Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing

Sarah Wiegreffe[*]
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Ana Marasović[*]
Allen Institute for AI
University of Washington
anam@allenai.org

Abstract

Explainable Natural Language Processing (ExNLP) has increasingly focused on collecting human-annotated textual explanations. These explanations are 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 a ground-truth to evaluate model-generated explanations. In this review, we identify 65 datasets with three predominant classes of textual explanations (highlights, free-text, and structured), organize the literature on annotating each type, identify strengths and shortcomings of existing collection methodologies, and give recommendations for collecting ExNLP datasets in the future.

1 Introduction

Interpreting supervised machine learning (ML) models is crucial for ensuring their reliability and trustworthiness in high-stakes scenarios. Models that produce justifications for their individual predictions (sometimes referred to as *local explanations*) can be inspected for the purposes of debugging, quantifying bias and fairness, understanding model behavior, and ascertaining robustness and privacy [83]. These benefits have led to the development of datasets that contain human justifications for the true label (overviewed in Tables 3–5). In particular, human justifications are used for three goals: (i) to aid models with additional training supervision [142], (ii) to train interpretable models that explain their own predictions [20], and (iii) to evaluate plausibility of model-generated explanations by measuring their agreement with human explanations [29].

Dataset collection is the most under-scrutinized component of the ML pipeline [93]—it is estimated that 92% of ML practitioners encounter data cascades, or downstream problems resulting from poor data quality [109]. It is important to constantly evaluate data collection practices critically and standardize them [13, 39, 95]. We expect that such examinations are particularly valuable when many related datasets are released contemporaneously and independently in a short period of time, as is the case with ExNLP datasets.

This survey aims to review and summarize the literature on collecting textual explanations, highlight what has been learned to date, and give recommendations for future dataset construction. It complements other explainable AI (XAI) surveys and critical retrospectives that focus on definitions, methods, and/or evaluation [33, 15, 77, 1, 103, 51, 42, 133, 26, 44, 82, 121, 12, 86, 54, 19], but not on datasets. We call such datasets ExNLP datasets, because modeling them for the three goals mentioned above requires NLP techniques. Datasets and methods for explaining fact checking [65] and reading comprehension [117] have been reviewed; we are the first to review all datasets with textual explanations regardless of task, comprehensively categorize them into three distinct classes, and provide critical retrospectives and best-practice recommendations.

[*] Equal contributions.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

Instance	Explanation
Premise: A white race dog wearing the number eight runs on the track. *Hypothesis:* A white race dog runs around his yard. *Label:* contradiction	(highlight) *Premise:* A white race dog wearing the number eight runs on the track . *Hypothesis:* A white race dog runs around his yard .
	(free-text) A race track is not usually in someone's yard.
Question: Who sang the theme song from Russia With Love? *Paragraph:* ...The theme song was composed by Lionel Bart of Oliver! fame and sung by Matt Monro... *Answer:* Matt Monro	(structured) *Sentence selection:* (not shown) *Referential equality:* "the theme song from russia with love" (from question) = "The theme song" (from paragraph) *Entailment:* X was composed by Lionel Bart of Oliver! fame and sung by ANSWER. ⊢ ANSWER sung X

Table 1: Examples of explanation types discussed in §2. The first two rows show a highlight and free-text explanation for an E-SNLI instance [20]. The last row shows a (partial) structured explanation from QED for a NATURALQUESTIONS instance [70].

Instance with Highlight	Highlight Type Clarification
Review: this film is extraordinarily horrendous and I'm not going to waste any more words on it. *Label*: negative	(¬comprehensive) *Review:* this film is [redacted] and I'm not going to waste any more words on it.
Review: this film is extraordinarily horrendous and I'm not going to waste any more words on it . *Label*: negative	(comprehensive) *Review:* this film is [redacted] and I'm not going to [redacted] .
Premise: A shirtless man wearing white shorts. *Hypothesis:* A man in white shorts is running on the sidewalk. *Label*: neutral	(¬sufficient) *Premise:* [redacted] *Hypothesis:* [redacted] man [redacted] running on the sidewalk.

Table 2: Examples of highlights differing in comprehensiveness and sufficiency (discussed in §2, §4).

We first define relevant EXNLP terminology (§2) and overview 65 existing datasets (§3), accompanied with a live version of the tables as a website accepting community contributions: `https://exnlpdatasets.github.io`. We next analyze what can be learned from existing data collection methodologies. In §4 and §5, we highlight two points that we expect to be particularly important to the current ExNLP research. Specifically, §4 discusses the traditional process of collecting explanations by asking annotators to highlight parts of the input, and its discrepancies with evaluating model-generated highlight explanations. We also draw attention to how assumptions made for collecting free-text explanations (introduced in §2) influence their modeling, and call for better documentation of explanation collection. In §5, we illustrate that not all template-like free-text explanations are incorrect, and call for embracing the structure of an explanation when appropriate. Unlike discussions in §4–5 that are motivated by EXNLP modeling and evaluation choices, the rest of this paper reflects on relevant points from a broader NLP research. In §6, we present a proposal for controlling quality in explanation collection, and in §7, gather recommendations from related subfields to further reduce data artifacts by increasing diversity of collected explanations.

2 Explainability Lexicon

An explanation can be described as a "three-place predicate: *someone* explains *something* to *someone*" [50]. The *something* being explained in machine learning systems are task labels: explanations are implicitly or explicitly designed to answer the question "why is [input] assigned [label]?". However, collected explanations can vary in format. We identify three types in the EXNLP literature: *highlights*, *free-text*, and *structured* explanations. An example of each type is given in Table 1. Since a consensus on terminology has not yet been reached, we describe each type below.

Highlights are subsets of the input elements (words, phrases, or sentences) that explain a prediction. Lei et al. [73] coin them *extractive rationales*, or subsets of the input tokens of a textual task that satisfy two properties: (i) *compactness*, they are short and coherent, and (ii) *sufficiency*, they suffice for prediction as a substitute of the original text. Yu et al. [141] introduce a third criterion, (iii) *comprehensiveness*, that all the evidence that supports the prediction is selected, not just a sufficient set. Since the term "rationale" implies human-like intent, Jacovi and Goldberg [55] argue to call this type of explanation *highlights* to avoid inaccurately attributing human-like social behavior to AI systems. They are also called *evidence* in fact-checking and multi-document question answering (QA) [65]—a part of the source that refutes/supports the claim. To reiterate, highlights should be sufficient to explain a prediction and compact; if they are also comprehensive, we call them *comprehensive*

Dataset	Task	Granularity	Collection	# Instances
MOVIEREVIEWS [142]	sentiment classification	none	author	1,800
MOVIEREVIEWS$_c$ [29]	sentiment classification	none	crowd	200$^{\ddagger\diamondsuit}$
SST [113]	sentiment classification	none	crowd	11,855$^{\diamondsuit}$
WIKIQA [136]	open-domain QA	sentence	crowd + authors	1,473
WIKIATTACK [22]	detecting personal attacks	none	students	1089$^{\diamondsuit}$
E-SNLI† [20]	natural language inference	none	crowd	∼569K (1 or 3)
MULTIRC [60]	reading comprehension QA	sentences	crowd	5,825
FEVER [118]	verifying claims from text	sentences	crowd	∼136K‡
HOTPOTQA [137]	reading comprehension QA	sentences	crowd	112,779
Hanselowski et al. [47]	verifying claims from text	sentences	crowd	6,422 (varies)
NATURALQUESTIONS [68]	reading comprehension QA	1 paragraph	crowd	n/a‡ (1 or 5)
COQA [104]	conversational QA	none	crowd	∼127K (1 or 3)
COS-E V1.0† [100]	commonsense QA	none	crowd	8,560
COS-E V1.11† [100]	commonsense QA	none	crowd	10,962
BOOLQ$_c$ [29]	reading comprehension QA	none	crowd	199$^{\ddagger\diamondsuit}$
EVIDENCEINFERENCE V1.0 [71]	evidence inference	none	experts	10,137
EVIDENCEINFERENCE V1.0$_c$ [29]	evidence inference	none	experts	125‡
EVIDENCEINFERENCE V2.0 [30]	evidence inference	none	experts	2,503
SCIFACT [123]	verifying claims from text	1-3 sentences	experts	995‡ (1-3)
Kutlu et al. [67]	webpage relevance ranking	2-3 sentences	crowd	700 (15)
SCAT [139]	document-level machine translation	none	experts	∼14K
ECTHR [24]	alleged legal violation prediction	paragraphs	auto + expert	∼11K
HUMMINGBIRD [48]	style classification	words	crowd	500
HATEXPLAIN [79]	hate-speech classification	phrases	crowd	20,148 (3)

Table 3: Overview of datasets with textual **highlights**. Values in parentheses indicate number of explanations collected per instance (if > 1). DeYoung et al. [29] collected or recollected annotations for prior datasets (marked with the subscript c). $\diamondsuit$ Collected > 1 explanation per instance but only release 1. † Also contains free-text explanations. ‡ A subset of the original dataset that is annotated. It is not reported what subset of NATURALQUESTIONS has both a long and short answer.

highlights. Although the community has settled on criteria (i)–(iii) for highlights, the extent to which collected datasets (Table 3) reflect them varies greatly, as we will discuss in §4. Table 2 gives examples of sufficient vs. non-sufficient and comprehensive vs. non-comprehensive highlights.

Free-text explanations are free-form textual justifications that are not constrained to the words or modality of the input instance. They are thus more expressive and generally more readable than highlights. This makes them useful for explaining reasoning tasks where explanations must contain information outside the given input sentence or document [20, 128]. They are also called *textual* [62] or *natural language explanations* [20], terms that have been overloaded [98]. Synonyms, *free-form* [20] or *abstractive explanations* [87] do not emphasize that the explanation is textual.

Finally, **structured explanations** are explanations that are not entirely free-form although they are still written in natural language. For example, there may be constraints placed on the explanation-writing process, such as the required use of specific inference rules. We discuss the recent emergence of structured explanations in §5. Structured explanations do not have one common definition; we elaborate on dataset-specific designs in §3. An example is given in the bottom row of Table 1.

3 Overview of Existing Datasets

We overview currently available EXNLP datasets by explanation type: highlights (Table 3), free-text explanations (Table 4), and structured explanations (Table 5). Besides SCAT [139], to the best of our knowledge, all existing datasets are in English. The authors of ∼66% papers cited in Tables 3–5 report the dataset license in the paper or a repository, and 45.61% use *common* permissive licenses; for more information see Appendix B. See Appendix C for collection details.

For each dataset, we report the number of instances (input-label pairs) and the number of explanations per instance (if > 1). The annotation procedure used to collect each dataset is reported as: crowd-annotated ("crowd"); automatically annotated through a web-scrape, database crawl, or merge of existing datasets ("auto"); or annotated by others ("experts", "students", or "authors"). Some authors perform semantic parsing on collected explanations (denoted with $*$); we classify them by the dataset type before parsing and list the collection type as "crow + authors". Tables 3-5 elucidate that the dominant collection paradigm ($\geq$90%) is via human (crowd, student, author, or expert) annotation.

Dataset	Task	Collection	# Instances
Jansen et al. [56]	science exam QA	authors	363
Ling et al. [76]	solving algebraic word problems	auto + crowd	∼101K
Srivastava et al. [115]*	detecting phishing emails	crowd + authors	7 (30-35)
BABBLELABBLE [46]*	relation extraction	students + authors	200‡‡
E-SNLI [20]	natural language inference	crowd	∼569K (1 or 3)
LIAR-PLUS [4]	verifying claims from text	auto	12,836
COS-E V1.0 [100]	commonsense QA	crowd	8,560
COS-E V1.11 [100]	commonsense QA	crowd	10,962
ECQA [2]	commonsense QA	crowd	10,962
SEN-MAKING [124]	commonsense validation	students + authors	2,021
CHANGEMYVIEW [10]	argument persuasiveness	crowd	37,718
WINOWHY [144]	pronoun coreference resolution	crowd	273 (5)
SBIC [111]	social bias inference	crowd	48,923 (1-3)
PUBHEALTH [64]	verifying claims from text	auto	11,832
Wang et al. [125]*	relation extraction	crowd + authors	373
Wang et al. [125]*	sentiment classification	crowd + authors	85
E-δ-NLI [18]	defeasible natural language inference	auto	92,298 (∼8)
BDD-X†† [62]	vehicle control for self-driving cars	crowd	∼26K
VQA-E†† [75]	visual QA	auto	∼270K
VQA-X†† [94]	visual QA	crowd	28,180 (1 or 3)
ACT-X†† [94]	activity recognition	crowd	18,030 (3)
Ehsan et al. [34]††	playing arcade games	crowd	2000
VCR†† [143]	visual commonsense reasoning	crowd	∼290K
E-SNLI-VE†† [32]	visual-textual entailment	crowd	11,335 (3)‡
ESPRIT†† [101]	reasoning about qualitative physics	crowd	2441 (2)
VLEP†† [72]	future event prediction	auto + crowd	28,726
EMU†† [27]	reasoning about manipulated images	crowd	48K

Table 4: Overview of EXNLP datasets with **free-text explanations** for textual and visual-textual tasks (marked with †† and placed in the lower part). Values in parentheses indicate number of explanations collected per instance (if > 1). ‡ A subset of the original dataset that is annotated. ‡‡ Subset publicly available. $*$ Authors semantically parse the collected explanations.

Highlights (Table 3) The granularity of highlights depends on the task they are collected for. The majority of authors do not place a restriction on granularity, allowing words, phrases, or sentences of the original input document to be selected. The coarsest granularity in Table 3 is one or more paragraphs in a longer document [68, 24]. We exclude datasets that include an associated document as evidence without specifying the location of the explanation within the document (namely document retrieval datasets). We exclude BEERADVOCATE [80] because it has been retracted.

Some highlights are re-purposed from annotations for a different task. For example, MULTIRC [60] contains sentence-level highlights that indicate justifications of answers to questions. However, they were originally collected for the authors to assess that each question in the dataset requires multi-sentence reasoning to answer. Another example is STANFORD SENTIMENT TREEBANK [SST; 113] which contains crowdsourced sentiment annotations for word phrases extracted from movie reviews [90]. Word phrases that have the same sentiment label as the review can be heuristically merged to get phrase-level highlights [23]. Other highlights in Table 3 are collected by instructing annotators. Instead of giving these instructions verbatim, their authors typically describe them concisely, e.g., they say annotators are asked to highlight words justifying, constituting, indicating, supporting, or determining the label, or words that are essential, useful, or relevant for the label. The difference in wording of these instructions affects how people annotate explanations. In §4, we discuss how one difference in annotation instructions (requiring comprehensiveness or not) can be important.

Free-Text Explanations (Table 4) This is a popular explanation type for both textual and visual-textual tasks, shown in the first and second half of the table, respectively. Most free-text explanations are generally no more than a few sentences per instance. One exception is LIAR-PLUS [5], which contains the conclusion paragraphs of web-scraped human-written fact-checking summaries.

Structured Explanations (Table 5) Structured explanations take on dataset-specific forms. One common approach is to construct a chain of facts that detail the reasoning steps to reach an answer

Dataset	Task	Explanation Type	Collection	# Instances
WorldTree V1 [57]	science exam QA	explanation graphs	authors	1,680
OpenBookQA [81]	open-book science QA	1 fact from WorldTree	crowd	5,957
Yang et al. [135]††	action recognition	lists of relations + attributes	crowd	853
WorldTree V2 [132]	science exam QA	explanation graphs	experts	5,100
QED [70]	reading comp. QA	inference rules	authors	8,991
QASC [61]	science exam QA	2-fact chain	authors + crowd	9,980
eQASC [58]	science exam QA	2-fact chain	auto + crowd	9,980 (∼10)
+ Perturbed	science exam QA	2-fact chain	auto + crowd	n/a‡
eOBQA [58]	open-book science QA	2-fact chain	auto + crowd	n/a‡
Ye et al. [138]*	SQuAD QA	semi-structured text	crowd + authors	164
Ye et al. [138]*	NaturalQuestions QA	semi-structured text	crowd + authors	109
R^4C [53]	reading comp. QA	chains of facts	crowd	4,588 (3)
StrategyQA [41]	implicit reasoning QA	reasoning steps w/ highlights	crowd	2,780 (3)
TriggerNER	named entity recognition	groups of highlighted tokens	crowd	∼7K (2)

Table 5: Overview of ExNLP datasets with **structured explanations** (§5). Values in parentheses indicate number of explanations collected per instance (if > 1). †† Visual-textual dataset. $*$ Authors semantically parse the collected explanations. ‡ Subset of instances annotated with explanations is not reported. Total # of explanations is 855 for eQASC Perturbed and 998 for eOBQA.

("chains of facts"). Another is to place constraints on the textual explanations that annotators can write, such as requiring the use of certain variables in the input ("semi-structured text").

The WorldTree datasets [57, 132] propose explaining elementary-school science questions with a combination of chains of facts and semi-structured text, termed "explanation graphs". The facts are individual sentences written by the authors that are centered around a set of shared relations and properties. Given the chain of facts for an instance (6.3 facts on average), the authors can construct an explanation graph by linking shared words in the question, answer, and explanation.

OpenBookQA [OBQA; 81] uses single WorldTree facts to prime annotators to write QA pairs. Similarly, each question in QASC [61] contains two associated science facts from a corpus selected by human annotators who wrote the question. Jhamtani and Clark [58] extend OBQA and QASC with two-fact chain explanation annotations, which are automatically extracted from a fact corpus and validated with crowdsourcing. The resulting datasets, eQASC and eOBQA, contain multiple valid and invalid explanations per instance, as well as perturbations for robustness testing (eQASC-perturbed).

A number of structured explanation datasets supplement datasets for reading comprehension. Ye et al. [138] collect semi-structured explanations for NaturalQuestions [68] and SQuAD [102]. They require annotators to use phrases in both the input question and context, and limit them to a small set of connecting expressions. Inoue et al. [53] collect R^4C, fact chain explanations for HotpotQA [137]. Lamm et al. [70] collect explanations for NaturalQuestions that follow a linguistically-motivated form (see the example in Table 1). We discuss structured explanations further in §5.

4 Link Between ExNLP Data, Modeling, and Evaluation Assumptions

All three parts of the machine learning pipeline (data collection, modeling, and evaluation) are inextricably linked. In this section, we discuss what ExNLP modeling and evaluation research reveals about the qualities of available ExNLP datasets, and how best to collect such datasets in the future.

Highlights are usually evaluated following two criteria: (i) *plausibility*, according to humans, how well a highlight supports a predicted label [133, 29], and (ii) *faithfulness* or *fidelity*, how accurately a highlight represents the model's decision process [6, 127]. Human-annotated highlights (Table 2) are used to measure the plausiblity of model-produced highlights: the higher the overlap between the two, the more plausible model highlights are considered. On the other hand, a highlight that is both sufficient (implies the prediction, §2; first example in Table 2) and comprehensive (its complement in the input does *not* imply the prediction, §2; second example in Table 2) is regarded as faithful to the prediction it explains [29, 23]. Since human-annotated highlights are used only for evaluation of plausibility but not faithfulness, one might expect that the measurement and modeling of faithfulness cannot influence how human-authored explanations should be collected. In this section, we show that this expectation might lead to collecting highlights that are unfitting for the goals (ii) and (iii) in §1.

Typical instructions for collecting highlights encourage sufficiency and compactness, but not comprehensiveness. For example, DeYoung et al. [29] deem MOVIEREVIEWS and EVIDENCEINFERENCE highlights non-comprehensive. Carton et al. [23] expect that FEVER highlights are non-comprehensive, in contrast to DeYoung et al. [29]. Contrary to the characterization of both of these work, we observe that the E-SNLI authors collect non-comprehensive highlights, since they instruct annotators to highlight only words in the hypothesis (and not the premise) for neutral pairs, and consider contradiction/neutral explanations correct if at least one piece of evidence in the input is highlighted. Based on these discrepancies in characterization, we first conclude that post-hoc assessment of comprehensiveness from a general description of data collection is error-prone.

Alternatively, Carton et al. [23] empirically show that available human highlights are not necessarily sufficient nor comprehensive for predictions of *highly accurate* models. This suggests that the same might hold for gold labels, leading us to ask: are gold highlights in existing datasets flawed?

Let us first consider insufficiency. Highlighted input elements taken together have to reasonably indicate the label. Otherwise, a highlight is an invalid explanation. Consider two datasets whose sufficiency Carton et al. [23] found to be most concerning: neutral E-SNLI pairs and no-attack WIKIATTACK examples. Neutral E-SNLI cases are not justifiable by highlighting because they are obtained only as an intermediate step to collecting free-text explanations, and only free-text explanations truly justify a neutral label [20]. Table 2 shows one E-SNLI highlight that is not sufficient. No-attack WIKIATTACK examples are not explainable by highlighting because the absence of offensive content justifies the no-attack label, and this absence cannot be highlighted. We recommend (i) avoiding human-annotated highlights with low sufficiency when evaluating and collecting highlights, and (ii) assessing whether the true label can be explained by highlighting.

Consider a highlight that is non-comprehensive because it is redundant with its complement in the input (e.g., a word appears multiple times, but only one occurrence is highlighted). Highlighting only one occurrence of "great" is a valid justification, but quantifying faithfulness of this highlight is hard because the model might rightfully use the unhighlighted occurrence of "great" to make the prediction. Thus, comprehensiveness is modeled to make faithfulness evaluation feasible. Non-comprehensiveness of human highlights, however, hinders evaluating plausibility of comprehensive model highlights since model and human highlights do not match by design. To be able to evaluate both plausibility and faithfulness, we should annotate comprehensive human highlights. We summarize these observations in Figure 2 in Appendix A.

Mutual influence of data and modeling assumptions also affects free-text explanations. For example, the E-SNLI guidelines have far more constraints than the COS-E guidelines, such as requiring self-contained explanations. Wiegreffe et al. [128] show that such data collection decisions can influence modeling assumptions. This is not an issue per se, but we should be cautious that EXNLP data collection decisions do not popularize explanation properties as *universally necessary* when they are not, e.g., that free-text explanations should be understandable without the original input or that highlights should be comprehensive. We believe this could be avoided with better documentation, e.g., with additions to a standard datasheet [39]. Explainability fact sheets have been proposed for models [114], but not for datasets. For example, an E-SNLI datasheet could note that self-contained explanations were required during data collection, but that this is not a necessary property of a valid free-text explanation. A dataset with comprehensive highlights should emphasize that comprehensiveness is required to simplify faithfulness evaluation.

Takeaways

1. It is important to precisely report how explanations were collected, e.g., by giving access to the annotation interface, screenshotting it, or giving the annotation instructions verbatim.
2. Sufficiency is necessary for highlights, and EXNLP researchers should avoid human-annotated highlights with low sufficiency for evaluating and developing highlights.
3. Comprehensiveness isn't necessary for a valid highlight, it is a means to quantify faithfulness.
4. Non-comprehensive human-annotated highlights cannot be used to automatically evaluate plausibility of highlights that are constrained to be comprehensive. In this case, EXNLP researchers should collect and use comprehensive human-annotated highlights.
5. Researchers should not make (error-prone) post-hoc estimates of comprehensiveness of human-annotated highlights from datasets' general descriptions.
6. EXNLP researchers should be careful to not popularize their data collection decisions as universally necessary. We advocate for documenting all constraints on collected explanations

in a datasheet, highlighting whether each constraint is necessary for explanation to be valid or not, and noting how each constraint might affect modeling and evaluation.

5 Rise of Structured Explanations

The merit of free-text explanations is their expressivity, which can come at the costs of underspecification and inconsistency due to the difficulty of quality control (stressed by the creators of two popular free-text explanation datasets: E-SNLI and COS-E). In this section, we highlight and challenge one prior approach to overcoming these difficulties: discarding template-like free-text explanations.

We gather crowdsourcing guidelines for the above-mentioned datasets in Tables 6–7 in Appendix and compare them. We observe two notable similarities between the guidelines for the above-mentioned datasets. First, both asked annotators to first highlight input words and then formulate a free-text explanation from them, to control quality. Second, template-like explanations are discarded because they are deemed uninformative. The E-SNLI authors assembled a list of 56 templates (e.g., "There is ⟨hypothesis⟩") to identify explanations whose edit distance to one of the templates is <10. They re-annotate the detected template-like explanations (11% in the entire dataset). The COS-E authors discard sentences "⟨answer⟩ is the only option that is correct/obvious" (the only given example of a template). Template explanations concern researchers because they can result in artifact-like behaviors in certain modeling architectures. For example, a model which predicts a task output from a generated explanation can produce explanations that are plausible to a human user and give the impression of making label predictions on the basis of this explanation. However, it is possible that the model learns to ignore the semantics of the explanation and instead makes predictions based on the explanation's template type [66, 55]. In this case, the semantic interpretation of the explanation (that of a human reader) is not faithful (an accurate representation of the model's decision process).

Despite re-annotating, Camburu et al. [21] report that E-SNLI explanations still largely follow 28 label-specific templates (e.g., an entailment template "X is another form of Y") even after re-annotation. Similarly, Brahman et al. [18] report that models trained on gold E-SNLI explanations generate template-like explanations for the defeasible NLI task. These findings lead us to ask: what are the differences between templates considered uninformative and filtered out, and those identified by Camburu et al. [21], Brahman et al. [18] that remain after filtering? Are *all* template-like explanations uninformative?

Although prior work indicates that template-like explanations are undesirable, most recently, structured explanations have been intentionally collected (see Table 5; §3). What these studies share is that they acknowledge structure as *inherent* to explaining the tasks they investigate. Related work [GLUCOSE; 85] takes the matter further, arguing that explanations should not be entirely free-form. Following GLUCOSE, we recommend running pilot studies to explore how people define and generate explanations for a task *before* collecting free-text explanations for it. If they reveal that informative human explanations are naturally structured, incorporating the structure in the annotation scheme is useful since the structure is natural to explaining the task. This turned out to be the case with NLI; Camburu et al. [21] report: "Explanations in E-SNLI largely follow a set of label-specific templates. This is a *natural consequence of the task* and dataset". We recommend embracing the structure when possible, but also encourage creators of datasets with template-like explanations to highlight in a dataset datasheet (§4) that template structure can influence downstream modeling decisions. There is no all-encompassing definition of explanation, and researchers could consult domain experts or follow literature from other fields to define an appropriate explanation in a task-specific manner, such as in GLUCOSE [85]. For conceptualization of explanations in different fields see Tiddi et al. [119].

Finally, what if pilot studies do not reveal any obvious structure to human explanations of a task? Then we need to do our best to control the quality of free-text explanations because low dataset quality is a bottleneck to building high-quality models. COS-E is collected with notably less annotation constraints and quality controls than E-SNLI, and has annotation issues that some have deemed make the dataset unusable [87]; see examples in Table 7 of Appendix A. As exemplars of quality control, we point the reader to the annotation guidelines of VCR [143] in Table 8 and GLUCOSE [84]. In §6 and §7, we give further task-agnostic recommendations for collecting high-quality ExNLP datasets, applicable to all three explanation types.

Takeaways

1. ExNLP researchers should study how people define and generate explanations for the task before collecting free-text explanations.
2. If pilot studies show that explanations are naturally structured, embrace the structure.
3. There is no all-encompassing definition of explanation. Thus, ExNLP researchers could consult domain experts or follow literature from other fields to define an appropriate explanation form, and these matters should be open for debate on a given task.

6 Increasing Explanation Quality

When asked to write free-text sentences from scratch for a table-to-text annotation task outside ExNLP, Parikh et al. [92] note that crowdworkers produce "vanilla targets that lack [linguistic] variety". Lack of variety can result in annotation artifacts, which are prevalent in the popular SNLI [16] and MNLI [129] datasets [97, 45, 120], among others [40]. These authors demonstrate the harms of such artifacts: models can overfit to them, leading to both performance over-estimation and problematic generalization behaviors.

Artifacts can occur from poor-quality annotations and inattentive annotators, both of which have been on the rise on crowdsourcing platforms [25, 7, 87]. To mitigate artifacts, both increased **diversity of annotators** and **quality control** are needed. We focus on quality control here and diversity in §7.

6.1 A Two-Stage Collect-And-Edit Approach

While ad-hoc methods can improve quality [20, 143, 84], an effective and generalizable method is to collect annotations in two stages. A two-stage methodology has been applied by a small minority of ExNLP dataset papers [58, 144, 143], who first compile explanation candidates automatically or from crowdworkers, and secondly perform quality-control by having other crowdworkers assess the quality of the collected explanations (we term this Collect-And-Judge). Judging improves the overall quality of the final dataset by removing low-quality instances, and additionally allows authors to release quality ratings for each instance.

Outside ExNLP, Parikh et al. [92] use an extended version of this approach (that we term Collect-And-Edit): they generate a noisy automatically-extracted dataset for the table-to-text generation task, and then ask annotators to edit the datapoints. Bowman et al. [17] use this approach to re-collect NLI hypotheses, and find, crucially, that having annotators edit rather than create hypotheses reduces artifacts in a subset of MNLI. In XAI, Kutlu et al. [67] collect highlight explanations for Web page ranking with annotator editing. We advocate expanding the Collect-And-Judge approach for explanation collection to Collect-And-Edit. This has potential to increase linguistic diversity via multiple annotators per-instance, reduce individual annotator biases, and perform quality control. Through a case study of two multimodal free-text explanation datasets, we will demonstrate that collecting explanations automatically without human editing (or at least judging) can lead to artifacts.

E-SNLI-VE [32] and VQA-E [75] are two visual-textual datasets for entailment and question-answering, respectively. E-SNLI-VE combines annotations of two datasets: (i) SNLI-VE [131], collected by replacing the textual premises of SNLI [16] with Flickr30k images [140], and (ii) E-SNLI [20], a dataset of crowdsourced explanations for SNLI. This procedure is possible because every SNLI premise was originally the caption of a Flickr30k photo. However, since SNLI's hypotheses were collected from crowdworkers who did not see the original images, the photo replacement process results in a significant number of errors [122]. Do et al. [32] re-annotate labels and explanations for the neutral pairs in the validation and test sets of SNLI-VE. However, it has been argued that the dataset remains low-quality for training models due to artifacts in the entailment and the neutral class' training sets [78]. With a full EDIT approach, we expect that these artifacts would be significantly reduced, and the resulting dataset could have quality on-par with E-SNLI. Similarly, the VQA-E dataset [75] converts image captions from the VQA v2.0 dataset [43] into explanations, but a notably lower plausibility compared to a carefully-crowdsourced VCR explanations is reported in [78].

Both E-SNLI-VE and VQA-E present novel and cost-effective ways to produce large ExNLP datasets for new tasks, but also show the quality tradeoffs of automatic collection. Strategies such as crowdsourced judging and editing, even on a small subset, can reveal and mitigate such issues.

6.2 Teach and Test the Underlying Task

In order to both create and judge explanations, annotators must understand the underlying task and label-set well. In most cases, this necessitates teaching and testing the task. Prior work outside of ExNLP has noted the difficulty of scaling annotation to crowdworkers for complex linguistic tasks [106, 35, 99, 85]. To increase annotation quality, these works provide intensive training to crowdworkers, including personal feedback. Since label understanding is a prerequisite for explanation collection, task designers should consider relatively inexpensive strategies such as qualification tasks and checker questions. This need is correlated with the difficulty and domain-specificity of the task, as elaborated above.

Similarly, people cannot explain all tasks equally well and even after intensive training they might struggle to explain tasks such as deception detection and recidivism prediction [89]. Human explanations for such tasks might be limited in serving the three goals outlined in §1.

6.3 Addressing Ambiguity

Data collectors often collect explanations post-hoc, i.e., annotators are asked to explain labels assigned by a system or other annotators. The underlying assumption is that the explainer believes the assigned label to be correct or at least likely (there is no task ambiguity). However, this assumption has been shown to be inaccurate (among others) for relation extraction [8], natural language inference [96, 88], and complement coercion [35], and the extent to which it is true likely varies by task, instance, and annotator. If an annotator is uncertain about a label, their explanation may be at best a hypothesis and at worst a guess. HCI research encourages leaving room for ambiguity rather than forcing raters into binary decisions, which can result in poor or inaccurate labels [108].

To ensure explanations reflect human decisions as closely as possible, it is ideal to collect both labels and explanations from the same annotators. Given that this is not always possible, including a checker question to assess whether an explanation annotator agrees with a label is a good alternative.

Takeaways

1. Using a Collect-And-Edit method can reduce individual annotator biases, perform quality control, and potentially reduce dataset artifacts.
2. Teaching and testing the underlying task and addressing ambiguity can improve data quality.

7 Increasing Explanation Diversity

Beyond quality control, increasing annotation diversity is another task-agnostic means to mitigate artifacts and collect more representative data. We elaborate on suggestions from related work (inside and outside ExNLP) here.

7.1 Use a Large Set of Annotators

Collecting representative data entails ensuring that a handful of annotators do not dominate data collection. Outside ExNLP, Geva et al. [40] report that recruiting only a small pool of annotators (1 annotator per 100–1000 examples) allows models to overfit on annotator characteristics. Such small annotator pools exist in ExNLP—for instance, E-SNLI reports an average of 860 explanations written per worker. The occurrence of the incorrect explanation "rivers flow trough valleys" for 529 different instances in COS-E v1.11 is likely attributed to a single annotator. Al Kuwatly et al. [3] find that demographic attributes can predict annotation differences. Similarly, Davidson et al. [28], Sap et al. [110] show that annotators often consider African-American English writing to be disproportionately offensive.[2] A lack of annotator representation concerns ExNLP for three reasons: explanations depend on socio-cultural background [63], annotator traits should not be predictable [40], and the subjectivity of explaining leaves room for social bias to emerge.

On most platforms, annotators are not restricted to a specific number of instances. Verifying that no worker has annotated an excessively large portion of the dataset in addition to strategies from Geva

[2] In another related study, 82% of annotators reported their race as white [111]. This is a likely explanation for the disproportionate annotation.

et al. [40] can help mitigate annotator bias. More elaborate methods for increasing annotator diversity include collecting demographic attributes or modeling annotators as a graph [3, 126].

7.2 Multiple Annotations Per Instance

HCI research has long considered the ideal of crowdsourcing a single ground-truth as a "myth" that fails to account for the diversity of human thought and experience [9]. Similarly, ExNLP researchers should not assume there is always one correct explanation. Many of the assessments crowdworkers are asked to make when writing explanations are subjective in nature, and there are many different models of explanation based on a user's cognitive biases, social expectations, and socio-cultural background [82]. Prasad et al. [98] present a theoretical argument to illustrate that there are multiple ways to highlight input words to explain an annotated sentiment label. Camburu et al. [20] find a low inter-annotator BLEU score [91] between free-text explanations collected for E-SNLI test instances.

If a dataset contains only one explanation when multiple are plausible, a plausible model explanation can be penalized unfairly for not agreeing with it. We expect that modeling multiple explanations can also be a useful learning signal. Some existing datasets contain multiple explanations per instance (last column of Tables 3–5). Future ExNLP data collections should do the same if there is subjectivity in the task or diversity of correct explanations (which can be measured via inter-annotator agreement). If annotators exhibit low agreement between explanations deemed as plausible, this can reveal a diversity of correct explanations for the task, which should be considered in modeling and evaluation.

7.3 Get Ahead: Add Contrastive and Negative Explanations

The machine learning community has championed modeling *contrastive explanations* that justify why a prediction was made *instead of* another, to align more closely with human explanation [31, 49, 82]. Most recently, methods have been proposed in NLP to produce contrastive edits of the input as explanations [107, 134, 130, 55]. Outside of ExNLP, datasets with contrastive edits have been collected to assess and improve robustness of NLP models [59, 38, 74] and might be used for explainability too.

Just as highlights are not sufficiently intelligible for complex tasks, the same might hold for contrastive input edits. To the best of our knowledge, there is no dataset that contains contrastive free-text or structured explanations. These could take the form of (i) collecting explanations that answer the question "why...instead of...", or (ii) collecting explanations for other labels besides the gold label, to be used as an additional training signal. A related annotation paradigm is to collect *negative explanations*, i.e., explanations that are invalid for an (input, gold label) pair. Such examples can improve ExNLP models by providing supervision of what is *not* a correct explanation [112]. A human JUDGE or EDIT phase automatically gives negative explanations: the low-scoring instances (former) or instances pre-editing (latter) [58, 144].

Takeaways

1. To increase annotation diversity, a large set of annotators, multiple annotations per instance, and collecting explanations that are most useful to the needs of end-users are important.
2. Reporting inter-annotator agreement with plausibility of annotated explanations is useful to known whether there is a natural diversity of explanations for the task and should the diversity be considered for modeling and evaluation.

8 Conclusions

We have presented a review of existing datasets for ExNLP research, highlighted discrepancies in data collection that can have downstream modeling effects, and synthesized the literature both inside and outside ExNLP into a set of recommendations for future data collection.

We note that a majority of the work reviewed in this paper has originated in the last 1-2 years, indicating an explosion of interest in collecting datasets for ExNLP. We provide reflections for current and future data collectors in an effort to promote standardization and consistency. This paper also serves as a starting resource for newcomers to ExNLP, and, we hope, a starting point for further discussions.

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018. doi: 10.1109/ACCESS.2018.2870052. URL https://ieeexplore.ieee.org/document/8466590.

[2] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.238. URL https://aclanthology.org/2021.acl-long.238.

[3] Hala Al Kuwatly, Maximilian Wich, and Georg Groh. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.21. URL https://aclanthology.org/2020.alw-1.21.

[4] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5513. URL https://aclanthology.org/W18-5513.

[5] Tariq Alhindi, Smaranda Muresan, and Daniel Preotiuc-Pietro. Fact vs. opinion: the role of argumentation features in news classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6139–6149, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.540. URL https://aclanthology.org/2020.coling-main.540.

[6] David Alvarez-Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://arxiv.org/abs/1806.07538.

[7] Antonio Alonso Arechar and David Rand. Turking in the time of covid. PsyArXiv, 2020. URL https://psyarxiv.com/vktqu.

[8] Lora Aroyo and Chris Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013), 2013.

[9] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015. URL https://ojs.aaai.org//index.php/aimagazine/article/view/2564.

[10] David Atkinson, Kumar Bhargav Srinivasan, and Chenhao Tan. What gets echoed? understanding the "pointers" in explanations of persuasive arguments. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2911–2921, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1289. URL https://aclanthology.org/D19-1289.

[11] Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1216. URL https://aclanthology.org/D18-1216.

[12] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*,

58:82–115, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2019.12.012. URL https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[13] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.

[14] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1908.05739.

[15] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI Workshop on Explainable AI (XAI)*, pages 8–13, 2017. URL http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf.

[16] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

[17] Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. New protocols and negative results for textual entailment data collection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8203–8214, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.658. URL https://aclanthology.org/2020.emnlp-main.658.

[18] Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *the AAAI Conference on Artificial Intelligence*, 2021. URL https://arxiv.org/abs/2012.08012.

[19] Nadia Burkart and Marco F. Huber. A survey on the explainability of supervised machine learning. *The Journal of Artificial Intelligence Research (JAIR)*, 70, 2021. doi: https://www.jair.org/index.php/jair/article/view/12228.

[20] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://papers.nips.cc/paper/2018/hash/4c7a167bb329bd92580a99ce422d6fa6-Abstract.html.

[21] Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.382. URL https://aclanthology.org/2020.acl-main.382.

[22] Samuel Carton, Qiaozhu Mei, and Paul Resnick. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3497–3507, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1386. URL https://aclanthology.org/D18-1386.

[23] Samuel Carton, Anirudh Rathore, and Chenhao Tan. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.747. URL https://aclanthology.org/2020.emnlp-main.747.

[24] Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.22. URL https://aclanthology.org/2021.naacl-main.22.

[25] Michael Chmielewski and Sarah C Kucker. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473, 2020. URL https://journals.sagepub.com/doi/abs/10.1177/1948550619875149.

[26] Miruna-Adriana Clinciu and Helen Hastie. A survey of explainable AI terminology. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 8–13. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-8403. URL https://aclanthology.org/W19-8403.

[27] Jeff Da, Maxwell Forbes, Rowan Zellers, Anthony Zheng, Jena D. Hwang, Antoine Bosselut, and Yejin Choi. Edited media understanding frames: Reasoning about the intent and implications of visual misinformation. 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 2026–2039, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.158. URL https://aclanthology.org/2021.acl-long.158.

[28] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3504. URL https://aclanthology.org/W19-3504.

[29] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

[30] Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.bionlp-1.13. URL https://aclanthology.org/2020.bionlp-1.13.

[31] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 590–601, 2018. URL https://proceedings.neurips.cc/paper/2018/file/c5ff2543b53f4cc0ad3819a36752467b-Paper.pdf.

[32] Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. e-SNLI-VE-2.0: Corrected Visual-Textual Entailment with Natural Language Explanations. In *IEEE CVPR Workshop on Fair, Data Efficient and Trusted Computer Vision*, 2020. URL https://arxiv.org/abs/2004.03744.

[33] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608, 2017. URL https://arxiv.org/abs/1702.08608.

[34] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. In *Proceedings of the Conference of Intelligent User Interfaces (ACM IUI)*, 2019. URL https://arxiv.org/abs/1901.03729.

[35] Yanai Elazar, Victoria Basmov, Shauli Ravfogel, Yoav Goldberg, and Reut Tsarfaty. The extraordinary failure of complement coercion crowdsourcing. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 106–116, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.insights-1.17. URL https://aclanthology.org/2020.insights-1.17.

[36] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL https://aclanthology.org/P19-1346.

[37] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.48. URL https://aclanthology.org/2020.emnlp-main.48.

[38] Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117.

[39] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2018. URL https://www.fatml.org/media/documents/datasheets_for_datasets.pdf.

[40] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1107. URL https://aclanthology.org/D19-1107.

[41] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 2021. URL https://arxiv.org/pdf/2101.02235.pdf.

[42] Leilani H. Gilpin, David Bau, B. Yuan, A. Bajwa, M. Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018. URL https://arxiv.org/pdf/1806.00069.pdf.

[43] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and D. Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, 2017. URL https://arxiv.org/abs/1612.00837.

[44] Riccardo Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51:1 – 42, 2019. URL https://dl.acm.org/doi/pdf/10.1145/3236009.

[45] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017.

[46] Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1175. URL https://aclanthology.org/P18-1175.

[47] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1046. URL https://aclanthology.org/K19-1046.

[48] Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. Does bert learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. URL https://arxiv.org/abs/2109.02738.

[49] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating counterfactual explanations with natural language. In *International Conference on Machine Learning (ICML)*, 2018.

[50] Denis J Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65, 1990. URL https://www.researchgate.net/profile/Denis_Hilton/publication/232543382_Conversational_processes_and_causal_explanation/links/00b7d519bd8fa613f1000000/Conversational-processes-and-causal-explanation.pdf.

[51] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608, 2018. URL https://arxiv.org/abs/1812.04608.

[52] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1243. URL https://aclanthology.org/D19-1243.

[53] Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.602. URL https://aclanthology.org/2020.acl-main.602.

[54] Alon Jacovi and Yoav Goldberg. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386.

[55] Alon Jacovi and Yoav Goldberg. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 2021. URL https://arxiv.org/abs/2006.01067.

[56] Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1278.

[57] Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1433.

[58] Harsh Jhamtani and Peter Clark. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 137–150, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.10. URL https://aclanthology.org/2020.emnlp-main.10.

[59] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/pdf?id=Sklgs0NFvr.

[60] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1023. URL https://aclanthology.org/N18-1023.

[61] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. QASC: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. URL https://arxiv.org/pdf/1910.11473.pdf.

[62] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John F. Canny, and Zeynep Akata. Textual Explanations for Self-Driving Vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. URL https://arxiv.org/abs/1807.11546.

[63] Hana Kopecká and Jose M Such. Explainable AI for Cultural Minds. In *Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction (DEXA HAI) at the 24th European Conference on Artificial Intelligence (ECAI)*, 2020. URL https://kclpure.kcl.ac.uk/portal/files/134728815/DEXA_aug_crc.pdf.

[64] Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.623. URL https://aclanthology.org/2020.emnlp-main.623.

[65] Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.474. URL https://www.aclweb.org/anthology/2020.coling-main.474.

[66] Sawan Kumar and Partha Talukdar. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.771. URL https://aclanthology.org/2020.acl-main.771.

[67] Mucahid Kutlu, Tyler McDonnell, Matthew Lease, and Tamer Elsayed. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research*, 69:143–189, 2020. URL https://www.ischool.utexas.edu/~ml/papers/kutlu_jair20.pdf.

[68] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc

Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, March 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

[69] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082.

[70] Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. QED: A Framework and Dataset for Explanations in Question Answering. *Transactions of the Association for Computational Linguistics*, 9:790–806, 08 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00398. URL https://doi.org/10.1162/tacl_a_00398.

[71] Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. Inferring which medical treatments work from reports of clinical trials. 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 3705–3717, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1371. URL https://aclanthology.org/N19-1371.

[72] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8769–8784, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.706. URL https://aclanthology.org/2020.emnlp-main.706.

[73] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1011. URL https://aclanthology.org/D16-1011.

[74] Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. Linguistically-informed transformations (LIT): A method for automatically generating contrast sets. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.12. URL https://aclanthology.org/2020.blackboxnlp-1.12.

[75] Qing Li, Qingyi Tao, Shafiq R. Joty, Jianfei Cai, and Jiebo Luo. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. URL https://arxiv.org/abs/1803.07464.

[76] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015.

[77] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018. URL https://dl.acm.org/doi/pdf/10.1145/3236386.3241340.

[78] Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2810–2829, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.253. URL https://aclanthology.org/2020.findings-emnlp.253.

[79] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*, 2021. URL https://arxiv.org/abs/2012.10289.

[80] Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, 2012. URL https://ieeexplore.ieee.org/document/6413815.

[81] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260.

[82] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019. URL https://arxiv.org/pdf/1706.07269.pdf.

[83] Christoph Molnar. Interpretable machine learning: A guide for making black box models explainable, 2019. https://christophm.github.io/interpretable-ml-book/.

[84] Lori Moon, Lauren Berkowitz, Jennifer Chu-Carroll, and Nasrin Mostafazadeh. Details of data collection and crowd management for glucose (generalized and contextualized story explanations). *Github*, 2020. URL https://github.com/ElementalCognition/glucose/blob/master/data_collection_quality.pdf.

[85] Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.370. URL https://aclanthology.org/2020.emnlp-main.370.

[86] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1900654116. URL https://www.pnas.org/content/116/44/22071.

[87] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. WT5?! Training Text-to-Text Models to Explain their Predictions. arXiv:2004.14546, 2020. URL https://arxiv.org/abs/2004.14546.

[88] Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.734. URL https://aclanthology.org/2020.emnlp-main.734.

[89] R. Nisbett and T. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84:231–259, 1977.

[90] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219855. URL https://aclanthology.org/P05-1015.

[91] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

[92] Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.89. URL https://aclanthology.org/2020.emnlp-main.89.

[93] Praveen Paritosh. Achieving data excellence. In *NeurIPS 2020 Crowd Science Workshop*, 2020. URL https://neurips.cc/virtual/2020/public/workshop_16111.html. Invited talk.

[94] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/papers/Park_Multimodal_Explanations_Justifying_CVPR_2018_paper.pdf.

[95] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily L. Denton, and A. Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *The ML-Retrospectives, Surveys & Meta-Analyses NeurIPS 2020 Workshop*, 2020. URL https://arxiv.org/abs/2012.05345.

[96] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, March 2019. doi: 10.1162/tacl_a_00293. URL https://aclanthology.org/Q19-1043.

[97] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023.

[98] Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. To what extent do human explanations of model behavior align with actual model behavior? arXiv:2012.13354, 2020. URL https://arxiv.org/abs/2012.13354.

[99] Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.224. URL https://aclanthology.org/2020.emnlp-main.224.

[100] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1487. URL https://aclanthology.org/P19-1487.

[101] Nazneen Fatema Rajani, Rui Zhang, Yi Chern Tan, Stephan Zheng, Jeremy Weiss, Aadit Vyas, Abhijit Gupta, Caiming Xiong, Richard Socher, and Dragomir Radev. ESPRIT: Explaining solutions to physical reasoning tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7906–7917, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.706. URL https://aclanthology.org/2020.acl-main.706.

[102] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

[103] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. *Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges*, pages 19–36. Springer International Publishing, Cham, 2018. ISBN 978-3-319-98131-4. doi: 10.1007/978-3-319-98131-4_2. URL https://doi.org/10.1007/978-3-319-98131-4_2.

[104] Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7: 249–266, March 2019. doi: 10.1162/tacl_a_00266. URL https://aclanthology.org/Q19-1016.

[105] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1020.

[106] Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.626. URL https://aclanthology.org/2020.acl-main.626.

[107] Alexis Ross, Ana Marasović, and Matthew Peters. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.336. URL https://aclanthology.org/2021.findings-acl.336.

[108] Nithya Sambasivan. Human-data interaction in ai. In *PAIR Symposium*, 2020. URL https://www.youtube.com/watch?v=cjRF5a4eo2Y&t=83s. Invited talk.

[109] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. " everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. URL https://storage.googleapis.com/pub-tools-public-publication-data/pdf/0d556e45afc54afeb2eb6b51a9bc1827b9961ff4.pdf.

[110] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

[111] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.486. URL https://aclanthology.org/2020.acl-main.486.

[112] Hendrik Schuff, Heike Adel, and Ngoc Thang Vu. F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7076–7095, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.575. URL https://aclanthology.org/2020.emnlp-main.575.

[113] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

[114] Kacper Sokol and Peter Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67, 2020. URL https://dl.acm.org/doi/pdf/10.1145/3351095.3372870.

[115] Shashank Srivastava, Igor Labutov, and Tom Mitchell. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1527–1536, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1161. URL https://aclanthology.org/D17-1161.

[116] Julia Strout, Ye Zhang, and Raymond Mooney. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4807. URL https://aclanthology.org/W19-4807.

[117] Mokanarangan Thayaparan, Marco Valentino, and André Freitas. A survey on explainability in machine reading comprehension. arXiv:2010.00389, 2020. URL https://arxiv.org/abs/2010.00389.

[118] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074.

[119] Ilaria Tiddi, M. d'Aquin, and E. Motta. An ontology design pattern to define explanations. In *Proceedings of the 8th International Conference on Knowledge Capture*, 2015. URL http://oro.open.ac.uk/44321/.

[120] Masatoshi Tsuchiya. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1239.

[121] Sahil Verma, John P. Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. arXiv:2010.10596, 2020. URL https://arxiv.org/abs/2010.10596.

[122] Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. Grounded textual entailment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2354–2368, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1199.

[123] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609.

[124] Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1393. URL https://aclanthology.org/P19-1393.

[125] Ziqi Wang, Yujia Qin, Wenxuan Zhou, Jun Yan, Qinyuan Ye, Leonardo Neves, Zhiyuan Liu, and Xiang Ren. Learning from explanations with neural execution tree. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/pdf/1911.01352.pdf.

[126] Maximilian Wich, Hala Al Kuwatly, and Georg Groh. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.22. URL https://aclanthology.org/2020.alw-1.22.

[127] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002.

[128] Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. Measuring association between labels and free-text rationales. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. URL https://arxiv.org/abs/2010.12762.

[129] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

[130] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving 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 6707–6723, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.523. URL https://aclanthology.org/2021.acl-long.523.

[131] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. arXiv:1901.06706, 2019. URL https://arxiv.org/abs/1901.06706.

[132] Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5456–5473, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.671.

[133] Fan Yang, Mengnan Du, and Xia Hu. Evaluating explanation without ground truth in interpretable machine learning. arXiv:1907.06831, 2019. URL https://arxiv.org/abs/1907.06831.

[134] Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6150–6160, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://www.aclweb.org/anthology/2020.coling-main.541.

[135] Shaohua Yang, Qiaozi Gao, Sari Sadiya, and Joyce Chai. Commonsense justification for action explanation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2627–2637, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1283. URL https://aclanthology.org/D18-1283.

[136] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1237. URL https://aclanthology.org/D15-1237.

[137] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

[138] Qinyuan Ye, Xiao Huang, Elizabeth Boschee, and Xiang Ren. Teaching machine comprehension with compositional explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1599–1615, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.145. URL https://aclanthology.org/2020.findings-emnlp.145.

[139] Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://arxiv.org/abs/2105.06977.

[140] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006.

[141] Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1420. URL https://aclanthology.org/D19-1420.

[142] Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York, April 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-1033.

[143] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://ieeexplore.ieee.org/document/8953217.

[144] Hongming Zhang, Xinran Zhao, and Yangqiu Song. WinoWhy: A deep diagnosis of essential commonsense knowledge for answering Winograd schema challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.508. URL https://aclanthology.org/2020.acl-main.508.

[145] Ye Zhang, Iain Marshall, and Byron C. Wallace. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1076. URL https://aclanthology.org/D16-1076.

[146] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1004. URL https://aclanthology.org/D17-1004.

A Complementing Information

We provide the following additional illustrations and information that complement discussions in the main paper:

- Details of dataset licenses in Appendix B.
- Details of dataset collection in Appendix C.
- An illustration of connections between assumptions made in the development of self-explanatory highlighting models (discussed in §4) is shown in Figure 2.
- Overviews of quality measures and outcomes in E-SNLI, COS-E, and VCR in Tables 6–8.
- A discussion of explanation and commonsense reasoning in Appendix D.

B Dataset Licenses

The authors of 33.96% papers cited in Tables 3–5 do **not** report the dataset license in the paper or a repository; 45.61% use *common* permissive licenses such as Apache 2.0, MIT, CC BY-SA 4.0, CC BY-SA 3.0, BSD 3-Clause "New" or "Revised" License, BSD 2-Clause "Simplified" License, CC BY-NC 2.0, CC BY-NC-SA, GFDL, and CC0 1.0 Universal. We overview the rest:

- WIKIQA: "Microsoft Research Data License Agreement for Microsoft Research WikiQA Corpus"
- MULTIRC: "Research and Academic Use License"
- Hanselowski et al. [47]: A data archive is under **Copyright**.
- COQA: "Children's stories are collected from MCTest [105] which comes with MSR-LA license. Middle/High school exam passages are collected from RACE [69] which comes with its own license." The rest of the dataset is under permissive licenses: BY-SA 4.0 and Apache 2.0.
- Wang et al. [125]: The part of the dataset that is built on on TACRED [146] cannot be distributed (under "LDC User Agreement for Non-Members") and the license for the rest of dataset is not specified.
- BDD-X: "UC Berkeley's Standard Copyright and Disclaimer Notice"
- VCR: "Dataset License Agreement"
- VLEP: "VLEP Dataset Download Agreement"
- WORLDTREE V1: "End User License Agreement"
- WORLDTREE V2: "End User License Agreement"
- ECQA: "Community Data License Agreement - Sharing - Version 1.0"

C Dataset Collection

To collect the datasets, we used our domain expertise, having previously published work using highlights and free-text explanations, to construct a seed list of datasets. In the year prior to submission, we augmented this list as we encountered new publications and preprints. We then searched the ACL Anthology (`https://aclanthology.org`) for the terms "explain", "interpret", "explanation", and "rationale", focusing particularly on proceedings from 2020 and onward, as the subfield has grown in popularity significantly in this timeframe. We additionally first made live the website open to public contributions 3.5 months prior to submission, and integrated all dataset suggestions we received into the tables.

D Explanation and Commonsense Reasoning

The scope of our survey focuses on textual explanations that explain *human decisions* (defined in the survey as task labels). There has recently emerged a set of datasets at the intersection of commonsense

reasoning and explanation (such as GLUCOSE [85]). We class these datasets as explaining *observed events or phenomena* in the world, where the distinction between class label and explanation is not defined. For an illustration of the difference between these datasets and those surveyed in the main paper, see Figure 1.

Unlike the datasets surveyed in the paper, datasets that explain *observed events or phenomena* in the world (often in the form of commonsense inferences) do not fit the three main goals of ExNLP because they do not lend themselves to task-based explanation modeling. These datasets generally do not use the term "explanation" [52, 36, 37, *inter alia*], with two exceptions: ART [14] and GLUCOSE [85]. They produce tuples of the form (input, label), where the input is an event or observation and the label can possibly be seen as an explanation, rather than (input, label, explanation).

Some datasets surveyed in the paper fit both categories. For instance, SBIC [110] contains both human-annotated "offensiveness" labels and justifications of why social media posts might be considered offensive (middle of Fig. 1). Other examples include predicting future events in videos [VLEP; 72] and answering commonsense questions about images [VCR; 143]. Both collect observations about a real-world setting as task labels as well as explanations. We include them in our survey.

A side-note on the scope. We discuss some necessary properties of human-authored explanations (e.g., sufficiency in §4) and conditions under which they are necessary (e.g., comprehensiveness if we wish to evaluate plausibility of model highlights that are constrained to be comprehensive; §4), as well as properties that are previously typically considered as unwanted but we illustrate they are not necessarily inappropriate (e.g., template-like explanations in §5). However, there might be other relevant properties of human-annotated explanations that we did not discuss since we focus on discussing topics most relevant to the latest ExNLP and NLP research such as sufficiency, comprehensivness, plausibility, faithfulness, template-like explanations, and data artifacts. Moreover, as we highlight in §5, there is no all-encompassing definition of explanation and thus there we do not expect that there is universal criteria for an appropriate explanation.

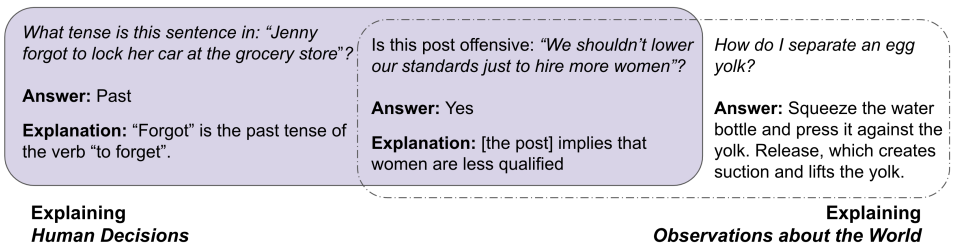


Figure 1: Two classes of EXNLP datasets (§D). The shaded area is our scope.

EXPLAINING NATURAL LANGUAGE INFERENCE (E-SNLI;Camburu et al. 20)
General Constraints for Quality Control Guided annotation procedure: • Step 1: Annotators had to highlight words from the premise/hypothesis that are essential for the given relation. • Step 2: Annotators had to formulate a free-text explanation using the highlighted words. • To avoid ungrammatical sentences, only half of the highlighted words had to be used with the same spelling. • The authors checked that the annotators also used non-highlighted words; correct explanations needs articulate a link between the keywords. • Annotators had to give self-contained explanations: sentences that make sense without the premise/hypothesis. • Annotators had to focus on the premise parts that are *not* repeated in the hypothesis (non-obvious elements). • In-browser check that each explanation contains at least three tokens. • In-browser check that an explanation is not a copy of the premise or hypothesis.
Label-Specific Constraints for Quality Control • For entailment, justifications of all the parts of the hypothesis that do not appear in the premise were required. • For neutral and contradictory pairs, while annotators were encouraged to state all the elements that contribute to the relation, an explanation was considered correct if at least one element is stated. • For entailment pairs, annotators had to highlight at least one word in the premise. • For contradiction pairs, annotators had to highlight at least one word in both the premise and the hypothesis. • For neutral pairs, annotators were allowed to highlight only words in the hypothesis, to strongly emphasize the asymmetry in this relation and to prevent workers from confusing the premise with the hypothesis.
Quality Analysis and Refinement • The authors graded correctness of 1000 random examples between 0 (incorrect) and 1 (correct), giving partial scores of k/n if only k out of n required arguments were mentioned. • An explanation was rated as incorrect if it was template-like. The authors assembled a list of 56 templates that they used for identifying explanations (in the entire dataset) whose edit distance to one of the templates was <10. They re-annotated the detected template-like explanations (11% in total).
Post-Hoc Observations • Total error rate of 9.62%: 19.55% on entailment, 7.26% on neutral, and 9.38% on contradiction. • In the large majority of the cases, that authors report it is easy to infer label from an explanation. • Camburu et al. [21]: "Explanations in e-SNLI largely follow a set of label-specific templates. This is a natural consequence of the task and the SNLI dataset and not a requirement in the collection of the e-SNLI. [...] For each label, we created a list of the most used templates that we manually identified among e-SNLI." They collected 28 such templates.

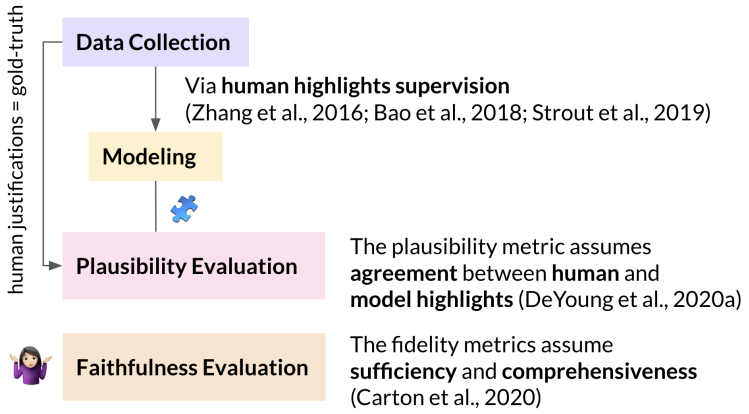
Table 6: Overview of quality control measures and outcomes in E-SNLI.

EXPLAINING COMMONSENSE QA (COS-E; Rajani et al. 100)
General Constraints for Quality Control Guided annotation procedure: • Step 1: Annotators had to highlight relevant words in the question that justifies the correct answer. • Step 2: Annotators had to provide a brief open-ended explanation based on the highlighted justification that could serve as the commonsense reasoning behind the question. • In-browser check that annotators highlighted at least one relevant word in the question. • In-browser check that an explanation contains at least four words. • In-browser check that an explanation is not a substring of the question or the answer choices without any other extra words.
Label-Specific Constraints for Quality Control (none)
Quality Analysis and Refinement • The authors did unspecified post-collection checks to catch examples that are not caught by their previous filters. • The authors removed template-like explanations, i.e., sentences "⟨answer⟩ is the only option that is correct obvious" (the only provided example of a template).
Post-Hoc Observations • 58% explanations (v1.0) contain the ground truth answer. • The authors report that many explanations remain noisy after quality-control checks, but that they find them to be of sufficient quality for the purposes of their work. • Narang et al. [87] on v1.11: "Many of the ground-truth explanations for CoS-E are low quality and/or nonsensical (e.g., the question "Little sarah didn't think that anyone should be kissing boys. She thought that boys had what?" with answer "cooties" was annotated with the explanation "american horror comedy film directed"; or the question "What do you fill with ink to print?" with answer "printer" was annotated with the explanation "health complications", etc.)" • Further errors exist (v1.11): The answer "rivers flow trough valleys" appears 529 times, and "health complications" 134 times, signifying copy-paste behavior by some annotators. Uninformative answers such as "this word is the most relevant" (and variants) appear 522 times.

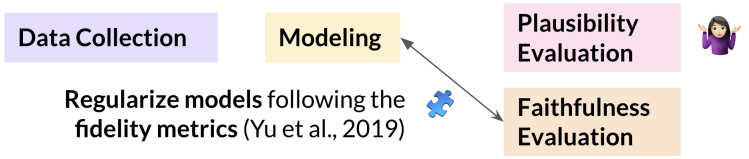
Table 7: Overview of quality control measures and outcomes in COS-E.

EXPLAINING VISUAL COMMONSENSE REASONING (VCR; Zellers et al. 143)
General Constraints for Quality Control • The authors automatically rate instance "interestingness" and collect annotations for the most "interesting" instances. Multi-stage annotation procedure: • Step 1: Annotators had to write 1-3 questions based on a provided image (at least 4 words each). • Step 2: Annotators had to answer each question (at least 3 words each). • Step 3: Annotators had to provide a rationale for each answer (at least 5 words each). • Annotators had to pass a qualifying exam where they answered some multiple-choice questions and wrote a question, answer, and rationale for a single image. The written responses were verified by the authors. • Authors provided annotators with high-quality question, answer, and rationale examples. • In-browser check that annotators explicitly referred to at least one object detected in the image, on average, in the question, answer, or rationale. • Other in-browser checks related to the question and answer quality. • Every 48 hours, the lead author reviewed work and provided aggregate feedback to make sure the annotators were proving good-quality responses and "structuring rationales in the right way". It is unclear, but assumed, that poor annotators were dropped during these checks.
Label-Specific Constraints for Quality Control (none)
Quality Analysis and Refinement • The authors used a second phase to further refine some HITs. A small group of workers who had done well on the main task were selected to rate a subset of HITs (about 1 in 50), and this process was used to remove annotators with low ratings from the main task.
Post-Hoc Observations • The authors report that humans achieve over 90% accuracy on the multiple-choice rationalization task derived from the dataset. They also report high agreement between the 5 annotators for each instance. These can be indicative of high dataset quality and low noise. • The authors report high diversity—almost every rationale is unique, and the instances cover a range of commonsense categories. • The rationales are long, averaging 16 words in length, another sign of quality. • External validation of quality: Marasović et al. [78] find that the dataset's explanations are highly plausible with respect to both the image and associated question/answer pairs; they also rarely describe events or objects not present in the image.

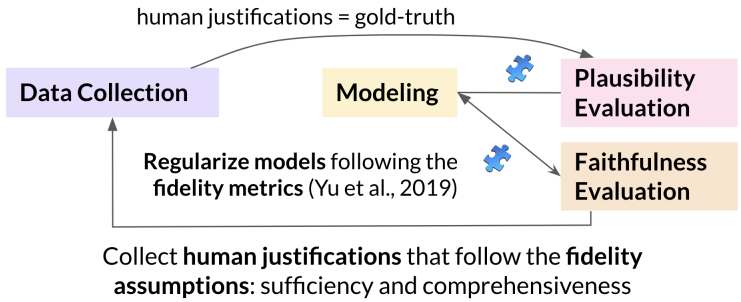
Table 8: Overview of quality control measures and outcomes for (the rationale-collection portion) of VCR. The dataset instances (questions and answers) and their rationales were collected simultaneously; we do not include quality controls placed specifically on the question or answer.



(a) Supervised models' development. When we use human highlights as supervision, we assume that they are the gold-truth and that model highlights should match. Thus, comparing human and model highlights for plausibility evaluation is sound. However, with this basic approach we do not introduce any data or modeling properties that help faithfulness evaluation, and that remains a challenge in this setting.



(b) Unsupervised models' development. In §4, we illustrate that comprehensiveness is not a necessary property of human highlights. Non-comprehensiveness, however, hinders evaluating plausibility of model highlights produced in this setting since model and human highlights do not match by design.



(c) Recommended unsupervised models' development. To evaluate both plausibility and faithfulness, we should collect comprehensive human highlights, assuming that they are already sufficient (a necessary property).

Figure 2: Connections between assumptions made in the development of self-explanatory **highlighting** models. The jigsaw icon marks a synergy of modeling and evaluation assumptions. The arrow notes the direction of influence. The text next to the plausibility / faithfulness boxes in the top figure hold for the other figures, but are omitted due to space limits. Cited: DeYoung et al. [29], Zhang et al. [145], Bao et al. [11], Strout et al. [116], Carton et al. [23], Yu et al. [141].