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.

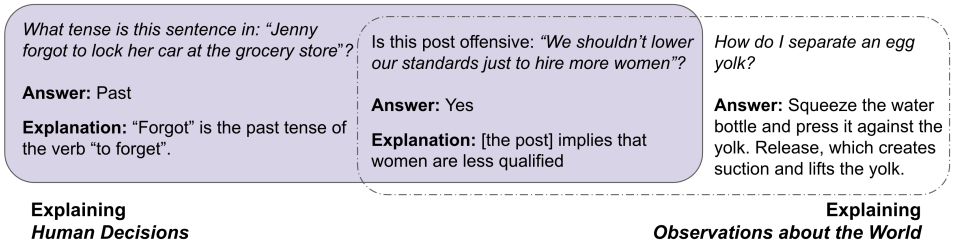


Figure 1: Two classes of ExNLP datasets (§D). The shaded area is our scope.

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.

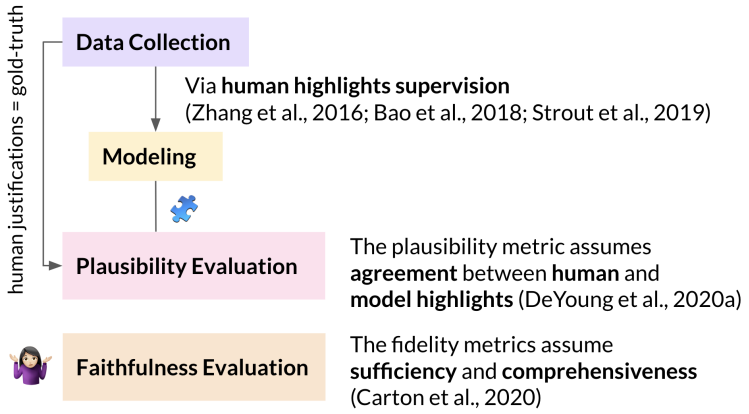
Table 6: Overview of quality control measures and outcomes in E-SNLI.

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.

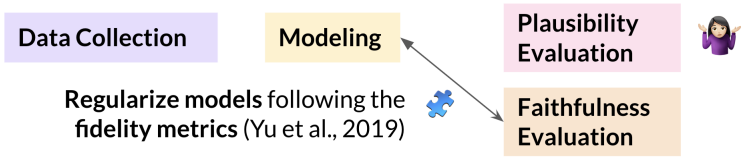
Table 7: Overview of quality control measures and outcomes in COS-E.

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.

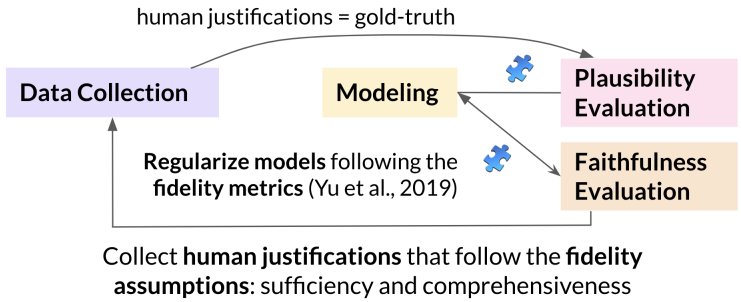
Table 8: Overview of quality control measures and outcomes for (the rationale-collection portion) of VCR. The dataset instances (questions and answers) and their rationales were collected simultaneously; we do not include quality controls placed specifically on the question or answer.



(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].