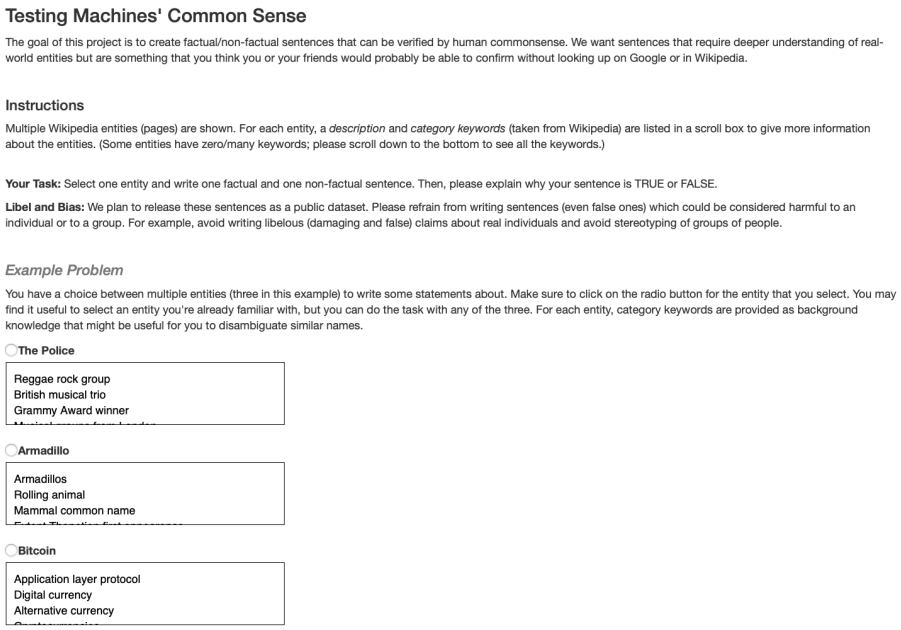
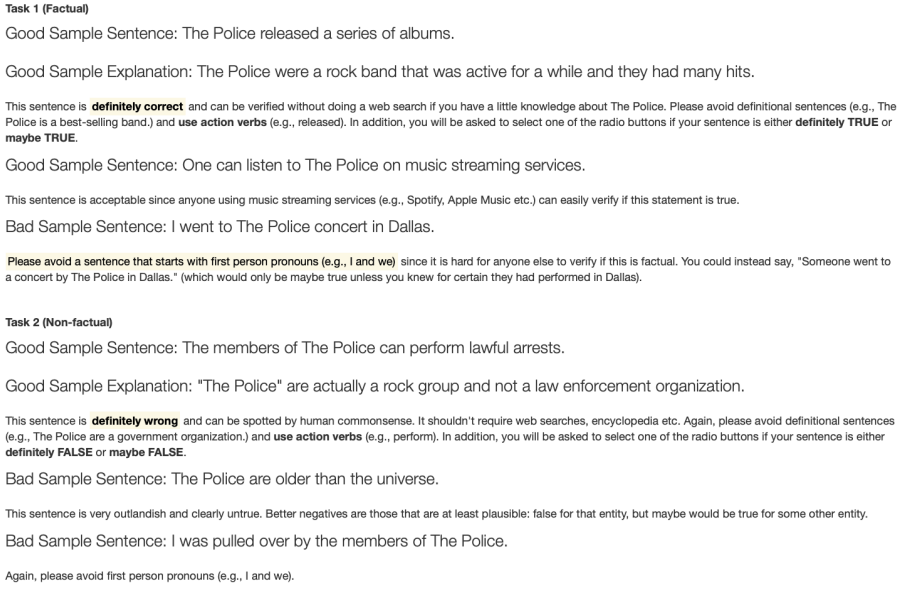
Appendix A: Annotation Interface

We compensate crowdworkers $0.60 USD per HIT. Each HIT is composed of generating one true and one false claim, along with a short explanation for each claim. Compensation was determined to approximate at least a $12 USD hourly wage. The total amount spent on compensating crowdworkers was roughly $4,000 USD. Our annotation instructions and interface are given in Figure 5.



(a) Instructions



(b) Examples

Figure 5: Annotation interface.

Appendix B: Examples

Table 5: CREAK claims with different reasoning types.

Claim	Reasoning Type	Label
Harry Potter can teach classes on how to fly on a broomstick.	Common Sense	TRUE
Grizzly bear live in danger of being hunted by other animals.	Common Sense	FALSE
The Atmosphere of Earth includes many types of gases.	Common Sense + Retrieval	TRUE
One can drive from La Jolla to New York City in less than two hours.	Common Sense + Retrieval	FALSE
J. P. Morgan restored the US Treasury surplus.	Retrieval	TRUE
François Mitterrand became a Texas Senator in 2001.	Retrieval	FALSE

Table 6: Unusable claims generated by crowdworkers.

Claim	Rejection Rationale	Label
It is alleged that a Nerd are computer geeks.	Subjective	True
Green Day radiates a folksy vibe.	Subjective	False
Dan Brown died in 2019 of heart failure.	Offensive	False
It is very fun to be audited.	Ambiguous	False
During the holidays people create performances.	Ambiguous	False
You can tell that a Goose is an alligator.	Outlandish	False

Table 7: Examples from contrast set.

True Claim	False Claim
U.S. Route 1 connects New York to Florida.	U.S. Route 1 connects New York to California.
The Beatles released their first album on vinyl.	The Beatles released their first album on Spotify.
Koi can cost someone hundreds of dollars.	Koi typically costs someone hundreds of dollars.
A nun takes a vow to remain unmarried and have no children.	A nun takes a vow to marry a priest and raise their children in the church.

Appendix C: Implementation Details

We train all models for a maximum of 10 epochs, with the exception of our T5-3B baseline finetuned on FEVER $_{\text{KILT}}$ which was trained for a maximum of 6. We select the best checkpoint, evaluated on development data after each epoch. All models are trained using the AdamW optimizer with no warmup steps. RoBERTa and T5-3B based models were trained with a learning rate of 5×10^{-6} and 3×10^{-5}, respectively. Our closed-book RoBERTa models and T5-3B model finetuned on FEVER $_{\text{KILT}}$ were trained with a batch size of 32 and our RoBERTa $_{\text{Large}}$ + DPR model with a batch size of 16. We use the `transformers` library Wolf et al. [2020][9] for our baseline implementations, and use DeepSpeed Rasley et al. [2020][10] for 16-bit floating point quantization on our T5-3B baselines. All experiments were run on four RTX 8000 GPUs, with our longest experiment taking three days. All our implementation details, including scripts for training/running each of our baselines are made available at `https://www.cs.utexas.edu/~yasumasa/creak`.

Appendix D: Datasheet for CREAK

A Motivation for Datasheet Creation

Why was the dataset created? 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

[9] transformers is licensed under the Apache-2.0 License

[10] DeepSpeed is licensed under the MIT License

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.

Has the dataset been used already? We require all papers reporting on our dataset to submit their results to our dataset website (`https://www.cs.utexas.edu/~yasumasa/creak`).

Who funded the dataset? This dataset was partially funded by the US National Science Foundation (NSF Grant IIS-1814522).

B Dataset Composition

What are the instances? 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.

How many instances are there? Our dataset consists of 13K claims, some of which form a small-scale contrastive evaluation set. A detailed breakdown of the number of instances can be seen in Table 1 of the main paper.

What data does each instance consist of? Each instance is a human-written claim about a given Wikipedia entity with an associated `TRUE` / `FALSE` label of its factually.

Does the data rely on external resources? No, all resources are included in our release.

Are there recommended data splits or evaluation measures? 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.

C Data Collection Process

How was the data collected? 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.

Who was involved in the collection process and what were their roles? We recruit crowdworkers from Amazon Mechanical Turk to perform the all the annotation steps outlined above.

Over what time frame was the data collected? The dataset was collected over a period of April to August 2021.

Does the dataset contain all possible instances? 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.

If the dataset is a sample, then what is the population? 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.

D Data Preprocessing

What preprocessing / cleaning was done? 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.

Was the raw data saved in addition to the cleaned data? 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.

Does this dataset collection/preprocessing procedure achieve the initial motivation? 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.

E Dataset Distribution

How is the dataset distributed? We make our dataset available at https://www.cs.utexas.edu/~yasumasa/creak.

When was it released? Our data and code is currently available.

What license (if any) is it distributed under? CREAK is distributed under the CC BY- SA 4.0 license.[11]

Who is supporting and maintaining the dataset? This dataset will be maintained by the authors of this paper. Updates will be posted on the dataset website.

F Legal and Ethical Considerations

Were workers told what the dataset would be used for and did they consent? 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.

If it relates to people, could this dataset expose people to harm or legal action? 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.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? 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

[11] https://creativecommons.org/licenses/by-sa/4.0/legalcode

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.