MOTIVATION FOR DATASHEET CREATION

[question1] Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

[answer1] Despite their impressive abilities, large-scale pretrained models often fail at performing simple commonsense reasoning. While most benchmark datasets target commonsense reasoning within the context of everyday scenarios, there is a rich, unexplored space of commonsense inferences that are anchored in knowledge about specific entities. We therefore create this dataset to benchmark how well current systems are able to perform this type of reasoning and to promote the development of systems that can handle these challenges.

[question2] Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

[answer2] We require all papers reporting on our dataset to submit their results to our dataset website (https://www.cs.utexas.edu/~yasumasa/creak).

[question3] What (other) tasks could the dataset be used for?

[answer3]

[question4] Who funded the creation dataset?

[answer4] This dataset was partially funded by the US National Science Foundation (NSF Grant IIS-1814522).

[question5] Any other comment?

[answer5] No

DATASHEET COMPOSITION

[question6] What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

[answer6] Each instance is a claim about an entity which may be either true or false. These claims are constructed such that validating them requires specific knowledge of each entity, with many also requiring commonsense reasoning incorporating these facts. All claims are written in English.

[question7] How many instances are there in total (of each type, if appropriate)?

[answer7] Our dataset consists of 13K claims, some of which form a smallscale contrastive evaluation set. A detailed breakdown of the number of instances can be seen in Table 1 of the main paper.

[question8] What data does each instance consist of ? “Raw” data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the in-stances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

[answer8] Each instance is a human-written claim about a given Wikipedia entity with an associated TRUE / FALSE label of its factually.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9]

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

[answer10]

[question11] Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11]
[question12] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
[answer12]
[question13] Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
[answer13] We include the recommended train, development, and test sets for our datasets. Each split is constructed such that there are no overlapping annotators nor entities between each set. We also include a small contrast set containing minimally edited pairs of examples with opposing labels of factually. The distribution of examples across splits can be seen in Table 1.
[question14] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
[answer14]
[question15] Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
[answer15] No, all resources are included in our release.
[question16] Any other comments?
[answer16] No

COLLECTION PROCESS

[question17] What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
[answer17] We use crowdsourcing to collect claims. Each worker is presented with 5 entities and are instructed to select one to generate two claims for, one true and one false. For each of these claims, workers are also instructed to provided a short explanation for why the claim is true or false.
[question18] How was the data associated with each instance ac-quired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
[answer18] We source our list of popular Wikipedia entities, as measured by number of contributors and backlinks, from Geva et al. [2021]. Annotators are also instructed to select one of five entities to construct an example for. Our sampling process, therefore, selects for popular entities that exist in Wikipedia.
While we do not cover the entire space of possible entity-centric claims, we promote diversity in our dataset by limiting the total number of claims a single worker can generate to 7% of any single split

and by sampling from a large pool of entities. In total, our dataset is comprised of claims that were generated from 684 total crowdworkers covering over 3,000 unique entities.

[question19] If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

[answer19] CREAK represents a subset of all possible entity-centric claims, including those which require commonsense in addition to retrievable facts to verify. Our dataset also only includes claims written in English.

[question20] Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

[answer20] We recruit crowdworkers from Amazon Mechanical Turk to perform the all the annotation steps outlined above.

[question21] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer21] The dataset was collected over a period of April to August 2021.

DATA PREPROCESSING

[question22] Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

[answer22] We do minimal preprocessing on the collected claims; however, we monitor crowdworker performance for sentence quality and remove repetitive examples produced by the same crowdworker. We also manually filter and clean our development and test sets for grammatically. This process removed roughly 18% of crowdsourced claims and high human performance (99% majority human performance) on 100 randomly sampled examples from our development set.

[question23] Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

[answer23] We maintain a record of all the original authored claims, as well as the explanations written by each claim’s author. This data will be made available upon request.

[question24] Is the software used to preprocess/clean/label the in-stances available? If so, please provide a link or other access point.

[answer24]

[question25] Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

[answer25] Our collection process indeed achieves our initial goals of creating a diverse dataset of entity-centric claims requiring commonsense reasoning. Using this data, we are able to evaluate how models that are trained on past data generalize to answering questions in the future, asked at the time of our data collection.

[question26] Any other comments

[answer26] No

DATASET DISTRIBUTION

[question27] How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)
[answer27] We make our dataset available at https://www.cs.utexas.
edu/~yasumasa/creak.
[question28] When will the dataset be released/first distributed? What license (if any) is it distributed under?
[answer28] Our data and code is currently available. CREAK is distributed under the CC BY- SA 4.0
license
[question29] Are there any copyrights on the data?
[answer29]
[question30] Are there any fees or access/export restrictions?
[answer30]
[question31] Any other comments?
[answer31] No

DATASET MAINTENANCE

[question32] Who is supporting/hosting/maintaining the dataset?
[answer32] This dataset will be maintained by the authors
of this paper. Updates will be posted on the dataset website.
[question33] Will the dataset be updated? If so, how often and by whom?
[answer33]
[question34] How will updates be communicated? (e.g., mailing list, GitHub)
[answer34]
[question35] If the dataset becomes obsolete how will this be communicated?
[answer35]
[question36] Is there a repository to link to any/all papers/systems that use this dataset?
[answer36]
[question37] If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?
[answer37]

LEGAL AND ETHICAL CONSIDERATIONS

[question38] Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer38]
[question39] Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.
[answer39]
[question40] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why
[answer40]

[question41] Does the dataset relate to people? If not, you may skip the remaining questions in this section.
[answer41]
[question42] Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
[answer42] We acknowledge that, because our dataset only covers English and annotators are required to be located in the US, our dataset lacks representation of claims that are relevant in other languages and to people around the world.
The data itself could possibly contain generalizations about groups of people; for example, one of the entities is Hopi people. As above, we audited all claims in the development and test set (20% of the data) and uniformly found claims to be respectful even when incorrect. However, incorrectly labeled claims in the training data could potentially teach false associations to trained models.
[question43] Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
[answer43]
[question44] Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
[answer44] Our dataset does not contain any personal information of crowd workers; however, our dataset can include incorrect information. We perform extensive quality control and error analysis to minimize the risk due to incorrect labels. We bear all responsibility in case of violation of rights.
Note that our dataset may, by design, contain false claims about real people or organizations. Most of the claims we saw are harmless in their incorrect nature rather than libelous; this includes all claims in the development and test data, which we manually inspected. However, there could be claims in the training set which are mislabeled and which could impart false "knowledge" to trained models.
We removed one entity from our dataset which was a deadname.
[question45] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
[answer45]
[question46] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
[answer46] Crowd workers informed of the goals we sought to achieve through data collection. They also consented to have their responses used in this way through the Amazon Mechanical Turk Participation Agreement.
[question47] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
[answer47] Crowd workers informed of the goals we sought to achieve through data collection. They also consented to have their responses used in this way through the Amazon Mechanical Turk Participation Agreement.
[question48] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
[answer48]

[question49] Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer49]
[question50] Any other comments?
[answer50] No