Diagnostic Evaluation of LLMs' Commonsense Reasoning Abilities

Ana Brassard

Graduate School of Information Sciences
Tohoku University

This dissertation is submitted for the degree of
Doctor of Information Science

March 2025

Acknowledgements

First and foremost, my deepest gratitude to Prof. Kentaro Inui for his generous trust when he first accepted me as a member of his research group, granting me an opportunity greater than I ever expected. Trough the years, his guidance, support and encouragement have been invaluable. I must extend similar thanks to Prof. Jun Suzuki for his always insightful and grounding advice, as well as to Prof. Keisuke Sakaguchi, who joined me in the final stages of my research, whose cheerful but sober feedback helped me overcome final hurdles and kept my spirits up in moments of doubt. Of closer collaborators, I wish to particularly thank Dr. Benjamin Heinzerling, whose cutting yet insightful comments became a constant inspiration and standard to which I align my own judgement. I also thank Keito Kudo, who graciously accepted the role of keeping me sane through weekly meetings with his helpful advice and curious, hardworking, and kind attitude. While not direct collaborators, I would also like to thank current and former members of the Tohoku NLP Group, who have particularly been a source of inspiration and a strong support network: Tatsuki Kuribayashi, Takumi Ito, Shota Sasaki, and many others. I must also extend thanks to RIKEN who provided me with more than ample resources allowing me to pursue my research without worry, as well as to Yoriko Isobe, Mayumi Sugawara, Haruka Aizawa, Asami Kato, and countless other administrative staff who gracefully let me push the limits of their patience. My deepest gratitude to my family, who have given me immeasurable support and understanding through a journey that they know is not easy. I stand here today thanks to uncountable moments, big and small, that paved my life, each step, decision, and effort returning to me in ways I could have never predicted. A particular thanks to Marie-Josée Brassard, whose presence and care has held me up in critical moments so far from home. I cannot possibly fit the full list of friends and other sources of inspiration throughout the years in this space; know that I am grateful for each and every one of you. Last but not least, thank you to Mirna and Bosco, who unexpectedly stumbled into my life but will never leave it. Many bugs, mice, and the warmest basking spots to you both.

Abstract

This dissertation explores the evaluation of commonsense reasoning through the lens of textual explanation evaluation. It begins by examining the current state of commonsense reasoning evaluation and the limitations of existing methods. A new framework is then proposed that leverages large language models (LLMs) to evaluate explanations for commonsense reasoning tasks over several fine-grained quality criteria. Then, two new datasets are introduced: one containing semi-structured explanations for a commonsense reasoning benchmark, and another containing aspect-wise judgments of quality of these explanations as well as from several other sources. The new data support several analyses of the reliability of LLMs as judges of explanation quality, finding them to be highly (but not perfectly) correlated with majority-voted human judgments but not outside of the range of human annotator agreement. Models were also found to be over-sensitive to prompting variations, as well as being over- or under-sensitive to explanation differences compared to humans. Finally, the utility of this approach is demonstrated by applying it to several model variants, such as through its training checkpoints, to show the progress of how these models learn to explain over time. The dissertation concludes with a discussion of the limitations of this approach and suggests future directions for commonsense reasoning evaluation as well as LLM-based explanation evaluation.

Table of contents

List of figures vii

List of tables viii

1 Introduction 1
- 1.1 Research Motivation and Goals . . . 1
- 1.2 Dissertation Structure . . . 3

2 Background 5
- 2.1 Commonsense Reasoning . . . 5
- 2.2 Evaluating Commonsense Reasoning . . . 6
- 2.3 Explanations . . . 7
- 2.4 Evaluating Free-Text Explanations . . . 8
- 2.5 LLMs as Judges . . . 8

3 Data Collection 10
- 3.1 COPA-SSE: Semi-Structured Explanations for Commonsense Reasoning . 10
  - 3.1.1 Design goals . . . 11
  - 3.1.2 Crowdsourcing . . . 12
  - 3.1.3 Data stats and examples . . . 16
- 3.2 ACORN: Explanations with Aspect-wise Quality Labels . . . 17
  - 3.2.1 Quality aspects . . . 17
  - 3.2.2 Source datasets . . . 18
  - 3.2.3 Crowdsourcing . . . 19
  - 3.2.4 Data stats and examples . . . 21

4 LLM-based Explanation Evaluation 25
- 4.1 Settings . . . 25
  - 4.1.1 Models . . . 25

4.1.2 Prompting Strategy . . . 26
4.1.3 Postprocessing: Label Extraction . . . 26
4.2 Alignment with human judges . . . 27
4.2.1 Inter-annotator Agreement . . . 27
4.2.2 LLMs As A Replacement for Human Evaluation . . . 28
4.2.3 LLMs As An Additional Rater . . . 30
4.2.4 Discussion . . . 31
4.3 Sensitivity to Explanation and Prompt Differences . . . 31
4.3.1 Explanation sensitivity . . . 32
4.3.2 Prompt sensitivity . . . 35
4.3.3 Discussion . . . 39
4.4 Human vs. LLMs: How Do They Disagree? . . . 39
4.4.1 Overall Bias in Human and LLM Ratings . . . 40
4.4.2 Calibrating LLM Ratings to Human Ratings . . . 41
4.5 Conclusions . . . 43

5 Application: *When* do models learn to explain? 45
5.1 Settings . . . 45
5.2 Results . . . 46
5.2.1 Size Comparison . . . 46
5.2.2 Training Checkpoints . . . 49
5.2.3 Proprietary Models . . . 49
5.3 Conclusions . . . 51

6 Discussion 52

7 Conclusion 55

References 56

List of Publications 62

List of figures

1.1 Motivational example: commonsense reasoning QA and output explanation 1
1.2 An evaluation method for textual explanations using LLMs. 2

3.1 A structured explanation manually extracted from ConceptNet 11
3.2 Form for collecting semi-structured explanations. 12
3.3 Form for rating semi-structured explanations. 13
3.4 Number of statements per explanation in COPA-SSE. 16
3.5 Rating distribution before and after re-collection of COPA-SSE 16
3.6 Data sources for ACORN 18
3.7 Explanation rating form for ACORN. 20

4.1 Impact of LLMs on inter-annotator agreement. 27
4.2 Correlation between LLMs' ratings and majority-voted humans' 29
4.3 Impact on correlation with five-way voted ratings when LLMs are added to a reduced pool of voters. 30
4.4 LLMs' sensitivity to explanation edits (heatmaps) 33
4.5 Average human *vs.* LLM ratings 40
4.6 Change in mean ratings after calibration 42
4.7 Correlation of human and LLM ratings after calibration 43

5.1 Size-wise comparison of explanation skill by Pythia models 46
5.2 Checkpoint-wise comparison of explanation skill (Pythia 70M & 6.9B) . . 48
5.3 Comparison of accuracy and explanation skill per GPT version. 50

List of tables

3.1 COPA-SSE examples 14
3.2 Top- and bottom-rated examples from COPA-SSE 15
3.3 Explanation rating criteria in ACORN 17
3.4 Samples per source dataset in ACORN 21
3.5 Ratin label distribution per criterion in ACORN 21
3.6 Top- and bottom-rated example explanations for CommonsenseQA in ACORN 22
3.7 Top- and bottom-rated example explanations for BCOPA in ACORN . . . 23
3.8 Mean ratings per data subset in ACORN 24

4.1 Full results of the experiments comparing the difference in inter-annotator agreement between humans and with a random rater replaced by an LLM (§4.2.1). All values represent Krippendorff's α averaged over 20 iterations. Extraction failures are excluded from analysis. Replace * with "Instruct" and # with "Turbo" in the model names. 28
4.2 Full results of the experiments measuring Spearman's rank correlation between majority-voted human labels and LLM-generated ones (§4.2.2). Extraction failures are excluded from analysis. Replace * with "Instruct" and # with "Turbo" in the model names. 29
4.3 Examples of original and edited explanations 34
4.4 Sensitivity to prompt differences (GPT-3.5) 38
4.5 LLM *vs.* human bias in explanation rating 41
4.6 Examples with the most extreme total NMAE values before and after calibration 43

Chapter 1

Introduction

1.1 Research Motivation and Goals

Large language models (LLMs) have been rapidly improving with the scaling up of their size, with complex skills emerging at specific size thresholds (Wei et al., 2022a). Modern LLMs display a wide range of reasoning abilities, including arithmetic, commonsense, and symbolic reasoning (Qiao et al., 2023). Reasoning, in general, is crucial for solving complex problems, and the pursuit of reasoning skills goes hand-in-hand with increasingly sophisticated testing methods. Commonly used benchmarks become saturated, models remain opaque while increasing in complexity, and there is an increasing imperative of human-interpretability as models are deployed and democratized. The way we analyze performance must therefore evolve in tandem. As a potential solution, an increasingly popular approach is to generate textual justifications for models' reasoning (Figure 1.1)—a practice even defined by Gurrapu et al. (2023) as a new field of Rational AI (RAI). One significant benefit of these explanations

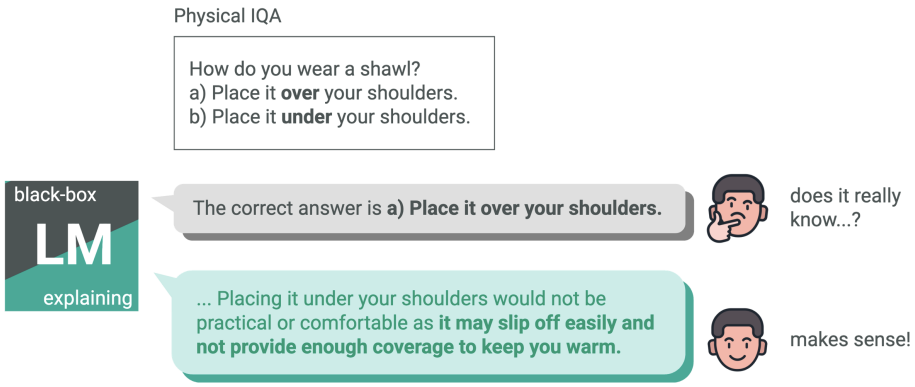


Figure 1.1 A commonsense reasoning question and an example output with the answer and a justification.

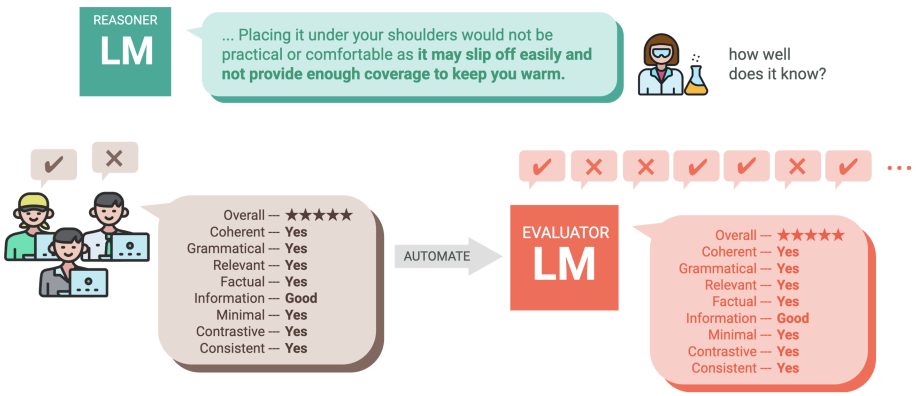


Figure 1.2 An evaluation method for textual explanations using LLMs.

is that they may improve transparency and interpretability since they are easy for human users to understand and can include richer reasoning than other explanation methods (Kunz et al., 2022). On the models' side, generating explanations has the added benefit of improving their performance. For example, Wei et al. (2022b) found that generating a chain of thought, a series of intermediate reasoning steps, significantly improves the ability of large language models to perform complex reasoning. Lampinen et al. (2022) confirmed that explanations could improve performance both with and without tuning. Additionally, prompting for explanations in addition to predictions has been shown to reduce the impact of superficial cues, i.e., models basing their reasoning on shallow statistical patterns instead of "truly" reasoning, in adversarial NLI (Kavumba et al., 2023). For this reason, several commonsense reasoning datasets have been enriched with natural language explanations. They are intended to be used downstream in three ways: as data augmentation to improve performance on a predictive task, as supervision to train models to produce explanations for their predictions, and as ground truth to evaluate model-generated explanations (Wiegreffe and Marasović, 2021). Here, a new issue arises: how does one evaluate the resulting explanations? At the time of writing, textual explanation evaluation is yet to be established with a consensus on criteria and a matching coherent body of work of benchmarks, approaches, and comparisons. Automatic evaluations are borrowed from machine translation (Clinciu et al., 2021) and only measure overlap with hand-written gold explanations. More detailed, explanation-specific evaluations are usually conducted by hand (Wiegreffe et al., 2022, e.g.,), which can be expensive and difficult to reproduce. As a potential solution, this dissertation explores a new approach to evaluating textual explanations: evaluating explanations with an LLM, and validating this aproach using a new diagnostically scored explanation dataset (Figure 1.2). In addition to the quality of LLM-generated explanations, there are concerns about their fidelity—language models notoriously produce fluent and seemingly plausible content, even

without grounding in truth (Bommasani et al., 2021). Lipton (2018) raised similar concerns earlier, stressing that post-hoc textual explanations are trained to maximize the likelihood of previously observed ground-truth explanations from human players and may not faithfully describe an agent's decisions, however plausible they may appear. To address this, as a starting step, one of the diagnostic criteria proposed in this dissertation will measure the consistency between the outputted label and which answer the explanation supports. This is a necessary (but not sufficient) criterion for faithful natural language explanations. In summary, the research questions addressed in this dissertation are (i) How can we evaluate free-text explanations in NLP? (ii) Are LLMs reliable judges for free-text explanations?, and (iii) When do models learn to explain? To answer these questions, this dissertation makes the following contributions:

- Two new datasets useful for evaluating free-text explanations in NLP.
- A new evaluation method based on LLMs.
- Meta-evaluation methods for analyzing the reliability of LLMs-as-judges.
- New insights into when models learn to explain, by comparing explanation quality across different checkpoints, sizes, and versions.

1.2 Dissertation Structure

This dissertation is structured as follows:

- **Chapter 1** introduced the problem of evaluating free-text explanations in NLP, the limitations of existing methods, and the motivation for using LLMs as judges. We defined the research questions and summarized the contributions of the work.
- **Chapter 2** provides an overview of related works. It covers commonsense reasoning in NLP, evaluating commonsense reasoning, free-text explanations, evaluation methods, and LLMs as judges.
- **Chapter 3** describes the creation of new data for evaluating free-text explanations. We collected new explanation data, defined a list of criteria for explanation quality, and collected aspect-wise quality judgments from five raters for 3,500 explanations.
- **Chapter 4** proposes a new evaluation method based on LLMs, exploring best practices for prompting, analyzing correlation with majority-voted judgments, and evaluating reliability and differences between LLM and human judgments.

- **Chapter 5** demonstrates the application of the proposed method, comparing explanation quality across different checkpoints, sizes, and versions.
- **Chapter 6** discusses the implications of the results, limitations of the method, and future work.
- **Chapter 7** provides a brief conclusion re-stating the contributions of this work.

Chapter 2

Background

Commonsense reasoning is a notoriously difficult-to-define task. It requires the ability to reason about the world in a way that is intuitive to humans, often involving implicit knowledge that is seldom explicitly mentioned in texts. From ancient philosophical debates to modern benchmarks, its history is long yet still ongoing. This chapter provides a brief overview of commonsense reasoning in natural language processing (NLP), focusing on the challenges of evaluating models' reasoning abilities, and the potential of LLMs as judges for free-text explanations.

2.1 Commonsense Reasoning

> *Annual income twenty pounds, annual expenditure nineteen nineteen and six, result happiness. Annual income twenty pounds, annual expenditure twenty pounds ought and six, result misery.*
>
> — Charles Dickens, *David Copperfield*

Johnson-Laird (2010) used this famous advice by Mr. Micawber to illustrate the futility of defining reasoning purely through logical formulae. In reality, it encompasses simple deductions to complex social interactions and scientific discovery. Our practical reasoning (Wallace and Kiesewetter, 2024) will decide on what we should do next, and moral reasoning (Richardson, 2018) may tell us if it is the right thing to do. Reasoning can take the form of logical deduction, induction, or abduction, and can involve knowledge about the world, logical rules, physical laws, social norms, moral principles, and more. Some schools of thought consider reasoning as a "simulation of the world fleshed-out with our knowledge" (Johnson-Laird, 2010), i.e., as being based on *mental models* rather than formal logic. Our pursuit of understanding (and advancing) reasoning started millenia ago, such as in the

works of Aristotle (Patzig, 2013), the Analects by Confucius (Waley et al., 2012), or in Nyāya Sūtras of Gautama (Vidyabhusana and Sinha, 1990), and continues to this day in fields such as psychology, cognitive science, and neuroscience. Today, artificial intelligence (AI) emerges as our newest medium, born from the goal of creating machines that can reason like humans. What this entails, however, is also a debated topic: Korteling et al. (2021) discussed several problems with the idea of artificial general intelligence, widely considered to be the ultimate goal of AI research, being "technology containing or entailing (human-like) intelligence." They argued points such as the definition being a tautology, the concept being anthropocentric, and the idea being a moving target. Nevertheless, this fuzzy goal still proved fruitful in practice, as even the simplest tasks seemed impossible at first to inflexible machines with no inherent knowledge of the world. Reasoning methods evolved from rule-based systems to statistical models, and now to deep learning models, which have shown impressive results on many tasks. Today, models can successfully answer questions about the world, generate coherent text, and even play games, all of which require some form of reasoning. In a comprehensive overview of model capabilities in various forms of reasoning, Huang and Chang (2023) (cf. Qiao et al., 2023) acknowledged reasoning as an emergent property in LLMs at specific size thresholds, with arithmetic, commonsense, and symbolic reasoning being among the skills that have been observed. Each such advancement in reasoning capabilities was met with new benchmarks and evaluation methods, as we will discuss in the following sections.

2.2 Evaluating Commonsense Reasoning

Can machines think? This question, posed by Alan Turing in 1950, has driven the development of evaluation methods for AI systems. The Winograd Schema Challenge (Levesque et al., 2012), for example, was proposed as a direct alternative to it. Its questions, trivial to humans, challenge models to determine the referent of a pronoun based on commonsense knowledge. Others include the CommonsenseQA dataset (Talmor et al., 2019), which was the first to build a benchmark at scale using a knowledge base, and the SuperGLUE benchmark (Wang et al., 2019), a challenging collection of eight language understanding tasks. These benchmarks often involve tasks such as question-answering, natural language inference, and text generation, and require models to reason about the world in order to produce the correct answer. Benchmarks have since scaled up, e.g., BIG-bench (BIG-bench authors, 2023) consists of 204 tasks with problems from linguistics, childhood development, math, commonsense reasoning, biology, physics, social bias, software development, and beyond, and increased their difficulty, such as in the Google-proof q&a benchmark (Rein et al., 2024)

with graduate-level scientific questions or the (dramatically named) Humanity's Last Exam (Phan et al., 2025) with domain expert-level questions. However, as models have become more powerful, it has become clear that measuring performance on these benchmarks is not enough to ensure that models are reasoning as intended. For example, models were found to rely on superficial cues in the data to make their predictions (Gururangan et al., 2018), rather than truly understanding the underlying concepts, and neural models are especially opaque making their analysis extremely challenging. The field of explainable AI (XAI) has emerged to address this issue, focusing on making models' reasoning transparent and interpretable to humans. This has led to a growing interest in explainability, particularly in human-friendly formats, as a way to aid evaluating models' reasoning abilities. With it, new evaluation methods have been developed, focusing on the quality of the explanations generated by models.

2.3 Explanations

Independently of commonsense reasoning evaluation, systems that not only generate correct output, but also provide an explanation (Miller, 2019) of why that particular output is correct, are desirable for several reasons, such as increasing trustworthiness (Floridi, 2019), compliance with "right to explanation" laws (e.g., GDPR, European Parliament and Council of the European Union, 2016), increasing interpretability (Jacovi and Goldberg 2020, but cf. Lipton 2018 on the caveats of post-hoc explanations), as well as system improvement and knowledge discovery (Adadi and Berrada, 2018). For example, Rajani et al. (2019) used CoS-E to train language models to automatically generate explanations that can be used during training and inference in a novel framework, improving the then state-of-the-art by 10% on the challenging CommonsenseQA task. Aggarwal et al. (2021) defined a set of characteristics for an explanation, constructed a new dataset (ECQA), and demonstrated retriever and generation systems' effectiveness. Wiegreffe et al. (2022) developed a pipeline that combines GPT-3 with a supervised filter that incorporates binary acceptability judgments from humans in the loop. On the human side, however, Chaleshtori et al. (2024) questioned the *utility* of explanations, i.e., whether they help humans make better decisions, and called for application-grounded development of systems with explanation capabilities. In this dissertation, we keep our focus on general commonsense reasoning, as the techniques contained are more easily adapted from a more general to specific setting. In any case, the question remains: how do we evaluate these explanations?

2.4 Evaluating Free-Text Explanations

Since explanations are typically free-form text, automatic evaluation of explanations suffers the well-known, but as of yet unresolved, weaknesses of automatic evaluation measures (Celikyilmaz et al., 2021), while human evaluation is characterized by low scalability, high costs, subjectivity, and inconsistency (Hartmann and Sonntag, 2022). consistency. In a broader sense, a rich source of evaluation methods for explanations can be found in Explainable AI (XAI) literature. For example, Nauta et al. (2023) surveyed existing works on the topic and produced the CO-12 criteria for explanation quality: correctness, completeness, consistency, continuity, contrastivity, covariate complexity, compactness, composition, confidence, context, coherence, controllability. For explanations in the form of textual justifications, various works often define their own criteria for evaluation. Automatic evaluation often borrows from machine learning and measures overlap with "gold standard" text using (a) word-overlap metrics, e.g., BLEU, METEOR and ROUGE; and (b) embedding-based metrics, e.g., BERTScore and BLEURT (Clinciu et al., 2021). Human-tagged measures are more diverse and explanation-specific. For example, Clinciu et al. (2021) measured Informativeness and Clarity; Wiegreffe et al. (2022), inspired by social sciences, measured Acceptability, Generality, Factuality, Grammar, New Info, Supports Label, and Amount of Information; while Aggarwal et al. (2021) defined the criteria of Refutation, Complete, Comprehensive, Minimal, and Coherent. More specifically to commonsense reasoning, Wiegreffe et al. (2022) found that while models often produce factual, grammatical, and sufficient explanations, they have room to improve along axes such as providing novel information and supporting the label. This analysis, however, was conducted by human annotators on a small scale. In this dissertation, we propose to use LLMs as judges for explanation evaluation among similar axes as a way to diagnostically evaluate the ability of models to reason, and investigate the reliability and validity of this approach.

2.5 LLMs as Judges

Outside of commonsense reasoning and explanation evaluation, LLM-as-a-judge is a growing area of interest. Models are used to evaluate helpfulness, harmlesseness, reliability, relevance, feasibility, and overall quality, among other aspects, and are applied beyond just evaluation but to help with alignment, retrieval, and reasoning (Li et al., 2024). For example, part of the great success of LLMs has been attributed to the practice of learning with human preferences such as was done with GPT-4 (OpenAI, 2024). A natural extension of this is to use LLMs as judges to generate equivalent feedback, coined as reinforcement learning from

AI feedback (RLAIF), which was shown to be equally successful (Zheng et al., 2023). At the same time, there are efforts to keep track of their reliability, with mixed results. Zheng et al. (2023), for example, showed that GPT-4 can match both controlled and crowdsourced human preferences of translations, with a same level of agreement as between humans. In contrast, Bavaresco et al. (2024) concluded LLMs are not yet ready to systematically replace human judges in NLP, as they found them to be highly variant. In a more rigorous analysis, Calderon et al. (2025) proposed the alternative annotator test as a way to justify replacing human annotators with LLMs: in a side-by-side comparison, models must achieve a winning rate of 0.5 or more. The work in this dissertation was conducted contemporaneously with these developments, and contributes to the growing body of work on LLM-as-a-judge.

Chapter 3

Data Collection

> And AC said, "LET THERE BE LIGHT!"
>
> *Isaac Asimov, The Last Question*

At the start of this project, high-quality commonsense reasoning explanation data was scarce, and no datasets with quality labels existed. Wiegreffe and Marasović (2021) listed only two free-text explanation datasets, both for CommonsenseQA: CoS-E (Rajani et al., 2019) and ECQA (Aggarwal et al., 2021). None existed for COPA, another popular commonsense reasoning task. As this dissertation aims to develop an automatic explanation evaluation system, we needed high-quality explanation data and quality labels. We thus start by creating two new datasets: COPA-SSE (§3.1) and ACORN (§3.2). COPA-SSE is a dataset of semi-structured explanations for the Balanced variant of the Choice of Plausible Alternatives (BCOPA) task (Kavumba et al., 2019), and ACORN is a dataset of explanations with aspect-wise quality labels. In this chapter, we describe the data collection process for both datasets, including the design goals, crowdsourcing setup, data statistics, and examples.

3.1 COPA-SSE: Semi-Structured Explanations for Commonsense Reasoning

We start by adding a new source of explanations for a different commonsense reasoning benchmark than existing explanation data, BCOPA. This is to reduce the risk of overfitting to a specific benchmark while still remaining in the realm of commonsense reasoning. Its design is inspired by the limitations of structured and unstructured explanations, aiming for a golden middle. COPA-SSE also contains rudimentary quality labels (1-5 star ratings), but

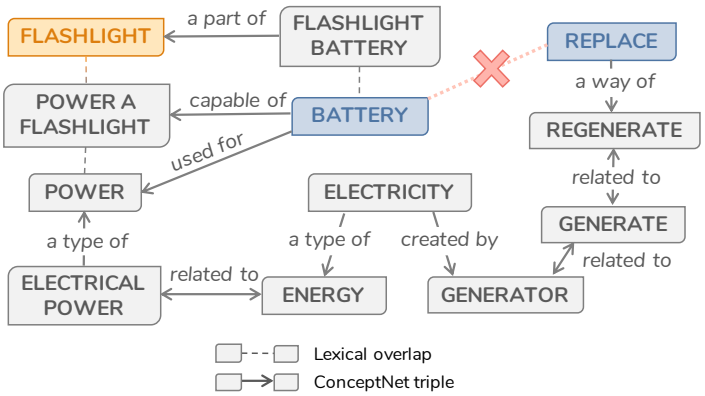


Figure 3.1 A manually extracted ConceptNet subgraph to illustrate the caveats of only using existing resources. The author attempted to find paths connecting concepts from the question *The flashlight was dead. Effect?* and the answer *I replaced the batteries.* and was unable to find a meaningful path between *battery* and *replace.* The two concepts are connected but the path contains irrelevant facts to the point of being meaningless.

they were only used to filter out low-quality explanations during development. COPA-SSE is available at: `a-brassard/COPA-SSE`.

3.1.1 Design goals

Since the nature of a good explanation is subject of debate (**?**), we adopt a working definition: A good explanation is a minimal set of relevant common sense statements that coherently connect the question and the answer. For example, the fact *Opening credits play before a film.* connects the question *The opening credits finished playing. What happened as a result?* and its answer *The film began.* Commonsense knowledge graphs (KGs) such as ConceptNet (Speer et al., 2017) provide such statements but have limited coverage (Hwang et al., 2021). For example, even if question and answer concepts are found in the KG, the paths between them can degenerate into long chains of statements that are neither minimal nor relevant (Figure 3.1). In contrast to structured approaches, unstructured free-form text is not limited by KG coverage. Previous work has elicited such free-form explanations from crowdworkers, but suffers from low quality. For example, in a manual inspection of 1,200 CoS-E samples most explanations were judged to be not relevant to the question and only a small fraction were deemed acceptable explanations. Aiming for a golden middle, we devise a semi-structured explanation scheme comprising a set of triple-like statements. Each statement consists of open-ended head text and tail text connected with a ConceptNet relation. In practice, crowdworkers created explanations by selecting a predicate from a list while providing free text for the two concept slots.This format encouraged workers to provide

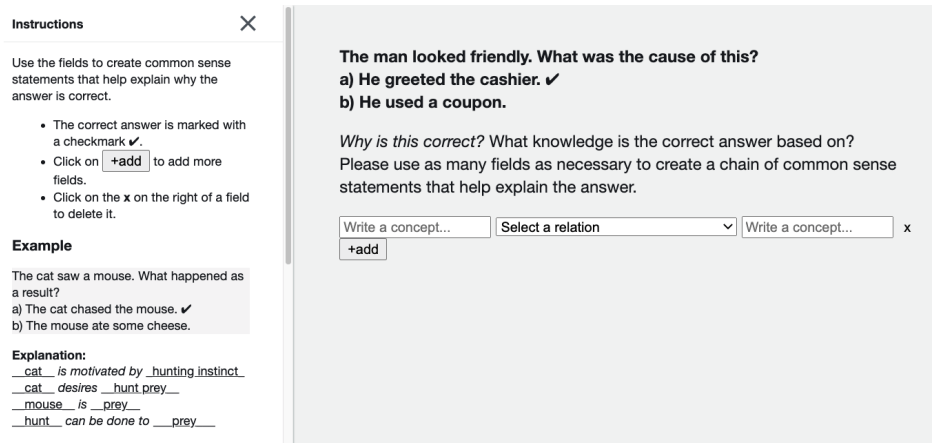


Figure 3.2 Form for collecting semi-structured explanations.

explanations close to our definition without being restricted to a pre-defined inventory of concepts. We refer to this combination of free text and ConceptNet predicates as *semi-structured explanations*.

3.1.2 Crowdsourcing

Crowdworkers were asked to provide one or more statements that connect the question and the answer in a triple format: a free-form head text, a selection of ConceptNet relations, and a free-form tail text, together forming a commonsense statement. Each set of statements was then rated by five different workers. To gather more high-quality explanations, we invited workers whose explanations were highly rated to provide additional explanations.

Collecting Explanations

Figure 3.2 shows our collection form. Workers were given a BCOPA question and two answer choices with the correct one marked. The input row below consists of two text fields for inputting concepts and a drop-down box for selecting the relation between them. Workers could increase the number of rows to provide explanations with multiple statements, as they were encouraged (but not forced) to do. The relations are a subset of ConceptNet predicates which we selected and translated into human-readable English for easier understanding by non-experts.[1] For example, the input *an apple* `is a` *fruit* corresponds to the statement *"An apple is a fruit."* and the triple (*"an apple"*, `IsA`, *"fruit"*). Free-form text guarantees neither consistent granularity nor chains of statements connected by matching concepts.

[1] E.g., `A HasSubevent B` is shown as `A` *happens during* `B`. The text-form explanations retained the original surface form, while in the triple format they are changed back to match ConceptNet.

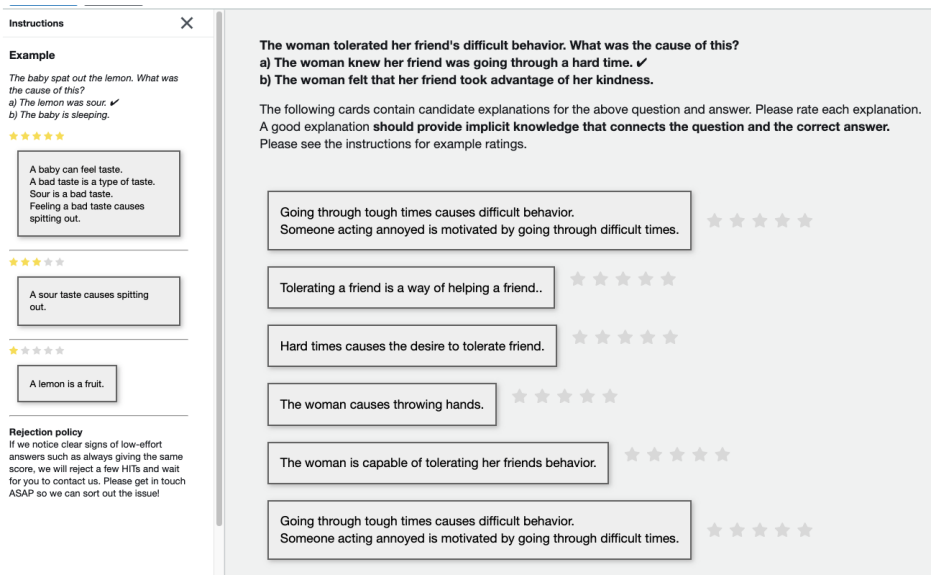


Figure 3.3 Form for rating semi-structured explanations.

For example, a phrase such as *"the act of eating a sweet fruit"* can be given as tail text, even though the next statement might not include that same phrase. We opted to leave this freedom as longer statements can still form coherent explanations, and, as we found in preliminary runs, introducing strict constraints might lead to unnatural and/or less informative explanations. Overly long statements were rare, as most workers followed the simple examples we provided.

Rating Explanations

Figure 3.3 shows our form for rating explanations. Each explanation was rated by five workers. Workers were shown a BCOPA instance and five explanations to rate with up to five stars. As a control, workers had to rate the first explanation again at the end of the HIT, totaling six ratings per HIT. We disregarded (but did not reject) ratings by workers who had more than a one-star difference in this control.[2] Workers were instructed to give a higher rating to explanations containing relevant and more detailed statements and low ratings to uninformative or nonsensical explanations. We observed that detailed, related statements were also low-rated if they did not explain why the answer is correct. Examples of high-rated and low-rated explanations are shown in Table 3.2. While these ratings serve as generic estimate of quality, we recommend against using them as measurements of any single characteristic such as relevance or thoroughness since they were not defined as such.

[2]We allowed a 1-star difference as one could change their opinion on the first seen explanation after seeing other examples. In case of such a difference, we only retain the last rating.

The documents were loose. Effect?	
✓ I paper clipped them together.	✗ I kept them in a secure place.
★★★★½	Paper clip is used for loose documents.
★★★★	Paper clips is used for keeping documents together. Paper clipping can be done to have the documents together.
★★★★	Paper clip is used for clipping paper together.
★★★½	Paper clip is used for organizing papers.
★★★½	Paper clip can be done to keep papers together.
★★★½	The paper clipped is a way of holding the papers together.
They lost the game. Cause?	
✗ Their coach pumped them up.	✓ Their best player was injured.
★★★★½	Game is a team work. Player is a part of a team. Player injured causes team not working properly. Team not working properly causes lose the game.
★★★★	Best player is a part of the team. Injury of the best player causes the team to lose.
★★★★	Their best player being injured causes the team to lose.
★★★★	Teams is made of players. Injuries is capable of causing losses.
★★★½	Injury is capable of causing loss.
★★½	The team causes the injury.

Table 3.1 Examples of collected and rated explanations for BCOPA questions.

Re-collection

To increase the number of higher-rated explanations, we invited workers who provided high-quality explanations to provide additional explanations for a higher fee. We collected four new explanations for questions that had all five explanations rated below 3.5-stars, two new explanations if one was above this threshold, and one new explanation if two were above this threshold. New explanations were then rated in the same way as the original ones.

Compensation and qualifications

Workers received \$0.30 per explanation in the first collection round and \$0.40 in the re-collection round. In the rating rounds, workers received \$0.30 for six ratings (five unique and one control). We restricted all our rounds to workers in GB or the US with a HIT approval rate of 98% or more and 500 or more approved HITs. For re-collection, we invited workers whose explanations averaged more than 3.5 stars over ten or more explanations. The total cost, including Amazon Machanical Turk fees and excluding trial runs, was \$8,651.16.

The woman sensed a pleasant smell. Effect? ✓ She was reminded of her childhood.	
★★★★★	Pleasant smell is a way of bring happiness. Happiness causes nostalgia. Nostalgia is related to a smell. Smell causes her to think her childhood.
The flashlight was dead. Effect? ✓ I replaced the batteries.	
★★★★★	Batteries is used for flashlights. Power is created by batteries. Replacing batteries is a way of restoring power.
The car looked filthy. Effect? ✓ The owner took it to the car wash.	
★★★★★	The owner desires clean car. Car wash is used for washing cars.
My favorite song came on the radio. Effect? ✓ I sang along to it.	
★	This is a symbol of simple.
The rain subsided. Effect? ✓ I went for a walk.	
★	The rain has a fresh smell.
The girl was not lonely anymore. Cause? ✓ She made a new friend.	
★	Making is motivated by loneliness.

Table 3.2 Examples of top-rated and bottom-rated explanations. Highly rated explanations tend to be detailed and explicitly connect the question and answer. Low rated ones are incoherent, completely irrelevant, or related facts but irrelevant as an explanation.

Post-processing: Aggregation

Free-form nodes occasionally contain very similar concepts expressed with different surface forms without being explicitly connected. Multiple explanations may also offer diverse information which, combined, results in a higher-quality explanation graph in terms of coverage. To aggregate the explanations, we scored the similarity between each node and merged similar nodes or connected them with a `RelatedTo` edge. Specifically, we computed the cosine similarity s of the node texts using Sentence-BERT (Reimers and Gurevych, 2019) and merged if $s > 0.85$ or connected if $0.60 > s \geq 0.85$.[3] Each edge also includes a weight calculated as the sum of average human ratings of the explanation the edge came from. Intuitively, these can be considered as the importance or relevance of the edge according to humans, at least in relation to all other given explanations for the sample. Post-processed versions of the graphs are also available in the repository, but were not used in this work.

[3]For example, *"sun"* and *"under the sun"* are connected ($s = 0.76$), *"shadow"* and *"shadows"* are merged ($s = 0.93$).

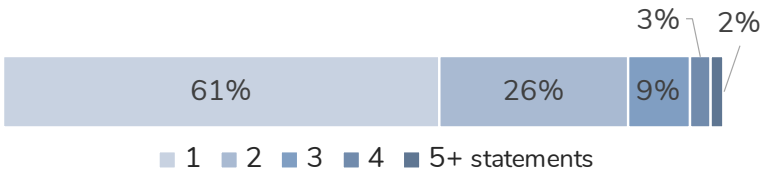


Figure 3.4 Number of statements per explanation in COPA-SSE.

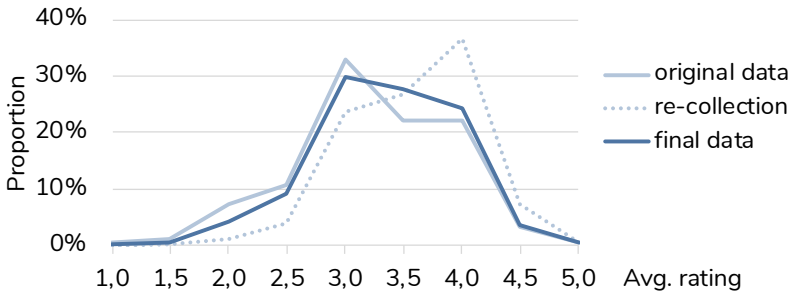


Figure 3.5 Average rating distribution before (original data) and after the re-collection round (final data). Values are rounded to the nearest half-star.

3.1.3 Data stats and examples

Dataset statistics. Table 3.1 shows examples of COPA-SSE explanations. COPA-SSE contains 9,747 commonsense explanations for 1,500 BCOPA questions. Each question has up to nine explanations given by different crowdworkers. We provide the triple-format described earlier, as well as a natural language version obtained by replacing ConceptNet relations with more human-readable descriptions. 61% of explanations are only one statement while the other 39% comprise two or more, with the longest explanation being ten statements (Figure 3.4). Each explanation has a quality rating on a scale of 1 to 5 as given by crowdworkers. Figure 3.5 shows the rating distribution after initial collection (original data). To guarantee that each BCOPA instance is explained by high-quality explanations, we collected additional explanations until most BCOPA instances (98%) had at least one explanation rated 3.5 or higher (final data). In other words, 98% of the questions have at least one highly-rated explanation. Initially, 38% of all explanation were over this threshold, which increased to 44% after the additional collection run. We kept the lower-quality explanations as they can be useful negative samples. Table 3.2 shows examples of the highest and lowest rated explanations in the dataset.

Criterion	Description	Label Choices
Supports	Which answer does it justify?	a), b), ..., none
Overall	How good is the given explanation, overall?	1 to 5 stars
Well-Written	Coherent, grammatically correct, fluent?	Yes, No
Related	Relevant to the Q and A?	Yes, No
Factual	Stated facts are generally true?	Yes, No, N/A
New Information	How much *new* information to support the ans.?	None, Some, Sufficient, Ample
Unnecessary Info.	Any unnecessary statements?	Yes, No
Contrastive	Clearly shows the difference between the ans.?	Yes, No

Table 3.3 Explanation rating criteria in ACORN

3.2 ACORN: Explanations with Aspect-wise Quality Labels

After creating COPA-SSE, we moved on to a more ambitious dataset, ACORN, which contains explanations for both BCOPA and CommonsenseQA. ACORN is unique in that it contains *aspect-wise* quality labels, allowing us to study the quality of explanations in more detail. In the following sections, we will define a list of quality aspects, describe the source datasets (including newly collected explanations, different from COPA-SSE) and the crowdsourcing setup, and provide data statistics and examples. ACORN is available at: `a-brassard/ACORN`.

3.2.1 Quality aspects

We defined a set of criteria to target *surface-level*, *information/content-level*, and *structural* aspects of explanations. We also included criteria to capture *(un)faithfulness* and an *overall* rating, the latter intended to implicitly capture any other aspects considered by the raters. We defined these criteria based on common practices in natural language generation evaluation (Howcroft et al., 2020), known challenges of free-text explanations (Lipton, 2018; Rawte et al., 2023), and insights from social sciences (Miller, 2019). Table 3.3 summarizes the criteria. The criteria largely aligns with the fine-grained analysis conducted by Wiegreffe et al. (2022).

Supports assesses *which* answer the explanation supports, intended to be cross-referenced with the predicted label. A mismatch between the predicted label and the supported answer indicates a lack of faithfulness, i.e., the explanation does not reflect the model's reasoning for the label. Labels choices are `a`, `b`, `c`, `d`, `e`,[4] or `none`.

[4]BCOPA is a binary classification task and CommonsenseQA a five-way classification.

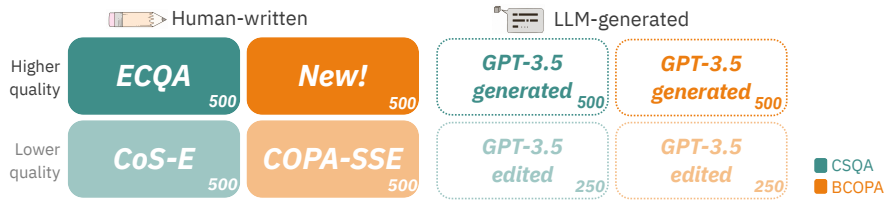


Figure 3.6 Data sources for ACORN

Overall is a holistic assessment of the explanation, capturing any potentially informative or useful aspects that we have not explicitly covered. We encouraged workers to consider this criterion *independently* of the other criteria, and provided general guidelines for each star rating to ensure a consistent understanding. Label choices are one to five stars.

Well-Written is a catch-all criterion to assess the surface-level quality of the explanation, combining criteria such as fluency, coherence, and grammaticality. (`Yes` or `No`)

Related assesses the relevance of the explanation to the question and answers. (`Yes` or `No`)

Factual evaluates the truthfulness of the statements in the explanation, if any, regardless of their relevance. Label choices are `Yes`, `No`, or `N/A` if no information is present.

New Information assesses the extent to which the explanation provides new information beyond the question and answers. Workers were given the choice of `none` for a complete lack of new information, `some` for a partial addition, `sufficient` for a satisfactory amount of new information, and `ample` for highly informative explanations.

Unnecessary Information assesses the extent to which the explanation includes irrelevant information. We included this criterion to capture the challenge of generating *minimal* explanations. (`Yes` or `No`, where `No` is the desired label)

Contrastive assesses whether the explanation contrasts the correct answer with the predicted answer. (`Yes` or `No`)

3.2.2 Source datasets

ACORN contains ratings for a diverse set of existing, newly-collected, and generated explanations. Our choice covers two commonsense reasoning benchmarks and their respective explanation datasets (Figure 3.6). From each, we selected a random subset of 500 explanations for rating, as well as an additional 500 samples of fluency-improved versions, resulting in a total of 3,500 explanations. The fluency-improved subset is included to prevent fluency from becoming a superficial signal, since well-written explanations typically also have high

scores in all other aspects. With five raters and eight criteria per sample, this amounts to 140k ratings in total. Specifically, as the target commonsense reasoning benchmarks, we selected BCOPA (Kavumba et al., 2019)[5] and CommonsenseQA (Talmor et al., 2019) based on the simplicity of their tasks and availability of large-scale explanation datasets (Wiegreffe and Marasović, 2021). Below are the respective datasets we used to source candidate explanations.

CoS-E (Rajani et al., 2019) A widely-used explanation dataset for CommonsenseQA, albeit notoriously uninformative to humans (Nauta et al., 2023). A subset is processed through GPT-3.5[6] for fluency improvement (500 samples + 250 fluency-improved versions)

ECQA (Aggarwal et al., 2021) An improved version of explanations for CommonsenseQA, aligning with our criteria for well-formed explanations. (500 samples)

Generated explanations for CommonsenseQA. Additional high-quality explanations generated by prompting GPT-3.5 to solve a subset of CommonsenseQA, though potential issues like irrelevant information were noted. (500 samples)

COPA-SSE (Brassard et al., 2022) Explanations for BCOPA, as described in Section 3.1, with a subset processed through GPT-3.5 for fluency improvement. Since COPA-SSE already contains overall quality ratings, we selected a random sample of 250 questions and used each question's top-rated and bottom-rated explanation. (500 samples + 250 fluency-improved versions)

Crowdsourced explanations for BCOPA. ECQA's counterpart for BCOPA; a new set of hand-written explanations, carefully crafted for contrastiveness and thoroughness. We collected this new subset of explanations from a hand-picked pool of highly-motivated workers, and instructed them to provide ample information and be explicitly contrastive, i.e., both supporting the correct answer and refuting the incorrect one. (500 samples)

Generated explanations for BCOPA. Similarly to CoS-E, we prompted GPT-3.5 to solve BCOPA questions. (500 samples)

3.2.3 Crowdsourcing

We crowdsourced ratings for the explanations in ACORN using Amazon Mechanical Turk (AMT). Each rater was required to pass a qualification test, after which they were asked to

[5] Balanced COPA; a superset of COPA (Gordon et al., 2012) with added "mirrored" questions that flip the correct label, i.e., the originally incorrect choice becomes correct. The goal is to nullify any annotation artifacts pointing to correct answers (Gururangan et al., 2018).

[6] `text-davinci-003`; The model was instructed to only improve the fluency and was not given any additional context that may encourage improving the content, e.g., by supplementing related information.

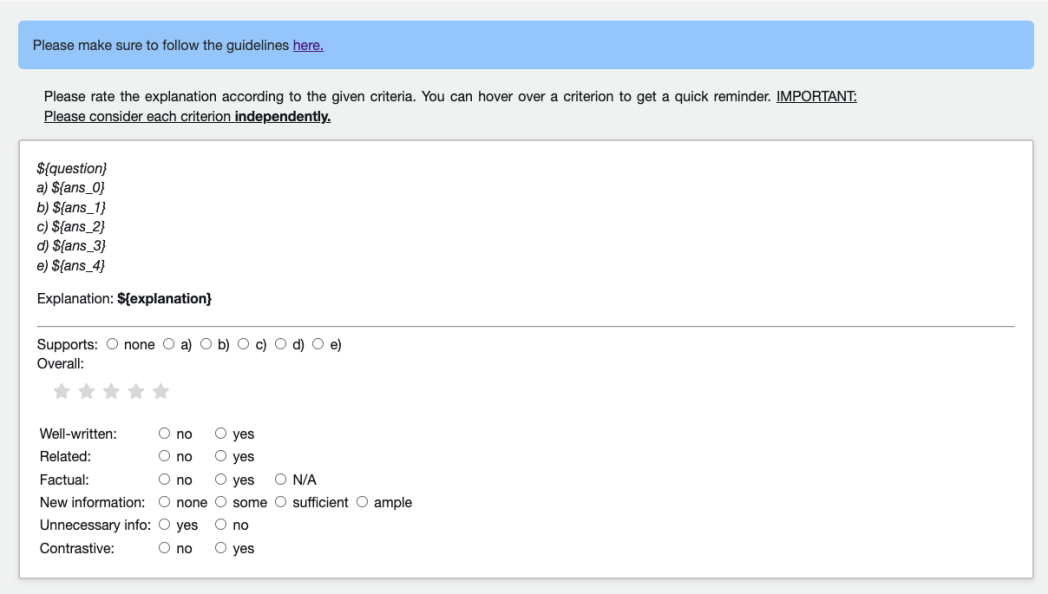


Figure 3.7 Explanation rating form for ACORN.

participate in trial rounds, during which we addressed several clarity issues in the guidelines. The final pool was hand-picked based on their responses, resulting in 28 participants. Our crowdsourcing protocol for label collection consisted of three phases: qualification rounds, trial rounds, and main collection rounds. We provided detailed guidelines showing general instructions, detailed information on each criterion and their respective labels, three examples, and a FAQ section based on questions we received from workers. The full document is available upon request to the first author. A copy of the guidelines is appended at the end of this document for reference.

Qualifications In the qualification rounds, we curated a question set of 6 explanations and manually tagged them with "acceptable" answers, focusing on overall alignment rather than exact matches. We included a dummy question with strict instructions for filtering. Out of 700 participants, the top 201 workers, with a match percentage of 59% or higher, proceeded to trial rounds. We addressed any concerns or clarifications through email or form feedback. We hand-picked a final group of 28 workers. Qualifications were open to workers with a HIT approval rate of 99% or more and 5,000 or more approved HITs. Note that the location requirement was removed as it was an unnecessary barrier for highly skilled and motivated workers.

Dataset	#samples
COPA-SSE (best & worst)	250 + 250
+ fluency fix	250
Generated (BCOPA)	500
Crowdsourced (BCOPA)	500
CoS-E	500
+ fluency fix	250
Generated (CommonsenseQA)	500
ECQA	500
Total	3,500

Table 3.4 Breakdown of explanation data in ACORN by source dataset.

Criterion	-1	0	1	2	3	4	5
Supports	11%	32%	31%	7%	9%	8%	
Overall			16%	13%	28%	28%	12%
Well-written			27%	72%			
Related			6%	93%			
Factual	10%	4%	84%				
New Info			30%	31%	35%	2%	
Unnecessary Info			82%	17%			
Contrastive			58%	41%			

Table 3.5 Label distributions per criterion of majority-voted human ratings in ACORN. -1 denotes "none" for *supports* and "N/A" for *factual*.

Main Rounds Each of the 3,500 explanations in the test set was rated by five workers. The ratings were aggregated to create the final gold labels used in our experiments. Figure 3.7 shows the crowdsourcing form.

Payment Information For qualifications, each worker was compensated $0.15 per HIT. For the main rounds, the fee was increased to $0.25 per HIT, roughly matching a payment of $20.00 per hour.

3.2.4 Data stats and examples

Table 3.4 shows a breakdown of the samples in the test set per source dataset as described in Section 3.2.2. Table 3.5 shows the label distributions of each criterion in our test set. Tables 3.6 and 3.7 show examples of best-rated and worst-rated explanations from each source dataset. Table 3.8 shows the mean ratings and standard deviations of majority-voted ratings per data subset, excluding the categorical aspect *supports*. Data labels are described

CoS-E (best) *Which effect of stress could cause death if not treated immediately? a) age grieving person b)* ***heart attacks*** *c) depression d) hair loss e) headache* *Explanation:* heart attacks can stop your heart from properly pumping blood, thus leading to death.	Over.: Supp.:	4 b)	Well-wr.: Rel.: Fact.: New i.: Unn. i.: Contr.:	yes yes yes ++ no no
CoS-E (worst) *The college kids weren't known for their apartment decor choices, they used a folding chair inside their what? a) bathroom b) closet c) school d)* ***room*** *e) church* *Explanation:* after years of being held captive in an isolated	Over.: Supp.:	1 no	Well-wr.: Rel.: Fact.: New i.: Unn. i.: Contr.:	no no n/a - yes no
CoS-E + fluency fixed (best) *What kind of driving leads to accidents? a) stressful b)* ***dangerous*** *c) fun d) illegal e) deadly* *Explanation:* Driving dangerously creates accidents.	Over.: Supp.:	4 b)	Well-wr.: Rel.: Fact.: New i.: Unn. i.: Contr.:	yes yes yes + no no
CoS-E + fluency fixed (worst) *What is a grumpy person likely to have? a)* ***headache*** *b) understand themselves c) compliments d) own house e) happiness* *Explanation:* A peculiar person.	Over.: Supp.:	1 none	Well-wr.: Rel.: Fact.: New i.: Unn. i.: Contr.:	no no n/a - yes no
Generated (best) *The cloth's are not yet placed where they belong, where should we put them? a)* ***linen closet*** *b) kitchen c) on the space heater d) fabric store e) bedroom* *Explanation:* linen closets are typically used to store items such as cloths.	Over.: Supp.:	5 a)	Well-wr.: Rel.: Fact.: New i.: Unn. i.: Contr.:	yes yes yes ++ no no
Generated (worst) *Sarah opened his chest and found a second heart. Her patient might not be what? a) istanbul b) animal c) antique shop d)* ***human being*** *e) bedroom* *Explanation:* a human being is the only creature known to have two hearts.	Over.: Supp.:	1 d)	Well-wr.: Rel.: Fact.: New i.: Unn. i.: Contr.:	yes yes no - no no
ECQA (best) *Where would one find a captive monkey? a) rain forest b) tree tops c) toy store d)* ***zoological gardens*** *e) great outdoors* *Explanation:* A monkey can be kept captive at a zoological garden as many animals are kept there for people to see. All the other options are incorrect as they do not have captive monkeys.	Over.: Supp.:	5 d)	Well-wr.: Rel.: Fact.: New i.: Unn. i.: Contr.:	yes yes yes ++ no yes
ECQA (worst) *Where are there lots of seats placed in rows surrounding a court? a) show b)* ***auditorium*** *c) movies d) soccer stadium e) hockey game* *Explanation:* In auditorium, there are lots of seats placed in rows. Auditorium seats are surrounding a court. In show, there are no surrounding court placement of seats. Movies is not a place. In soccer stadium, seats are not placed in rows. In hockey game, there are no proper seats.	Over.: Supp.:	2 b)	Well-wr.: Rel.: Fact.: New i.: Unn. i.: Contr.:	no yes no - no yes

Table 3.6 Top- and bottom-rated example explanations for CommonsenseQA in ACORN. New info. labels are denoted as follows: +++ (ample), ++ (sufficient), + (some), - (none).

COPA-SSE (best) *The vehicles travelled slowly. Cause?* *a)* ***It was foggy.*** *b) It was cloudy.* *Explanation:* Fog is a thick mist which prevents you from driving quickly as you cannot see far in front of your vehicle. Driving slow through is motivated by the desire not to crash.	Over.: 5 Supp.: a)	Well-wr.: yes Rel.: yes Fact.: yes New i.: ++ Unn. i.: no Contr.: no
COPA-SSE (worst) *The boy found the television show interesting. Result?* *a) He shut it off.* *b)* ***He recorded it.*** *Explanation:* The result causes the record shut off.	Over.: 1 Supp.: none	Well-wr.: no Rel.: yes Fact.: n/a New i.: - Unn. i.: yes Contr.: no
COPA-SSE + Fluency fixed (best) *The woman tolerated her friend's difficult behavior. Cause?* *a)* ***The woman knew her friend was going through a hard time.*** *b) The woman felt that her friend took advantage of her kindness.* *Explanation:* Going through a hard time can cause people to act in a difficult way. A friend can be motivated to support their friend when they're having a tough time.	Over.: 5 Supp.: a)	Well-wr.: yes Rel.: yes Fact.: yes New i.: ++ Unn. i.: no Contr.: no
COPA-SSE + Fluency fixed (worst) *The shirt did not have a hole anymore. Cause?* *a)* ***I patched the shirt.*** *b) I bleached the shirt.* *Explanation:* I desire to look good.	Over.: 1 Supp.: none	Well-wr.: no Rel.: no Fact.: n/a New i.: - Unn. i.: yes Contr.: no
Generated (best) *The boy was struggling to fall asleep. Cause?* *a)* ***He turned on his nightlight.*** *b) He watched a horror movie.* *Explanation:* Watching a horror movie can create fear and anxiety that can make it difficult for someone to fall asleep. The nightlight is not likely to be the cause since it is meant to help someone sleep.	Over.: 5 Supp.: b)	Well-wr.: yes Rel.: yes Fact.: yes New i.: ++ Unn. i.: no Contr.: yes
Generated (worst) *The woman awakened. Cause?* *a) She took sleeping pills.* *b)* ***Her husband was snoring.*** *Explanation:* Taking sleeping pills is a direct cause of awakening, whereas a husband snoring is an indirect cause.	Over.: 1 Supp.: a)	Well-wr.: yes Rel.: yes Fact.: no New i.: + Unn. i.: no Contr.: yes
Crowdsourced (best) *The man was bitten by mosquitoes. Cause?* *a) He fell asleep on his couch.* *b)* ***He went camping in the woods.*** *Explanation:* Sleeping on a couch would mean you are indoors where there are rarely mosquitoes. Mosquitoes are prevalent in wooded areas, so the man would be more likely to be camping in the woods, if he was bit by mosquitoes.	Over.: 5 Supp.: b)	Well-wr.: yes Rel.: yes Fact.: yes New i.: +++ Unn. i.: no Contr.: yes
Crowdsourced (worst) *I refilled my water bottle. Cause?* *a)* ***I drank all the water in it.*** *b) I kept it in the fridge.* *Explanation:* If you kept your water in the fridge, it would not need to be refilled.	Over.: 2 Supp.: a)	Well-wr.: yes Rel.: yes Fact.: yes New i.: + Unn. i.: no Contr.: yes

Table 3.7 Top- and bottom-rated example explanations for BCOPA in ACORN. New info. labels are denoted as follows: +++ (ample), ++ (sufficient), + (some), - (none).

Data source	Ovr. 1-5	Well-wr. 0, 1	Rel. 0, 1	Fact. -1, 0, 1	New i. 0-3	Unn. i. 1, 0*	Cntr. 0, 1
CoS-E	1.89 $_{0.95}$	0.34 $_{0.47}$	0.77 $_{0.42}$	0.23 $_{0.95}$	0.32 $_{0.51}$	0.41 $_{0.49}$	0.01 $_{0.09}$
CoS-E + fl. fix	2.06 $_{1.01}$	0.72 $_{0.45}$	0.72 $_{0.45}$	0.39 $_{0.90}$	0.42 $_{0.67}$	0.44 $_{0.50}$	0.01 $_{0.09}$
CSQA generated	3.20 $_{0.83}$	0.97 $_{0.16}$	1.00 $_{0.04}$	0.96 $_{0.26}$	1.00 $_{0.74}$	0.04 $_{0.19}$	0.01 $_{0.09}$
ECQA	3.05 $_{0.89}$	0.57 $_{0.49}$	1.00 $_{0.00}$	0.90 $_{0.31}$	1.29 $_{0.71}$	0.11 $_{0.32}$	0.96 $_{0.19}$
COPA-SSE	2.38 $_{1.13}$	0.45 $_{0.50}$	0.91 $_{0.28}$	0.58 $_{0.74}$	0.84 $_{0.80}$	0.33 $_{0.47}$	0.02 $_{0.13}$
COPA-SSE + fl. fix	2.81 $_{1.02}$	0.82 $_{0.39}$	0.97 $_{0.18}$	0.78 $_{0.60}$	0.85 $_{0.78}$	0.18 $_{0.38}$	0.00 $_{0.00}$
BCOPA generated	4.21 $_{0.79}$	1.00 $_{0.04}$	1.00 $_{0.00}$	0.97 $_{0.20}$	1.75 $_{0.63}$	0.00 $_{0.06}$	0.95 $_{0.21}$
BCOPA crowdsourced	4.32 $_{0.76}$	0.97 $_{0.17}$	1.00 $_{0.00}$	1.00 $_{0.00}$	1.94 $_{0.55}$	0.01 $_{0.12}$	0.95 $_{0.21}$
All	3.07 $_{1.27}$	0.72 $_{0.45}$	0.93 $_{0.25}$	0.75 $_{0.63}$	1.11 $_{0.87}$	0.17 $_{0.38}$	0.42 $_{0.49}$

Table 3.8 Mean ratings and $_{\text{standard deviations}}$ per data subset. Higher is better for all criteria except for *unnecessary information*, marked with an asterisk (*), where lower is better.

in Section 3.2.1. All ratings are the higher the better, except for *unnecessary information* which is the lower the better. The human raters seemed to find generated explanations to be most well-written on average, however, the higher quality human-written explanations (ECQA, BCOPA crowdsourced) had a higher amount of new information. The generated explanations, in turn, had the least amount of *unnecessary* information. Interestingly, even though they were not contrastive, the generated explanations for CommonsenseQA (CSQA generated) had a higher average overall rating than ECQA explanations which are explicitly contrastive. A Random Forest Regressor, achieving a mean squared error of 0.37, deemed the most important predictive feature to be *new information* (58%), followed by *factual* (20%), *unnecessary information* (9%), *well-written* (7%), *contrastive* (4%), *supports* (2%), and *related* (0%). Note that for *supports*, we defined a binary feature of whether the label matches the answer label.

Chapter 4

LLM-based Explanation Evaluation

Evaluating explanations, especially across fine-grained aspects of quality, is a manually intensive task and existing automatic measures are not well-suited to the task. Yet, in a world where models must provide explanations, we cannot ignore the need for a scalable and reliable method. In this chapter, we use the newly-constructed dataset containing human aspect-wise judgments of explanation quality to determine whether LLMs can serve for this purpose. Specifically, we consider whether LLMs' ratings correlate highly with human majority votes and, considering the subjectivity of the task, how they impact inter-annotator agreement (§4.2). In other words, we consider the LLM as a judge from three perspectives: (i) as an *individual* rater, (ii) to replace *collective* rating, and (iii) as an *additional* rater. Then, we will consider the overall reliability of LLMs as evaluators by measuring their sensitivity to variations in target explanations and prompt format (§4.3). Finally, we will conclude by exploring whether LLMs can be brought closer to humans by matching their overall strictness/leniency bias with individual raters, and reporting an ultimately unsuccessful attempt to do so with few-shot prompting (§4.4). Since all experiments follow the same inference setup unless otherwise stated, we start the chapter with a section describing these settings (§4.1).

4.1 Settings

4.1.1 Models

We compared four contemporary API-enabled LLMs, namely GPT-4o (OpenAI, 2024), Llama-3.1 (405B) (Llama Team, 2024), Gemma-2 (27B) (Gemma Team, 2024), and Mixtral (8x22B) (Jiang et al., 2024). Each have reported high performance in diverse tasks including text-based reasoning and represent the sate-of-the-art in general-

ist LLMs at the time of writing. Specifically, we used the following model versions: `gpt-4o-2024-05-13`, `Meta-Llama-3.1-405B-Instruct-Turbo`, `gemma-2-27b-it`, and `Mixtral-8x22B-Instruct-v0.1`. Temperature is set to 0.0 with all other parameters left at their default values.

4.1.2 Prompting Strategy

LLMs are notoriously oversensitive to prompt format (Wadhwa et al., 2023). For the purpose of our analyses, we explored several prompting strategies and selected the most successful one as measured by correlation with majority-voted human ratings. Specifically, we compared *single* and *compound* calling, where the former prompts the model for a single criterion at a time, and the latter prompts the model for all criteria at once. We also compared *default*, *averaged*, and *verbatim* prompt formats, corresponding to a simple prompt with the explanation and the rating criteria, a voting mechanism over several prompt variants, and an input identical to the human annotation guidelines, respectively. For single *verbatim* calls, only the relevant sections (guidelines and examples) for the target criterion were included. Default and averaged prompts were further compared in zero-shot and three-shot settings, where the latter contained the same examples as shown in the human guidelines. Most models worked best with *verbatim* prompts corresponding to a word-by-word copy of the guidelines given to humans (see Section 3.2.3), to which they responded with a structured list of criteria and their assigned labels for the given target explanation. Therefore, we opted to use *verbatim* prompts to collect explanation quality ratings for all analyses, with the *default* prompt occasionally used for comparison.

4.1.3 Postprocessing: Label Extraction

Using free-text generation models for a classification task introduces the problem of extracting said ratings, and presents an information extraction challenge in itself. This phenomenon, inherent to generative approaches (Wadhwa et al., 2023), is a source of additional noise that affects all evaluation pipelines necessitating a non-trivial solution in real applications. In our experiments, we used a rule-based extraction method backed up with LLM-based extraction in case of failure. We also manually inspected the remaining failures,[1] and excluded them from our experiments to maintain a fair comparison. The final extraction failure rates were <0.2% for all models but Gemma-2 (2.7%).

[1] Mostly due to non-compliance to the task format, such as responding verbosely with new labels instead of following the given choice: *"... Related: Somewhat. ..."* (instead of Yes or No)

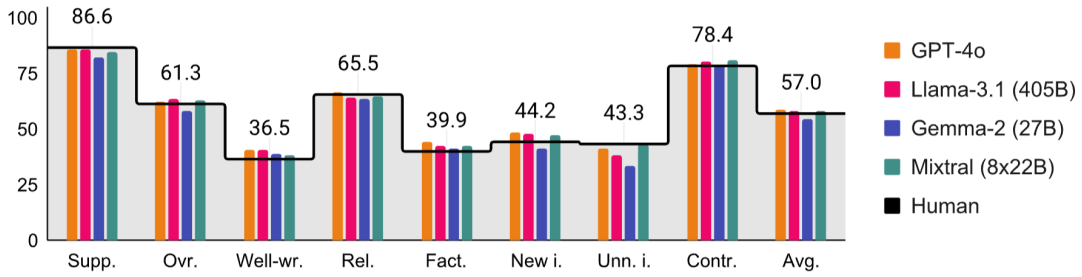


Figure 4.1 Inter-annotator agreement (Krippendorff's α, $*100$ for legibility) between human raters (shaded area) and with the LLM's rating replacing a random rater.

4.2 Alignment with human judges

4.2.1 Inter-annotator Agreement

In subjective tasks, some degree of label variance is expected. leading to lower inter-annotator agreement. This disagreement is not necessarily noise but can be a feature of the data, reflecting the diversity of human opinions (Aroyo and Welty, 2015). Regardless of absolute agreement, we posit that a successful LLM-based rater should be *harmonious* with the range of human labels rather than deviate from it. To measure this, we compared the inter-annotator agreement (Krippendorff's α) between human raters and when a random rater is replaced by an LLM. There are three possible outcomes: (i) agreement *decreases*, indicating that the LLM deviates from human judgments; (ii) agreement *remains the same*, indicating that it is harmonious with human judgments; or (iii) agreement *increases*, meaning that the LLM is both harmonious and biased towards a majority.[2]

Results. In Figure 4.1, the shaded area shows the agreement between human raters, while the bars show the agreement when the respective LLM's ratings replace a random human rater ($*100$ for legibility). Note the overall lower agreement between the human raters; despite careful selection and instruction, there still seems to be a high level of variance between the workers.[3] All values are averaged over twenty iterations. Mixtral, GPT-4o, and Llama-3.1 maintained or improved agreement in most cases, with slight decreases in *supports* with Mixtral, *related* with Llama-3.1, and with *unnecessary information* with GPT-4o and Llama-3.1. Gemma-2, on the other hand, decreased agreement in all but three criteria. However, the latter is also significantly smaller than the others, which illustrates the

[2]The latter may be desirable in use cases that rely on majority-voted labels as the ground truth, but comes with the trade-off of losing potentially useful label diversity.

[3]One of the reasons for this is that some workers lean stricter and some more lenient, with more than one point difference between their respective average *overall rating* ratings (2.66 to 3.72). This is further discussed in Section 4.4.1.

Model version	Supp.	Ovr.	W-wr.	Rel.	Fact.	New i.	Un. i.	Cntr.	Avg.
`gpt-3.5-turbo-0613`	0.861	0.561	0.327	0.644	0.413	0.458	0.386	0.756	0.551
`gpt-4-0613`	**0.873**	0.624	0.346	**0.673**	0.427	0.476	0.397	**0.812**	**0.578**
`gpt-4o-2024-05-13`	0.859	0.624	0.409	0.666	**0.441**	**0.485**	0.416	0.794	0.587
`gpt-4o-mini-2024-07-18`	0.852	0.619	0.411	0.648	0.423	0.474	0.393	0.791	0.576
`text-bison-001`	0.846	0.588	0.344	0.629	0.407	0.468	0.411	0.649	0.543
`gemini-1.0-pro`	0.834	0.597	0.359	0.619	0.404	0.474	0.408	0.714	0.551
`gemma-2-9b-it`	0.836	0.553	0.360	0.634	0.403	0.407	0.368	0.748	0.539
`gemma-2-27b-it`	0.823	0.582	0.387	0.633	0.414	0.411	0.336	0.788	0.547
`Meta-Llama-3.1-8B-*-#`	0.802	0.593	0.396	0.630	0.366	0.439	0.415	0.717	0.545
`Meta-Llama-3.1-70B-*-#`	0.860	0.623	**0.412**	0.647	0.416	0.480	0.378	0.792	0.576
`Meta-Llama-3.1-405B-*-#`	0.860	**0.637**	0.405	0.644	0.424	0.482	0.385	0.804	0.580
`Mixtral-8x7B-*-v0.1`	0.834	0.586	0.356	0.622	0.397	0.455	0.358	0.703	0.539
`Mixtral-8x22B-*-v0.1`	0.848	0.629	0.386	0.645	0.428	0.472	**0.428**	0.808	0.580
`Human`	0.866	0.613	0.365	0.655	0.399	0.442	0.433	0.784	0.570

Table 4.1 Full results of the experiments comparing the difference in inter-annotator agreement between humans and with a random rater replaced by an LLM (§4.2.1). All values represent Krippendorff's α averaged over 20 iterations. Extraction failures are excluded from analysis. Replace $*$ with "`Instruct`" and # with "`Turbo`" in the model names.

trade-off between model size and performance. All numerical results, including for several smaller and older models, are provided in Table 4.1. Overall, the larger models maintained or improved inter-annotator agreement in most criteria, suggesting that they do not deviate from an expected range of human ratings. The *unnecessary information* was the most challenging aspect, with most models decreasing agreement. Next, we ask whether they can then *replace* human evaluation. Specifically, we measured the degree to which the models' predictions align with majority-voted human labels.

4.2.2 LLMs As A Replacement for Human Evaluation

To fully replace human annotation, we expect the model's judgments to lead to the same outcome as with human annotation. In this case, we consider the common practice of majority voting as the ground truth,[4] and measure Spearman's ranking correlation between the majority-voted human labels and the model's predictions. Here, the higher the correlation, the more similar the outcome, with a perfect correlation of 1.0 indicating that the model's predictions are identical to the majority-voted human labels.

[4]Whether this is best practice is left to future work, as discussed in Section 6. See Aroyo and Welty (2015) for a discussion on common misconceptions of human annotation, including the assumption of having a single ground truth.

Model version	Supp.	Ovr.	W-wr.	Rel.	Fact.	New i.	Un. i.	Cntr.	Avg.
`gpt-3.5-turbo-0613`	0.846	0.627	0.409	0.696	0.621	0.611	0.383	0.786	0.622
`gpt-4-0613`	**0.900**	0.740	0.515	**0.778**	0.675	0.691	0.528	**0.946**	**0.722**
`gpt-4o-2024-05-13`	0.844	**0.758**	0.611	0.750	**0.688**	**0.714**	0.457	0.896	0.715
`gpt-4o-mini-2024-07-18`	0.853	0.708	0.620	0.700	0.628	0.678	0.463	0.888	0.692
`text-bison-001`	0.773	0.685	0.456	0.654	0.653	0.640	0.515	0.607	0.623
`gemini-1.0-pro`	0.787	0.681	0.497	0.613	0.549	0.671	0.451	0.648	0.612
`gemma-2-9b-it`	0.798	0.645	0.492	0.674	0.584	0.616	0.388	0.742	0.617
`gemma-2-27b-it`	0.769	0.626	0.574	0.668	0.603	0.537	0.191	0.867	0.604
`Meta-Llama-3.1-8B-*-#`	0.632	0.638	0.562	0.639	0.535	0.555	0.469	0.662	0.586
`Meta-Llama-3.1-70B-*-#`	0.847	0.732	**0.632**	0.703	0.652	0.680	0.311	0.886	0.680
`Meta-Llama-3.1-405B-*-#`	0.833	0.745	0.618	0.699	0.631	0.686	0.351	0.916	0.685
`Mixtral-8x7B-*-v0.1`	0.747	0.624	0.487	0.616	0.543	0.604	0.314	0.640	0.572
`Mixtral-8x22B-*-v0.1`	0.799	0.733	0.588	0.688	0.642	0.681	**0.540**	0.928	0.700

Table 4.2 Full results of the experiments measuring Spearman's rank correlation between majority-voted human labels and LLM-generated ones (§4.2.2). Extraction failures are excluded from analysis. Replace `*` with `"Instruct"` and `#` with `"Turbo"` in the model names.

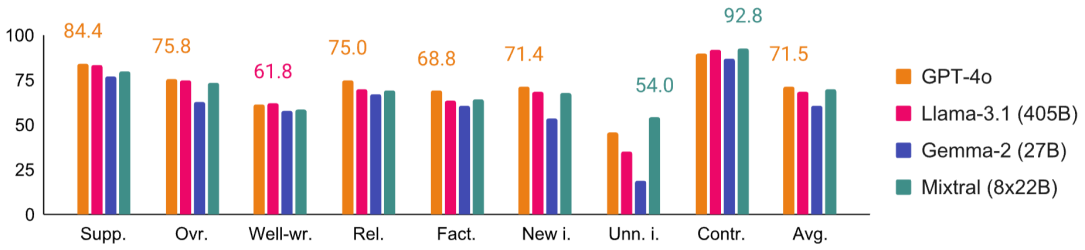


Figure 4.2 Spearman's ranking correlation between majority-voted human labels and LLM-generated ratings (∗100 for legibility).

Results. Figure 4.2 shows Spearman's rank correlation (∗100 for legibility) between aggregated human labels and LLM predictions. The best-performing model's values are annotated for each criterion. Table 4.2 shows the full results, including several smaller and older models. The correlation in *supports* and *contrastive* was particularly strong: 84.4 and 92.8, respectively. The *unnecessary information* criterion, however, stands out with a much lower correlation in all models (54.0 and less). Others ranged from 61.8 to 75.8, indicating a moderately high correlation. GPT-4o was the best-performing model in five out of seven criteria, with an average correlation of 71.5. Mixtral, the second-best model, followed closely, particularly outperforming GPT-4o in *unnecessary information* and *contrastive*. Overall, the larger models are relatively well-aligned with humans and could be considered effective depending on the use case. However, the *unnecessary information* criterion seems to be an outlier, calling for caution when interpreting results. Finally, considering the small change in

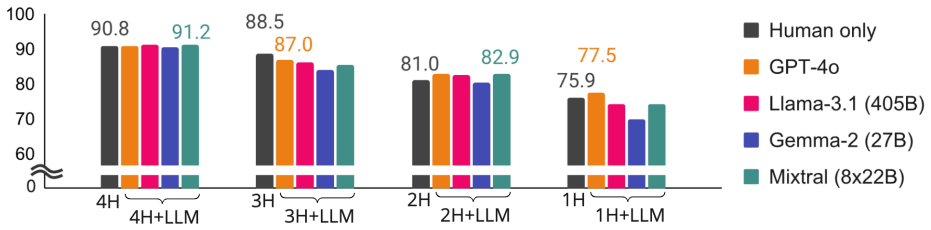


Figure 4.3 A comparison of Spearman's rank correlation with the original gold labels when using fewer raters (`*H`) and when an LLM is added as an additional rater (`*H+LLM`). From left to right, the number of human raters decreases from four to one (randomly selected). All values are multiplied by 100 for legibility.

inter-annotator agreement (§4.2.1), models could potentially be used as an *additional* rater instead of completely replacing human annotation, which we explore in the next section.

4.2.3 LLMs As An Additional Rater

The results so far hint towards LLMs acting similarly to an average human rater, potentially useful as an additional data point when human raters are difficult to access. Here, we verify this potential by measuring whether adding a model's rating improved the outcome—whether the majority-voted labels became more highly correlated to the original (five-way) voted majorities when the LLM was added as a rater. Specifically, we compared Spearman's rank correlation between the majority-voted labels with all available raters and in two alternative scenarios: one where the model is added as an additional rater and one where it is not. If the correlation *increases* when the model is added, it suggests that its predictions are in line with the original majority-voted labels, and it is useful as an additional rater. Otherwise, it would indicate a harmful or negligible effect, and thus its inclusion should be avoided. We repeated this comparison for scenarios with four, three, two, and one randomly selected human rater per sample.

Results. Figure 4.3 shows Spearman's rank correlation with the original gold ratings obtained by aggregating the labels of all five raters. `Humans only` (`*H`, e.g., `4H` for a four-way vote) denotes the correlation between human majority-voted labels only, and others when including the respective LLM as an additional rater (`*H+LLM`). Each column cluster represents a different number of human raters from four in the leftmost to one in the rightmost. With four humans, the correlation between their majority-voted labels and the original gold labels was 0.91. Adding a model as a fifth vote raised this by only 0.004 points. With three humans, we observed a slight *decrease* in correlation when adding a model. With two or one human rater, the correlation increased by 0.019 and 0.016 points, respectively. Overall,

the results suggest that LLMs can be useful as additional raters when the number of human raters is less than three. However, even in the best case, the voted labels with an added LLM rater still scored lower than with an equivalent number of human raters (0.83 with `2H+LLM` vs. 0.89 with `3H`). When there is a high number of human raters, in this case three or more, the model's inclusion as an additional rater does not improve the majority-voted labels' alignment with the original gold labels, and may even harm it.

4.2.4 Discussion

To summarize, while correlation between LLMs and human judges was high but not perfect, they did not negatively impact inter-annotator agreement when mixed with human judges. This suggests that LLMs act similarly to an average human rater, with some potential benefit when complementing a smaller number of human raters. Overall, we consider this a promising result for using LLMs to evaluate explanations, and we encourage referring to our results (or conducting similar analyses) to determine the reliability of the candidate LLM as a rater. Even without perfect alignment, explicitly quantifying its biases enables adjusting analyses and contextualizing the results.

4.3 Sensitivity to Explanation and Prompt Differences

Taking a step back from direct comparison with human ratings, we argue that any method for evaluating explanation quality should exhibit appropriate *equivariance* and *invariance* properties. If two explanations vary in a particular aspect that is relevant for expanation quality, then the evaluation method should be *equivariant* with respect to that aspect. For example, if we edit an explanation to be less informative, then the rating of this aspect should be reduced. Conversely, if two explanations vary in an aspect that is irrelevant for explanation quality, then the evaluation method should be *invariant* to this aspect. In LLMs, we also include variations in prompt format without changes in the target explanation as a form of meaning-preserving alteration. We operationalize our invariance and equivariance requirements by varying prompts and manipulating specific aspects of explanation quality and then test if the LLM output conforms to the expected outcome. Concretely, in Section 4.3.2, we vary the order and wording of the prompt instructions and expect that the LLM's rating of explanation quality should not be sensitive to these variations, i.e., be invariant. To test for equivariance of LLM ratings, in Section 4.3.1, we manipulate six specific aspects of explanation quality in a targeted manner and then record if the LLM's rating of the

manipulated aspect and the unmanipulated aspects changes, which we then compare to human counterparts.

4.3.1 Explanation sensitivity

Consider evaluator LLMs as a measuring instrument; to gauge their usefulness, one must know their level of sensitivity, i.e., *equivariance* to important differences. In terms of explanation evaluation, we translate this to sensitivity to differences in specific aspects: are LLMs able to discern, for example, that a negation can flip the factuality of a statement? In turn, will this affect unrelated ratings (*invariance*)? To test this, we constructed a small challenge set comprising pairs of original and edited explanations, and then measured the change between ratings for each. Ideally, LLMs should match the way the labels changed for humans. We applied automatic edit functions, described below, on 100 candidate samples each, then kept only those where the target criterion successfully changed according to human ratings, totaling 540 test pairs. Table 4.3 shows examples for each edit function used in Section 4.3.1.

Well-written We retrieved samples labeled as *not* well-written, but related and having some new information to ensure the candidates are not entirely incoherent. Then, we prompted GPT-3.5 to create fluency-improved counterparts without adding new information. (98 samples)

Related We fetched explanations with the same ratings but for different questions, creating related and unrelated pairs of otherwise similar quality. Sometimes, these were accidentally related; these cases were excluded from the test set. (81 samples)

Factual COPA-SSE contains triple-form explanations, where each sentence is expressed with a head concept, a relation, and a tail concept. To create non-factual versions, we replaced the relations with negated counterparts (e.g., 'is a'→'is not a'), reconstructed the sentences, and passed them through GPT-3.5 to ensure fluency. (92 samples)

New info We collected a set of explanations containing more than three sentences, then dropped all but the first one to create versions with a reduced amount of information. To further ensure a low amount of information, we only kept samples where the first sentence had less than 150 characters. (94 samples)

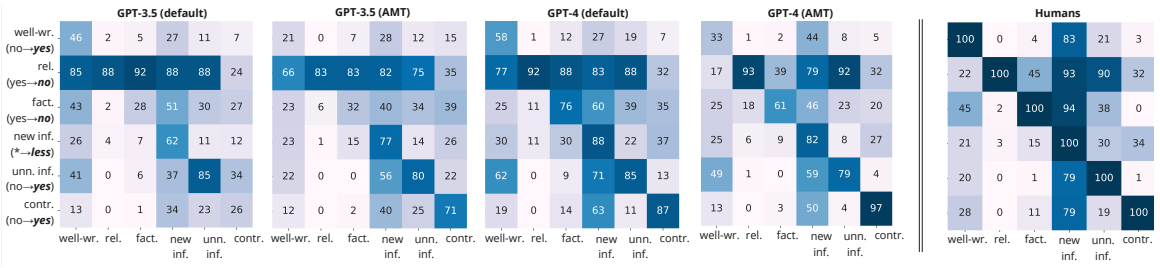


Figure 4.4 Explanation sensitivity expressed through change percentage between original predictions and with edited explanations targeting each criterion. From left: GPT-3.5 with the default prompt, GPT-3.5 with the same prompt as humans, GPT-4 with the default prompt, GPT-4 with the same prompt as humans, and humans.

Unnecessary info We selected samples without unnecessary information but an overall high amount of information. Then, we added extra sentences from a knowledge graph generated by Kogito (Ismayilzada and Bosselut, 2023) and used GPT-3.5 to maintain fluency. (81 samples)

Contrastive ECQA provides lists of positive and negative (contrastive) statements for each sample. We prepared non-contrastive versions of the explanations by concatenating only the positive statements. (94 samples)

Results. Figure 4.4 shows the percentage of changed labels between the original explanation's ratings and its transformed versions in the same settings. Note that here we compare GPT-3.5 and GPT-4, as these experiments were conducted prior to the release of the newer models in the previous section. `AMT` refers to the *verbatim* prompts as used in the previous experiments, and `Default` refers to a simple rubric with the same three examples as in the *verbatim* prompt (§4.1.2). The rightmost heat map shows the change in human ratings for comparison. In an ideal scenario, the corresponding transformations and criteria should have the maximum difference (equivariance), while the others should remain unchanged (invariance). In the heat map, this would result in a distinct diagonal with low values otherwise. However, in reality, some criteria are influenced by others; for example, editing the explanations consistently impacted the amount of new information for humans.[5] Replacing the explanations with completely unrelated ones (second row) seemed to influence all ratings but *contrastive* in GPT-3.5. I.e., the model seems to be oversensitive to whether the explanation is related, unable to distinctly rate other criteria. In turn, the model did not pick up on many of the targeted differences (diagonal). The situation seems to improve with GPT-4

[5]We suspect this to be due to the rating being interpreted as the amount of *useful* information, which could be reasonably inferred from the guidelines.

Well-written	
[Original] Crater happens during collisions with comets. Craters on moon is located at place where comet collided. Impact is created by comet. Impact is cause of crater.	Well-written: No; Related: Yes; Factual: Yes; New info: 1.8; Unn. info: Yes; Contrastive: No
[Edited] Craters occur when comets collide, creating impacts that are the direct cause of the craters on the moon, where they are located.	Well-written: Yes; Related: Yes; Factual: Yes; New info: 1.2; Unn. info: No; Contrastive: No
Related	
[Original] Not paying a bill can create a debt, and having a debt can lead to more bills.	Well-written: Yes; Related: Yes; Factual: Yes; New info: 1.4; Unn. info: No; Contrastive: No
[Edited] Cleaning is motivated by friend staying overnight.	Well-written: No; Related: No; Factual: N/A; New info: 0.4; Unn. info: Yes; Contrastive: No
Factual	
[Original] Blowing air is capable of extinguishing candle flame. Extinguished candle flame causes smoke. Smoke is capable of entering eye.	Well-written: Yes; Related: Yes; Factual: Yes; New info: 2.2; Unn. info: No; Contrastive: No
[Edited] Blowing air *does not have the ability to* extinguish a candle flame. The flame of a candle *is extinguished by* smoke. However, it is important to note that smoke *is unable to* enter the eye.	Well-written: Yes; Related: Yes; Factual: No; New info: 1.0; Unn. info: Yes; Contrastive: No
New info	
[Original] The mailbox has a defect. *Wind is capable of blowing paper. An open mail box is susceptible to wind.*	Well-written: Yes; Related: Yes; Factual: Yes; New info: 1.8; Unn. info: No; Contrastive: No
[Edited] The mailbox has a defect.	Well-written: Yes; Related: Yes; Factual: N/A; New info: 0.6; Unn. info: No; Contrastive: No
Unnecessary info	
[Original] Feeling cold usually requires some kind of physical action to warm up, such as sipping a hot drink like coffee. Chuckling would not be a logical response in this case.	Well-written: Yes; Related: Yes; Factual: Yes; New info: 2.6; Unn. info: No; Contrastive: Yes
[Edited] Feeling cold usually requires [...] would not be a logical response in this case. *Sometimes, sipping a drink may lead to sipping coffee. Beforehand, it is necessary to drink coffee. Coffee consists of coffee beans, and the act of sipping comes into play here. This action was [...]*	Well-written: No; Related: Yes; Factual: Yes; New info: 1.6; Unn. info: Yes; Contrastive: Yes
Contrastive	
[Original] In a new york city, there is the Met. As Sean wanted to go to the Met, so he took a subway stop in new york city. *There is no Met in large city or toronto or metropolis.*	Well-written: No; Related: Yes; Factual: Yes; New info: 1.0; Unn. info: No; Contrastive: Yes
[Edited] In a new york city, there is the Met. As Sean wanted to go to the Met, so he took a subway stop in new york city.	Well-written: Yes; Related: Yes; Factual: Yes; New info: 1.4; Unn. info: No; Contrastive: No

Table 4.3 Examples of original and edited explanations using edit functions targeting each fine-grained aspect (§4.3.1). Targeted labels are underlined.

and *verbatim* prompts, but the model is still under-sensitive to *well-written* (33% changed) and *factual* edits (61% changed). The *well-written* ratings seemed to have also been overly influenced by the *unnecessary information* edits (49% changed).

4.3.2 Prompt sensitivity

Sensitivity to prompt format is a well-known phenomenon, albeit often in a positive light guiding research aiming to optimize performance by prompt-tuning (Liu et al., 2021). However, here, we heed Webson and Pavlick (2022)'s warning of models' limited understanding of tasks, as evidenced by "good" predictions despite irrelevant prompts. Thus, we adopt an ideal of robust predictions, *invariant to meaningless differences* and, (along with equivariance to important differences §4.3.1) indicative of deeper understanding of the task. In the following experiment, we investigated the impact of prompt variations on the ratings: the same explanation should be rated in the same way, no matter the wording of the form. We designed six prompt variations, as described below, that all follow the same rating schema. We then calculated the rate of label changes compared to the *Default* prompt (§4.1.2).

Prompt Variations

Parts identical to the default prompt are omitted and marked with '. . .' for brevity. Indented linebreaks are added for legibility and were not present in the actual prompt. In few-shot settings, the initial guidelines are only shown once, followed by *N* examples with a simple corresponding answer format, unless otherwise specified. E.g., for the default prompt:

```
QUESTION: <question>
a) <answer choice>
b) <answer choice>

EXPLANATION: <explanation>

1. a
2. 2
3. no
4. yes
5. yes
6. sufficient
7. no
8. yes
```

Default Prompt The prompt presents the question, answer choices, and the explanation, as described in Section 4.1.2.

```
Evaluate the given explanation according to the following criteria:

1. Which answer does it support? (a, b, c, d, e, none)
2. Overall rating? (1, 2, 3, 4, 5)
3. Is it well-written? (no, yes)
4. Is the explanation related to the question and answers? (no, yes)
5. Are all contained facts correct? (N/A, no, yes)
6. How much new information is given? (none, some, sufficient, ample)
7. Any unnecessary information? (no, yes)
8. Is it contrastive? (no, yes)

QUESTION: <question>
a) <answer choice>
b) <answer choice>

EXPLANATION: <explanation>
```

Nonsensical Criteria Adds three non-sensical criteria as distractors: the presence of "quibberfluff," whether it's "fizzlewopped," and how much "drizzlewhisk" there is. Labels for these criteria are ignored when extracting the ratings. In n-shot settings, the examples are assigned random labels for those criteria. The model's answers do not count towards performance.

```
Evaluate the given ...

...
9. Does it have quibberfluff? (yes, no)
10. Is it fizzlewopped? (yes, no)
11. How much drizzlewhisk is there? (none, some, good)

QUESTION: ...
```

Shuffled Criteria Maintains the same criteria as the default prompt but presents them in a different order.

```
Evaluate the given ...

1. How much new information is given? (none, some, sufficient, ample)
2. Is it contrastive? (no, yes)
3. Is the explanation related to the question and answers? (no, yes)
4. Is it well-written? (no, yes)
5. Overall rating? (1, 2, 3, 4, 5)
6. Which answer does it support? (a, b, c, d, e, none)
7. Are all contained facts correct? (N/A, no, yes)
8. Any unnecessary information? (no, yes)

QUESTION: ...
```

Slightly Paraphrased The overall format is not changed, but some sections are slightly paraphrased.

```
Evaluate the given explanation based on these
   criteria:

1. Which answer does it support? (a, b, c, d, e, none)
2. Rate the overall explanation (1, 2, 3, 4, 5)
3. Is it well-written? (no, yes)
4. Is the explanation relevant to the question and answers?  (no, yes)
5. Are all provided facts accurate? (N/A, no, yes)
6. How much new information is given? (none, some, sufficient, ample)
7. Does it contain unnecessary information? (no, yes)
8. Is it contrastive? (no, yes)

QUESTION: ...
```

Verbose Provides more elaborated instructions for each criterion but does not introduce any additional information compared to the default prompt.

```
Evaluate the explanation provided based on the following criteria:

1. Which answer choice does the explanation support? (a, b, c, d, e, none)
2. Rate the overall quality of the explanation. (1, 2, 3, 4, 5)
3. Is the explanation well-written? (no, yes)
4. Is the explanation related to the question and answer choices? (no, yes)
```

```
5. Verify the accuracy of all facts, if any, mentioned in the explanation.
   (N/A, no, yes)
6. How much new information is given in the explanation? (none, some,
   sufficient, ample)
7. Does the explanation contain any unnecessary information? (no, yes)
8. Is the explanation contrastive? (no, yes)

QUESTION: ...
```

Repeated Instructions Identical to the default prompt, except for the example formatting—instead of simple successions of questions, explanations, and ratings, the full instructions are repeated every time.

Swapped Instructions This prompt maintains the same content as the default prompt but presents the question, answer choices, and ratings in a different order. Similarly to the repeated instructions prompt, the full instructions are repeated in every example.

```
QUESTION: ...

EXPLANATION: ...

Evaluate the given explanation according to the following criteria:

1. Which answer does ...
```

Results

Prompt	Supp.	Ovr.	W-wr.	Rel.	Fact.	New i.	Unn. i.	Cntr.	Avg.
Nonsense	2.0	16.2	5.7	1.3	4.1	10.9	7.5	11.6	7.4
Shuffled	8.0	40.6	4.9	2.0	6.1	19.7	6.7	21.4	13.7
Paraphrased	1.9	22.4	4.4	2.5	4.7	15.6	7.4	8.0	8.4
Verbose	3.2	26.4	5.8	1.8	13.1	12.8	7.3	11.5	10.3
Repeated	3.8	35.4	7.8	5.3	5.9	28.6	8.6	8.8	13.0
Swapped	6.2	45.8	18.6	4.0	11.1	31.2	16.0	13.5	18.3
Average	4.2	31.1	7.9	2.8	7.5	19.8	8.9	12.5	11.8

Table 4.4 Per-criterion and average label change rates compared to the *Default* prompt for GPT-3.5. All values are percentages.

Table 4.4 shows the per-criterion and average percentages for each prompt variant compared to the *Default* prompt for GPT-3.5.[6] That is, each value represents the percentage of labels that changed compared to when it was prompted with *Default* prompts.Ideally, the model should be invariant to prompt changes, meaning 0% change in all cases. In reality, changes ranged from 7.4% to 18.3%, averaging 11.8%. Prompts that changed the order of instructions, i.e., *Shuffled*, *Repeated*, and *Swapped*, had the highest rates of change. In turn, the *overall rating* and *new information* criteria had higher rates of changes (31.1% and 19.8%, respectively) compared to other criteria.

4.3.3 Discussion

We defined two key properties that an explanation evaluation method should exhibit: equivariance and invariance. We tested these properties by manipulating specific aspects of explanation quality and prompt format and observing the LLM's response. The best-performing model, GPT-4, still was less sensitive to explanation differences than humans, and GPT-3.5 was overly sensitive to prompt differences. While not entirely disqualifying LLMs as evaluators, these results suggest that humans may still be better judges for edge cases where a higher sensitivity is required. In turn, their sensitivity to prompt differences confirm both a strength and weakness of LLMs—this property can a boon if leveraged to calibrate their ratings or a bane if not accounted for in the evaluation setup. The former, however, may not be straightforward: averaging over multiple prompts did not yield better results (§4.1.2), and we will see in the next section an unsuccessful attempt to calibrate LLMs by prompting with examples sourced from individual raters.

4.4 Human vs. LLMs: How Do They Disagree?

In Section 4.2, we found that LLMs performed similarly but not identically to humans. While the difference may be tolerable in certain applications, understanding the reason for the difference will help guide improving the LLM-based evaluation setup should the setting require higher alignment with humans. In this section, we focus on the differences in *overall bias* in humans and LLM ratings, and report the results of our calibration attempts to align the two. While ultimately unsuccessful, the results add to our understanding of the characteristics of LLMs as explanation evaluators.

[6]This was the newest version at the time when these experiments were run; the exact values may be outdated, but the evaluation methodology remains applicable to newer models.

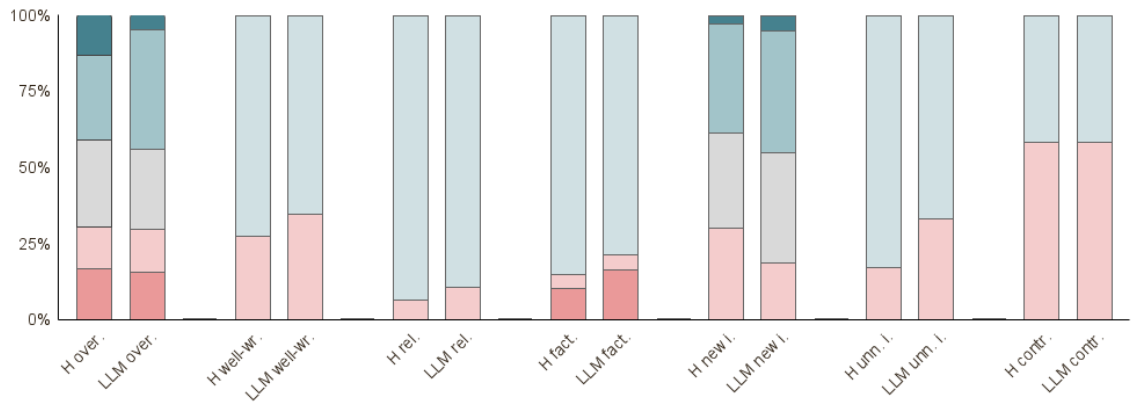


Figure 4.5 Distribution of majority-voted human and LLM ratings per criterion across all explanations. Every left bar represents human and right bar represents LLM ratings. The y-axis represents the ratio of samples given each rating, where lower segments are worse ratings and higher segments are better ratings.

4.4.1 Overall Bias in Human and LLM Ratings

One natural characterization of a rater is their overall bias—the tendency to rate more strictly or leniently compared to others. For example, in ACORN, the average mean *overall rating* of all workers was 3.15 (1-5 stars); of the twenty-seven crowdworkers that rated the 3,500 explanations in ACORN, the most lenient had a mean of 3.72 stars while the strictest had 2.66 stars, a more than one-star difference. Interestingly, there was a weak positive correlation between the number of samples rated by a worker and their mean ratings ($r = 0.15$). The reason for this bias is unclear and left to future work, but it is a plausible hypothesis that being stricter requires more effort, and hence, workers who rated more samples were more lenient, or vice-versa, more lenient workers were able to rate more samples. However, this hypothesis is challenged by the fact that there were workers who rated a large number of samples and were still strict. Where do LLMs stand in this spectrum? Figure 4.5 compares the distribution of human and LLM[7] ratings across all explanations. In all aspects but one (*new information*), the LLM tended to give **stricter** ratings than human voters, with a mean *overall rating* of 3.03 stars and a mean absolute error (MAE) of 0.53 stars, as shown in Table 4.5.

In terms of normalized MAE (NMAE), where the MAE values are normalized over the range of possible ratings for each criterion to allow direct comparison, the LLM performed best on the *contrastive* criterion, with a NMAE of 0.03, followed by *related* and *well-written* with 0.05 and 0.16, respectively. The LLM's bias was most pronounced in *new information*,

[7] `gpt-4o-2024-08-06` with *verbatim* prompts, following the best-performing setup in Section 4.2.1.

Criterion	MAE	NMAE	μ_{human}	μ_{LLM}	LLM bias
Contrastive	0.03	0.03	0.42	0.42	0.00
Related	0.05	0.05	0.93	0.89	-0.04
Factual	0.19	0.10	0.75	0.62	-0.12
Overall	0.53	0.13	3.07	3.03	-0.04
New Info	0.44	0.15	1.11	1.31	+0.20
Well-Written	0.16	0.16	0.72	0.65	-0.07
*Unnecessary Info	0.21	0.21	0.17	0.33	+0.16

Table 4.5 MAE, NMAE, mean (majority-voted) human rating, mean LLM rating, and LLM bias per criterion, sorted by NMAE ascending. NMAE is normalized over the range of possible ratings for each criterion. * indicates that the criterion has inverted labels, i.e., lower is better and a positive bias is more strict.

and *unnecessary information*, the former also being the only criterion where the LLM was more lenient than humans.[8] If the goal is to reproduce human ratings as closely as possible, shifting this biases closer to humans' may be a potential solution. In the following section, we will explore a way to calibrate the LLM's ratings by prompting it with subsets of examples corresponding to individual humans.

4.4.2 Calibrating LLM Ratings to Human Ratings

A straight-forward approach to influencing LLM outputs is to guide it with carefully selected examples. Specifically, we hypothesized that by prompting the LLM with examples rated by a specific human, we could influence the LLM to rate more similarly to that human. Focusing on five workers who rated the most samples in ACORN, we repeated the explanation evaluation process with one LLM per worker, each prompted with the explanations rated by the corresponding worker. To observe the magnitude of the calibration effect, we compared the mean ratings of LLMs prompted in 0-shot, 10-shot, 100-shot, and 200-shot settings with a simple explanation of the rubric, as well as the original *verbatim* prompts with three hand-written examples and extensive guidelines. Due to resource constraints, we limit this experiment to 500 test samples per run.

Results

As shown in Figure 4.6, the results revealed a surprising tendency. The workers are ordered from most lenient to most strict from left to right, with the mean human ratings shown as

[8]Recall that *unnecessary information* has inverted labels, i.e., lower is better, so a positive bias value indicates a stricter bias.

Figure 4.6 Mean ratings of LLMs prompted with explanations rated by individual humans, compared to the original *verbatim* prompts. The y-axis represents the mean ratings, where a higher value indicates a more lenient rating.

the shaded area under the gray dashed line. Each bar represents the mean ratings resulting from each run, i.e., using N examples from the target worker denoted by the x-axis. For context, compared to the means of all other workers averaged over all criteria, workers `3`, `16`, `1`, `19`, and `12` have a +1.15, +1.09, +1.08, +1.06, and -0.11 bias, respectively. With `verbatim` prompts (`amt_verbatim`), we can once again see that it results in a stricter bias than the more lenient workers. However, it was more lenient than the strictest worker, `worker_12` (with some variance due to re-running the LLM). The zero-shot setting dropped the biases, after which we see a surprising effect as the number of examples increases: the biases were amplified *far beyond the original workers*, even with the strictest worker.

There seems, however, to be a sweet spot for each worker of how many examples should be provided to nudge the model appropriately. While the exact optimal value could be determined by fitting a curve to the data, we started by selecting the settings that had the closest result, e.g., 10 examples for `worker_3` and 100 examples for `worker_16`. If this modification improves the alignment with humans, we can conclude that the gap between LLMs and humans is in rating tendencies. In reality, this was not the case—while the modification somewhat improved correlation between the model's ratings and the workers, the original *verbatim* prompts still outperformed most calibrated LLMs (Figure 4.7), with a similar pattern seen in NMAE. A manual inspection reveals the source of this decrease in performance—with higher n-shot examples, the model seems to completely lose its ability to evaluate explanations, causing increasingly extreme NMAE. Table 4.6 shows two examples with the most extreme total NMAE values[9] compared to `worker_3`, the most lenient worker, with 0-shot prompts (total NMAE = 4.83) and 200-shot calibration (total NMAE = 6.33). The first case reads as a genuine difference in leniency, where some might argue the human worker was overly forgiving. The second case, however, cannot be reasonably seen as

[9]Summed over all criteria but *supports*, creating a possible range of 0 to 7 per sample.

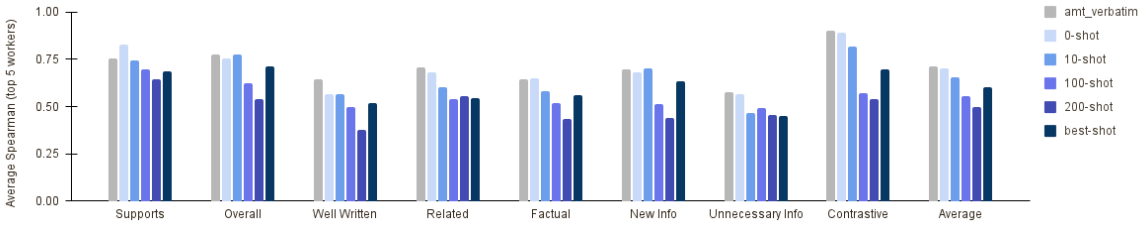


Figure 4.7 Correlation between human and LLM ratings per criterion, comparing the original *verbatim* prompts with the best-performing calibration settings.

a possible rating for the explanation, indicating a complete loss of the model's ability to evaluate explanations with a pathological positive bias.

	Over.	Well-wr.	Rel.	Fact.	New I.	Unn. I.	Contr.
Lasso is capable of being missed. Horse entering barn is obstructed by lasso.							
Worker_3	3	yes	yes	yes	some	no	yes
0-shot	2	no	yes	yes	none	yes	no
Cleaning is a part of work. Not cleaning causes dirty house. Work is used for dirty house.							
Worker_3	1	no	no	N/A	yes	yes	no
200 Shot	5	yes	yes	yes	sufficient	no	yes

Table 4.6 Examples with the most extreme total NMAE values compared to `worker_3`, with *verbatim* prompts (total NMAE = 2.58) and 200-shot calibration (total NMAE = 6.33).

We conclude this section with a few observations. First, humans and LLMs can have a different overall bias, with a recent LLM leaning towards a stricter rating. Second, while it is possible to calibrate LLMs to rate more similarly to humans, the calibration process is not straightforward and can lead to catastrophic failure. Finally, the calibration process was not a successful solution to improve alignment with humans, however, whether that is a desirable outcome to begin with remains an open question.

4.5 Conclusions

In Section 4.2, we found that LLMs can be used as additional raters, but they are not a perfect replacement for human judges. LLMs can also be overly sensitive to prompt differences as well as explanation differences that should not affect unrelated scores (Section 4.3). The best-performing model was found to be overall stricter than human raters, and an attempt to calibrate LLM ratings to align them with human ratings showed that this process

was not straightforward and could lead to catastrophic failure (Section 4.4) Overall, these results suggest that while LLMs can be useful for evaluating explanations, they are not a perfect substitute for human judges, and care must be taken when using them in this capacity. Especially, there should be particular caution in defining the ideal outcome of the alignment between LLMs and humans, as the two may have different biases that are not easily reconcilable or even desirable to align. For example, increased strictness may be desireable in safety-critical applications, but not in creative writing or other more subjective tasks. While individualized calibration proved challenging within the scope of this work, it is an important area for future research to explore further to enable more adapted and appropriate applications. Nonetheless, the results of this chapter suggest a promising use of LLMs during model evaluation, which will be explored in the next chapter.

Chapter 5

Application: *When* do models learn to explain?

The analyses in the previous chapter confirmed that pre-trained LLMs are imperfect but still useful tools for evaluating explanations, since they can provide aspect-wise evaluations while being more scalable and less expensive than human evaluations. Their ratings were slightly different from a human majority vote with a correlation coefficient of 0.72 on average (Figure 4.2), but did not negatively impact agreement between annotators (Figure 4.1). The best-performing model overall, GPT-4o, specifically seems to tend to over-estimate the amount of new information, which is also the one with the lowest correlation with human majorities, but be stricter on all other aspects except for *contrastive* (Figure 4.5, Table 4.5). With these findings in mind, we can now practically apply this method to evaluate explanations at scale, and interpret the results appropriately—It now becomes possible to track the progress of a model's commonsense reasoning explanation skill development, and to identify specific weaknesses that may have been missed with more shallow evaluation or would have been impractical to investigate manually. Specifically, this chapter demonstrates findings gained from evaluating explanations *over time,* across different model sizes, and between different model versions.

5.1 Settings

As the evaluator model, we use GPT-4o prompted with the same guidelines as the human evaluator. This was the most successful setting in the previous analyses in terms of correlation with human majorities, impact on inter-annotator agreement, and sensitivity to explanation differences. Using *verbatim* prompts, the model is given extensive guidelines, three examples,

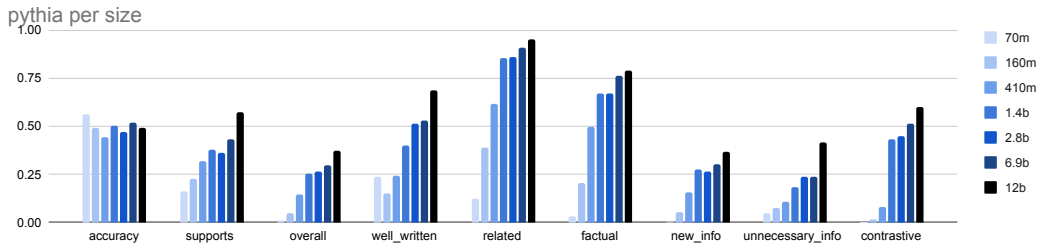


Figure 5.1 Size-wise comparison of explanation skill by Pythia models

and a FAQ based on the most common issues encountered by workers in test rounds (§4.1.2). As for the test set, we use a random subset of 200 questions from the BCOPA dataset. Models are prompted to predict the answer to each question and then explain their reasoning. All models and model versions (henceforth "solver models") are prompted with the same 10-shot predict-and-explain prompts, and the generated explanations are evaluated by the evaluator model. The ten example explanations are sourced from the newly collected high-quality explanations for BCOPA (§3.2.2). We use the following variants of Pythia (Biderman et al., 2023) and GPT (Brown et al., 2020; OpenAI, 2023, 2024; Radford et al., 2018) as solver models:

- Pythia × 70M, 160M, 410M, 1.4B, 2.8B, 6.9B, and 12B parameters;
- Pythia 70M × 5 training checkpoints;
- Pythia 6.9B × 5 training checkpoints; and
- GPT × 5 versions: 2, 3.5, 4, 4o, and 4o-mini.

To enable direct comparison of explanation quality scores, the raw labels are converted to a value normalized between 0 and 1, and *unnecessary information* is further inverted to have a consistent display of higher scores indicating better quality. All results that follow are reported in percentage scores following this normalization.

5.2 Results

5.2.1 Size Comparison

First, let us compare Pythia across different sizes: 70M, 160M, 410M, 1.4B, 2.8B, 6.9B, and 12B parameters. Each was prompted on 200 BCOPA questions with ten examples as described previously, and its labels and explanations are extracted from its generations automatically. Prediction success is measured in accuracy, and the quality of explanations

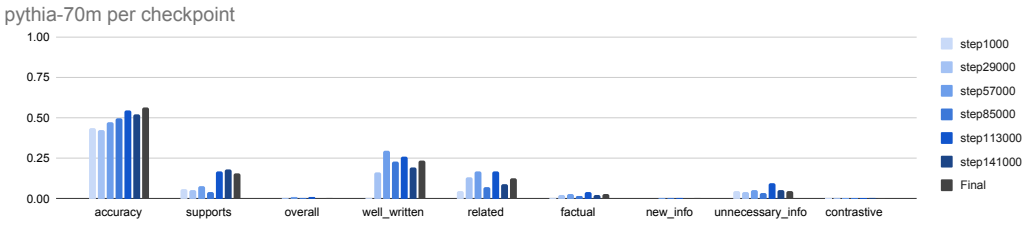
are evaluated by GPT-4o. Figure 5.1 shows the prediction accuracies and aspect-wise quality of explanations, normalized to values between 0 and 1 (higher is better). Model sizes are ordered from smallest to largest from left (lightest) to right (darkest).

If we consider only accuracy, it would seem like none of the models are able to achieve more than random performance (50%) in this setting. However, the explanation quality ratings tell a different story: explanation quality clearly improves with larger sizes. The smallest model, Pythia-70M, has the lowest scores across all but one aspect, with the largest model, Pythia-12B, having the highest scores across all. Further, even though overall scores stay relatively low, fine-grained aspect-wise evaluations reveal more nuanced trends. The largest improvements are seen in the more shallow aspects: *well-written*, *related*, and *factual*. The more challenging aspects, such as *new information* and *unnecessary information*, show less improvement. We can also observe more dramatic jumps in quality at certain size thresholds, such as between 6.9B and 12B in *supports*, *well-written*, and *unnecessary information*, or between 410M and 1.4B for *contrastive*.

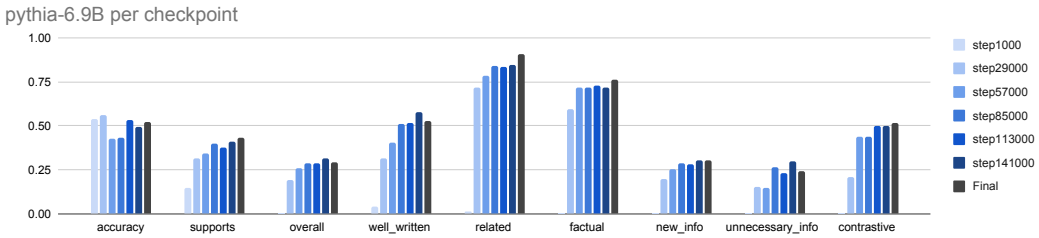
Overall, the results suggest that at within this size range, models have gained the ability to generate highly related and somewhat well-written and factual explanations, but still struggle with adding new information,[1] avoiding unnecessary information, and having overall high-quality explanations. However, the consistent upwards trend suggests that larger models may potentially continue to improve in these aspects, and future work may be done to explore extrapolation methods to predict the required parameters for a desired explanation quality level. As for why the accuracy stays low, a manual inspection of samples that were incorrectly answered revealed more specific weaknesses. For example, some explanations seemed fluent but contained falsehoods, possibly pointing to a lack of factual knowledge leading to lower performance. More interestingly, sometimes the model provided perfectly reasonable explanations for the correct answer but *still stated the wrong answer*. For example, consider the following question, response by Pythia-12B, and evaluation by the evaluator model:

- **Question:** The patient was a close friend of the physician. What is the more plausible effect of this?
  A: The patient disclosed confidential information to the physician.
  B: The patient filed a malpractice lawsuit against the physician.
- **Response:** B; If a patient has a close relationship with the doctor, they're more likely to disclose confidential information to him, but the lawsuit is not a result of the relationship.

[1] Recall that the evaluator model was even more lenient on this aspect than human evaluators, §4.4.1.



(a) Pythia-70M.



(b) Pythia-6.9B.

Figure 5.2 Prediction accuracy and explanation skill of Pythia models at different training checkpoints, ordered left (1000th step, lightest) to right (final, darkest).

- **Evaluation:** supports—`A`, overall—`4/5`, well-written—`yes`, related—`yes`, factual—`yes`, new info—`sufficient`, unnecessary info—`no`, contrastive—`yes`.

While the model incorrectly answered `B`, its explanation actually supports the answer `A` with high quality ratings across all aspects. This not only proves that the explanation and the answers are not necessarily consistent (a well-known issue of post-hoc explanations (Lipton, 2018)), but also that the model may be more capable than the accuracy score would suggest. Out of 200 samples, 55 (28%) were flagged as having such mismatches, of which 48 had an incorrect prediction but an explanation that supported the correct explanation, accounting for 24% of the test samples. If these were correct predictions, it would theoretically bring the accuracy up from 49% to 73%. This inconsistency is reflected in the *supports* aspect, where only 58% of the explanations were rated as supporting the predicted answer.[2] Thanks to this analysis, we can now see that a larger size is likely needed for higher-quality explanations, and that we should pay particular attention to improving its self-consistency if we want to improve the accuracy of the model as well.

[2] 57.5% of explanations by Pythia-12B supported its predicted answer, 27.5% supported the opposite answer than the predicted one, and 15.0% supported neither.

5.2.2 Training Checkpoints

Next, let us consider how a single model gains explanation skill over time during its training, comparing a smaller (70M) and a larger (6.9B) model. Note that this does not include any task-specific fine-tuning, but only the pre-training process. For each training checkpoint, we evaluate the model's accuracy and explanation quality in the same way as before. Figure 5.2a shows the results for Pythia-70M and and Figure 5.2b for Pythia-6.9B, with the rightmost columns corresponding to the final iteration of training, i.e., matching the values in the size-wise comparison (§5.2.1). The difference between the two models is immediately obvious—the 70M model hardly gained any explanation skill from start to finish, while the 6.9B model demonstrated a similar trend to the size comparison, with a steady increase in explanation quality across all aspects. The accuracy of the 70M version does show some improvement, and seems like it may have continued with further training. However, the other aspects show no such trend, suggesting that even continued training would not have resulted in the model learning to explain. The 6.9B model, on the other hand, shows a clear trend of improvement throughout its training path, with the largest jumps in *well-written*, *related*, and *factual* aspects, and more modest improvements in *new information* and *unnecessary information*. This suggests that the model is gaining some ability to explain, but further training is required to learn deeper aspects of explanation. Overall, the results suggest that Pythia-6.9B has not yet reached its full potential in explanation quality, and that further training may continue to improve its explanation quality. The 70M model, on the other hand, does not show any signs of learning to explain, and further training would likely not improve its explanation quality. As with the size comparison, aspect-wise analysis reveals more nuanced trends, and can be used as useful metrics to direct setting choices. It may also be an interesting avenue for future research to investigate the interplay between prediction and explanation skills, and how they can be improved in tandem.

5.2.3 Proprietary Models

Finally, we turn our attention to the GPT model family, comparing different versions of the model, namely GPT-2, GPT-3.5, GPT-4, GPT-4o-mini, and GPT-4o.[3] Figure 5.3 shows accuracy and aspect-wise explanation skill, as in the previous comparisons. The closed nature of these models make this analysis less valuable for development purposes, but it serves well as a historic reflection—the jump between GPT-2 and GPT-3.5 mirrors the emergence of successful few-shot prompting without additional finetuning, with nearly maxed-out scores

[3]The exact model versions are as follows: `gpt-3.5-turbo-0125`, `gpt-4-0125-preview`, `gpt-4o-mini-2024-07-18`, and `gpt-4o-2024-08-06`. GPT-2 is loaded from the Hugging Face repository `openai-community/gpt2`.

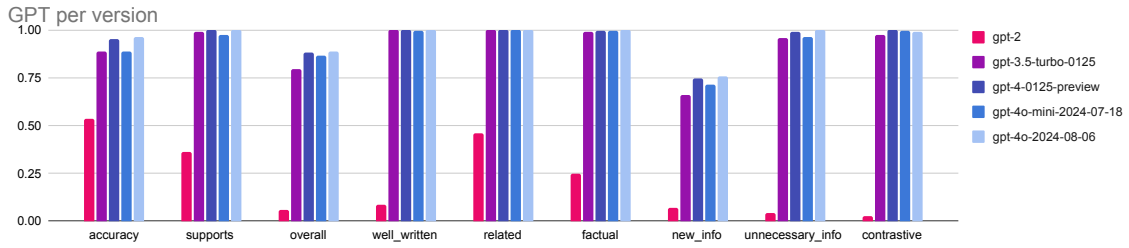


Figure 5.3 Comparison of accuracy and explanation skill per GPT version.

in all aspects. The *new information* aspect stands out as an outlier, even though the evaluator model tends to be more lenient than humans on average. The reason for this is that this criterion is defined in such a way that the second-best rating (i.e., a 2 on a scale of 0 to 3) is still a good rating.[4] Therefore, we can interpret any score above 67% as successful, and anything above it to be a reflection of the level of detail the model provides. All but GPT-2 cross that threshold, and cross-referencing with the *unnecessary information* aspect, we can see that the additional verbosity did not come at the cost of adding irrelevant information. Taking GPT-4 as an example, its 76% normalized average score actually results from 28% of explanations being rated as having *ample* information and the remaining 72% as *sufficient*, meaning the model was 100% successful in providing enough information to explain the answer.

Needless to say, the newer models also excel in terms of accuracy, as BCOPA is a relatively simple task for models of this size. It is particularly encouraging that GPT-4o, the model with the highest *but not perfect* accuracy (96.5%), had a perfect score in *supports*, meaning that its few incorrect predictions came with explanations that supported it. While its explanations are still post-hoc and therefore not necessarily faithful to the model's reasoning by definition, inconsistency seems less prevalent in practice. In contrast, its `mini` variant had five samples with mismatches between labels and explanations, and all five were also incorrect predictions, meaning that its explanations were 0.3% more accurate than the score measured with its prediction labels. This does not fill the gap between the two models' accuracy (96.5% vs. 88.5%). Thus, we meet a limit of this analysis method: we can be confident that a model is capable of explaining skillfully and consistently, but not necessarily that it internally reasons correctly. Determining *why* a model answered incorrectly is still a vast and open research question, and one that is not, and will never be, addressed by this method.

[4] Sufficient (2): There is sufficient additional information to explain the answer. Ample (3): The given information is highly detailed; there is more than enough information. (Section 3.2.3)

5.3 Conclusions

In this chapter, we applied LLMs as explanation evaluators and explored how models develop the ability to explain commonsense reasoning across model sizes, training checkpoints, and versions, reaffirming their utility as diagnostic tools on a large scale. We observed that explanation quality improves with increasing model size and throughout training progressions, demonstrating how larger and more extensively trained models provide richer and more accurate explanations. We discussed in detail how these improvements manifest across different aspects of explanation quality, revealing nuanced trends that can guide future model development. For example, smaller models seemed to have limited potential to learn to explain despite continued training, while larger models showed consistent improvement in explanation quality over time. We particularly highlighted a lack of consistency between predicted labels and explanations, which both obscured the true reasoning abilities of the models in this setting and underscored the potential for further research into the relationship between prediction and explanation skills. Finally, with the strongest models available to us today, we confirmed their high explanation and prediction skill, but also discussed how it reveals the limitations of our evaluation method in determining the internal reasoning of the models. Evaluating the quality and consistency of explanations will tell us how well the model can *explain* a reasoning, but not necessarily how well it can *reason* in the first place. Answering this question will require a different approach, and still remains an open question.

Chapter 6

Discussion

Commonsense reasoning, as a task within NLP, has evolved through the years. From explicit mechanisms to more implicit, data-driven approaches, the field has seen a variety of benchmarks and models. Approaches moved from classification to open-ended responses, to now elements of reasoning having become seemingly ever-present in both research and outside of it. How can evaluation keep up with this evolution? In this chapter, we discuss the implications of our findings, the limitations of our approach, and suggest future directions for commonsense reasoning evaluation, particularly that with the help of LLM-based judges.

When this work was at its infancy, models were not as skilled at commonsense reasoning as they are today. The focus was on classification tasks, where models were given a prompt and asked to choose the correct answer from a set of options. The models were not required to provide explanations for their answers, and the benchmarks were designed to test the models' ability to reason about common sense. However, as models improved, it became clear that they were not always reasoning correctly. They were often relying on superficial cues in the data to make their predictions, rather than truly understanding the underlying concepts. This led to the development of new benchmarks that required models to provide explanations for their answers, in order to better evaluate their reasoning abilities. The hope was that by requiring models to explain their reasoning, it would reveal the model's weaknesses and help create more robust systems as we correct their mistakes. With generative models, explanations became easier to obtain than ever before, as they are a natural extension of the task of generating text. Meanwhile, as models started getting released to the public, the ability to provide justifications for critical decisions became increasingly urgent from legal and practical perspectives. Even from a performance perspective, the added task of explanation generation seemed to improve the model's reasoning abilities such as chain-of-thoughts prompting. The field now faced a new challenge: how to evaluate these explanations?

Early approaches to explanation evaluation focused on criteria which were easy to measure automatically, such as word overlap-based methods using gold references (which were few and far between), or human evaluation. From that starting point, in this work, we added to the resources with a full new dataset of higher quality explanations, as well as human-annotated quality labels, with a dream of helping create an automated yet fine-grained approach to explanation evaluation, which, in turn, we hoped would propel reasoning evaluation with a newly found precision, depth, and flexibility. As LLMs developed, so did their capabilities, and the target of the evaluation itself became a potential tool for it. In this we find the culmination of this work—a thorough analaysis of the feasibility and reliability of such an approach, as well as a demonstration of new insights that can now be gained. We introduced several novel perspectives on the evaluation of the evaluators themselves, taking into consideration the inherent subjectivity of the task.

The field, however, continues to march on, as well as societal, ethical, and technological demands. Generating a plausible explanation consistent with a model's prediction is an important goal, but still a only a performative one. Today, we enter the realm of digging deeper into models—understanding not just their *how* but also their *why*. We want to understand the internal mechanisms of the model's reasoning, and there is an increasing interest in *faithful* textual explanations, which could be defined as a traceable product of some sort of effective reasoning mechanism in the internals of the model. This brings us to the first fundamental limitation of this work, which is that there is no external way to evaluate said "faithfulness." Verifying whether the explanation supports the outputted label is a step in the right direction, but can only prove the lack thereof, not in any way indicate or explain its presence. Future work interested in this direction must start by defining what kind of mechanism we are looking for, before we can start to design methods to evaluate it. Such methods will necessarily have to look beyond only the outputs but without abandoning them entirely: while it's a well-known limitation of generated explanations that they are not faithful representations of the model's inner workings and therefore not a reliable diagnostic tool for reasoning on their own, they are still a valuable and increasingly necessary product of the model's reasoning process.

Even keeping to the goal of helping build high-performing systems, the question remains of what the *ideal* explanation is, and with it who is best qualified to judge its success. In this work, we followed the age-old tradition of using crowdsourcing to collect "average" human judgments, with a research-focused list of criteria to fulfill building upon previous literature on the subject. We measured LLM raters against these human judgments (among other analyses), and declared LLMs to be strong but imperfect judges. Attempts to elaborate said imperfection pushes us to face the fact that the human judgments themselves are not perfect,

and those familiar with any kind of human annotation will be well aware of the slippery concept of an "average" human. Anything from culture and demographics to mood can affect a human's judgment, which in turn is shaped by the context of the task at hand. While LLMs showed promise throughout our analyses, we believe to have only completed the work of characterizing them generally and recording the methodologies themselves. Future work interested in preparing or using LLM-based judges for any kind of reasoning task should start from a practical human perspective, clearly designing the ideal judge for the evaluation at hand, and then working to make the LLMs align with that ideal, with the help of the insights contained within this dissertation.

Finally, we may conclude by presenting a resumé of the concrete insights gained through this work. First, commonsense reasoning, as reflected in explanation skill, can be broken down into sub-skills involving surface-level, factuality, and more abstract skills such as minimality and completeness. Sucess in each of these skills can be separately traced as models become more powerful, either throughout their training, or through sizes and versions. This progress is not linear, and even fluent and plausible explanations may sometimes come paired with the wrong label, betraying a lack of internal consistency or deeper reasoning. These insights are not newly discovered, but we can now empirically illustrate it by enabling scalable, model-agnostic, and fine-grained evaluation. As for LLMs as evaluators, they behave closely but not "perfectly" as humans, causing a slight drift in inter-annotator agreement when they are introduced to a human annotator group. In turn, their ratings were also not completely matching an individual human, as demonstrated by a much smaller gain towards restoring the original voted labels by adding it as an additional vote. When tested on a targeted subset of edited examples, some were vulnerable to changes in labels that humans were not sensitive to, as well as being unstable with even negligible changes in prompt format. We were unable to confirm the hypothesis that the remaining gap between models and humans is in overall rating tendencies, i.e., strictness or leniency, by calibrating the models to rate more similarly to humans. However, we did observe a difference in overall bias between humans and LLMs, with the model evaluated leaning towards a stricter rating in all but *new information*. As mentioned previously, the solution to this is not straightforward, and should begin with a reexamination of how we define the ideal judge for the task at hand. For commonsense reasoning evaluation, the flexibility of LLMs allow us to move past set benchmarks, and may allow a more involved process of dynamically coupling challenges and fine-grained evaluations. This is a promising direction for future work, as we continue to strive for more robust and reliable models in the field of commonsense reasoning.

Chapter 7

Conclusion

This dissertation has explored the evaluation of commonsense reasoning through the lens of textual explanation evaluation. We have introduced a new dataset of explanations for commonsense reasoning tasks, and annotated them with quality labels. We have demonstrated the feasibility and reliability of using large language models as judges for explanation evaluation, and have shown that they can provide valuable insights into the reasoning capabilities of these models. We have also explored the limitations of this approach, and have suggested future directions for commonsense reasoning evaluation.

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Aggarwal, S., Mandowara, D., Agrawal, V., Khandelwal, D., Singla, P., and Garg, D. (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.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Mag.*, 36(1):15–24.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., and Testoni, A. (2024). LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. *arXiv preprint arXiv:2406.18403*.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

BIG-bench authors (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brassard, A., Heinzerling, B., Kavumba, P., and Inui, K. (2022). COPA-SSE: Semi-structured Explanations for Commonsense Reasoning. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3994–4000, Marseille, France. European Language Resources Association.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Calderon, N., Reichart, R., and Dror, R. (2025). The Alternative Annotator Test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. *arXiv preprint arXiv:2501.10970*.

Celikyilmaz, A., Clark, E., and Gao, J. (2021). Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Chaleshtori, F. H., Ghosal, A., Gill, A., Bambroo, P., and Marasović, A. (2024). On evaluating explanation utility for human-ai decision making in nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7456–7504.

Clinciu, M.-A., Eshghi, A., and Hastie, H. (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*, Stroudsburg, PA, USA. Association for Computational Linguistics.

European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council.

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1(6):261–262.

Gemma Team (2024). Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Gordon, A., Kozareva, Z., and Roemmele, M. (2012). SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Gurrapu, S., Kulkarni, A., Huang, L., Lourentzou, I., and Batarseh, F. A. (2023). Rationalization for explainable nlp: a survey. *Frontiers in artificial intelligence*, 6:1225093.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hartmann, M. and Sonntag, D. (2022). A survey on improving NLP models with human explanations. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 40–47, Dublin, Ireland. Association for Computational Linguistics.

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., and Rieser, V. (2020). Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Huang, J. and Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065.

Hwang, J. D., Bhagavatula, C., Bras, R. L., Da, J., Sakaguchi, K., Bosselut, A., and Choi, Y. (2021). COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Ismayilzada, M. and Bosselut, A. (2023). kogito: A Commonsense Knowledge Inference Toolkit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 96–104, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacovi, A. and Goldberg, Y. (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.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of Experts. *arXiv preprint arXiv:2401.04088*.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250.

Kavumba, P., Brassard, A., Heinzerling, B., and Inui, K. (2023). Prompting for explanations improves adversarial nli. is this true?{Yes} it is {true} because {it weakens superficial cues}. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2165–2180.

Kavumba, P., Inoue, N., Heinzerling, B., Singh, K., Reisert, P., and Inui, K. (2019). When Choosing Plausible Alternatives, Clever Hans can be Clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Korteling, J. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., and Eikelboom, A. R. (2021). Human-versus artificial intelligence. *Frontiers in artificial intelligence*, 4:622364.

Kunz, J., Jirenius, M., Holmström, O., and Kuhlmann, M. (2022). Human ratings do not reflect downstream utility: A study of free-text explanations for model predictions. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 164–177.

Lampinen, A., Dasgupta, I., Chan, S., Mathewson, K., Tessler, M., Creswell, A., McClelland, J., Wang, J., and Hill, F. (2022). Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563.

Levesque, H. J., Davis, E., and Morgenstern, L. (2012). The winograd schema challenge.

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., et al. (2024). From generation to judgment: Opportunities and challenges of LLM-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queueing Syst.*, 16(3):31–57.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv preprint arXiv:2107.13586*.

Llama Team (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.*, 55(13s):1–42.

OpenAI (2023). GPT-4 Technical Report. *ArXiv preprint arXiv:2303.08774*.

OpenAI (2024). GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/. Accessed: 2024-8-10.

Patzig, G. (2013). *Aristotle's theory of the syllogism: A logico-philological study of book A of the prior analytics*, volume 16. Springer Science & Business Media.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Shi, S., Choi, M., Agrawal, A., Chopra, A., et al. (2025). Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., and Chen, H. (2023). Reasoning with language model prompting: A survey. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rajani, N. F., McCann, B., Xiong, C., and Socher, R. (2019). Explain yourself! Leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rawte, V., Sheth, A., and Das, A. (2023). A survey of hallucination in Large Foundation Models. *ArXiv preprint arXiv:2309.05922*.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. (2024). GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Richardson, H. S. (2018). Moral Reasoning. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2018 edition.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Vidyabhusana, S. and Sinha, N. (1990). *The Nyâya Sûtras of Gotama*. Motilal Banarsidass.

Wadhwa, S., Amir, S., and Wallace, B. (2023). Revisiting relation extraction in the era of large language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Waley, A. et al. (2012). *The analects of Confucius*. Routledge.

Wallace, R. J. and Kiesewetter, B. (2024). Practical Reason. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Webson, A. and Pavlick, E. (2022). Do Prompt-Based Models Really Understand the Meaning of Their Prompts? In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022a). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022b). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M., and Choi, Y. (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.

Wiegreffe, S. and Marasović, A. (2021). Teach me to explain: A review of datasets for explainable natural language processing. In Vanschoren, J. and Yeung, S., editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

List of Publications

International Conference Papers (Refereed)

1. Ana Brassard, Benjamin Heinzerling, Keito Kudo, Keisuke Sakaguchi and Kentaro Inui. ACORN: Aspect-wise Commonsense Reasoning Explanation Evaluation. 1st Conference on Language Modeling (COLM 2024).

2. Ana Brassard, Benjamin Heinzerling, Pride Kavumba, and Kentaro Inui. COPA-SSE: Semi-structured Explanations for Commonsense Reasoning. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022).

Explanation Rating HIT Guidelines

Index

1. Introduction
2. Task description
3. Detailed descriptions of scoring criteria
4. Examples
5. Policies
6. FAQs

---

1. Introduction

Welcome! We are the Riken AIP Natural Language Understanding Team—we conduct research in natural language processing, a subfield of AI. Our account is shared by several subteams, each handling their respective projects. This particular one is managed by Ana, who will address all queries concerning this HIT. Thank you for your continued support.

Feel free to contact us at nlu.team.amt@gmail.com.

2. Task description

In this HIT, we collect detailed evaluations of how "good" explanations are for given commonsense questions and answers. Some explanations were written by humans in earlier HITs; others were generated by AI systems.

We defined a list of criteria for the quality of explanations ranging from surface-level to more high-level concerns. We ask you to carefully **consider each criterion separately** and additionally give an overall score for the explanation, independently of our criteria.

3. Detailed descriptions of the scoring criteria

Below are more detailed descriptions for each criterion. Make sure to consider each criterion independently. For example, if an explanation is contrastive but contains false information, mark "Contrastive" as "yes" and "Factual" as "no".

1. **“Supports”**

Which answer does the explanation attempt to justify—a), b), ..., or none? Note that this may not be the same as the correct answer to the question.

2. **“Overall rating”**

Independently of the other criteria, how good is the given explanation? Use your own judgment here; imagine an AI system is offering it to you as a justification for its answer to the question. Are you convinced that its reasoning is sound?

★☆☆☆☆ Terrible, utterly insufficient.
★★☆☆☆ A decent attempt but has major issues.
★★★☆☆ Okay, could be better. Not terrible, but not great either.
★★★★☆ Very good, but not quite amazing.
★★★★★ Excellent. You are convinced; wouldn’t add or remove anything.

3. **“Well-written”**

Is the explanation a generally coherent, grammatically correct, and/or fluent sentence? Ignore all other factors, such as being related to the question or the amount of given information.

No → Some issues present, e.g., grammatical errors, not fluent, incoherent.
Yes → A clear and correct sentence; no issues.

4. **“Related”**

Is the explanation relevant to the question and answer? We consider an explanation that mentions the same concepts as still relevant even if it does not form a coherent argument. Irrelevant explanations mention completely unrelated concepts.

No → The explanation is an unrelated statement; mentions completely irrelevant things.
Yes → The explanation is related to both the question and answer.

5. **“Factual”**

Explanations should provide some facts, i.e., background information, about the world that help justify the answer. E.g., “*It cannot be dark if the sun is rising because the sun is a light source.*” Here we ask whether those mentioned facts, regardless of their relevancy, are generally true statements.

No → There is a false statement.
Yes → All given facts, if any, are generally true.
N/A → No general facts are mentioned.

6. **“New information”**

Regardless of whether the information is true, how much new information does the explanation provide to support the answer? This can be a subjective measure; imagine the explanation is given to you by a student or an AI. How much information would you be satisfied with?

> **None** → No new information provided, e.g., restating the question or answer.
> **Some** → Some new information, but insufficient to fully explain the answer.
> **Sufficient** → There is sufficient additional information to explain the answer.
> **Ample** → The given information is highly detailed; there is more than enough information.

7. **“Unnecessary information”**

Some explanations are highly elaborate and detailed, but not all information is necessarily needed. Are there any statements that do not belong? E.g., mentioning that a mouse has four legs to explain why it ate some cheese is not a necessary piece of information.

> **Yes** → There is some unnecessary information mixed in the explanation.
> **No** → All information, if any, supports the answer.

8. **“Contrastive”**

Does the explanation clearly show the *difference* between the answers? Does it explain why the alternative answer is incorrect or less likely?

> **No** → The difference between the answers is not explained. Select this choice if the explanation only supports the correct answer.
> **Yes** → The explanation clarifies both why the answer is correct and why the alternative is incorrect.

4. Examples

Below are some example HITs and expected responses.

The woman sensed a pleasant smell. What happened as a result?
a) She was reminded of her childhood.
b) She remembered to take her medication.

Explanation: **Pleasant smell is a way of bring happiness. Happiness causes nostalgia. Nostalgia is related to a smell. Smell causes her to think her childhood.**

Supports:	[ ] neither	[X] a)	[ ] b)	
Overall rating:	★★★★☆ **(4/5)**			
Well-written:	[X] no	[ ] yes		
Related:	[ ] no	[X] yes		
Factual:	[ ] no	[X] yes	[ ] N/A	
New information:	[ ] none	[ ] some	[X] sufficient	[ ] ample
Unnecessary information:	[ ] yes	[X] no		
Contrastive:	[X] no	[ ] yes		

It got dark outside. What happened as a result?

a) Snowflakes began to fall from the sky.
b) The moon became visible in the sky.

Explanation: **The darkness is filled with snowflakes.**

Supports:	[X] neither	[ ] a)	[ ] b)	
Overall rating:	★☆☆☆☆ **(1/5)**			
Well-written:	[X] no	[ ] yes		
Related:	[ ] no	[X] yes		
Factual:	[ ] no	[ ] yes	[X] N/A	
New information:	[X] none	[ ] some	[ ] sufficient	[ ] ample
Unnecessary information:	[X] yes	[ ] no		
Contrastive:	[X] no	[ ] yes		

The woman tolerated her friend's difficult behavior. What was the cause of this?

a) The woman knew her friend was going through a hard time.
b) The woman felt that her friend took advantage of her kindness.

Explanation: **Going through tough times can lead to challenging behavior, such as someone appearing to be irritable or annoyed. This kind of behavior is often a result of the struggles they are going through.**

Supports:	[ ] neither	[X] a)	[ ] b)
Overall rating:	★★★★★ **(5/5)**		

Well-written:	[ ] no	[X] yes		
Related:	[ ] no	[X] yes		
Factual:	[ ] no	[X] yes	[ ] N/A	
New information:	[ ] none	[ ] some	[ ] sufficient	[X] **ample**
Unnecessary information:	[ ] yes	[X] no		
Contrastive:	[X] no	[ ] yes		

---

5. Policies

Rejections
We carefully select workers to avoid rejections and combat fraud attempts. If answers are unsatisfactory, we will revoke the qualification. We may reverse our decision, but we try to keep a limited but reliable, motivated, and skilled workforce. Thank you.

6. FAQs

*> How should we rate completely **unrelated** explanations?*

A: Each criterion should be considered independently. E.g., if the explanation is contrastive, contains generally true facts, and provides ample new information, mark it as such even if it lacks in other aspects. Assign a low overall rating if you feel it's generally not a satisfactory justification for the answer.

*> How should we mark "**New Information**" and "**Unnecessary Information**" if the given information is completely false?*

A: Disregard the truthfulness of the information. Focus on whether the explanation provides new information and whether the provided information is necessary. Whether the information is true should be stated under "Factual".

*> **How strict** should we be in evaluating the explanations? What is considered a sufficient or ample **amount of information**?*

A: Consider the purpose of this data—to train AI systems to explain answers well to humans. Evaluate the explanations based on your judgment of satisfactory information and factual claims.