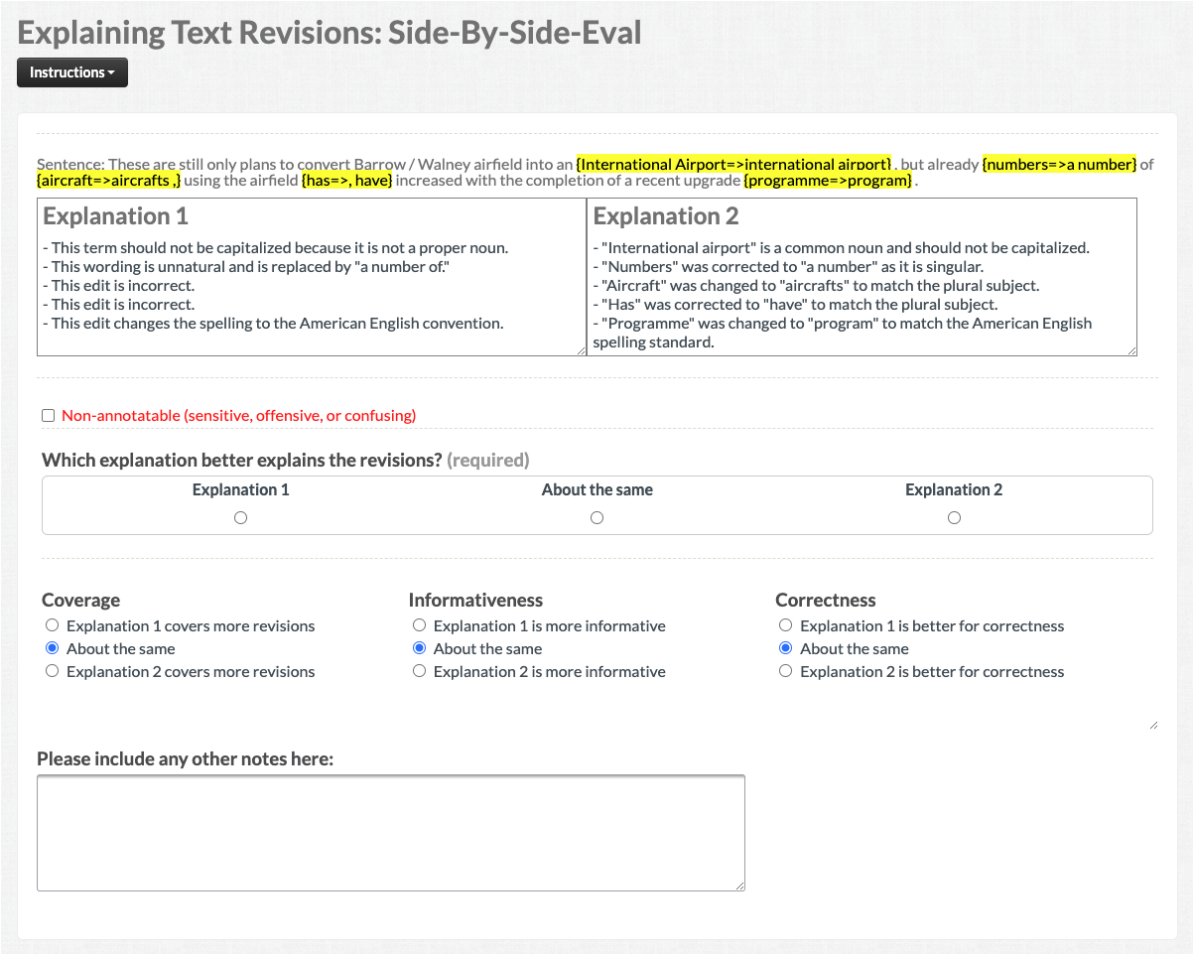

Figure 4: Sample item for the SBS task

Correctness: Whether the explanations are correct/accurate.				
Aligned	**Correct/accurate**	**Incorrect/inaccurate**	**Comments**	**id**
Bradley has three sons : Kevin , Kyle , and Allyn **{!!=>!}**	- Though sometimes more exclamation points are used for stylistic effect, using only one is considered grammatical.	- **The exclamation marks are unnecessary and were removed** to improve the overall fluency of the sentence.	The incorrect explanation says that the exclamation marks were removed, but actually, only one of them was removed. This is **not a true description** of the revision.	`wiki_test_492`
He started to shine in the 2003-04 season with Persepolis, where he was used as the defensive midfielder, and moved to Esteghlal Ahvaz, where he stayed for 4 seasons, and had excellent **{season=>seasons}** in 2007-08, where he played the most minutes of any player in the league.	- "Season" was substituted with "seasons". The player had excellent performances in multiple seasons, not just one.	- "Season" was changed to "seasons" **to match the plural form of "excellent."**	In this case, "excellent" (and, in general, adjectives in English) does not have a plural form. So this is **not a valid reason** for the change.	`wiki_dev_379`
Later, a new law was proposed in 1998, **{specifing=>specifying}** that a lawyer who was once a judge or a prosecutor must not serve as a lawyer **{=>,} within two years {=>,}** at criminal trials of a court of which he was a member directly before.	- The spelling of 'specifying' is corrected. - This parenthetical is set off with commas.	- "Specifing" is a misspelling of "specifying," so it was substituted for the correct spelling. - A comma was inserted after "lawyer" to separate the introductory clause from the main clause. - A comma was inserted after "two years" to separate it from the rest of the sentence. - **The preposition "at" was deleted because it is not needed in this context.**	In this case, the last explanatory statement in the incorrect/inaccurate explanation is about **a revision that did not occur in the sentence**.	`wiki_test_627`

Table 12: Examples for the Correctness category in the SBS evaluation.