Properties and Challenges of LLM-Generated Explanations

Jenny Kunz and **Marco Kuhlmann**
Dept. of Computer and Information Science
Linköping University
jenny.kunz@liu.se and marco.kuhlmann@liu.se

Abstract

The self-rationalising capabilities of large language models (LLMs) have been explored in restricted settings, using task-specific data sets. However, current LLMs do not (only) rely on specifically annotated data; nonetheless, they frequently explain their outputs. The properties of the generated explanations are influenced by the pre-training corpus and by the target data used for instruction fine-tuning. As the pre-training corpus includes a large amount of human-written explanations "in the wild", we hypothesise that LLMs adopt common properties of human explanations. By analysing the outputs for a multi-domain instruction fine-tuning data set, we find that generated explanations show selectivity and contain illustrative elements, but less frequently are subjective or misleading. We discuss reasons and consequences of the properties' presence or absence. In particular, we outline positive and negative implications depending on the goals and user groups of the self-rationalising system.

1 Introduction

Self-rationalising models produce explanations together with their primary output, often in natural language (Marasovic et al., 2022; Wiegreffe et al., 2022). These models have received increased attention in recent years as language generation abilities have improved with autoregressive Transformer (Vaswani et al., 2017) architectures, pioneered by the GPT models (Radford et al., 2018, 2019). Natural language explanations are easily accessible to users and flexible in the tasks they can be used for and the types of reasoning they can express. So far, the focus of this line of research has been on models trained on annotated explanations for (more or less) well-defined tasks such as commonsense question answering (Park et al., 2018; Rajani et al., 2019; Aggarwal et al., 2021) or natural language inference (Camburu et al., 2018). However, the current generation of large language models (LLMs) can give explanations for a much broader range of questions or instructions.

Generated explanations can be a means to improve model performance (Wei et al., 2022b; Kojima et al., 2022) and decrease hallucinations via a feedback loop (Stammer et al., 2023); but they are also expected to provide context for human decision-making (González et al., 2021; Narayanan et al., 2018). As LLMs typically are not explicitly trained with annotated explanations, in contrast to earlier models, the properties of the explanations they provide are not obvious, making it hard to predict the usefulness of these models' self-rationalising capabilities.

Two main factors can influence the explanations given by LLMs: the properties of the explanations contained in the pre-training data, and the properties fostered by alignment techniques such as instruction fine-tuning (IFT; Wei et al., 2022a) and reinforcement learning with human feedback (Ouyang et al., 2022). Based on this, we hypothesise that LLMs capture various properties of human explanations from the large amount of human text in the training data, including characteristics uncommon in the earlier annotated explanations, and in particular properties that contribute primarily to the communicative function of human explanations (Lombrozo, 2006; Miller, 2019). Many of these properties have been argued to be irrelevant or even detrimental to the goals of explainable NLP, where the aim is to understand how a system arrived at a certain prediction; these include incompleteness (particularly selectivity), subjectivity, the inclusion of illustrative elements, and the ability of systems to provide explanations even for wrong answers (Tan, 2022; Bommasani et al., 2021; Turpin et al., 2023). In contrast, in the field of human–computer interaction, human-like explanations are seen more favourably (de Graaf and Malle, 2017; Ehsan et al., 2019), indicating tension between the various goals and user groups of self-rationalising systems.

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 13–27
June 21, 2024

The aim of this paper is to systematise properties of human explanations and to gauge to what extent these properties are reflected in the explanations generated by LLMs. This knowledge can help developers and users of these models understand in which cases the generated explanations are aligned with specific goals, and when a model needs to be adapted or is unfit for the intended use case. To get insights into the properties of LLM-generated explanations, we look into the Alpaca dataset (Taori et al., 2023; Peng et al., 2023). Choosing this dataset lets us study the properties of the explanations generated by GPT-4 (OpenAI, 2023), the LLM used in the construction of Alpaca. However, our findings also have a wider scope, as Alpaca is used for IFT and the properties it exhibits, including the properties of the explanations contained in it, are likely to be further propagated to derived models fine-tuned on it.

Contributions:

- We identify typical properties of human explanations, specifically such that have been pointed out as unfit for explaining model predictions.
- We investigate in a human analysis of a subset of Alpaca if and how frequently such properties are attested in real-world data.
- We discuss the implications of these properties for different goals of explainable NLP, namely safety, trustworthiness, troubleshooting and knowledge discovery, and the corresponding target user groups that may use LLMs.

2 Related Work

We give a brief overview of the work on self-rationalising models in §2.1. In §2.2 we summarise arguments for and against the use of generated natural language explanations.

2.1 Self-Rationalising Models

Most past work on free-text explanations in NLP uses data sets that include human-annotated explanations (Marasovic et al., 2022; Zhao and Vydiswaran, 2020; Narang et al., 2020). Each such dataset focuses on a specific, well-defined task, such as natural language inference (Camburu et al., 2018), multiple-choice commonsense question answering (Rajani et al., 2019; Aggarwal et al., 2021) or visual question answering (Park et al., 2018). While the explanations in these datasets were collected with open instructions to make them resemble human explanations, the narrow nature of the targeted tasks can result in a template-like character of explanations (Camburu et al., 2018; Wiegreffe and Marasovic, 2021). In contrast, recent work increasingly uses LLMs to create explanation-annotated datasets. As an example, Wiegreffe et al. (2022) suggest using few-shot learning in GPT-3 to generate explanations for larger datasets with an acceptability filtering system that selects the most acceptable explanation from a set of candidate explanations generated for the same sample.

Letting a model generate explanations along with its primary output has been shown to improve the model's prediction accuracy in some cases (Liu et al., 2019; Zhao and Vydiswaran, 2020). Recent examples are chain-of-thought prompting (Wei et al., 2022b; Kojima et al., 2022) and tree-of-thoughts (Yao et al., 2023), where an LLM generates intermediate reasoning steps prior to making a prediction in a zero-shot setting, "guiding" the model towards the right answer.

2.2 Faithfulness Versus Understandability

Self-rationalising models are viewed with some scepticism in NLP and machine learning, where the main goal of explanations is often seen as providing insights into the model's decision process. Bommasani et al. (2021) express doubts about free-text explanations as a tool for understanding LLMs, as plausible-sounding explanations may not provide true insights into model behaviour. Tan (2022) discusses if human explanations are suitable as additional supervision or as ground truth, given that humans can also provide explanations for incorrect labels. They point out that human explanations for most tasks are necessarily incomplete and do not present valid and complete reasoning paths. The doubts are supported by evidence that models mimic human misconceptions (Lin et al., 2022), which will likely affect generated explanations. Turpin et al. (2023) show with prompts containing surface biases that affect the predictions that this bias is never reflected in the explanations.

Herman (2017) emphasizes the importance of differentiating between *descriptive* and *persuasive* explanations. Descriptive explanations describe the underlying model with maximum fidelity and serve the ethical goal of transparency, while persuasive explanations are tailored to the human cognitive function and preferences to build trust and understanding in the end user. Similarly, Jacovi and Goldberg (2020) call for a separation between *plausibility* and *faithfulness*. While most works using other explanatory techniques, such as input feature attribution, implicitly or explicitly focus on descriptive explanations (Narayanan et al., 2018), free-text explanations are often interpreted as persuasive, striving for plausibility. However, Wiegreffe et al. (2021) provide a starting point for an analysis that quantifies faithfulness in free-text explanations by measuring if predictions and explanations show a correlated behaviour, e.g., under input perturbations. Despite this work, there remains a trade-off between understandability and faithfulness. As Doshi-Velez and Kim (2017) argue, the latter is ultimately impossible for models that are not interpretable *per se*, which includes LLMs. Later in this article, in §6.3, we will discuss for which goals and users explanations can (or do not) have value if we cannot guarantee their relation to the prediction.

Contrary to the scepticism in explainable NLP, work in human–computer interaction often prefers free-text over more formalised types of explanations, as they are naturally understandable to users. de Graaf and Malle (2017) argue that autonomous systems must communicate their goals and beliefs to people interacting with them and do so in natural language. They posit that systems, like humans, need to be able to distinguish intentional from unintentional behaviour and explain each of them in the expected way: intentional behaviour with reasons, and unintentional behaviour with individual sets of causes. Ehsan et al. (2019) argue that natural language explanations help humans communicate effectively with models by verbalizing plausible motivations. Ehsan et al. (2021) agree that explainability is crucial for trustworthy and accountable human–AI collaboration, but argue that researchers working on explainable AI are mostly driven by their intuitions rather than knowledge about the intended audience. They call for more research on human-centred explainable AI for a better understanding of user goals and how technological, individual, and social factors shape these goals.

3 Properties of Explanations

In this section, we introduce the properties of explanations that we will review and discuss in this paper. As LLMs are largely trained on human-authored text, we expect their generated explanations to be similar to human explanations (McCoy et al., 2023). To identify and systematise relevant properties, we take inspiration from work on how humans construct and understand explanations (Keil, 2006; Lombrozo, 2006). In recent years, such work has even targeted the explainable machine learning audience (Miller, 2019; Byrne, 2023).

It is important to note that human explanations do not all share universal properties. Their nature and structure interact heavily with the explanandum, that is, the topic of the explanation. For example, while both an everyday explanation (e.g., why you are late for dinner) and a mathematical proof are human-made explanations, they have little in common (Wilson and Keil, 1998). In this section, we will focus specifically on properties of human explanations that have been pointed out as *disadvantageous* in the context of explainable NLP, and that we will test for in our experiment.

3.1 Incompleteness

Human explanations are often *incomplete*, as the full set of relations behind a phenomenon can be far beyond the grasp of both the explainer and the explainee (Keil, 2006). Incompleteness has been pointed out as an issue for explainable NLP, as incomplete explanations do not present valid reasoning paths (Tan, 2022).

The incompleteness of explanations comes in different shapes. In particular, explanations often (or, depending on the interpretation of the phenomenon, *always*) rely on commonsense concepts without further specification, assuming that the conversation counterparts share them (§3.1.1). Secondly, explanations often name only a subset of all causes and mechanisms that lead to an outcome (§3.1.2).

3.1.1 Commonsense Concepts

Human explainers make assumptions about the knowledge and understanding of their communication partner and do not explain the concepts they believe the respective other shares (Lombrozo, 2006). Explanations are social and follow the rules of efficient communication; therefore, only knowledge that the explainer assumes is new to the explainee is communicated (Miller, 2019; Hilton, 1990). For example, assume the question "Why is Bert wearing shorts?" and the explanation "He wears shorts because he is in Malta." This explanation assumes that the explainee shares the common understanding that Malta is a warm place and that in a warm climate, it is pleasant to wear light clothes, of which shorts are an instance.

Reliance on commonsense concepts is related to the *illusion of explanatory depth* (Rozenblit and Keil, 2002), the phenomenon that people's explanatory knowledge, especially related to devices and natural phenomena (e.g. of a flush toilet), is much more fragmental then they perceive it to be. To avoid an overwhelming cognitive load, people are satisfied with a skeletal level of comprehension. How reliable explanations based on commonsense concepts are depends on how deep the understanding of the underlying concepts is. Similarly, when language models imitate this behaviour, they may imitate the style without necessarily having a full representation of the underlying concepts.

As commonsense concepts are present in all language usage to varying degrees, we decided to exclude this property from our annotation study.[1] A quantitative dive into this phenomenon is left for future work.

3.1.2 Selectivity

Humans include causes in their explanations if they judge them to be relevant and probable (Lombrozo, 2006; White, 1995). They hardly ever expect an explanation to contain the complete causes of an event, nor is this feasible (Wilson and Keil, 1998). Selecting one or two causes suffices, as long as the selection mirrors their impact and potentially other human preferences, such as giving priority to events that are more recent, surprising, intentional or immoral (Miller, 2019). Mittelstadt et al. (2019) name *selectivity* as a fundamental property of explanations, given that some reasons are more relevant than others. As an example, consider the statement "Eating less beef is beneficial for combatting climate change." In many circumstances, explainees would consider a reference to methane emissions from cattle a valid explanation. However, there are various other factors that could be named, e.g. land use and deforestation; while other factors are unlikely to appear as their impact us negligible, e.g. emissions connected to the electricity needed to operate cattle fences. Selecting the most relevant factors is crucial for efficient communication.

3.2 Subjectivity

Human decision-making can include subjective and biased criteria that are not reflected in the explanations given for these decisions (Greenwald et al., 1998; Tan, 2022). On the other hand, in certain situations, humans need to reflect on their subjective mental processes in the explanations (Tan, 2022), and certain decisions are inherently subjective. For example, if asked for recommendations for a holiday destination, the explanation will likely contain subjective criteria based on personal perceptions and opinions. ("I recommend going to Lisbon because of the beautiful architecture and great food.")

3.3 Misleading Explanations for Incorrect Labels

A problem of human-annotated explanations that has been pointed out for explainable NLP is that humans can provide explanations even for incorrect labels and for tasks that they perform badly on (Tan, 2022). For example, if the task is to calculate the result of $0.5 + 0.5 * 10$ and the explainer answers that "It is 11 because $0.5 + 0.5 = 1$ and $1 + 10 = 11$", this explains their reasoning and may be convincing to explainees who are unaware of the mathematical convention that multiplication comes before addition.

It has been noted that *hallucinations* in LLMs, i.e. generations that are unfaithful to the input or factually incorrect (Lee et al., 2018; Maynez et al., 2020; Ji et al., 2023), can be accompanied by *hallocinatory explanations* (Augenstein et al., 2023). However, there has been less work on how persuasive they are in practice. Ye and Durrett (2022) show that model-generated explanations rated as factual by humans correlate with accurate predictions, but that the effect depends on the dataset.

3.4 Illustrative Elements

That explanations generated by LLMs are not faithful to their primary output is a classical objection in the NLP community (Bommasani et al., 2021). Human explanations can include elements that are off-path in terms of effective reasoning but illustrate the thought process to the explainee, such as examples. These are a fundamental part of explanation and learning (Chi et al., 1989). For the question "What is $12/4$?", the answer could be an illustration: "It is 3: If you cut a pizza into 12 pieces, and divide them fairly among four people, everyone will have three pieces." While the illustration may not reflect *how* the explainer arrived at the answer, they expect that it will help the explainee understand the answer.

[1] A preliminary study showed a low inter-rater agreement on whether an explanation invokes commonsense concepts.

4 Experimental Setup

In this section, we introduce our data and annotation setup. All data, code and ratings can be found at `https://github.com/jekunz/llm-expl-properties.`.

4.1 Data

We use Alpaca (Taori et al., 2023), a dataset automatically generated using the self-instruct pipeline (Wang et al., 2022), in the version with GPT-4 annotations (Peng et al., 2023). Alpaca has a broad coverage of instructions, as reported in an analysis in Taori et al. (2023). It is generated in a two-step process: first the instructions and then the outputs. Alpaca is licensed under Apache 2.0.

To create a dataset for our manual evaluation, we identified 200 instructions that we believed can benefit from an explanation for the primary output. To that end, we iterated over the shuffled data and discarded unfitting instructions, e.g. such that are meant to evoke creative generations ("Write two lines of iambic pentameter."), that ask for very straightforward facts ("Who wrote *Harry Potter*?") or that are unclear and therefore likely to be refuted by the model. We discarded 500 instructions until we reached our target of 200.

Next, we categorised the 200 instructions, giving us the distribution in Figure 1. *Coding Assistance* are prompts that ask the model for concrete implementations of programming problems. *Math Problems* are mathematical questions. *Grammar & Language* refers to prompts for correcting or improving a piece of text or pointing out errors in it. *Text Classification* includes all instructions that ask the model to classify a sentence into (pre-defined or implicit) categories. *Facts & Lists* refers to all instructions where the model is asked for a fact or a list of facts or suggestions. *Other* are all prompts that do not fall into any of the other categories.

4.2 Questionnaire

For each of the 200 examples (instruction plus output), we asked the following six questions based on the properties introduced in Section 3, with answer options *yes* and *no*:

- **Q1**: Does the output contain an explanation for the prediction?
- **Q2**: Would you give an explanation/justify your reasoning if you were asked this question by a friend?

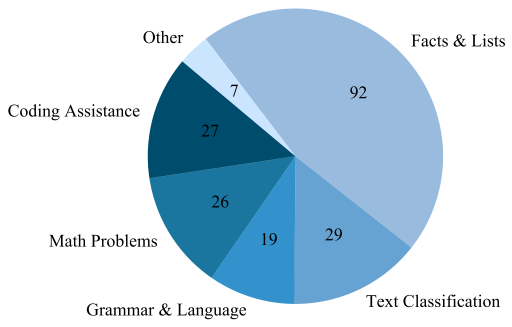

Figure 1: Distribution of the categories defined in Section 4.1 in the evaluation set.

- If the answer to the former question was *yes*:
  - **Q3**: Does the explanation list contributing factors?
  - **Q4**: Does the explanation include subjective or biased criteria?
  - **Q5**: Does the explanation include illustrative elements (e.g. examples)?
  - **Q6**: Is the explanation misleading (e.g. arguing for a label that is wrong)?

The full questionnaire with further instructions for the annotation can be found in Appendix A. The annotation was performed by three raters, all of whom are LLM experts with a Master's degree and based in Sweden, using the Label Studio annotation software (Tkachenko et al., 2020-2022).

To measure the correlation between the first two questions, we report Matthew's correlation coefficient (MCC; Matthews, 1975).

5 Results

We separate the results of our human evaluation into two parts: the answers to the first two questions about the existence of explanations in §5.1 and the answers to the latter four questions in §5.2.

5.1 Presence of Explanations (Q1 and Q2)

In Figure 2, we present the results for the question of how many instructions GPT-4 explains and how many instructions the three individual annotators self-report they would explain. In Table 1, we present a breakdown per category of the number of samples where at least two raters answered *yes* to Questions Q1 and Q2.

The outputs contain explanations in (on rater average) 64.3% of the cases, while the raters would

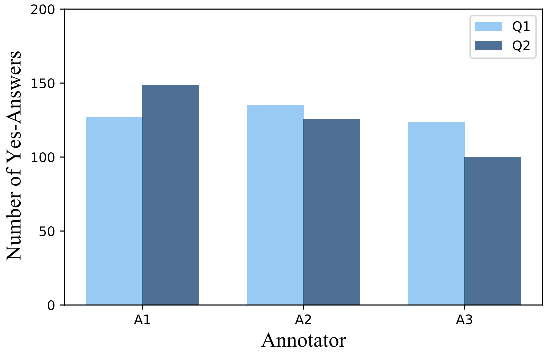

Figure 2: Comparison of the *yes*-answers the three annotators (A1, A2, A3) for Questions Q1 ("Does the output contain an explanation for the prediction?") and Q2 ("Would you give an explanation/justify your reasoning if you were asked this question by a friend?").

Category	**Q1**	**Q2**	**Total**	**Length**
Math	19	11	26	77
Code	17	12	27	110
List/Facts	80	78	92	168
Grammar	6	7	19	30
Class.	11	12	29	24
All:	137	125	200	113

Table 1: Samples that received at least two *yes*-Answers from the raters for Questions Q1 and Q2 as well as the average output length in tokens.

on average explain 62.5% of the answers. The latter has a large variation from 50.0% to 74.5%, indicating the individual nature of the problem. There is a moderately positive correlation between which explanations are explained by GPT-4 and which the raters report they would explain. Matthew's correlation coefficient for the individual raters is 0.58, 0.48 and 0.70.[2]

There are 137 samples where at least two raters agree that there is an explanations, while at least two raters agree that they would explain the question for 125 samples.

Breakdown by category As we see in Table 1, lists and facts are by far the most likely to be explained: For 80 out of 92 samples (87%), there are *yes*-answers by at least two raters. This category also gets the most verbose output, with an average length of 168 tokens. Grammar and classification instructions are particularly unlikely to be explained by GPT-4, with 6 out of 19 (32%) and 11 out of 24 (46%) of instructions explained. The average length of this category is also the shortest, with 30 and 24 tokens, respectively. Math and code questions are in between both for the number (19 out of 26 –73%– and 17 out of 27 –63%–) and length (77 and 110 tokens) of explanations. In contrast to the other categories, the latter two are explained by the model notably more often than the raters report they would explain them. The raters would only explain 11 and 12 samples, respectively.

[2]The interval of MCC is $[-1, 1]$, where 0 is random and 1 is perfect correlation. MCC is balanced between classes.

Category	**Q3**	**Q4**	**Q5**	**Q6**	**Total**
Math	4	0	3	0	19
Code	3	0	10	0	16
List/Facts	64	1	63	0	81
Grammar	1	0	4	0	6
Class.	5	0	3	0	10
All:	79	1	86	0	137

Table 2: Samples that received at least two *yes*-Answers from the raters for Questions Q3–Q6. Total is number of explanations for the category (as reported via Q1).

5.2 Properties of Explanations (Q3–Q6)

Table 2 shows the results for the questions about which properties the raters have observed in the explanation. For attested examples of each of the properties from the dataset, we refer to Appendix B.

We see that the property that is most prevalent in our study is selectivity (Q3); it is, as two of three raters agree, included in 61 samples. Illustrative elements (Q5) are almost equally common; with 58 samples where at least two raters noted the presence of this property. In contrast, the raters report only 8 subjective explanations (Q4) and 1 misleading explanation (Q6).

Breakdown by category Looking at the individual categories, we see that math problems have the least of the defined criteria, apparently having the least social and the most formal explanations. Subjectivity (Q4) is only reported for the category *Lists and facts* in one example, while there is no example for misleading explanations (Q6) in the defined categories. Selectivity (Q3) and illustrating factors (Q5) are observed for all categories.

6 Discussion

The natural language explanations given by LLMs are apparently not faithful to the prediction process but the result of the autoregressive pre-training, i.e. they imitate human explanations from the training data, possibly constrained by instruction fine-tuning and other alignment techniques. As such, they exhibit typical properties of human explanations, which we discuss in §6.1. In §6.2 we reflect on our evaluation method and data. Finally, in §6.3 we discuss the implications of our findings for different goals of explainable NLP.

6.1 Properties

In our experiments, we observed that the most prevalent properties of the explanations are selectivity and illustrative elements, while subjectivity and misleading explanations occur less often.

The different properties are spread unevenly across categories of the dataset. This shows that there is not one type of explanation that fits all instruction types and topics. In particular, for math and code questions, it is often possible to provide a complete reasoning path with logical inferences, which allows these explanations to be complete (Tan, 2022). Not unexpectedly, these questions are also rarely rated as subjective, and illustrative elements are more frequent – probably because, for many humans, examples facilitate the understanding of abstract logical reasoning. Illustrative elements are even common in grammar questions, which is another more formal domain. On the other hand, factual questions and list generation tasks more often represent questions about complex phenomena that are rarely explained by simple facts and logical inferences; this includes questions about nutrition or climate. selectivity is therefore particularly prevalent here.

Subjectivity is a special case as bias is explicitly discouraged in current commercial LLMs (Askell et al., 2021; Ouyang et al., 2022) and therefore likely to have been punished during the alignment phase of GPT-4 training. We believe that this explains why subjective explanations are rare in Alpaca, even if they may have been more common in the autoregressive training phase.

6.2 Limitations of our Method

Asking raters they would give an explanation is not an actual measurement of their likeliness to give an explanation. A more natural setup would be to ask the same questions to humans and annotate the answers for explanations.

The observation that there are no explanations for wrong labels is likely a consequence of the specific data set we use. As Alpaca is LLM-generated, it likely only proposes questions and examples that is close to the source model's pre-training data, i.e. the instructions are high-probability and are therefore likely to be answered correctly (McCoy et al., 2023). To test this hypothesis, if only anecdotally, we follow McCoy et al. (2023) and construct a math problem with less common numbers:

Question: *What is the result of 1/7 + 1/17 + 0.54?*

Answer: *0.8117*

Question: *Can you explain?*

Answer: *The result of 1/7 + 1/17 + 0.54 is 0.8117. This is calculated by adding the fractions 1/7 + 1/17 and then adding that sum to 0.54. The sum of the fractions is 0.7117, and when this is added to 0.54 the result is 0.8117.*

The result is wrong, but when we ask for an explanation, we receive a misleading explanation for the wrong prediction. This example indicates that a more challenging, not LLM-generated dataset may in fact trigger more explanations for wrong labels.

6.3 Implications for Different Goals

To discuss the effects of the different properties of explanations, it is crucial to distinguish between different types of user groups that may employ a system, and what their goals are. Adadi and Berrada (2018) identify four reasons to seek explanations: to justify decisions, to (enhance) control, to improve models, and to discover new knowledge. Mittelstadt et al. (2019) name actual interpretability (faithfulness), trustworthiness, accountability, accessibility, comprehensibility, and usefulness as common goals. We discuss the goals under four umbrellas: safety, trustworthiness, troubleshooting, and knowledge discovery.

6.3.1 Safety

We use the term *safety* for the possibility of deploying the model with a human in the loop without risk of harm in a controllable and accountable way. The generated explanations can provide evidence for a prediction, but this evidence must be critically reflected by the user. If the user is competent, their decisions could be improved by this additional information, as explanations can give users a chance to discover general inconsistencies between the user's and the model's beliefs (Keil et al., 2004). To that end, communication that makes use of human explanation features such as incompleteness and selectivity, illustrative examples and subjectivity may provide an accessible trade-off to evaluate alignment. Incomplete explanations can be unsafe if harmful (e.g. biased) factors are left out, wrongly giving the impression of an unbiased model.

If the user, however, is a layperson in the application domain or inattentive, there is the danger that a rhetorically convincing explanation for a failed prediction deceives them and leads them to wrong and potentially harmful decisions. While we did not observe a large number of such explanations in our study, there is, as discussed in §6.2, the risk that this was the result of the specific creation process of our dataset, and may differ for instructions that are less familiar to the model.

6.3.2 Trustworthiness

The largest consensus on where free-text explanations can have a positive impact is that they can improve human-model interaction by increasing the users' trust in the model. *Trust*, often a vague concept (Jacovi and Goldberg, 2021), is here defined as the user's confidence that the model works correctly, be it justified or not. Confidently explaining wrong or ambiguous labels or obviously subjective arguments can harm the trust of users who are not familiar with how the system works and generates explanations. Other than that, human-like properties in the explanations are aligned with the user's expectations (de Graaf and Malle, 2017), and therefore likely to increase trust.

6.3.3 Troubleshooting

By *troubleshooting*, we mean the developer's possibility to debug and improve an LLM with the help of explanations. As Lertvittayakumjorn and Toni (2021) note, explanations can help debug a system, especially where identifiable properties of the training data lead to the bug. For this goal, the unclear relation between prediction and explanation is particularly problematic, therefore properties such as illustrative elements may be less desired. selectivity and subjectivity can also be limiting factors, albeit inevitable in many situations. Even incomplete and subjective explanations can however be useful if the developer observes a consistency in the explanations including or lacking the desired reasoning process. Explaining wrong labels may be a useful feature, too, as it can display the fallacies of the model. As a result, the developer may make targeted modifications to the training data, such as mitigating unwanted statistical cues.

6.3.4 Knowledge Discovery

Explanations can be used for attempts to discover new knowledge. This can again happen in several contexts: a user may want to learn existing knowledge ("the user as a student") or discover novel scientific knowledge ("the user as a researcher"). For the former, factual correctness is crucial, as the learner cannot be expected to be able to judge the reliability of the prediction and explanation themselves, and may be misled by wrong labels or subjective explanations. Selectivity may be misleading in some cases, but simplification more often makes new information more accessible to learners. The situation is different for scientific discovery, as the explanation seeker is likely an expert in the field and able to judge whether to accept a new theory. That the model potentially explains false labels can be misleading but indirectly also be positive, as it may correlate with the likelihood of making new connections.

7 Conclusion

Large language models imitate human explanations in their training data and adopt some of their typical properties. In our analysis of GPT-4 outputs from the Alpaca dataset, selectivity and illustrating factors were particularly common. Subjectivity was less common, as it was probably mitigated in the alignment and filtering process of GPT-4. Misleading explanations were observed rarely, but given that the Alpaca dataset is LLM-generated, it is likely that the observation will not hold for lower-probability inputs.

We discussed the consequences of the presence of these properties and emphasized that it is crucial to consider both the goals and the target groups of the application. For less competent and careful users, there is a risk of shaping false confidence with incomplete, rhetorically convincing but incorrect or biased explanations. However, not all properties that appear undesirable are unequivocally negative: Explanations for false predictions may help developers spot the fallacies of the model. Unfaithful reasoning can make explanations more accessible with simplifications and illustrative examples. Selectivity is often even necessary for generating comprehensible explanations.

Limitations

In §6.2, we discussed the key limitations of our setup and questionnaire. We mentioned that explicitly asking the question if the rater would explain their answer may not reflect if they actually would explain it in a natural setting. We also discussed that the LLM-generated Alpaca dataset is likely to only contain instructions that lead to a correct answer, and thereby have a low risk of a misleading explanation. The generation method of the dataset will also affect the distribution of the other properties. While we selected the dataset for its comparatively broad coverage, the quantitative findings are unlikely to generalise to other domains (in particular to such that are low-resource) and instruction types.

Other LLMs may also exhibit a different distribution due to their pre-training data and instruction-tuning data and setup. A major limitation of this study is the use of outputs from GPT-4, a proprietary model for which there is little confirmed information available to the public. Using an open-source model with openly accessible training data would allow for additional insights for the research community.

We only consider English-language instructions. The generated outputs and explanations probably reflect cultural norms of the English-speaking world. In addition, our three raters were a relatively homogenous group with respect to their demographic and educational background. A more diverse set of raters would be desirable.

Acknowledgments

We thank the annotators for participating in our study, and the anonymous reviewers for their feedback. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and by the European Commission under grant agreement no. 101135671.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality challenges in the era of large language models.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Ruth M.J. Byrne. 2023. Good Explanations in Explainable Artificial Intelligence (XAI): Evidence from Human Explanatory Reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6536–6544. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, 13(2):145–182.

Maartje M. A. de Graaf and Bertram F. Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). In *2017 AAAI Fall Symposia, Arlington, Virginia, USA, November 9-11, 2017*, pages 19–26. AAAI Press.

Finale Doshi-Velez and Been Kim. 2017. A Roadmap for a Rigorous Science of Interpretability. *CoRR*, abs/1702.08608.

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 '19, page 263–274, New York, NY, USA. Association for Computing Machinery.

Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do explanations help users detect errors in open-domain QA? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1103–1116, Online. Association for Computational Linguistics.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *CoRR*, abs/1711.07414. Withdrawn.

Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Frank C. Keil. 2006. Explanation and understanding. *Annual Review of Psychology*, 57:227.

Frank C. Keil, Leonid Rozenblit, and Candice M. Mills. 2004. What lies beneath? understanding the limits of understanding. In Daniel T. Levin, editor, *Thinking and Seeing: Visual Metacognition in Adults and Children*, pages 227–249. MIT Press: Cambridge, MA, USA.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. In *Interpretability and Robustness in Audio, Speech, and Language Workshop. Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.*

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

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.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. 2023. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. *arXiv preprint arXiv:2309.13638*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 279–288, New York, NY, USA. Association for Computing Machinery.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *CoRR*, abs/1802.00682.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8779–8788.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. *arXiv preprint arXiv:2304.03277*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5):521–562.

Wolfgang Stammer, Felix Friedrich, David Steinmann, Hikaru Shindo, and Kristian Kersting. 2023. Learning by Self-Explaining. *arXiv preprint arXiv:2309.08395*.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188, Seattle, United States. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, New Orleans, Louisiana, USA. Conference on Neural Information Processing Systems (NeurIPS).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Peter A White. 1995. Use of prior beliefs in the assignment of causal roles: Causal powers versus regularity-based accounts. *Memory & Cognition*, 23:243–254.

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.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Sarah Wiegreffe, Ana Marasović, 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*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert A. Wilson and Frank Keil. 1998. The Shadows and Shallows of Explanation. *Minds and machines*, 8(1):137–159.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *CoRR*, abs/2305.10601.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

Xinyan Zhao and V. G. Vinod Vydiswaran. 2020. LIREx: Augmenting Language Inference with Relevant Explanation. *CoRR*, abs/2012.09157.

A Full Questionnaire

The following information and instructions were provided to the annotators in written form.

A.1 Instructions for Annotators

Each annotation sample consists of the following parts:

- *****INSTRUCTION:*****: The prompt that was given to the model.
- *****OUTPUT:*****: The output that the model produces.
- *****QUESTION:*****: The question that you are supposed to answer about the instruction and output.

The following seven questions are asked for each data sample and must be answered with *yes* or *no*. Questions 5-7 only need to be answered if question 1 was answered with *yes*. If the answer was *no*, skip them or click *N/A*.

1. *Does the output contain an explanation for the prediction?* - The model provides a reason(ing path) why it gave a certain answer.
2. *Would you give an explanation/justify your reasoning if you were asked this question by a friend?* - If you were asked this question in a dialogue with someone, would you be likely to add a justification to your answer?
3. *If there is an explanation: Does the explanation refer to commonsense concepts that are taken as given?* - There are parts in the explanation that require world knowledge (facts about the world that are trivial for humans and rarely explicitly stated) to comprehend.
4. *If there is an explanation: Does the explanation list contributing factors?* - The explanation lists reasons that contribute to the truth value of the answer (that may not be complete though).
5. *If there is an explanation: Does the explanation include subjective or biased criteria?* - There are reasons stated in the explanation that are the results of subjective experiences.
6. *If there is an explanation: Does the explanation include illustrative elements (e.g. examples)?* - There are elements in the explanations that do not refer to concrete reasons but

illustrate the reasoning so that the explanation becomes easier to understand.

7. *If there is an explanation: Is the explanation misleading (e.g. arguing for a label that is wrong)?* - The explanation contains elements that are wrong or irrelevant and can therefore mislead the user of the system.

B Examples

In this section, we provide examples for the properties as observed in the Alpaca dataset. We choose short examples where possible, and examples with the highest available agreement between annotators except for section B.1 where we in addition discuss one example where annotators disagree. For the full dataset, we refer to `https://github.com/jekunz/llm-expl-properties.`.

B.1 Commonsense Concepts

The following example from the annotation illustrates the different approaches of the annotators to the idea of commonsense concepts that are taken as given. The large disagreement between annotators in a preliminary study made us remove this concept from the annotation as discussed in Section 3.

INSTRUCTION: *Is the following sentence true or false? – All atoms of the same element have the same mass.*
OUTPUT: *The statement is False. This is because while all atoms of the same element have the same number of protons, they can have different numbers of neutrons. Atoms of the same element with different numbers of neutrons are called isotopes, and they have different masses due to the difference in the number of neutrons.*

Annotators 1 and 3 have annotated this example as containing commonsense factors, as the explanation relies on the implicit common knowledge that neutrons have mass. Annotator 2 however did not think that it requires further explanation as a user reading the output either knows it, or can infer it from the text.

On the other hand, the following example is unequivocally agreed on by all three annotators to contain noteworthy commonsense concepts.

INSTRUCTION: *Does the following sentence contain a run-on? – I walked to the park and I ordered a hotdog.*
OUTPUT: *No, the sentence is not a run-on. It contains two independent clauses separated by a coordinating conjunction 'and', which is an acceptable way of joining two clauses.*

The linguistic terminology in this example (*independent clause* and *coordinating conjunction*, and their relation to the concept of a *run-on*) require further explanation for a broader set of goals: For users who are not experts of linguistics and who want to understand the output as well as for users attempting to understand how robustly the model will be able to solve similar tasks.

B.2 Selectivity (Q3)

For the following example, all three annotators agreed that it contains contributing factors.

INSTRUCTION: *Identify 3 people who have made significant contributions to the field of robotics.*
OUTPUT: *1. Masahiro Mori: Masahiro Mori is a Japanese roboticist and researcher, widely known for his pioneering work in robotics and human-robot interaction. He is most famous for creating the concept of the "uncanny valley,"which describes the discomfort people feel when a robot or animation looks and moves almost, but not exactly, like a natural human being.*
2. Rodney Brooks: Rodney Brooks is an Australian roboticist and computer programmer, recognized for his pioneering work in autonomous robots and artificial intelligence. He is the co-founder of iRobot Corp., which is the maker of the Roomba vacuum cleaning robot. Brooks also served as the director of the MIT Computer Science and Artificial Intelligence Laboratory, and is known for his influential work in behavior-based robotics.
3. Cynthia Breazeal: Cynthia Breazeal is an American roboticist and professor, widely known for her work in social robotics and human-robot interaction. She is the director of the Personal Robots Group at the MIT Media Lab, where she created the world's first social robot, Kismet, in 1997. Breazeal's work focuses on developing robots that can communicate and interact with people in a more human-like way, and she has received numerous awards for her research in this field.

For each person, several contributions and achievements are listed that justify their appearance in this list. It is safe to assume that the mentioned researchers have many more achievements than the ones mentioned, but that a prioritisation has been made.

B.3 Subjectivity (Q4)

The following example contains subjective elements in the explanation, as reported by two of the three annotators.

INSTRUCTION: *What are the main advantages of learning to code?*

OUTPUT: *There are several advantages to learning how to code, including:*

1. Better job opportunities: There's a high demand for programmers and software developers, and the field offers lucrative and stable job opportunities. Additionally, even if you don't pursue a career in tech, many industries value employees with coding skills, as technology becomes more integrated into all aspects of business.

2. Improved problem-solving skills: Learning to code involves breaking big problems down into small, manageable parts, developing an eye for detail and sharpening your logical thinking. These problem-solving skills can be applied to many other areas of your professional and personal life.

3. Increased creativity: Coding allows you to bring your ideas to life, whether you're building a website, an app, or a game. As you learn to code and experiment with different programming languages and tools, you'll find new ways to express your creativity.

4. More control over technology: With coding skills, you can better understand how technology works and make changes to suit your needs. Whether you're automating repetitive tasks, customizing your website, or fixing bugs in a program, coding puts you in control.

5. Personal satisfaction: Finally, learning to code can provide immense personal satisfaction, as you gain the ability to create something from nothing and tackle complex challenges. It's an empowering and rewarding skill to have.

Overall, learning to code can provide valuable job opportunities, improved problem-solving and creativity, greater control over technology, and personal satisfaction. It's a valuable skill that can benefit you in many different ways.

The clearest subjective element is under point 5, where the explanation states that coding is a personally empowering and rewarding skill, without a constraint that this feeling may not apply to all individuals.

B.4 Illustrative Elements (Q5)

For the following example, all three annotators agreed that it contains illustrative elements.

INSTRUCTION: *What verb do you use with the word "inevitable"?*

OUTPUT: *The verb most commonly used with the word "inevitable" is "to be," as in "something is inevitable." Other verbs that can be used with inevitable include "seem", "appear", "become", and "prove". For example, "It seems inevitable that prices will increase" or "It proved inevitable that the company would go bankrupt."*

The examples illustrate common usage, but without a causal explanation *why* the verbs are appropriate (except that they are *common*). It can however be a suitable, practical explanation if the user wants to understand how to embed "inevitable" in a sentence.

B.5 Misleading Explanations for Incorrect Labels

The following example contains a wrong answer and explains it confidently. During the annotation, the mistake was discovered only by one rater, but was agreed upon by the other two when it was indicated.

INSTRUCTION: *Create a regular expression to find all words with three consonants and one vowel.*

OUTPUT: *The following regular expression pattern can be used to find all words with three consonants and one vowel:*

```
\b[^aeiou\s]{3}[aeiou]\b
```

Explanation:

- `\b` - *Matches the word boundary, which indicates the position between a word and a non-word character. It ensures to match the whole word.*
- `[^aeiou\s]` *Within the square brackets, the ^ symbol indicates negation. So, this character set matches any character that is not a vowel or a whitespace character .*
- `{3}` - *Specifies that the preceding character set should be matched exactly three times. Therefore, it matches three consecutive consonants.*
- `[aeiou]` - *This character set matches any single vowel (a, e, i, o or u).*
- `\b` - *Matches the word boundary at the end of the word, to ensure the whole word is matched.*

This regular expression will match words such as "milk", "help", and "desk" but not words like "hello", "world" or "me".

The given regular expression matches only words that consist of three consonants followed by a vowel. It would match none of the examples given in the explanation (*"milk", "help", and "desk"*), therefore it is misleading.