Overview: We introduce the CODAH dataset, an adversarially-constructed evaluation dataset for testing common sense. CODAH forms a challenging extension to the recently-proposed SWAG dataset, which tests answering systems. Workers are rewarded for submissions that models fail to answer correctly both before and after fine-tuning. We create 2.8k questions via this procedure and evaluate the performance of multiple state-of-the-art question answering systems on our dataset. We observe a significant gap between human performance, which is 95.3%, and the performance of the best baseline accuracy of 69.6% by the BERT model.

**Dataset Construction**

- **CODAH dataset** for common sense question-answering:
  - **CODAH: CommonSense Dataset** Adversarially-authored by Humans
  - Multiple Choice sentence completion in the style of SWAG
  - Tagged with different types of commonsense reasoning
  - Over 25% gap between model & human expert accuracy

- **Novel method for adversarial question generation:**
  - Annotators are educated on SOTA QA models
  - Submissions are credited for questions that the model fails to answer correctly

**Results**

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idioms</td>
<td>Including phrases whose meaning cannot be readily interpreted from the meaning of constituent parts</td>
<td>A man on his first date wanted to break the ice. He drank all of his water. threw the ice at the wall. looked at the menu. made a corny joke.</td>
</tr>
<tr>
<td>Negation</td>
<td>Including negatives to dictate the meaning of the sentence</td>
<td>The man’s rebuttal was clearly not nonsensical. The rebuttal has nothing to do with sense. had some reasons associated with it. did not make any sense. was funny.</td>
</tr>
<tr>
<td>Polysemy</td>
<td>Testing the understanding of multiple meanings of a single word</td>
<td>An architect retrieves his compass. He computes the area of a circle. explores the open sea. draws building dimensions on a canvas. uses his compass to find the north cardinal direction.</td>
</tr>
<tr>
<td>Reference</td>
<td>Requiring understanding of reference to one of multiple subjects</td>
<td>Rose is walking the dog while Joseph cooks dinner. Rose is following a new recipe. enjoys the fresh air. wags her tail with joy. cuts tomatoes for the soup.</td>
</tr>
<tr>
<td>Quantitative Reasoning</td>
<td>Involving basic arithmetic calculations or comparisons</td>
<td>A woman is walking two dogs and carrying a cat on her way to her car. She puts all three animals in the back seat before driving off. puts all four animals in the back seat before driving off. puts both animals in the back seat before driving off. puts all nine animals in the back seat before driving off.</td>
</tr>
</tbody>
</table>

**Model Accuracy by Category**

- **Idioms**: BERT > GPT-1 > Human
- **Reference**: BERT > GPT-1 > Human
- **Polysemy**: BERT > GPT-1 > Human
- **Negation**: BERT > GPT-1 > Human
- **Quantitative**

**Discussion**

**Annotation Artifacts**

- Authors are incentivized against writing questions with artifacts which are learnable by the model in CV
- Artifacts do not provide sufficient signal for models to approach human-level accuracy

**Dataset Size:**

- CODAH is a challenging extension to the SWAG dataset
- Finetuned models with human-level SWAG performance still struggle on CODAH in validation
- Distinct from and complementary to SWAG questions

**Acknowledgements**

We thank the students of Northwestern EECS 349 Fall 2018, whose creativity and insight made this work possible. This work was supported in part by NSF Grant IIS-1351029 and the Allen Institute for Artificial Intelligence. We also thank Yiben Yang for helpful advice and Jay Alammar for the BERT graphic.