CommonsenseQA 2.0: Exposing the limits of AI through Gamification - Supplementary Material

Alon Talmor[1] **Ori Yoran**[1,2] **Ronan Le Bras**[1] **Chandra Bhagavatula**[1]
Yoav Goldberg[1] **Yejin Choi**[1,3] **Jonathan Berant**[1,2]
[1]The Allen Institute for AI, [2]Tel-Aviv University, [3]University of Washington
{ronanlb,chandrab,yoavg,yejinc,jonathan}@allenai.org
{alontalmor,oriy}@mail.tau.ac.il

1 Dataset Additional Information

1.1 Dataset documentation and intended uses

Documentation Our dataset and code are available at `http://allenai.github.io/csqa2`.

The dataset is provided in *jsonl* format (`https://jsonlines.org/`), such that each line is a single example.

Each example contains the following fields:

- `id`: Unique identifier for the example
- `answer`: *"yes"* or *"no"*
- `question`: Natural language question or assertion to which the answer is yes or no (for assertions: yes is considered true, and no is considered false)
- `confidence`: A number between 0 and 1.0 related to the quality of the question as produced by the Automatic question verification model (see section 2.2 in the main paper)
- `relational_prompt`: The relational prompt as displayed to the player (see section 2.1 in the main paper for details)
- `relational_prompt_used`: True/False, indicates whether the composing player has chosen to use the relational prompt.
- `topic_prompt`: The topic prompt as displayed to the player (see section 2.1 in the main paper for details)
- `topic_prompt_used`: True/False, indicates whether the composing player has chosen to use the topic prompt.
- `validations`: A list of player validations for the question that can take the values *"yes"*,*"no"* (answering the question directly), or *"bad question"*, *"sensitive"* indicating the question should be filtered out.

Intended uses We constructed this dataset to help researchers improve current natural language understanding models, by way of benchmark evaluation or as a training-set for other tasks. This dataset may also be used for probing models' commonsense and reasoning skills.

Personally identifiable information We do not include the AMT WorkerIDs or any personal information about our players in the public version of the dataset. Nor did we store any information except the AMT WorkerIDs while collecting the dataset.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

Potential negative societal impacts and offensive content Our methodology involving gamification that utilizes human players should be used with caution. We took extra measures to mark sensitive questions using validating players and we prevented the use of words that may be found offensive when players compose questions. Collecting data without these measures may result in unsafe questions that could cause model biases when trained on.

Estimated hourly wage As mentioned at the end of section 2.2, players received 4.4$ upon reaching 300 points, whereas the average session length was 18 minutes. To maintain a fair to high hourly wage we qualified players that have a high AI beat rate - thus they gain more points in a shorted time span. This resulted in an average hourly wage of 14.6$ per hour.

Structured metadata As the data is not in final publication state at the time of submission, the structured metadata to a dataset like schema.org, DCAT, and DOI will be added and maintained by The Allen Institute for AI website uppon adding the dataset to `https://allenai.org/data`.

Hosting and maintenance plan The dataset and code are hosted and fully maintain by The Allen Institute for AI. It will be permanently available under the link `http://allenai.github.io/csqa2`.

Dataset and code license We license our work using Creative Commons Corporation ("Creative Commons") 4.0. The exact licence can be found here `http://creativecommons.org/licenses/by/4.0` and in our website `https://github.com/allenai/csqa2/blob/master/LICENSE`.

We the authors will bear all responsibility in case of violation of rights.

2 Data Collection through Gamification

2.1 External Assets Licensing

For dataset construction as well as baseline analysis we used the following assets.

ConceptNet 5 This work includes data from ConceptNet 5 [1], which was compiled by the Commonsense Computing Initiative. ConceptNet 5 is freely available under the Creative Commons Attribution-ShareAlike license (CC BY SA 4.0) from `https://conceptnet.io`. The included data was created by contributors to Commonsense Computing projects, contributors to Wikimedia projects, Games with a Purpose, Princeton University's WordNet, DBPedia, OpenCyc, and Umbel.

GPT-3 The GPT-3 predictions used in this work were generated using OpenAI Beta API `https://beta.openai.com` that was licensed and paid for by The Allen Institute for AI.

Google Snippets The Google snippets were queries using a service called Zenserp `https://zenserp.com` that was licensed and paid for by The Allen Institute for AI.

T5 and Unicorn For our baselines we used T5 [2] `https://github.com/google-research/text-to-text-transfer-transformer` and Unicorn [3] `https://github.com/allenai/rainbow`.

2.2 Quality Assurance and Dataset Construction

Automatic question verification - input features example We provide an example for the input features to the question validation model defined in section 2.2 in the main paper.

Given a question such as *"Can a month ever have 5 Sundays? "*that was marked by a `Qualified` player as having the answer *"yes"*, were the composing player had `Medium` experience (determined by number of validations done by the player) and `High` validation accuracy. In addition, the question was answered *"no"* by the model-in-the-loop and answered *"yes"* by two validating players with `High` experience and `High` accuracy:

Question/Phrase
Q: You can see some light from ten feet under the water 1. Sunlight entering the ocean may travel 3,280 feet (long-tail knowledge) 2. 10 < 3,280 (comparison)
Q: None had ever reached the top of Mount Everest before 1977? 1. The first person reached the top of Everest in 1953 (long-tail knowledge) 2. May 23 1953 is before 1977 (comparison) 3. No one have reached the the top of Everest before 1977 (plausibility)
Q: A tea made of two cups of milk will be less darker than tea made of one cup? 1. Adding white substances to dark solutions causes them to be brighter (physical) 2. Two cups of milk is greater than one cup of milk (comparison)

Table 1: Examples of manually-annotated questions, with the required reasoning skill breakdown needed to arrive at the answers if they are not explicitly known. Each question is manually decomposed to phrases with the required commonsense skills required to arrive at the question.

- The feature `Qualified` will be assigned a value of 1, all other qualification features (`Unqualified`, `Expert level-1`, `Expert level-2`) will be assigned a value of 0.

- The feature that corresponds to the entry (`Composer_Ans:Yes`, `Composer_Acc:High`, `Composer_Exp:Medium`, `AI_Ans:No`) will be assigned a value of 1. All the other features that are conjunctions of the composer answer, composer accuracy, composer experience, and model-in-the-loop answer will be assigned a value of 0.

- The validation feature that corresponds to the entry (`Label:True,Acc:High,Exp:High`) will be assigned a value of 2, because 2 validating players independently matched this entry with their validation; all other validation features will be assigned a value of 0.

3 Dataset Analysis

Additional statistics The vast majority of our players were from the USA as we only opened the AMT task to USA users. However, some users have been identified to be from India and UK. Figure 1 provides a heat map of the amount of players from each state of the USA. The majority of our players are from California, Texas, Florida and New York.

Figure 2 shows breakdown of returning players vs. new players as well as player preference of device. We adapted our UI to support mobile usage, which some players found preferable.

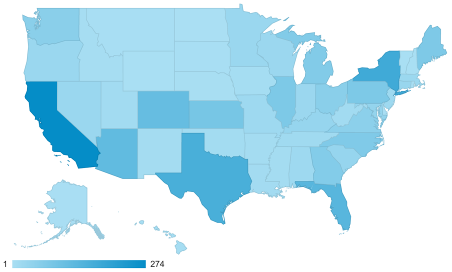


Figure 1: Heat map of players in the USA.

Reasoning Skills In the supplementary zip file we included the file *qualitative_reasoning_skills.csv* that contains the information used in the reasoning skills qualitative analysis in section 4 of the main paper. For each of the 110 random questions that were annotated, we provided the following fields: the ID, question, relational prompt, topic prompt, and answer to the question; boolean fields indicating whether the following reasoning skills are necessary to correctly answer the question: *capable of*, *long-tail knowledge*, *plausibility*, *comparison*, *physical*, *causality*, *temporal*, *negation*, *strategy* and *event chain*.

Table 1 provides three example annotations. Each annotation includes a question alongside a manual breakdown of the subquestions and skills required to answer the question.

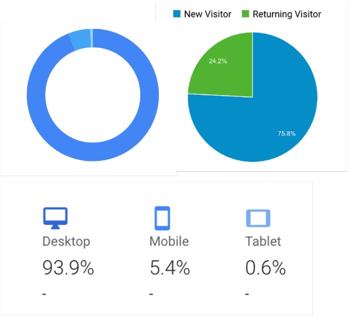


Figure 2: Session statistics.

4 Experimental Evaluation

4.1 GPT-3

4.1.1 How do we select few-shot examples?

We tried a variety of strategies for picking the few-shot examples in the prompt given to GPT-3.

K-Random We randomly select K examples from the training set. The order of these K examples is shuffled for each test instance and the prompt is formatted as shown above. We experiment with $K = \{5, 10, 15\}$.

K-NN We embed the questions in the training data using a SentenceTransformer [4]. Next, given a test instance, we find the K nearest neighbors as the training examples in the prompt. We experiment with $K = \{5, 10\}$

1-Random-Per-Relation We randomly select one example for each relation type in the training set. There are 34 unique relation types. Therefore, in this setting, our prompt contains 34 examples, one for each relation type.

5-Random-Per-Label We randomly select five examples per label – i.e. five examples whose answer is “yes" and five whose answer is “no". This results in a total of 10 examples in the prompt.

K-Random-Pair-For-Target-Relation For a given test instance, we pick K random *pairs* of training instances whose relation type is the same as that of the test instance. The *pair* of training instances is chosen such that one of the questions answer is “yes" and the other one is “no". Thus, we have a balanced set of $2 * K$ questions in the prompt. We experiment with $K = \{1, 5\}$

4.1.2 How are few-shot examples formatted?

In all cases, a question-answer pair is formatted as shown in Figure 3. For instances in the CSQA2 dataset that are formatted as statements, we convert them into a question by adding the prefix “Is it true that ...". This ensures a standard format across all questions in the prompt and the question being evaluated. However, in preliminary experiments, we did not notice any significant difference in performance when we add the prefix.

```
Question: Is it true that drinking milk causes bones to become weaker?
Answer: no
##
...
##
Question: When water freezes, does it get softer?
Answer:
```

Figure 3: Format of prompt, including few-shot examples, provided to GPT-3 to evaluate its performance on CSQA2

4.1.3 How do we select the best system?

We first evaluate 100 dev-set examples using all strategies mentioned above with five random seeds. Table 2 shows the results. Using five random examples achieves the best result.

Next, for the best strategy–i.e. 5-Random– we evaluate the full dev set using three random seeds. The maximum accuracy achieved across the three seeds is 0.547, while the mean is 0.512. We report the max performance in the main paper and use the corresponding output for analysis.

4.1.4 Other Hyperparameters

We use the OpenAI API to conduct these experiments. In all our experiments, we use the `davinci` engine with the following settings: $top_p = 1.0$, $temperature = 0.0$, $max\_tokens = 1$, $best\_of = 1$ and $stop = [“.”, “\backslash n”]$.

Strategy	Run 1	Run 2	Run 3	Run 4	Run 5	Mean	Std. Dev
5-NN	0.53	0.53	0.62	0.55	0.59	0.564	0.04
10-NN	0.60	0.58	0.56	0.48	0.52	0.548	0.04
5-Random	0.49	0.53	0.63	0.66	0.60	**0.582**	0.06
10-Random	0.52	0.48	0.67	0.53	0.56	0.552	0.06
15-Random	0.51	0.48	0.57	0.55	0.55	0.532	0.03
1-Random-Per-Relation	0.52	0.66	0.54	0.55	0.61	0.576	0.05
5-Random-Per-Label	0.50	0.52	0.52	0.57	0.56	0.534	0.04
Random-Pair-For-Target-Relation	0.46	0.49	0.59	0.56	0.52	0.524	0.05

Table 2: Results of experiment to pick the best strategy for selecting few shot examples for GPT-3[5]

4.2 T5 and Unicorn

4.2.1 Which pre-trained weights do we use to initialize the models?

For T5-Large and T5-11B, we use the pre-trained weights available at `gs://t5-data/pretrained_models/large` and `gs://t5-data/pretrained_models/11b`, respectively.

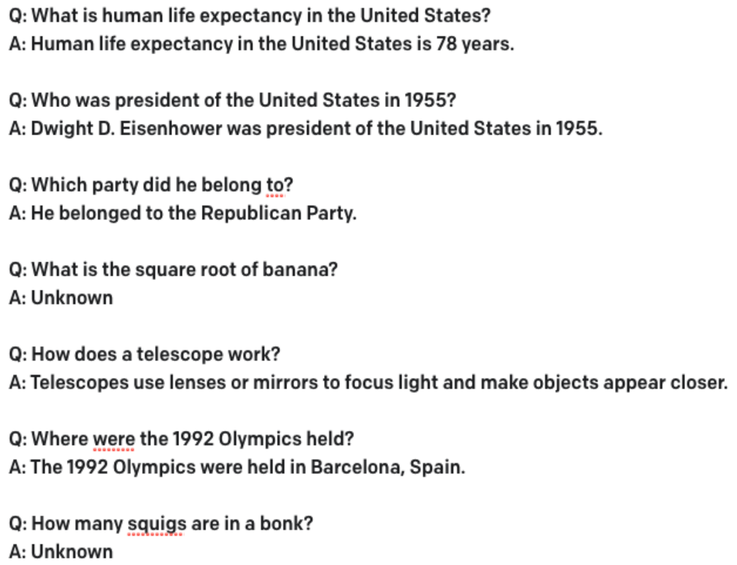
Similarly, the weights of the pre-trained Unicorn models are publicly available at `gs://ai2-mosaic-public/projects/rainbow/v1.0/unicorns-large/lr-4e-3` (Large) and `gs://ai2-mosaic-public/projects/rainbow/v1.0/unicorns/lr-2e-3_batch-size-32` (Unicorn-11B).

4.2.2 What are the hyper-parameters of the models?

For the T5 models (both Large and 11B), the learning rate is set to 1e-3 with a batch size of 4 and an number of epochs of 4. The number of gradient updates is set to 10, 000.

The Unicorn models (both Large and 11B) share the same learning rate (1e-3) and number of epochs (4) as the T5 models, while the batch size is set to 2 and the number of gradient updates is set to 20, 000.

5 Model Analysis



Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: Which party did he belong to?
A: He belonged to the Republican Party.

Q: What is the square root of banana?
A: Unknown

Q: How does a telescope work?
A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

Q: Where were the 1992 Olympics held?
A: The 1992 Olympics were held in Barcelona, Spain.

Q: How many squigs are in a bonk?
A: Unknown

Figure 4: The prompt used for the GPT-3 generative experiments. In order to allow the model to indicate when it is not sure of any specific answer, two questions in the prompt have an *Unknown* answer.

Contrast set We included in the supplementary zip the file *CSQA2_contrast_set.csv* containing the questions and annotations of the contrast set analysis described in section 6 of the main paper. Each row in the file relates to a question that appears in the dev set of CSQA2. For each question, we provide the original question ID (original_question_id), the original question (original_question), and the relational (original_question_relational_prompt) and topic (original_question_topic_prompts) prompts that were shown to the player when composing the question. The accuracy for each of our models on the original question is provided in the t5_11b_original_correct, unicorn_11b_original_correct, and gpt3_original_correct fields.

Additionally, each row features all the contrast set questions which are minor perturbations of the original questions. We provide the perturbed questions and answers in the perturbed_question_i, perturbed_answer_i fields, such that each question has between 2 and 6 perturbations.

We also provide fields that indicate whether each of our models is correct on every perturbation: t5_11b_correct_perturbation_i, unicorn_11b_correct_perturbation_i, and gpt_correct_perturbation_i. The *consistency* accuracy, which indicates whether each model is correct on the original question and all the perturbations is provided in the t5_11b_consistency, unicorn_11b_consistency and gpt3_consistency fields.

GPT-3 analysis The prompt used for the GPT-3 free-form predictions can be seen in Fig. 4. In addition, in the supplementary zip file we included the file *GPT-3_generative_predictions.csv* file with more information about the predictions in the GPT-3 analysis in section 6. The file contains the following fields:

- `id`: the ID of the question in the dataset
- `question`: the original question from the dataset
- `relational_prompt`: the relational prompt shown to the player when composing the question
- `topic_prompt`: the topic prompt shown to the player when composing the question
- `answer`: the correct answer to the question
- `GPT-3_prediction`: the prediction of the yes/no prompt GPT-3 model (main GPT-3 results)
- `GPT-3_correct`: whether the prediction of the yes/no prompt GPT-3 model (main GPT-3 results) is correct
- `GPT-3_generative_prediction`: the prediction of GPT-3 when presented with a default prompt for generating free-from answers
- `GPT-3_generative_prediction_consistent`: a boolean flag indicating whether the generative prediction of GPT-3 is consistent with our GPT-3 prediction. This flag was manually annotated by expert annotators.

Of the questions where GPT-3's generative answer agrees with the original prediction, we add the word *"why"* as a prefix to the question, and present the new question to the GPT-3 free-form prompt (see section 6, GPT-3 analysis, in the main paper for details). Expert annotators then annotated whether the generated explanation makes sense. For example, when asked to predict *Why does a cat not always have a tail?*, GPT-3 generated the following answer: *"A cat does not always have a tail because it is a mammal"*, that was annotated as a bad explanation to the original prediction. In the supplementary material zip file we added the file *GPT-3_why_predictions.csv* that includes the following fields in addition to the fields in the previously introduced *GPT-3_generative_predictions.csv*:

- `GPT-3_why_question`: the question with the *"why"* prefix that was presented to GPT-3
- `GPT-3_why_prediction`: the prediction generated by GPT-3 when presented with the question at the GPT-3_why_question field
- `GPT-3_why_accuracy`: a boolean flag indicating whether the prediction at the GPT-3_why_prediction field makes sense. This flag was manually annotated by expert annotators.

References

[1] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2018.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[3] Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *AAAI*, 2021.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.